CPS 844 Lab 4: Model Overfitting

Classification is the task of predicting a class label using other attributes.

You are provided a dataset of 1500 labeled two-dimensional records. The file ‘Xdata.npy’ has the attribute values of the 1500 records, while the file ‘Ydata.npy’ has the corresponding class labels of these 1500 records. The assigned labels consist of 0 or 1 instances. You are given this dataset without further explanation about the data, its provenance, or any description of the relevant characteristics. You are tasked to do your level-best at developing and training a decision tree that will accurately classify unseen records.

You will grow several decision trees with different depths. As the model becomes more complex, you will notice that the training accuracy will improve. Notwithstanding, the test accuracy will initially improve, up to a maximum depth (that you will have to determine) before decreasing due to model overfitting.

Write a Python script that performs the tasks described in lab4.py. Submit the .py file on D2L. Please note that if you submit your file in some other format besides .py or (.txt should you meet an issue), then your mark will at most be 60%.