# CS 6220 Data Mining — Assignment 8

## Decision Tree

This assignment will require you to implement and interpret the concepts of decision tree. Keep in mind that the main objective of this assignment is to highlight the insights that we can derive from applying these techniques—the coding aspect is secondary. Accordingly, you are welcome to consult any online documentation and/or code that has been posted to the course website, so long as all references and sources are properly cited. You are also encouraged to use code libraries, so long as you acknowledge any source code that was not written by you by mentioning the original author(s) directly in your source code (comment or header).

**Objectives:**

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

**Submission:**

Through the assignment submission portal on Canvas, submit your ipynb with a pdf of your assignment solution; no need to zip the files.

**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.

# Assignment Description

**The Data**

The given dataset contains information on the Banknote Authentication dataset. The given dataset contains information on images of genuine and forged banknotes, represented by 4 numeric attributes (variance of Wavelet Transformed image, skewness of Wavelet Transformed image, curtosis of Wavelet Transformed image, and entropy of image) and a binary target attribute indicating whether the banknote is genuine or forged. You can know more about the dataset from here.
The dataset can be downloaded using
https://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt

**What to Do**

First, download the Banknote Authentication Dataset from here https://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt
You can load this dataset into a dataframe using the pandas library.

1. Split the dataset into 60% training set and 40% test set.
2. Using scikit-learn's DecisionTreeClassifier, train a supervised learning model that can be used to generate predictions for your data. A reference to how you can do that can be found here.
3. Report the tree depth, number of leaves, feature importance, train score, and test score of the tree.
4. Now you will generate decision trees on the same training set using fixed tree depths. The tree depth can be set using $max{=}d$, where $d$ is the depth of the tree.
5. Decrease depth from the decision tree in Step 2, and for every depth (from max depth to depth 1), report tree depth, number of leaves, feature importance, train score, and test score of the tree.
6. Show the visual output of the decision tree from Step-2.
7. Show the visual output of the decision tree with highest test score from Step-5.

   To visualize the decision tree, use Graphviz library. You can find details in this link. Show the feature names and class names in the visualization.

**What to Provide**

Your output should contain the following:

- Report of tree depth, number of leaves, feature importance, train score, and test score of the decision tree.

- Report of tree depth, feature importance, train score, and test score of the decision tree using different tree depths.
- Visual output of the decision trees. There should be two separate decision trees as mentioned previously.