Northeastern University, College of Computer and Information Science

CS 6220 Data Mining — Assignment 8A

Decision Tree

To complete this assignment, you will download a classification dataset of your choice from the UCI machine learning repository.Choose something that seems interesting to you!
On previous assignments, we have learned how to load two-dimensional data into pandas dataframes, which can be manipulated to serve as input data to scikit-learn models. Use that as a reference.

Objectives:

1. Implement a decision tree using scikit learn.
2. Display the final decision tree.
3. Visually interpret generated during the training process.

Submission:
Submit your assignment through Blackboard as before.

Grading Criteria:
Follow the instructions in the pdf, and complete each task. You will be graded on the application of the modules topics, the completeness of your answers to the questions in the assignment notebook, and the clarity of your writing and code.

WHAT TO DO

1. Write a small paragraph describing the dataset that you choose, its features, number of instances, nature of the data, and anything else that you found to be interesting.

2. Provide a brief analysis of the dataset you downloaded. Does it have missing data? Are the features numeric/discrete/categorical? Create some histograms/boxplots/other visualizations to illustrate the content o the dataset.

3. Using scikit-learns DecisionTreeClassifier, train a supervised learning model that can be used to generate predicions for your data. A reference to how you can do that can be found on scikit-learn.

4. The link above explains how you can generate a visual output for the tree you just trained. Use that code snippet to create a visualization of your tree.

5. Create a new instance with your choice of values for each of the features. Use your trained model to generate a prediction for it. Using your tree illustration as a reference, write a short paragraph describing how your model went about generating that specific prediction. Does it make sense to you? Can it be improved? Go back and play with the parameters that you used for training your tree and see if you can obtain better results.