Machine Learning
CS456- 2022

**Notes:**

**This Project to be solved in group of 3.**
**Students from different groups are ok if the taught by same TA.**
**Deadline for submission 17/12/2022.**

**Problem 1**

The data set available for this assignment is based on the U.S. congress voting record from 1984. The data set consists of the votes (*yes* or *no*) on sixteen issues for each of the 435 members of congress. from the voting record. You will use this data to learn a decision tree that predicts the political party of the representative based on his/her vote.

Dataset Link
https://github.com/mikeizbicki/datasets/blob/master/csv/uci/house-votes-84.data

Use the voting data to train a decision tree to predict political party (Democrat or Republican) based on the voting record.

• Measure the impact of training set size on the accuracy and the size of the learned tree. Consider training set sizes in the range (50-80%). Because of the high variance due to random splits repeat the experiment with five different random seeds for each training set size then report the mean, maximum and minimum accuracies at each training set size. Also measure the mean, max and min tree size.

- Start with training data size 50% , 60% .... Until you reach 80%.

- Rerun this experiment ***five times*** and notice the impact of different random splits of the data into training and test sets. Report the sizes and accuracies of these trees in each experiment.
- Turn in two plots showing how accuracy varies with training set size and how the number of nodes in the final tree varies with training set size.

- The data set contained many *missing values*, i.e., votes in which a member of congress failed to participate. To solve those issue insert—for each absent vote—the voting decision of the majority.

Machine Learning
CS456- 2022

**Problem 2**

- Implement your own simple KNN classifier using Python , (Don't use any build in functions)

- Use provided train and test file pendigits_test.txt, pendigits_training.txt
- Use follow link to download Dataset and understand it's description http://vlm1.uta.edu/~athitsos/courses/cse6363_spring2017/assignments/uci_datasets/
-

- Each record in dataset contain feature values are separated by commas, and the last value on each line is the class label

- If there is a tie in the class predicted by the *k*-nearest neighbors, then

  among the classes that have the same number of votes, the tie should

  be broken in favor of the class comes first in the Train file.
- Each dimension should be normalized, separately from all other dimensions.

  Specifically, for both training and test objects, each dimension should be transformed using function: F(v) = (v - mean) / std, using the mean and std of the values of that dimension on the TRAINING data.

- Use Euclidean distance to compute distances between instances.

- Report accuracy on testing data when k=1,2,3....9.

- As output, your programs should print the value of *k* used for the test

  set on the first line, each output line should list the predicted class

  label, and actual class label.

- Also output the number of correctly classified test instances, and the total number of instances in the test set &Accuracy.