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

# Assignment 1

September 7, 2022

**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, Images and Model files.
5. Report, models and code in .py format should be submitted in the classroom in a zip folder with the name 'A1_RollNumber1_RollNumber2.zip'.
6. You **can use any library** for pre-processing, training, doing experiments and post-processing in all questions.
7. One member should submit on google classroom while other members can mark turn in without the attachment.
8. In case of doubts, please comment on the classroom.
9. The data will have inconsistencies and outliers please handle them as per your understanding and mention them in the readme. Split dataset in 80-20 ratio while maintaining equal class distribution in both train and test set.

**Extension and Penalty clause:** You can submit the assignment till t+3 day with penalty. Submitting within t+2 day will attract 10% penalty and submissions after t+2 and before t+3 will attract 20% penalty. Any submissions after t+3 will be not be considered. Even a 1 minute late submission on google classroom will be considered as late. Please turn-in your submissions atleast 5 minutes before the deadline.

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: 100
Deadline t: October 9, t+2 day: October 11 (10% penalty), t+3 day: October 12 (20% penalty)

### (A) Training (30 points)

**Q1: (40 points)** Train the Decision Tree classifier on the 3 datasets (all three) by using Logistic regression as the algorithm/function to split the node (Hint: Look into **Weight of Evidence**). You can adapt scikit-learn or Weka to implement this. Report the metrics precision, recall, accuracy and AUC-ROC curve. The most relevant paper in this topic can be found at LogitTree. Please read the paper in the initial week. (25 points)
Do this by choosing:

- One attribute for the split. (5 points)
- Pair of attributes at each node for the split. (5 points)

Q2: **(25 points)** Interpret the rules output from the decision tree. You can visualize the tree and the split criteria for this. Compare these rules to the rules output when you fit a normal decision tree from scikit-learn. Comparison can be a list of your observations from visualizing the splits. Do this for both single-attribute split and multi-attribute split models.

Q3: **(35 points)** Perform 5 fold cross-validation and report the performances (P,R,F1,accuracy). Also look into statistical tests (example: student-t tests) or other relevant tests using tools like Weka or scipy

and compare the single attribute (Hypothesis 1) and two attribute models (Hypothesis 2) and check which hypothesis is better.

Q4:

**Deliverables**

1. Detailed explanation of assumptions made for solving the mentioned problems.
2. Provide the various parameters asked in the each question like accuracy, comparisons or visualizations in readme.pdf file.
3. Your zip file should contain a folder ”visualizations”, ”code”, ppt, ”models” and a readme.pdf file.
4. The folder ”visualizations” should contain all the DT images and folder ”code” should contain all the codes including notebook if used.
5. You will have to upload your model for Part C Q1 in the model folder named as DT_C_1.pkl/zip/tar/xxx (any extension).
6. In the readme.pdf file, you will have to mention steps to recreate all your experiments and results.
7. Please provide references to all the sources including but not limited to libraries, GitHub Repositories, Research Articles, Blog Posts used in completing this assignment.