Feature Encoding & Scaling

Objective

The objective of this task is to convert categorical data into numerical form and scale numerical features so that the dataset becomes suitable for machine learning models.

Dataset Used

- Adult Income Dataset

Tools Used

- Python
- Pandas
- NumPy
- Scikit-learn
- VS Code
- Jupyter Notebook

Task Performed

1. Data Understanding

- Loaded the Adult Income dataset using Pandas.
- Identified numerical and categorical features.
- Numerical features included age, education number, capital gain, capital loss, and hours per week.
- Categorical features included workclass, education, marital status, occupation, gender, and native country.

2. Feature Encoding

- Applied **Label Encoding** on the target column (income) because it has an order.

- Applied **One-Hot Encoding** on categorical features where no order exists.

- Converted categorical data into numerical format so that machine learning models can understand it.

3. Feature Scaling

- Used **StandardScaler** to scale numerical features.

- Scaling transformed the data so that features have similar ranges.

- After scaling, numerical features had mean close to 0 and standard deviation close to 1.

4. Dataset Comparison

- Compared dataset before and after encoding and scaling.

- The processed dataset is now suitable for machine learning algorithms.

5. Dataset Saving

- Saved the final preprocessed dataset as a CSV file for future model training.

Conclusion

This task helped in understanding the importance of feature engineering in machine learning. Encoding and scaling are essential steps before training any model. Proper preprocessing improves accuracy, efficiency, and overall model performance.