International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Early Prediction of Student Academic Performance Using Machine Learning Algorithms for Educational Intervention

Nevedha K
Assistant Professor, Department of Information Technology, Velalar College of Engineering and Technology, Erode, Tamil Nadu

Abstract
Early identification of academically at-risk students is essential for improving learning outcomes and enabling timely educational intervention. Traditional evaluation methods detect weak students only after examinations, which limits the ability of instructors to provide corrective support. This study proposes a machine learning based prediction model to identify student academic performance at an early stage using behavioural and academic activity data.
The xAPI-Edu-Data dataset was used for experimentation. Several supervised machine learning algorithms including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors and Support Vector Machine were implemented and compared. The dataset was preprocessed using label encoding and divided into training and testing subsets. Model performance was evaluated using accuracy and classification metrics.
Experimental results show that the Random Forest classifier outperformed other models with an accuracy of 86.45%, followed by Logistic Regression (79.16%), Decision Tree (77.08%), Support Vector Machine (64.58%) and K-Nearest Neighbors (63.54%). The results demonstrate that ensemble learning methods provide better prediction capability for educational datasets.
The proposed model can help institutions detect low-performing students early and enable instructors to provide targeted academic support, thereby improving overall student success rates.

Keywords: Machine Learning, Student Performance Prediction, Educational Data Mining, Random Forest, Learning Analytics.

INTRODUCTION

Education analytics has become an important research area due to the increasing availability of student behavioral and academic data in digital learning environments. Educational institutions continuously seek methods to improve student performance and reduce failure rates. One of the major challenges faced by educators is identifying academically weak students at an early stage so that appropriate interventions can be provided.
Traditional student evaluation methods rely primarily on examinations and periodic assessments. These approaches detect poor performance only after the learning process has progressed significantly, leaving limited opportunity for corrective measures. Early prediction of student performance can help instructors provide personalized guidance, improve retention rate and enhance overall academic success.

Machine learning techniques have recently been widely applied in educational data mining to analyze student learning patterns and predict academic outcomes. By analyzing attributes such as attendance, participation in discussions, resource usage and parental involvement, predictive models can identify performance trends before final examinations occur.

In this study, multiple supervised machine learning algorithms are implemented and compared to predict student academic performance using the xAPI-Edu-Data dataset. The objective of this research is to evaluate the effectiveness of different classifiers and determine the most accurate model for early performance prediction. The performance of Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors and Support Vector Machine classifiers are analyzed.

The main contribution of this work is the identification of an effective predictive model that can assist educational institutions in detecting at-risk students and enabling timely academic intervention.

LITERATURE REVIEW

Educational Data Mining (EDM) and Learning Analytics have been widely used to analyze student behaviour and predict academic outcomes. Romero and Ventura (2010) discussed the application of data mining techniques in educational environments and highlighted the importance of predictive models in improving learning performance. Their study emphasized that machine learning algorithms can discover hidden patterns in student interaction data.

Kotsiantis et al. (2013) applied classification algorithms such as Decision Tree and Naïve Bayes to predict student academic success. The study reported that classification models can effectively categorize students into performance groups, enabling instructors to identify weak learners. However, the accuracy varied significantly depending on dataset characteristics.

Ayesha et al. (2010) implemented clustering and classification techniques for predicting student performance and found that attendance and classroom participation were strong predictors of academic achievement. Their research demonstrated that behavioural attributes play a crucial role in performance prediction.

Amrieh, Hamtini and Aljarah (2016) introduced the xAPI-Edu-Data dataset and applied machine learning algorithms including Support Vector Machine and Decision Tree classifiers. Their results showed that educational activity features such as raised hands, resource visits and discussion participation significantly affect prediction accuracy.

More recently, ensemble learning methods such as Random Forest have been widely used due to their robustness and high accuracy. Studies indicate that ensemble classifiers outperform individual models in handling complex educational datasets and reducing overfitting problems.

Although previous research has applied various machine learning techniques, identifying the most suitable algorithm for early intervention remains an open challenge. Therefore, this study performs a comparative analysis of multiple supervised learning models to determine the most effective classifier for student performance prediction.

METHODOLOGY

A. Dataset Description

The study uses the xAPI-Edu-Data dataset which contains student academic and behavioural attributes collected from a learning management system. The dataset includes demographic information, academic background and interaction-based features such as raised hands, visited resources, discussion

participation, announcements viewed and parental involvement. The target variable represents student performance categorized into three classes: Low (L), Medium (M) and High (H).

B. Data Preprocessing

The dataset contains categorical attributes that cannot be directly processed by machine learning algorithms. Therefore, label encoding was applied to convert categorical values into numerical form. Missing values were not present in the dataset. After preprocessing, the dataset was divided into feature set (X) and target variable (y). The data was then split into training and testing subsets using an 80:20 ratio.

C. Classification Algorithms

To evaluate predictive performance, five supervised machine learning algorithms were implemented:

1. Logistic Regression: A statistical classification model used for predicting categorical outcomes based on probability estimation.
2. Decision Tree: A tree-structured model that splits data based on feature conditions to classify student performance.
3. Random Forest: An ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.
4. K-Nearest Neighbors (KNN): A distance-based classifier that assigns class labels based on the majority class among the nearest neighbors.
5. Support Vector Machine (SVM): A classifier that identifies the optimal hyperplane separating different classes in high-dimensional space.

D. Performance Evaluation

Model performance was evaluated using classification accuracy and confusion matrix analysis. Each algorithm was trained on the training dataset and tested on unseen data. The algorithm achieving the highest accuracy was considered the most effective model for early student performance prediction.

