Decision Tree and Ensemble Learning

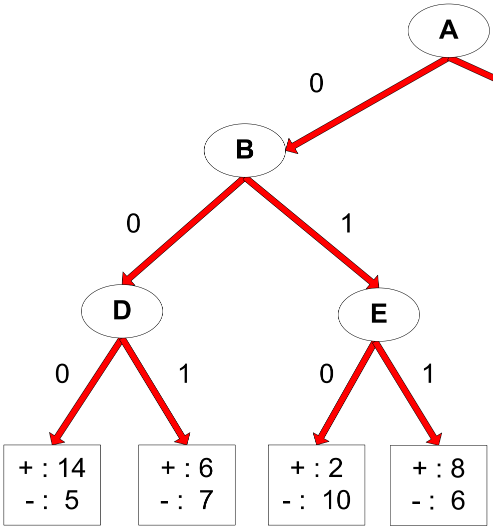
Task 1

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;
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?

Task 2

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.

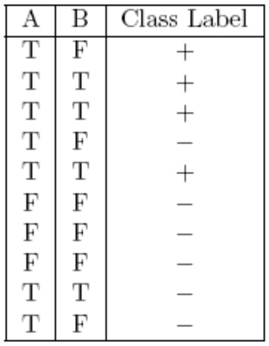


(a) What is the training error rate for the tree? Explain how you get the answer?

(b) 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?

Task 3

Consider the following data set for a binary class problem.



A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Q1: What is the overall entropy before splitting?

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

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

Q4: Which attribute would the decision tree choose?

Q5: Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations. (We want to split first on the variable which maximizes the information gain until there are no nodes with two class labels.)

Task 4: Please answer and explain.

Q1: Are decision trees a linear classifier?
Q2: What are the weaknesses of decision trees?
Q3: Is Misclassification errors better than Gini index as the splitting criteria for decision trees?

Please 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., Github, Dropbox, OneDrive, GoogleDrive) that can allow the TA to read your code.