DMG Assignment 2

Hello everyone, after the first assignment of sampling, its time to learn classification which is one of the most important parts of Data Mining as well as Machine Learning.

1. [25 Marks]
   In this question, you need to classify the given dataset (attached) using the Decision Tree Classifier. You need to classify this on different depths (at least 5) of the decision tree and plot the accuracy vs depth graph. This accuracy plot will be made for training data as well as for the testing data. For training data, use cross-validation to compute the accuracy. (Take the average of 5 accuracies if you use 5 fold cross-validation. Find the optimal decision tree and use that in the next question(Q2)).

2. [25 Marks]
   In this question, you need to classify the dataset (same dataset) using
   a. Naive Bayes Classifier
   b. KNN Classifier
   c. Decision Tree Classifier

   For each classifier, you need to draw a decision boundary and show the classification over the graph. To reiterate, please choose the optimal classifier for the decision tree that you found in question-1.

Description of the Dataset:

- Dataset is a by-product of information from customer clicks monitored by an online fashion portal. (It is a real-world dataset and hence might require cleansing).
- There are about 275 attributes [Categorical in Nature] (the last one being the binary classification label- True/False).
- Goal: Predict whether the visitor will view another page on the site or will he leave. (True/False)
- Note that Dataset is in the **arff file format** (a frequent one in data mining), to get you started a **short snippet** has been provided which will help you visualize the data as well as explore the categories *(please understand it and read the comments for a better understanding)*
  PS: Before Running the script please install liac-arff (pip3 install liac-arff)

Test and Train Dataset along with snipped has been attached