Data Driven Teaching Strategies for CS1 at TU Wien

SS24, 194.147 @ TU Wien

Dr. Stefan Podlipnig, Senior Lecturer (main-supervisor)
Dr. Martin Riener, Senior Lecturer (co-supervisor)

Yahya Jabary (11912007)

Abstract

This interdisciplinary project aims to apply educational data mining techniques to analyze student performance in the CS1 course (185.A91 Introduction to Programming 1) at TU Wien. By examining student data, we seek to develop insights that can inform data-driven teaching strategies. This project bridges higher education didactics – particularly computer science education – and data science, to improve student outcomes and retention in introductory programming. Our approach leverages the TUWEL API to collect and analyze student data, focusing on developing predictive models to identify at-risk students and provide actionable insights for course instructors. The project will follow the CRISP-DM methodology, focusing on feature significance and actionable insights. The expected outcomes include a comprehensive set of visualizations and statistical analyses that provide insights into student performance in CS1, as well as actionable recommendations for teaching strategies and course design.

Motivation

Introductory programming courses like CS1 often have high failure and dropout rates, with worldwide pass rates averaging around 67% [1]. At TU Wien, the computer science program is particularly large, with 5,076 students enrolled in the winter semester of 2022 / 23 [2]. Given this substantial student population, data-driven approaches to teaching and course management are crucial to improve outcomes. Educational data mining offers promising approaches to leverage the rich data generated in programming courses to inform instructional strategies and course design.

This project aims to answer the following key research questions:

- **RQ1:** What student attributes are most associated with performance in CS1? (feature significance)
- **RQ2:** What insights can be derived to inform teaching strategies and course design? (actionable insights)
- **RQ3:** How can we use data to predict and support at-risk students in CS1? (predictive modeling)

In our context student performance is defined as a binary variable, where students either pass or fail the course.

Predictive modeling would require historical (time-series) data on student performance and engagement, which may be available from previous semesters. This data could then be used to train learning models such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) to predict student performance and identify at-risk students.

Given the exploratory nature of this project, we will start by focusing on feature significance and actionable insights, with predictive modeling as a stretch goal depending on data availability and feasibility. This will allow us to provide immediate value to CS1 instructors.

Methodology

The project will adopt the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. In the business understanding phase, collaboration with CS1 instructors will help define key performance indicators and areas of interest. Data will be collected from multiple sources, including the learning management system, automated assessment tools, and student information systems (such as TUWEL and TISS). Data preparation will involve cleaning the data, handling missing values, and engineering relevant features that capture student engagement and performance.

In the analysis phase, basic statistics, data visualization, and exploratory data analysis techniques will be employed to uncover patterns and relationships in the data. Correlation analyses will identify factors most strongly

[1] Watson, C., & Li, F. W. (2014, June). Failure rates in introductory programming revisited. In Proceedings of the 2014 conference on Innovation & technology in computer science education (pp. 39-44).

[2] TU Wien (2022). Facts and Figures / TU Wien in numbers. https://www.tuwien.at/en/tu-wien/about-tu-wien/facts-and-figures (accessed 2024-07-10)

associated with student performance. Machine learning models, such as logistic regression, decision trees, and random forests, will be used to classify students as failing or passing CS1 based on their data. This will help identify key factors that predict student success and inform teaching strategies. For evaluation, the relevance and actionability of findings will be assessed through discussions with CS1 instructors, focusing on generating insights that can be directly applied to improve teaching strategies and course design.

In the deployment phase, visualizations and reports will be developed to effectively communicate findings to CS1 instructors. Collaboration with instructors will be essential to interpret results and discuss potential applications to teaching practice.

Additionally, ethical considerations and data privacy regulations will be addressed throughout the project to ensure responsible use of student data. Finally, on the operations side, we want to ensure the reproducibility of our results, by providing a Docker container, Conda environment, or similar setup that can ensure reproducibility of our results with compiled pip requirements.

Expected Results

We aim to achieve the following key outcomes:

1. Development of a data pipeline that integrates data from multiple sources and prepares it for analysis.
2. Creation of comprehensive visualizations that provide insights into student performance in CS1.
3. Identification of key factors associated with student success in CS1.
4. Provision of actionable insights to instructors on how to improve teaching strategies and course design based on data analysis.
5. Compilation of a report summarizing the project findings and their potential applications to CS1 teaching.

Given the exploratory nature of this project and a budget of 125 hours / 3 weeks of full-time work (5 ECTS), the depth and breadth of analyses will depend on data availability and quality. Unlike the current approach, which relies primarily on instructor intuition and experience, this project aims to provide data-driven insights to complement and enhance teaching practices in CS1. The effectiveness of any changes implemented based on these insights will need to be evaluated in future work. Ultimately, the project aims to provide a framework for future research in this area, including the necessary data infrastructure and analysis tools.

Domain-specific Lectures

In addition to the 5 ECTS for the project, I must take 3 ECTS worth of domain-specific lectures to deepen my understanding of the subject matter. Prof. Emanuel Sallinger has kindly approved extending the list of electives in 194.147 to include the following courses:

- 194.042 Informatics didactics (3 ECTS)
- 184.228 Didactics in computer science education (3 ECTS)

These courses will provide me with the necessary background in educational theory and computer science education to understand the context of the CS1 course and the challenges faced by students. I plan to take one of these courses in parallel to the project to deepen my understanding of the domain and apply the knowledge gained directly to the project.