Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Predictive Models for Forecasting Learner Achievement: A Data Mining and Machine Learning Approach

Ean Teng KHOR* **& David NG**
Nanyang Technological University, Singapore
***eanteng.khor@nie.edu.sg**

Abstract: The study aims to identify key factors influencing learner achievement and develop an early detection model for at-risk students. After the process of exploratory data analysis and feature engineering, predictive models using three machine learning algorithms: logistic regression (LR), decision tree (DT), and support vector machine (SVM) were developed and evaluated. Results show that SVM outperformed the others across all performance metrics. SEMESTER_GPA, PROGNAME, and INTAKESEM emerged as the most significant predictors.

Keywords: Educational data mining, machine learning, predictive analytics, models, learners' achievement

1. Introduction

Predictive analytics leverages data mining and machine learning to build models that forecast student outcomes. For example, such models have predicted academic performance using data from intelligent tutoring systems, log files, and MOODLE usage. Despite numerous interventions, early identification of at-risk students remains a key challenge (Wang, 2021). Timely detection is essential, as delays can lead to disengagement and poor performance. Effective early warning systems can enhance both individual outcomes and overall educational quality (Jokhan et al., 2019). While predictive analytics is growing globally, its use in Singapore's teacher education remains limited. Existing models often fail to transfer due to differences in grading, culture, and student profiles (Alamri & Alharbi, 2021). Hence, the study aims to develop a predictive model tailored to the Singapore context using data from the National Institute of Education (NIE), the nation's premier teacher-training institute. The model's feature selection incorporates Singapore-specific variables such as INTAKESEM and PROGNAME, enhancing its local relevance and effectiveness in early detection.

2. Research Methods

This study uses anonymized learner records from 16 NIE programmes from 2007 to 2020. Data was processed in Python and analyzed the following 4 stages adapted from Khor (2022): (1) exploratory data analysis, (2) data pre-processing and feature engineering, (3) model development, and (4) model evaluation. Learners were classified into 3 achievement groups based on CGPA: Class 0 (≤ 3.5), Class 1 (3.6–4.0), and Class 2 (> 4.0). Pre-processing involved removing noisy data, imputing missing values with column means, and encoding categorical variables. Feature selection excluded redundant attributes and CGPA (to prevent data leakage), retaining features with absolute correlation > 0.05. LR, DT, and SVM machine learning algorithms were selected for their effective balance of accuracy and interpretability, fitting Singapore's data-driven education context. An 80-20 train-test split was applied, and model performance was evaluated using a confusion matrix, with metrics including accuracy, precision, recall, and F-measure.

3. Analysis and Results

SEMESTER_GPA, PROGNAME, and INTAKESEM emerged as key predictive features. SVM outperformed other machine learning algorithms across all performance metrics, achieving 75% classification accuracy (Table 1). The performance of the models was evaluated by comparing their predictions against the actual CGPA classifications or grades, to determine their accuracy. These scores are shown in Table 2.

Table 1. *Classification Analysis Result*

Classifier	Accuracy	Precision	Recall	F-measure
LR	0.69	0.70	0.70	0.70
DT	0.72	0.72	0.73	0.73
SVM	0.75	0.78	0.73	0.75

Table 2. *Actual CGPA Correlation Score*

Actual CGPA category	Probability of Category 0 Prediction	Probability of Category 1 Prediction	Probability of Category 2 Prediction
0	0.56	0.08	0.36
1	0.29	0.18	0.52
2	0.14	0.45	0.41
Grand Total	**0.32**	**0.23**	**0.45**

4. Conclusions

This study developed predictive models using LR, DT, and SVM machine learning algorithms. Among them, SVM achieved the highest accuracy at 0.75. SEMESTER_GPA emerged as the most predictive feature. The model can support early identification of at-risk students, enabling timely, tailored interventions from instructors and program offices to prevent academic decline. However, since the model is based on data from pre-service teachers in higher education, its generalizability to other educational levels in Singapore is limited. Future research could explore predictive models for different student groups and deploy ensemble techniques such as soft voting classifiers to enhance performance.

References

Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: a systematic review. *IEEE Access*, *9*, 33132-33143. https://doi.org/10.1109/ACCESS.2021.3061368

Jokhan, A., Sharma, B., & Singh, S. (2019). Early warning system as a predictor for student performance in higher education blended courses. *Studies in Higher Education, 44*(11), 1900–1911. https://doi.org/10.1080/03075079.2018.1466872

Khor, E. T. (2022). A data mining approach using machine learning algorithms for early detection of low-performing students. *The International Journal of Information and Learning Technology*, *39*(2), 122-132. https://doi.org/10.1108/IJILT-09-2021-0144

Wang, L. Y. (2021). Levelling up academically low-performing students in student-centric education in Singapore: global trend, local policies and future directions. *Educational Review, 73*(3), 374–390. https://doi.org/10.1080/00131911.2019.1642304