Title

Predictive Modeling of Student Academic Performance Using Multivariate Statistical Techniques and Machine Learning Algorithms

Authors

Arthur Nduwayo *(Corresponding Author)* Department of Actuarial Science and Finance, University of Burundi, Bubanza, Burundi Email: nduwayoarthur@gmail.com

Medard Bivuzeyezu Department of Actuarial Science and Risk Management, University of Moulay Ismaël, Meknes, 50000, Morocco Email: bivuzemeda1998@gmail.com

Abstract

Introduction. Student attrition and academic underperformance remain critical challenges in higher education institutions worldwide. Early identification of at-risk learners enables timely interventions that can significantly improve student outcomes. This study compares traditional multivariate statistical techniques with modern machine learning algorithms to predict student academic performance and identify learners requiring support.

Methods. We collected comprehensive data from 525 undergraduate students over four academic years (2020-2024) through administrative records, online questionnaires (87.3% response rate), and institutional databases using stratified random sampling. The dataset comprised 18 predictor variables including prior academic achievement, behavioral factors (attendance rate, weekly study hours), socio-demographic characteristics, and engagement indicators. We systematically compared six predictive models: Multiple Linear Regression, Classification and Regression Trees, Random Forests, Gradient Boosting Machines, XGBoost, and Support Vector Regression. Model performance was evaluated using 5-fold cross-validation with R^2 and RMSE metrics.

Results. Feature importance analysis identified five dominant predictors: previous academic performance (34.2%), class attendance rate (21.8%), weekly study hours (15.6%), entrance examination scores (12.4%), and parental education level (8.9%). Ensemble machine learning methods substantially outperformed traditional approaches.

XGBoost achieved the highest performance (R^2=0.761, RMSE=0.378), representing 12-17% improvement over Multiple Linear Regression (R^2=0.634, RMSE=0.487). The optimal model identified at-risk students with 84.3% accuracy, segmenting learners into four risk categories: high-achievers (23%), on-track (42%), at-risk (28%), and critical-risk (7%).

Conclusions. This research demonstrates that advanced machine learning algorithms significantly outperform traditional statistical methods in predicting student academic performance, explaining over 76% of variance. A web-based dashboard implementing the XGBoost model enables real-time risk assessment and automated alerts for academic advisors. These findings illustrate the transformative potential of data-driven approaches for evidence-based early intervention strategies and educational planning initiatives.