PREDICTING LEARNING STATUS OF INFORMATICS STUDENTS AT DONG THAP UNIVERSITY BASED ON MACHINE LEARNING MODELS

Le Minh Thu and Nguyen Quoc Anh*

Faculty of Mathematics - Informatics Teacher Education, School of Education,

Dong Thap University, Cao Lanh 870000, Vietnam

**Corresponding author: Nguyen Quoc Anh, Email: nqanh@dthu.edu.vn*

Article history

Received: 01/7/2023; Received in revised form: 15/7/2023; Accepted: 20/7/2023

Abstract

For effective student support, it is essential to forecast student academic performance status. This study explores student academic performance and applies four machine learning (ML) models to predict student learning alert status. The research collects input data to build a model including parameters such as entrance grades, information about accommodation, learning environment, and first-semester result from undergraduates of the Faculty of Mathematics and Computer Education, Dong Thap University. The effective ML techniques such as Logistic regression (LR), Support vector machine (SVM), Decision trees (DTs), and Random Forest (RF) are suggested to predict student learning alert status. The outcomes are evaluated on metrics like Accuracy, Precision, Recall, and F1 score. The results show that the forecasting ability of logistic regression models outperforms in classifying student performance compared to other methods by yielding optimal classification results like high accuracy and Sensitivity followed by RF, SVM, and DTs.

Keywords: *Decision trees, learning alert status, logistic regression, random forest, support vector machine.*

DOI: https://doi.org/10.52714/dthu.13.5.2024.1287

Cite: Le, M. T., & Nguyen, Q. A. (2024). Predicting learning status of informatics students at Dong Thap University based on machine learning models. *Dong Thap University Journal of Science, 13*(5), 45-52. https://doi.org/10.52714/dthu.13.5.2024.1287.

DỰ BÁO TÌNH TRẠNG HỌC TẬP CỦA SINH VIÊN TIN HỌC TRƯỜNG ĐẠI HỌC ĐỒNG THÁP ỨNG DỤNG CÁC MÔ HÌNH MÁY HỌC

Lê Minh Thư và Nguyễn Quốc Anh*

Khoa Sư phạm Toán - Tin, Trường Sư phạm, Trường Đại học Đồng Tháp, Việt Nam

**Tác giả liên hệ: Nguyễn Quốc Anh, Email: nqanh@dthu.edu.vn*

Lịch sử bài báo

Ngày nhận: 01/7/2023; Ngày nhận chỉnh sửa: 15/7/2023; Ngày duyệt đăng: 20/7/2023

Tóm tắt

Để hỗ trợ sinh viên hiệu quả, việc dự báo tình trạng kết quả học tập của sinh viên là vô cùng cần thiết. Nghiên cứu này tìm hiểu kết quả học tập của sinh viên và áp dụng bốn mô hình máy học (ML) để dự đoán trạng thái cảnh báo học tập của sinh viên. Nghiên cứu thu thập dữ liệu đầu vào để xây dựng mô hình bao gồm các thông số như điểm đầu vào, thông tin về nơi ở, môi trường học tập và kết quả học kỳ đầu tiên của sinh viên đại học Khoa Sư phạm Toán-Tin, Trường Đại học Đồng Tháp. Các kỹ thuật ML hiệu quả như Hồi quy logistic (LR), Máy vectơ hỗ trợ (SVM), Cây quyết định (DT) và Rừng ngẫu nhiên (RF) được đề xuất để dự đoán trạng thái cảnh báo học tập của sinh viên. Kết quả được đánh giá dựa trên các chỉ số như Accuracy, Precision, Recall và F1 Score. Kết quả cho thấy khả năng dự báo của mô hình hồi quy logistic vượt trội hơn hẳn trong dự báo tình trạng cảnh báo của sinh viên so với các phương pháp khác nhờ mang lại kết quả phân loại tối ưu như độ chính xác và độ nhạy cao, tiếp theo là RF, SVM và DTs.

Từ khóa: *Cây quyết định, hồi quy logistic, máy vector hỗ trợ, rừng ngẫu nhiên, trạng thái cảnh báo học tập.*

1. Introduction

Student performance prediction is an important research topic in educational data mining that is of interest to many researchers. Early prediction of academic result can help students choose courses suitable to their individual abilities thereby improving the student learning performance (Liu et al., 2019; Anand, 2019; Wang et al., 2020), and help administrators and teachers identify students who need more attention and support to successfully complete their studies, reducing the number of academic warnings or expulsion due to poor academic results, thereby saving time and educational costs for students, families, schools and society (Reeves, 2018).

