CSE 482 Exercise 5 (Date: November 23, 2022)

The purpose of this exercise is to help you learn how to use some of the data summarization and visualization functions as well as the decision tree method available in Python. Follow the instructions below to complete the exercise. Save your IPython notebook as exercise5.ipynb.

1. Download the data vehicle.csv from the class Web page (from lecture 2). The data contains information about different vehicles along with their respective types (saab, opel, bus, van).
2. In this exercise, you will first compute summary statistics about the data. To do this, write an ipython notebook that implements the following steps:
   a. Use pandas to load the CSV file into a DataFrame object named “data”. After the file is successfully loaded, type data.head() to display the first 5 rows of the table.
   b. Compute the correlation between the first 4 attributes of the data (compactness, circularity, distance_circularity, and radius_ratio). You should obtain a correlation matrix of size 4 x 4.
   c. Plot a histogram for the compactness attribute.
   d. Draw a boxplot for the first four attributes (the notebook should display 4 different boxplots on the same diagram).
3. For this question, you will apply a decision tree classifier to the data set.
   a. Create a Series object named Y by extracting the last column of the DataFrame named “data”. The Series will contain the true class labels of the data.
   b. Create a DataFrame named X by dropping the last column of the DataFrame named “data”.
   c. Split the data so that 2/3 of the examples are in the training set and 1/3 of the examples are in the test set. Set the random seed to be 1.
   d. Create a list named maxdepth = [2, 5, 10, 15, 20]
   e. Repeat the following steps by iteratively changing the maximum depth of the tree to be one of the values given in the “maxdepth” list above
      i. Train a decision tree with the given maximum depth from the training data.
      ii. Apply the tree to the training data and calculate the training accuracy.
      iii. Apply the tree to the test data and calculate the test accuracy.
   f. Plot the training and test accuracies as a function of the maximum depth (maxdepth) of the tree. To do this, you should store the training and test accuracies in two separate lists, named trainAcc and testAcc. You can then plot them as follows:

      ```
      import matplotlib.pyplot as plt
      plt.plot(maxdepths,trainAcc,'ro-',maxdepths,testAcc,'bv--')
      ```

      This will plot the training and test accuracies on the same plot.

Deliverables: Submit (via D2L) the file exercise5.ipynb created.