Machine Learning for Higher Education: Analyzing and Forecasting Student Enrollment and Degrees Awarded

Thomas Simmons

Austin Peay State University

Department of Computer Science and Information Technology

Dr. Jiang Li

February 4, 2025

## Introduction

Higher education institutions rely heavily on enrollment and degrees awarded to maintain and increase funding, whether they are funded through government or tuition. There are many socioeconomic, demographic, and situational factors that fall outside a higher ed institution and many of these have already been studied, however more research can be done to study how the institution operates internally. This paper seeks to study key performance indicators such as student to faculty ratios, credit hours to faculty ratios, degrees awarded, and student dropout rates.

Accurate predictions for enrollment and the awarding of degrees are of high interest to higher education management as they grapple to maintain high enrollment, especially in a post-COVID environment. The primary challenge in college enrollment and degree analysis is producing accurate predictions while considering multiple factors. Traditional analysis methods use historical trends and statistical modeling, therefore failing to capture complex, nonlinear patterns between enrollment or degrees awarded and higher education parameters. Without accurate predictions, institutions may face challenges such as underutilized resources, budgetary inefficiencies, and ineffective strategies.

Advancements in Machine Learning and Data Mining in the last decade, and particularly in the last few years, provide strong analysis tools to detect these complex, nonlinear patterns that may have been missed before using traditional techniques. A newer data-driven approach enables higher education institutions to anticipate enrollment fluctuations, implement targeted interventions, and enhance decision-making processes.

This project will attempt to accomplish the following target goals:

1. Collect meaningful higher educational data and utilize necessary data preprocessing and cleaning techniques.
2. Conduct Exploratory Data Analysis to identify and loud trends or correlations.
3. Implement Feature Selection, Scaling, and other Data Engineering techniques to further clean the data and construct meaningful relationships and features.
4. Develop several different Machine Learning models such as Regression, Neural Networks, and Decisions Trees.
5. Interpret the models and predictions to determine the best model for use in deriving actionable insights.
6. Design a predictive framework to assist in making data-driven decisions using the predictions from the model.

Ultimately, by leveraging machine learning, this project aims to enhance the accuracy of predictions, providing educational institutions with valuable insights to improve planning, recruitment, and retention strategies.

## Theoretical Framework & Literature Review

There are not many papers currently that provide analysis on higher education institutional policy and decision-making using machine learning. Current papers on the subject that do use machine learning focus heavily on external factors outside a university’s control such as a student’s socioeconomic status, race, and previous test scores. [2-4] This project seeks to contribute to the extensive Educational Data Mining community by filling gaps in where machine learning has not been utilized.

Concepts and theories that provide this project with a theoretical framework and strong foundation include Predictive Analytics Theory, Machine Learning Theories, the Educational Data Mining subfield, and decision theory. By building off these heavily researched and tested theories, this project will have a strong launching off point and require less assumptions.

## System Design

This project will use Google Collab as the Integrated Development Environment. The primary language will be python so that powerful machine learning and statistical libraries such as Scikit-learn, Keras, TensorFlow, NumPy, and Pandas can be leveraged. A linear regression model will be trained on the data and then used to predict future results and gain insights into the weights. Results are based off performance metrics such as RMSE, and R-Squared.

## Procedures of Data Collection & Analysis

The scope of data collected has not yet been determined. For now, the current datasets that have been chosen include Enrollment Statistics from Tennessee Higher Education Institutions such as Ausitn Peay State University, University of Tennessee Knoxville, University of Memphis, and Middle Tennessee State University. Scope can potentially be increased to all Schools in Tennessee or The United States of America by using datasets from other Tennessee Higher Education Institutions or the National Center for Education Statistics and the Education Data Initiative respectively.

As discussed previously, this project will rely heavily on Machine Learning techniques to do much of the data analysis. Fine-tuning of the model may include hyperparameter tuning, cross-validation, Feature Scaling and Selection, and Regularization. Other techniques that may be used include Association Rule Mining, Exploratory Data Analysis, Support Vector Machines, K-means Clustering, Monte Carlo Simulations, and various Statistic Metrics. Performance will be analyzed based off RMSE, R-Squared, and classification metrics such as a Confusion Matrix or Area under Receiver Operating Characteristics Curve where applicable.

## Timeline

* Weeks 1-2: Explore datasets and decide on Project to Propose
* Week 3: Finish Project Proposal and Meet with Dr. Li.
* Weeks 4-5: Literature Review.
* Weeks 6-7: Collect Data and Clean as needed.
* Week 8: Exploratory Data Analysis and Feature Selection.
* Weeks 9-12: Model Developments and fine-tuning.
* Week 13: Develop interpretation framework for results.
* Week 14-15: Writing the final project paper and gathering all development documents and artifacts.
* Week 16: Finish and turn in Project

References

1. <https://www.sciencedirect.com/science/article/abs/pii/S027277570100036X#preview-section-cited-by>
2. <https://www.emerald.com/insight/content/doi/10.1108/k-12-2020-0865/full/html>
3. <https://link.springer.com/chapter/10.1007/978-3-031-65522-7_5>
4. <https://www.tandfonline.com/doi/pdf/10.1080/00221546.1978.11780411>
5. <https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide>
6. <https://irsa.utk.edu/reporting/fact-book/>
7. <https://sas.utk.edu/SASVisualAnalytics/?reportUri=%2Freports%2Freports%2F3cf1e321-2ca0-4f4e-b0eb-d3e7948c3117&sectionIndex=0&sso_guest=true&sas-welcome=false>
8. <https://nces.ed.gov/surveys/SurveyGroups.asp?group=2>
9. <https://educationdata.org/college-enrollment-statistics#:~:text=College%20Enrollment%20%26%20Student%20Demographic%20Statistics&text=Report%20Highlights.,0.4%25%20from%202021%20to%202022>.