CS 4038 – Data Mining

Spring 2024 – Course Project

Team size: 3-4 students	**Total marks:** 15 absolutes (Phase-I=5, Phase-II=10)

Dataset: The dataset contains students' assessment scores including <Assignments, Quizzes, Mid-I, Mid-II>, and a predictor variable <Grade>. The data has been anonymized to hide identities of the students and course(s). The data is shared on seven sheets (D1 to D7), where each sheet contains a different number of assignments and quizzes. However, only the best 5 assignments and quizzes are included for each student before calculating their grades. Also note that total marks for assignments and quizzes are given on the top along their corresponding weights.

Problem: To predict students' grade as "pass" or "fail" before: (a) Mid-II, and, (b) Final exams. For Mid-II grade prediction, use the following features: first four assignments, first four quizzes and Mid-I scores; and, for grade prediction before final exam, use all the features (take best 5 assignments and quizzes).

Objectives: To answer the following two research questions.

- RQ-1: How accurately can we predict students' grades before the Mid-II exam?
- RQ-2: How accurately can we predict students' grades before Final exam?

Project Phase-I

Total marks	05 absolutes
Objective	Perform exploratory data analysis (EDA) of the given dataset for understanding and preprocessing the data that might help you in the second phase of the project.
Methods	Topics covered in the course
Languages/Tools	Python, R, Weka, Orange (as per the team's preference)
Evaluation mode	Code/Tool + PDF Report [3-4 pages]
Deadline	**April 28, 2023**

Project Phase-II

Total marks	10 absolutes
Objective	Model training and results reporting using the three classifiers (Nearest neighbor, Decision tree)
Methods	Topics covered in the course (only use the two classifiers taught in the course)
Languages/Tools	Python, R, Weka, Orange (as per the team's preference)

Evaluation mode	Oral viva + Code/Tool presentation + PDF Report
Deadline	**May 10, 2023**

Expected outputs:

Phase-I	• Report data analyses that you performed through charts/tables. • Report issue(s) that you identified and the corrective measure taken in pre-processing phase. • Paste screenshots if a tool is used, submit the code otherwise.
Phase-II	• Details of data preprocessing steps (if performed). • Model's baseline accuracy. • Results reporting: confusion matrix, performance evaluation metrics (accuracy, sensitivity, specificity, etc.), tables and/or charts.

Evaluation criterion: See details above for each phase.

Additionally, to score above 50% marks in any phase, there has to be unique elements in a team's work. That is, if the same analysis or methodology is followed by another team then all teams will get 50% marks at maximum. **There shall be no collaboration/discussion about the project across teams of both sections**. Teams can only discuss their ideas/methodology with CS4038 teaching staff (course instructor and teaching assistant).

Reference material:

- For EDA, you may follow this Python notebook: https://www.kaggle.com/code/spscientist/student-performance-in-exams/notebook
- For R language: https://cran.r-project.org/web/packages/dlookr/vignettes/EDA.html
- To learn Weka (video lectures are available at): https://www.youtube.com/user/WekaMOOC
- To learn Orange (video lectures are available at): https://www.youtube.com/channel/UClKKWBe2SCAEyv7ZNGhIe4g