Capstone Session 6

Machine Learning for Modeling

Session 6 : Dataset Description

Adultcensusincome.csv

Variable	Description	Variable	Description
Age	Age of the person	workclass	Workclass of the person
fnlwgt	Weighted tally of specified socio-economic characteristics of the population	Education	Education level of the person
education.num	Number of years of education	marital.status	Marital status of the person
occupation	Occupation of the person	relationship	Relationship status of the person
race	The race of the person	sex	The person’s sex (Male/ Female)
hours.per.week	Number of working hours per week	native.country	The native country of the person
Income	The income category of the person	-	-

Session 6: Building an Income Classification Model

Task: Build a classification model for predicting the income using the Adult Census Income Dataset.

- Load the dataset
- Check for null values and ? in any columns and handle those values. Check the distribution of target variable income and identify if the dataset is balanced.
- Perform the following Univariate analysis: Create a distribution plot for columns income, age, education.num, and education
- Create a pie chart for Marital status. Use column marital.status
- Perform the following Bivariate analysis
    - Using Plotly, build a scatter plot for age and education.num. Add income and hours per week for color and size.
    - Using Seaborn, build a violin plot of education and age. Add income for hue and use split=True and inner='quart'

Note: If time allows, add a sentence or two to provide your observations for each visual.

Session 6: Building an Income Classification Model

- Perform the following Bivariate analysis
  - Create a countplot of income across columns age, education, marital status, race, sex
  - Draw a heatmap of data correlation and find out the columns to which income is highly correlated
- Prepare the dataset for modeling
  - Encode all the categorical columns
  - Prepare independent variables X and dependent variable Y (Income).
  - Perform feature scaling using StandardScaler and fix the imbalance in the dataset using any one of the techniques like SMOTE or RandomOverSampler
  - Perform a train test split in the ratio 80:20 and random_state 42.
- Perform Data Modeling
  - Train Logistic Regression Model, KNN Classifier Model, SVM Classifier, Naive Bayes Classifier, Decision Tree Classifier and Random Forest Classifier
  - Perform model evaluation on Accuracy and F1 score and identify the best model. Hint: build a table to compare all results side-by-side

Note: If time allows, add a sentence or two to provide your observations for each visual/evaluation.

Thank You