CENG313 Introduction to Data Science
Fall 2022-2023
Lecturer: Dr. Duygu Sarıkaya
Teaching Assistant: Berrin İşlek
Gazi University, Department of Computer Engineering
Assignment 3 is due 19th of December 2022, Monday 23:59

A random forest is a meta-estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Iris Dataset:

The Iris flower data set or Fisher's Iris data set is a multivariate data set. The data is collected to quantify the morphologic variation of Iris flowers of three related species. The dataset consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor, where total number of samples is 150). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. You can load the Iris dataset using scikit-learn (sklearn.datasets. load_iris)

For this assignment, you will need these libraries: pandas, NumPy, scikit-learn (sklearn), GraphViz,.

1. Split the Dataset and Build a Random Forest Classifier

In this assignment, you will build a Random Forest Classifier which consists of 10 trees to predict the species of Iris Flowers. You will build the model with gini index criterion. **First you will split your data into training and test sets.** For this step you can use %75 of your data for training and the remaining %25 for testing. Do not forget to first **randomize your data**! Finally, fort his step, train your model using the training data.

2. Evaluate your model on the Test Set

Then use the model you have trained to **predict the species of each flower in the test set**. Create a **confusion matrix** for your predictions (https://en.wikipedia.org/wiki/Confusion_matrix) Report the **accuracy, precision, recall** (https://en.wikipedia.org/wiki/Precision_and_recall) **and f-1 score** (https://en.wikipedia.org/wiki/F-score) of your model on your test set.

3. Visualize a Single Tree

Now, extract a single tree from your Random Forest. You can pick this tree at random, it should be one of the 10 trees used to build the Random Forest Classifier. Now, please use Graphviz library to visualize this tree.

For this assignment, **you are allowed to use scikit-learn library's built in functions** to create Random Forest Classifier. For visualization purposes, you can use Graphviz library. As usual, you may use other fundamental libraries we use in class (matplotlib, numpy, scipy, pandas).

Submission:

You will submit a jupyter notebook (ipynb file) with executable Python script and comments (explanations). The file will be uploaded on lms (guzem). You can upload a zip file that contains the

jupyter notebook (ipynb file). **Important Note: Please submit your file name with this format: Studentno_StudentName_StudentSurname**

Grading:

The total is 100 points. You will receive points only if your script executes and works correctly, if you have covered each point mentioned and answered these questions, and written comments that explain each main step.

Course Rules and Expectations

All work on programming assignments must be done individually unless stated otherwise. You are encouraged to discuss with your classmates about the given assignments, however, everything that is turned in for each assignment must be your own work. In particular, it is not acceptable to: submit another person's assignment as your own work (in part or in its entirety), get someone else to do all or a part of the work for you, submit a previous work that was done for another course in its entirety (self- plagiarism), submit material found on the web as is etc. **Important Note: Material found online and used as is will lead to your code being similar to many others.** These acts are in violation of academic integrity (plagiarism), and these incidents will not be tolerated. Homeworks, programming assignments, exams and projects are subject to Turnitin and Moss (Measure of Software Similarity) checks. Use sources to learn from only, and write your own code from scratch.