CI6227 Data Mining – Assignment 1

In this assignment you will do a hands-on, practical familiarization with classification models. There are two variants of this assignment and you are free to choose either one. Both variants are graded in the same way on a 0–100 points scale. Assignment is individual. Use this dataset for the assignment: https://archive.ics.uci.edu/ml/datasets/Census+Income

Variant 1

In this variant, you will write your own implementation of any one classification algorithm of your choice. It can be a decision tree, a rule-based, a kNN, *etc.* – anything that was discussed in class. In this variant you are not allowed to use existing implementations of the algorithms in any form; you are supposed to fully implement it on your own. Apply your implemented algorithm to the dataset. Estimate performance of your model using cross-validation.

Reporting

Your submission for this assignment is a single PDF file with a report on assignment. Your report should be no longer than **two pages**. Somewhere at the top of the first page should be your matric number, full name, and a line "CI6227-2020-Assignment-1.1". The only requirement for report formatting is that it is readable, otherwise you are free to arrange information in any way you prefer.

Make sure that you provide performance and speed metrics for your final trained model. Explain all design decisions that you made along the way, e.g., did you do any data pre-processing, what was the similarity matric you chose, how you work with missing values, what is the stopping criterion, etc. Only mention the ones that are applicable to your classification model.

Variant 2

In this variant, you will compare two existing implementations of classifiers. You can choose any two existing implementations of classification models. Train and test them on the dataset linked in the beginning. Compare the two models using techniques for classification model comparison.

Reporting

Your submission for this assignment is a single PDF file with a report on assignment. Your report should be no longer than **two pages**. Somewhere at the top of the first page should be: your matric number, full name, and a line "CI6227-2020-Assignment-1.2". The only requirement for report formatting is that it is readable, otherwise you are free to arrange information in any way you prefer.

Make sure to provide full performance comparison for two models including time it took to train and apply the model. Explain all decisions you make along the way, e.g., model parameters. If you do any data pre-processing, please explain what and why was done.

Submission

Submission should be done in NTULearn. Access the assignment submission page through the left navigation bar by selecting "Assignments". Submit a **single PDF file**. Submissions are accepted up to **Thursday, September 30th 23:59:59**.