School of Computer Science Engineering and Technology

Course- BTech	Type- Core
Course Code- 301	Course Name-AIML
Year- 2022	Semester- Even
Date- 21-02-2022	Batch- 4th Sem (SPL)

6 - Lab Assignment No. 6.1

Objective: To Implement Random forest classifier

Problem Statement: Build a Random forest classifier using Sklearn that will detect if the mushroom is edible or poisonous by its specifications like cap shape, cap color, gill color, etc.

About Dataset: This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy..

Data Set Characteristics:	Multivariate	**Number of Instances:**	8124	**Area:**	Life
Attribute Characteristics:	Categorical	**Number of Attributes:**	22	**Date Donated**	1987-04-27
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	711016

Steps

1. **Dataset:** Download the dataset from the link
   https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/
   Refer .names file to add the column name
2. Check the shape of dataset
3. Print the count of edible and poisonous mushrooms
4. Check the presence of Null values if present handle those null values
5. Remove unwanted features
6. Convert the categorical features into numerical data using lable encoding.
7. Extract Idependent variale and dependent variable
8. Split the dataset into training and testing using 70-30 division
9. Build a Random Forest classification model using Sklearn with n_estimators=2, random_state=42.
10. Predict the target values in the testing set.
11. Create the confusion matrix.
12. Check the accuracy
13. 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{"gini", "entropy"}
    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

14. Compare the performance of the Random Forest model with other classification model such as logistic Regression.

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.