IN6227 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. The assignment is individual. The dataset used in this assignment is available at the following link: 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 training dataset. Estimate the performance of your model on the testing dataset.

Reporting

Your submission for this assignment is a single PDF file with a report on the 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 "IN6227-2023-Assignment-1.1".

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 metric 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.

Please **upload your source code to GitHub** and provide the repository link in the report.

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 provided 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 the 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 "IN6227-2023-Assignment-1.2".

Make sure to provide full performance comparison for the two models including the time it took to train and apply the model. Explain all decisions you make along the way, e.g., how you fine-tune model hyper-parameters, how you work with missing values, what is the stopping criterion, etc. If you do any data pre-processing, please explain what and why was done.

Please **upload your source code to GitHub** and provide the repository link in the report.

Grading

The assignment will be graded based on the two variants:

- **Variant 1**: The coding component will account for 80% of your grade, while the report will contribute 20%.
- **Variant 2**: The coding component will account for 20% of your grade, with the report making up the remaining 80%.

Please choose the variant that best aligns with your strengths and interests.

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 Wednesday, **2024-Oct-09, 23:59:59**.