Course ID: CSE-506
Course Name: Data Mining (DMG)

Assignment 4

November 13, 2023

Instructions

1. The assignment is to be attempted in groups.
2. Programming Language: Python
3. For Plagiarism, institute policy will be followed
4. You need to submit the readme.pdf, Code files, **PPT** and Images. No need to prepare a report.
5. You **can use any library** for pre-processing, training, doing experiments and post-processing in all questions.
6. One member should submit on google classroom while other members can mark turn in without the attachment.
7. In case of doubts, please comment on the classroom.
8. The data will have inconsistencies and outliers please handle them as per your understanding and mention them in the readme and ppt. Split dataset in 80-20 ratio while maintaining equal class distribution in both train and test set.

You have to work on the following three datasets:

Dataset1: Link Target class column: The biopsy results "Healthy" or "Cancer".
Dataset2: Link; Target class column: "fetal_health".
Dataset3: Banking dataset link; Target column: last column

Total Marks: **40**

Q1: (10 points) Train a decision tree classifier on the 3 datasets and report the precision, recall, F1, accuracy and AUC-ROC curve. You need to do this performing 5 fold cross validation. Provide visualizations for the decision boundaries accordingly in the ppt.

Q2: (10 points) Do the task in **Q1** by choosing Random Forest as the classifier.

Q3: (10 points) Do the task in **Q1** by choosing XGboost as the classifier.

Q4: (10 points) Do the task in **Q1** by choosing Adaboost as the classifier.