Project II

Programming for Data Analysis

This document contains the Project II instructions for Programming for Data Analysis. Please be advised that all students are bound by the Quality Assurance Framework [1] at ATU which includes the Code of Student Conduct and the Policy on Plagiarism. The onus is on the student to ensure they do not, even inadvertently, break the rules. A clean and comprehensive git history (see below) is the best way to demonstrate to the examiner that your submission is your own work. It is, however, expected that you draw on works that are not your own to build your submission and you should systematically reference those works to enhance your submission.

Problem Statement

This project will investigate the Wisconsin Breast Cancer dataset. The following list presents the requirements of the project

- Undertake an analysis/review of the dataset and present an overview and background.
- Provide a literature review on classifiers which have been applied to the dataset and compare their performance
- Present a statistical analysis of the dataset
- Using a range of machine learning algorithms, train a set of classifiers on the dataset (using SKLearn etc.) and present classification performance results. Detail your rationale for the parameter selections you made while training the classifiers.
- Compare, contrast and critique your results with reference to the literature
- Discuss and investigate how the dataset could be extended – using data synthesis of new tumour datapoints
- Document your work in a Jupyter notebook.
- As a suggestion, you could use Pandas, Seaborn, SKLearn, etc. to perform your analysis.
- Please use GitHub to demonstrate research, progress and consistency.

References

[1] ATU. Quality assurance framework. https://www.atu.ie/sites/default/files/2022-08/Student%20Code_Final_ August_2022.pdf.