Lab Assignment 7 – Random Forests

Build **Random Forest Classifier** using **Sklearn** for predicting Online News Popularity.

Steps

1. **Dataset:** Download the dataset from the link https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity.
2. **Preprocessing**:
   - Remove unwanted features such as url, timedelta.
   - Convert the categorical features into numerical using one hot encoding if any.
   - Perform range normalization on numerical features not in the range of 0 to 1.
3. **Data Splitting:** Split the dataset into training and testing using 70-30 division.
4. **Random Forest Modelling:** Build a Random Forest regression model using Sklearn with default parameters. Predict the target values in the testing set. Apply regression metrics and visualize the results as graphs.
5. **Playing with Random Forest:** Change the following parameters of the random forest and analyze their performance for training and testing using the evaluation measures.
   - n_estimators
   - criterion{“mse”, “mae”}
   - max_depth
   - min_samples_split
   - bootstrap
   - n_jobs
   - min_samples_leaf
   - max_features
   - random_state
   - max_leaf_nodes
6. **Comparison:** Compare the performance of the Random Forest model with other regression models such as linear regression, polynomial regression, decision tree regression etc.
7. **Performance improvement:** You can try different strategies to see whether testing error comes down or not. Strategies can be different 1. removal of some features, 2. normalization methods, 3. Shuffling of training samples. Check the model error for the testing data for each setup.
8. **Random Forest Classifier:** Pick a Regression dataset of your choice and perform training testing similar to above. Play with model parameters and analyse the results using regression measures.

Suggested Packages: Numpy, Pandas, Sklearn