Assignment - 1 (ML / DL / NLP) - Junayed
Mahdy Research Academy
Dhaka, Bangladesh

Machine Learning Assignment: Adult Income Dataset

In this assignment, you will apply various machine learning techniques to the Adult income dataset as shown in class, which contains information about individuals from the 1994 Census database, such as age, education, occupation, and whether they make more than $50K per year. The goal is to predict the income class of each individual based on the given features.

You will follow the general steps below to complete each experiment:

1. **Data loading:** Load the dataset from the given URL or file path using pandas or any other library of your choice. The dataset has 15 columns and ~49000 rows. The column names and descriptions are as follows:

Column Name	Description
age	Age of the individual
workclass	Working class of the individual
fnlwgt	Final weight, which is the number of units in the target population that the individual represents
education	Level of education of the individual
education-num	Number of years of education of the individual
marital-status	Marital status of the individual
occupation	Occupation of the individual
relationship	Relationship status of the individual
race	Race of the individual
sex	Sex of the individual
capital-gain	Capital gain of the individual
capital-loss	Capital loss of the individual
hours-per-week	Number of hours worked per week by the individual
native-country	Native country of the individual
income	Income class of the individual (<=50K or >50K)

2. **Data inspection:** Inspect the dataset using methods such as head, tail, info, describe, etc. to get a sense of the data types, ranges, distributions, and missing values of each column.
3. **Data preprocessing:** Perform the following data preprocessing tasks on the dataset:
    a. **Handling missing/null values:** The dataset contains some missing values, which are denoted by a question mark (?). You will explore different methods to handle these missing values.

b. **Data splitting:** You will split the dataset into training and testing sets, using a ratio of your choice.

c. **Data cleaning:** You will clean the dataset by removing any irrelevant, redundant, or noisy columns or rows. For example, you will remove any columns that have too many null values, or any rows that have outliers or incorrect values.

d. **Data encoding**: You will encode the categorical columns of the dataset, such as workclass, education, occupation, etc. using methods such as label encoding. You will interpret and understand the advantages and disadvantages of each method, and how they affect the machine learning models later.

e. **Data balancing**: You will balance the dataset by addressing the class imbalance problem in the target variable (income). You will use methods such as upsampling, downsampling, or both to create a balanced dataset. You will use techniques such as SMOTE and TOMMEK. You will compare the performance of your machine learning models on the balanced and unbalanced datasets later.

f. **Feature selection:** You will select the most relevant and informative features for your machine learning models, using methods such as correlation analysis. You will use a threshold value to filter out the features that have low or high correlation or significance with the target variable. You will interpret and understand the impact of feature selection on the model performance and complexity later.

4. **Data visualization:** You will visualize the dataset using various plots and charts, such as histogram, bar plot, scatter plot, pairwise plot, etc. You will use these plots to explore the univariate, bivariate, and multivariate relationships among the features and the target variable. You will also use these plots to identify any patterns, trends, outliers, or anomalies in the data.
5. **Machine learning models:** You will apply various machine learning models to the dataset,. In our class, we used logistic regression, random forest, support vector machine, and decision tree. You will report each and every experiment on the following additional models: lightgbm, naive bayes, gradient boosting, ada boost, xg boost, K Nearest Neighbors, and Bagging classifier.
6. **Performance evaluation:** You will evaluate the performance of your machine learning models using various metrics, such as accuracy, precision, recall, f1-score, etc. You will use methods such as classification report, or confusion matrix, to visualize and interpret these metrics.

It is not necessary that you follow the above steps in the given order. You may follow the order shown in the session and in the github repository. Hints and sample codes can be found at:
https://github.com/junayed-hasan/Adult-Income-Prediction-Machine-Learning

Differently from what has already been shown in class, you will conduct the following set of experiments, and compare the results:

Experiment #	**Change**
1	In our original experiment of the class, we did preprocessing steps like removing null value columns and rows, handling outliers, normalization, balancing, and removing features using correlation. In this experiment, you will do none of the pre-processing steps other than removing null value rows. This will serve as the **baseline experiment** to compare your other results with.
2	Handling missing values - In class, we have removed the null columns using a threshold method, and we removed the null rows completely. Other methods like replacing with mean, median or mode can be done. You have to conduct the experiment by replacing with **mean values** and report the final results.

3	Data splitting - In class, we split the dataset into train and test sets in the ratio 80:20. You will split the data into **70:30** and report the final results.
4	Data balancing - In class, we used SMOTE to upsample the data to the majority class. In this experiment, you will use **both the SMOTE and the TOMMEK libraries** to upsample the minority class and downsample the majority to the average of their total samples, and report the results.
5	Feature elimination - In class, we used correlation and null values to drop features/ columns. In this experiment, you will also use RFECV to eliminate features along with the previous two techniques. For correlation, use >=80% as the threshold. For null values, use >=70% as the threshold. For RFECV, Implementation details can be found at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html
6	Handling outliers - In class, we handled the outliers and replaced them with the median of the feature if they lie more than 2 standard deviations away. In this experiment, you will handle those outliers by replacing them with the mean instead of median.
7	Cross-validation - In class, we did simple cross validation on the data. In this experiment, you will perform stratified K-Fold (for K = 10) cross validation and report the data.
8	Hyperparameter optimization - Perform randomized search CV on all the models keeping at least 5 hyperparameters in the search space such that the total number of combinations is greater than 1000, and perform at least 50 combinations to choose the hyperparameters, and report the results.
9	Ensemble modeling: In class, we used hard voting to make the ensemble. Here, report the results using **soft** voting.
10	Best combination: Observe the experiments for which you get the best results. Combine them and report the highest performance you can achieve. For example: you can use all the preprocessing steps, handle missing values using mean, use 10-Fold stratified cross validation, all the feature elimination techniques (setting threshold), handle outliers with median and 2 standard deviation rule, do hyperparameter optimization and hard voting ensemble modeling.

You will submit 10 notebooks, containing the results of these 10 experiments on each of the 11 models (logistic regression, random forest, support vector machine, decision tree, lightgbm, naive bayes, gradient boosting, ada boost, xg boost, K Nearest Neighbors, and Bagging classifier).

—----------------- Best of luck! —-----------------