Assignment 3

(10 points)

(1) Choose three arbitrary datasets; you can use scikit learn toy datasets as well (https://scikit-learn.org/stable/datasets/toy_dataset.html)

(2) Apply a prediction on these datasets, which all weak learners, including kNN, SVM, Naïve Bayes, and at least two decision trees (ID3, CART, C4.5, CHAID). Report the accuracy of each algorithm. The train test split should be 70/30, and you should also use 10 fold cross-validation.

(3) Augment two arbitrary datasets (from step one) and increase the number of their records. How to augment them is something you need to search and realize on your own. The augmentation result should include five times more data than the original dataset. In particular, you need to build five datasets as follows:

Dataset 1 = original dataset with no changes.
Dataset 2 = original dataset no changes + augmented dataset with equal number of records to the original one
Dataset 3 = original dataset no changes + augmented dataset with 2x number of records to the original one
Dataset 4 = original dataset no changes + augmented dataset with 3x number of records to the original one
Dataset 5 = original dataset no changes + augmented dataset with 4x number of records to the original one

(4) Measure and report the execution time and accuracy of applying XGBoost, CATBoost, and LightGBM on all five datasets for each sample. You must report them in a readable table and compare them in your explanations.

- If you have any trouble during your work, please contact the TA before the deadline. The TA will help you progress in your homework, but she will not provide you with the answer.

- Late homework delivery results in a penalty on your grade, every late day 10%.

- Sharing this assignment with a party outside this course is a Boston University code of conduct violation, and violators will be reported to the dean for further prosecution