TU Wien Informatics

Data Driven Teaching Strategies for CS1 at TU Wien

Yahya Jabary - Data Science

TU Wien Informatics
Theory and Logic, E192-05
Supervisor: Dr. Stefan Podlipnig
Co-Supervisor: Dr. Martin Riener
Contact: stefan.podlipnig@tuwien.ac.at

Abstract

This study applies educational data mining techniques to identify key factors associated with student success in the introductory programming course "Introduction to Programming 1" at TU Wien. Using the TUWEL API, we extracted and analyzed student data to provide actionable insights for course instructors. Our analysis revealed that the number of years a student has been enrolled at TU Wien is the most significant predictor of course performance, followed by course load and student gender. A binary classifier achieved 68% accuracy in predicting student outcomes using only inferred attributes.

Methodology

- Data source: TUWEL API (Moodle fork with custom endpoints)
- Data extraction: Asynchronous Python script using asyncio and aiohttp
- Features extracted:
  - Passed (target variable)
  - Gender (inferred using pre-trained transformer model)
  - Years enrolled (derived from matriculation number)
  - Current courses (course load)
  - Points from quizzes, assignments, checkmarks
- Preprocessing: Standardization of numeric features, one-hot encoding of categorical features
- Models: Traditional Binary Classifiers such as Decision Trees.

Key Findings

Conclusion:

- The number of years a student has been enrolled is the most significant predictor of course performance
- There's a need for more comprehensive and structured data collection methods in educational settings

Best inferred student success predictors:

- **Years enrolled** previous to EP1 (88.4% importance)
- Course load in parallel to EP1 (9.3% importance)
- Predicted student gender (2.3% importance)

Model performance:

- **Accuracy: 68%**
- Precision: 70%
- Recall: 92%

Limitations

The project faced significant limitations due to data quality issues, lack of historical information, and time constraints, with 75% of the time spent on data extraction and preprocessing from a partially malfunctioning API with inadequate documentation. Future work could focus on manual data validation, implementing time-series analysis and incorporating expert knowledge, but as the saying goes, "garbage in, garbage out" – the quality of insights is ultimately limited by the quality of available data.