Lab 1: Tree-Based Classification (May 21—24, 2024)

Evaluation: 5% of the total grade

You will be working on a real world "Titanic" dataset in this Lab exercise. We will fetch the titanic (version 1) dataset from openml.org that contains a target feature ("survived") for each passenger along with a set of descriptive features. More details of this dataset are given here.

A word of caution: This is a real-world, somewhat "messy" dataset requiring data exploration and cleaning before you proceed to model development.

Use this dataset to build and evaluate decision classification models to predict whether or not the passengers survived the sinking of the Titanic. Your models will be based on "features" like passengers' gender, age, class, etc. You can also create/engineer new features.

Perform the following tasks:

1. Load and explore the dataset for data quality issues – check for missing values, outliers, feature correlations, etc.
2. Prepare the data for model development – data cleaning, feature selection, train-test-split, etc.
3. Train and evaluate a decision tree classifier to predict whether or not the passengers survived the tragic incident
4. Report and briefly discuss the performance of the model on an independent test set based on the following measures of performance:
    a. Classification report
    b. Confusion matrix
    c. Overfitting / underfitting
5. Examine the decision tree structure in terms of depth of the tree and number of nodes.
6. Briefly discuss the performance of each model on an independent test/validation set as well as a brief summary at the end comparing the performance of these models.
7. Perform tree pruning and retrain the model
8. Evaluate and discuss the performance of the retrained/pruned model

Submit your lab report along with the source code by May 24, 2024.

Lab Grading Criteria

Fraction / Category	0.8 – 1	0.4 – 0.7	0 – 0.3	Max Score
Model development workflow	Followed all steps in the modeling process	Followed most steps in the modeling process	No order in the modeling process	1
Tasks and observations	Performed all the tasks correctly, results properly presented and well explained	Performed most of the tasks correctly, results fairly presented and explained	Performed some of the tasks correctly, poorly presented results with little explanation	3
Code	Well organized, easy to follow and clean code with meaningful comments	Somewhat organized, a little hard to follow and fairly clean code with some comments	Not organized, very hard to follow and messy code with little or no comments	1
			Total	**5**