Data Mining

Classification II

1 Decision Tree - DT

Select two datasets with labels available to perform the following experiments:

- Analyze and pre-process the data (if any).
- Divide the original dataset into two subsets: one for training (80%) and one for testing (20%).
- Build a DT for the training subset and test the built model for data from the testing subset. Note: Try the "tree" package from sklearn in Python or the function fitctree() in Matlab.
- Calculate the error of classification.

2 Random Forest - RF

Select two datasets with labels available to perform the following experiments:

- Analyze and pre-process the data (if any).
- Create K = 100 training set (using cross-validation or bagging technique), and build 1 testing set.
- Build an RF for each training set. Classify data from the testing set using RF and calculate the error of classification.
- Compare the results of RF and DT for the testing dataset. Comment.