# CAP 5610: HW3: Decision Tree and Ensemble Learning

# **Question 1. (10 points) Understanding Training Error and Testing**

Consider the decision tree shown in the diagram below. The counts shown in the leaf nodes correspond to the number of training records associated with the nodes.



1. What is the training error rate for the tree? Explain how you get the answer?
2. Given a test instance T={A=0, B=1, C=1, D=1, E=0}, what class would the decision tree above assign to T? Explain how you get the answer?

# **Question 2 (16 points) Understand Splitting Process**

Consider the following data set for a binary class problem.

A close up of a logo

Description automatically generated

Q1: What is the overall gini before splitting?

Q2: What is the gain in gini after splitting on A?

Q3: What is the gain in gini after splitting on B:

Q4: Which attribute would the decision tree choose?

# **Question 3: (15 points) Please answer and explain.**

Q1: Are decision trees a linear classifier? Why?

Q2: What are the weaknesses of decision trees? Why?

Q3: Is Misclassification error better than Gini index as the splitting criteria for decision trees? Why?

# **Question-4 (35 points) Build decision tree and random forest using Scikit Learn （**[**https://scikit-learn.org/stable/**](https://scikit-learn.org/stable/)**）**

For the Titanic challenge (https://www.kaggle.com/c/titanic), we need to guess whether the individuals from the test dataset had survived or not. Please:

1. Preprocess your Titanic training data; Please briefly describe what preprocess you have done.
2. Select a set of important features. Please show your selected features and explain how you perform feature selection.
3. Learn and fine-tune a decision tree model with the Titanic training data, plot your decision tree;
4. Apply the five-fold cross validation of your fine-tuned decision tree learning model to the Titanic training data to extract average classification accuracy;
5. Apply the five-fold cross validation of your fine-tuned random forest learning model to the Titanic training data to extract average classification accuracy;
6. Which algorithm is better, Decision Tree or Random Forest?
7. What are your observations and conclusions from the algorithm comparison and analysis?

# **Question-5 (20 points) Build a bagging classifier using Scikit Learn for the above Titanic challenge.**

# **Question-6 (20 points) Build an Adaboost classifier using Scikit Learn for the above Titanic challenge.**

**Please write down your answers and submit a PDF report. In your report, please answer each question with your explanations, plots, results in brief. DO NOT paste your code or snapshot into the PDF. At the end of your PDF, please include a website address (e.g. Google colab) that can allow the TA to read your code (for Questions 4-6). Or you can submit your Python or Jupyter Notebook code together with the PDF report.**