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

# Assignment 1

September 22, 2021

**Instructions**

1. The assignment is to be attempted in pairs.
2. Programming Language: Python
3. For Plagiarism, institute policy will be followed
4. You need to submit the readme.pdf, Code files, 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 member 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.
10. Use "data_1" to answer all questions, use "data_1" and "data_2" to answer Part C Q1. 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.

Dataset: Link; Target class column: "fetal_health".
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: (5 points)** Classify the dataset using Decision Tree classifier and report precision, recall, accuracy and AUC-ROC curve on the test set. Visualize the Decision Tree (DT) and save the visualization as image (image name: DT_A_1.png/pdf/jpg).

**Q2: (5 points)** Train the Decision Tree classifier on different depths (Minimum 5 different depths) of tree and plot the accuracy vs depth graph.

**Q3: (20 points)** Taking the Decision Tree formulated in Q1, vary the following hyperparameters:- Criterion, Splitter, min samples split, max depth, min samples leaf, max features (sqrt/log2), class weight and max leaf nodes and report the observations (precision, recall, accuracy and AUC-ROC curve) on each. You just have to change one hyperparameter in one experiment keeping others fixed and observe the performance. Minimum observations to take should be 8. Report your best intuition/reasoning behind positive/negative performance scores against the base model (Q1).
Keep the best obtained hyperparameters (DT referred as DT-A) from Q3 and carry on to Part B. Do not influence class weight in further experiments and set it as default ("none").

**(B) Post Pruning (40 points)**

**Q1: (5 points)** Remove a random node from the DT-A and observe the changes. Report your observations in terms of performance (precision, recall, AUC-ROC curve and accuracy) along with the tree diagram. Save the diagram as DT_B_1.png/pdf/jpg.

**Q2: (15 points)** Apply the Cost Complexity pruning technique and any other pruning technique of your choice on the DT obtained from part A (DT-A). Report the value of alpha for Cost Complexity Pruning and parameter values for the second pruning technique (ex. alpha, beta values if using Alpha-Beta pruning) and precision, recall and accuracy between the DT-A and the pruned trees. Visualize the DT-B-2-CC and DT-B-2-XX and save the image as DT_B_2_CC.png/pdf/jpg and DT_B_2_XX.png/pdf/jpg.

**Q3: (20 points)** Apply the Hybrid SLIQ pruning technique on DT-A and observe the performance difference. Compare the tree size, precision, recall and accuracy for DT-A, DT-B-2-CC, DT-B-2-XX and DT-B-3 and report your observations. Save the best performing tree as DT_B_3.png/pdf/jpg.

**(C) Experiments (30 points)**

**Q1: (15 points)** Train your decision tree on the training data provided "data_1" and obtain the classifier DT-C-1. You are provided with additional data "data_2" similar to the training data "data_1". You have to make DT-C-1 augment the new data and formulate DT-C-1-X which will be a new DT for both data_1 and data_2. You can not train the DT from scratch on both the datasets. Obtain the DT-C-1-X and compare its performance (precision, recall, accuracy and AUC-ROC curve) with the DT-C-1 on "data_1" test set and "data_2" test set (test set is the remaining 20% split from each data source) and report precision, recall, accuracy and AUC-ROC curve.

**Q2: (15 points)** Visualize the decision surface boundaries of the DT obtained from Part B Q3. Find the distance of a sample to the nearest decision boundary of decision tree. (DT doesnt have samples on the boundary so you will have to search for the region where the label changes. That will be considered as the decision boundary for this question).

**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", "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. Make a section "Learning's" in the readme.pdf and describe your learning's from this assignment.
7. Please provide references to all the sources used in completing this assignment.