Lab Assignment 6 – Ensembles

Objective: To create classification models using RandomForest, Bagging Classifier and Logistic regressor for the census dataset which consists of 32651 observations and 15 columns, 14 of them being features and one the target variable (**high_income**). Use this dataset to classify if the potential income of people into 2 categories: people who make less or equal to $50K a year and people who make more than $50K a year.

1. Load dataset
2. Data Pre-processing:
   - Encode target column "high_income" : {encode values "<=50K" as 0 and ">50K" as 1}
   - Also encode the other categorical features
   - Impute the missing values present in categorical columns and numerical columns
   - remove any undesirable feature from the dataset.
   - Check for the outliers in the columns and treat the outliers if present.
3. Split the dataset into train and test.
4. Construct classification model using RandomForest classifier with different hyper-parameters values. Use hyper-parameters such as: **n_estimators**, **criterion**: {"entropy", "ginni"}, **max_depth, min_samples_split, bootstrap, min_samples_leaf, max_features, max_leaf_nodes.**
5. Construct Bagging classifier models using **base_estimator** such as {logistic regressor, Naïve Bayes} and with different values for n_estimators.
6. Compare the performance of these models constructed in step 4 and 5, using various evaluation metrics such as accuracy, precision, recall.