School of Computer Science Engineering and Technology

Course- BTech
Course Code- 301
Year- 2022
Date- 21-02-2022

Type- Core
Course Name-AIML
Semester- Even
Batch- 4th Sem (SPL)

6 - Lab Assignment No. 6.1

Objective: To Implement Random forest classifier

Problem Statement: Build a Decision Tree Classifier using Sklearn for predicting the number of shares in social networks (popularity).

About Dataset: This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years.

Data Set Characteristics:	Multivariate	**Number of Instances:**	39797	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	61	**Date Donated**	2015-05-31
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	355284

Steps

1. **Dataset:** Download the dataset from the link https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity
2. Remove unwanted features such as url, timedelta
3. Convert the categorical features into numerical using one hot encoding if any.
4. Perform range normalization on numerical features not in the range of 0 to 1.
5. Split the dataset into training and testing using 75-25 division
6. 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.
7. Playing with Random Forest: Change the following parameters of the random forest and analyze their performance for training and testing using the evaluation measures.
    a. n_estimators
    b. criterion{“mse”, “mae”}
    c. max_depth
    d. min_samples_split
    e. bootstrap
    f. n_jobs
    g. min_samples_leaf
    h. max_features
    i. random_state
    j. max_leaf_nodes
8. Compare the performance of the Random Forest model with other regression models such as linear regression, polynomial regression, decision tree regression etc.

Suggested Platform: Python: Azure Notebook/Google Colab Notebook, packages such as numpy, nltk, regular expression package re.

Additional (Not a part of the evaluation**)**

*** 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.

***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.