RESULTS AND DISCUSSION

The performance of the implemented machine learning algorithms was evaluated using classification accuracy on the testing dataset. The confusion matrix of the best performing model is shown in Fig. 1.

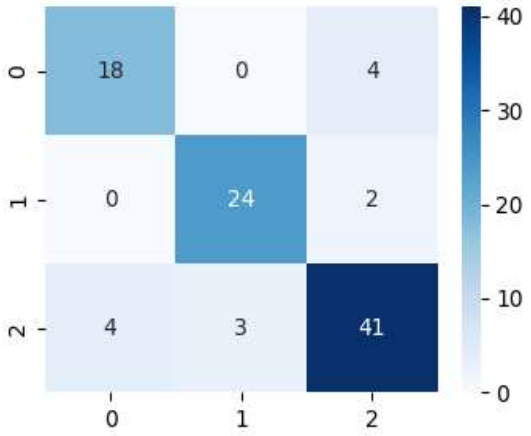

Fig. 1. Confusion matrix of Random Forest classifier

Table I and Fig. 2 present the comparative performance of all classifiers.

Table 1. Accuracy Comparison Of Classifiers

Algorithm	Accuracy (%)
Logistic Regression	79.16
Decision Tree	77.08
Random Forest	86.45
K-Nearest Neighbors	63.54
Support Vector Machine	64.58

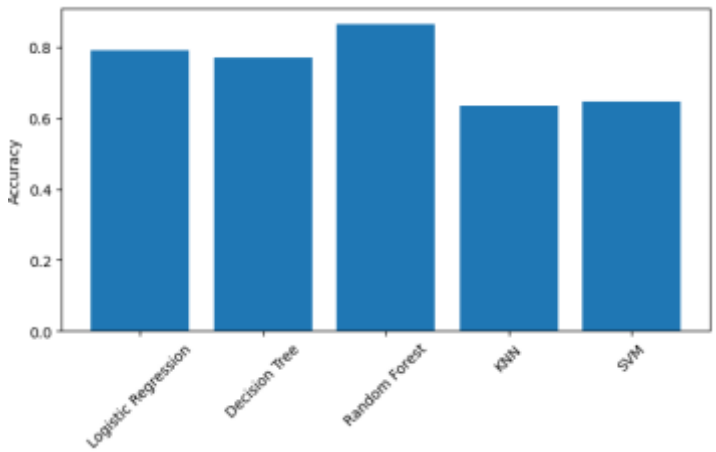

Fig. 2. Accuracy comparison of machine learning algorithms

From Table I and Fig. 2, the Random Forest classifier achieved the highest accuracy of 86.45%, outperforming all other models. Logistic Regression and Decision Tree produced moderate performance, while K-Nearest Neighbors and Support Vector Machine showed comparatively lower prediction accuracy.

The superior performance of Random Forest can be attributed to its ensemble learning mechanism. By constructing multiple decision trees and aggregating their outputs, the model reduces overfitting and improves generalization capability. Educational datasets often contain complex nonlinear relationships between behavioural attributes such as participation, attendance and resource usage, which are effectively captured by ensemble methods.

Logistic Regression performed reasonably well due to partially separable feature patterns, whereas Decision Tree exhibited overfitting when handling multiple categorical attributes. KNN showed reduced accuracy due to sensitivity to irrelevant features and distance scaling. Similarly, SVM performance was limited due to dataset size and mixed attribute characteristics.

The results indicate that behavioural attributes significantly contribute to student performance prediction. Features such as raised hands, visited resources and discussion participation demonstrated strong predictive capability, highlighting the importance of learning analytics in academic monitoring. Therefore, Random Forest is identified as the most suitable model for early prediction of student academic performance and can assist institutions in identifying at-risk students before final examinations.

CONCLUSION AND FUTURE WORK

This study presented a machine learning based approach for early prediction of student academic perfor-

mance using behavioural and academic attributes from the xAPI-Edu-Data dataset. Multiple supervised learning algorithms including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors and Support Vector Machine were implemented and compared.

Experimental analysis showed that the Random Forest classifier achieved the highest accuracy of 86.45%, demonstrating superior predictive capability compared to other models. The results confirm that ensemble learning methods are more effective in handling educational datasets with complex relationships between student activities and academic outcomes. The proposed model can assist educators in identifying academically at-risk students at an early stage and enable timely intervention strategies to improve learning outcomes.

Future work can focus on integrating deep learning models and hybrid ensemble approaches to further enhance prediction accuracy. Additionally, incorporating real-time learning management system data and longitudinal student records may improve model generalization and support adaptive personalized learning systems.

REFERENCES

1. C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man, and Cybernetics, Part C, vol. 40, no. 6, pp. 601–618, 2010.
2. S. B. Kotsiantis, C. J. Pierrakeas and P. E. Pintelas, "Predicting Students' Performance in Distance Learning Using Machine Learning Techniques," Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004.
3. S. Ayesha, T. Mustafa, A. R. Sattar and M. I. Khan, "Data Mining Model for Higher Education System," European Journal of Scientific Research, vol. 43, no. 1, pp. 24–29, 2010.
4. E. A. Amrieh, T. Hamtini and I. Aljarah, "Mining Educational Data to Predict Student's Academic Performance using Ensemble Methods," International Journal of Database Theory and Application, vol. 9, no. 8, pp. 119–136, 2016.
5. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
6. T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, 2009.
7. I. H. Witten, E. Frank and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2011.