Title: Student Programming Performance Prediction using Competency-based Clustering

Proposer: Vincent Zhang

Supervisor: A/Prof. Irena Koprinska and Dr Bryn Jefferies

Background: Predicting student achievement in programming using competency clusters

Educational data mining (EDM) presents a significant opportunity to gain deeper and more personalised insights into student learning patterns and behaviours. In particular, applying EDM techniques to data from online coding platforms such as Grok Academy allows researchers to draw inferences about student programming behaviours and predict a student's level of achievement in programming. These inferences can be used to positively reinforce advantageous programming behaviours and practices for students, identify negative behaviours that can detriment programming performance, and inform interventions for struggling students.

Problem Statement:

To determine how specific programming skills and competencies affect student achievement and performance in programming.

Method:

There are five key stages in this project:

1. **Building skills-based learning profiles:** Learning profiles focusing on specific programming skills and competencies will be created for individual students. These profiles will use features related to individual submissions (e.g. complexity and quality of submitted code, activity time, number of test cases passed, type of test cases passed) and student interactions with course content (e.g. time spent viewing course material).

2. **Measuring problem difficulty:** Difficulty levels will be estimated for each coding problem, based on several features such as the number of students who passed all test cases, the total time spent on each problem, and the total number of submissions. The results of this analysis will be compared with the level of difficulty of the problem as indicated on the platform.

3. **Clustering students based on learning profiles:** Clustering analysis will be performed on students using their learning profiles to identify different competency clusters of students. The membership features of each cluster, such as the underlying skills, competencies and levels of achievement of students, and the performance of students within each cluster on different types of questions, will be examined.

4. **Time-based analysis:** The movement of students between competency clusters will be analysed, and time-based trends indicating improvement or decline in performance for individual students and competency clusters across time will be measured. Results from the time-based analysis will be synthesised with analysis done for problem difficulty to determine the effect of problem difficulty on student performance.

5. **Student performance prediction:** The student learning profiles, and the competency clusters will be combined to form the feature set for the student performance regressor, which will predict the student's overall programming performance

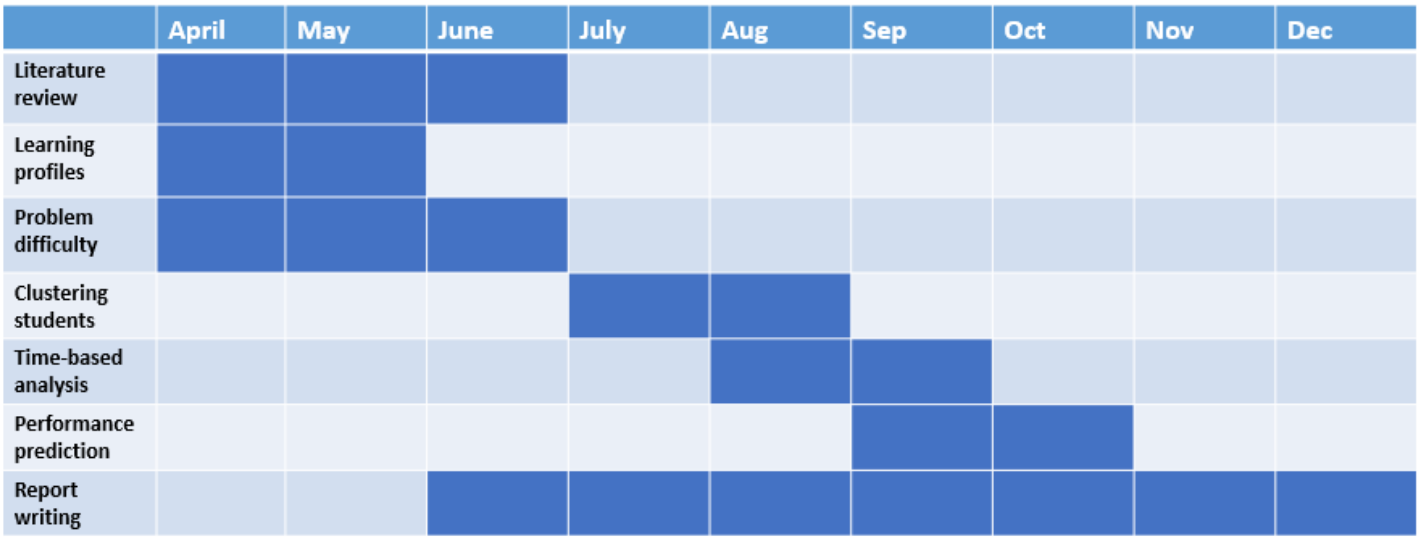

Fig 1: Time schedule for the thesis