In recent years, artificial intelligence (AI) has emerged as evidence of the fourth industrial revolution (4.0). Artificial intelligence has become a core component in high-tech systems. It creeps into most areas of our lives without us even realizing it. Self-driving cars from Google and Tesla, Facebook's self-tagging system of faces in photos; Apple's Siri virtual assistant, Amazon's product recommendation system, Netflix's movie recommendation system, Google Translate's multilingual translation system, the AlphaGo go player and recently Google DeepMind's AlphaGo Zero, etc., These are just a few of the outstanding applications of artificial intelligence. Machine Learning (ML) is a specialized area of artificial intelligence. This is a small area of computer science that is capable of self-learning based on input without having to be specifically programmed.

Several machine learning techniques can be used to predict student performance and identify student dropout risk so that students can improve their academic performance as soon as possible: Artificial neural network (ANN), decision trees (DTs), linear regression (LiR), logistic regression (LoR), Naïve Bayes,…One of the most important steps when applying these machine learning algorithms is to select attributes or descriptive features to be used as input when implementing any algorithm in machine learning. Attributes can be categorized into GPA, demographics, background knowledge, learning attitudes, motivations, personal preferences, etc.

The main goal of this research is the application of ML algorithms in forecasting academic results of informatics students at Dong Thap University. This is necessary research with scientific and practical significance in order to identify and early detect students with low academic performance and need support to avoid academic warnings. At the same time, it also aims to identify good students to foster training to help students, their families and society.

2. Related works

Educational data about student living and learning when exploited by a combination of computational methods and psychological methods will be effective for the purpose of understanding more about student learning behavior (Tokan & Imakulata, 2019; Romero & Ventura, 2020). Educational data mining will help (1) predict future learning behavior by creating a model based on a combination of information such as knowledge, attitude, motivation, perception of learner; (2) identify important content to learn and optimize the teaching sequence; (3) study the influence of teaching forms on the learning process of learners; and (4) promote scientific research on the learning process through building computational models based on educational data (Aldowah et al., 2019).

The research objectives of educational data mining are as follows:

- Learning performance of students will be in the future

- What process should students follow to achieve the best effect?

- Does the learning environment affect student learning outcomes?

- What factors can predict the alert studying level of student will go down?

Understanding and solving the above problems, the school will have plans to help students over those risks.

Many research studies have been conducted on predicting the performance or result of students based on certain parameters using different machine learning algorithms. Chui et al. (2020) have built a model to predict student learning outcomes based on the training vector-based support vector machine so that the most optimal training vector reduction and predictive values can be obtained by calculating approximate multiple times to improve prediction accuracy. The research results help teachers and educational administrators with appropriate solutions to improve the learning outcomes of students with unstable learning processes. Tsiakmaki et al. (2020) believe that predicting student learning outcomes is an

important work in educational data mining; Student knowledge can be improved and accumulated over time. From this idea, an approach using automated Machine Learning (autoML) has been proposed for early identification of students' performance in three compulsory courses. This method serves as a significant aid in the early estimation of students' performance, and thus enabling timely support and effective intervention strategies.

Dewan et al. (2016) propose promising technique was proposed for predicting the risk of dropout at early stages in online courses, a serious problem where dropout rate is high for online courses at university. In this study, the model is based on a parallel combination of three ML techniques (K-Nearest Neighbor (KNN), RBN, and SVM), which make use of 28 attributes per student. Sultana et al. (2017) present a study using DT which aims to improve the existing prediction mechanism by exploiting both cognitive and non-cognitive features of students for predicting their results.

The applying of machine learning and deep learning techniques to predicting learning outcomes is continuing to be of interest and research. Iqbal et al. (2017) use the techniques of Collaborative Filtering (CF), matrix (MF) and Restricted Boltzmann Machines (RBM) techniques to systematically analyze the collected data from a university. The results show that the RBM technique predicts student learning outcomes better than the other techniques. In study of Tanuar et al. (2020), Machine learning techniques: general linear models, deep learning, decision trees are used to predict the student's final year result based on their first semester result.

Chung et al. (2019) uses of Random Forests (RF) for predicting school dropout, the result showed excellent performance in terms of various performance metrics for binary classification. Finally, ANN, SVM, LR, NB, and DT were analyzed for similar purposes by using the data recorded by e-learning tools, the result shows the high accuracies (Hussain et al. 2018).

These above studies show that the power of ML model for predicting student result by using the previous results. In this study, however, the facilities and services of school also are used as attributes for training model. The contribution of study will help school in student consultation, supporting the best conditions for students to study well in Dong Thap University.

3. Methodology

3.1. Data collection and processing

Data to build the model was collected from information of 218 students in the 2018, 2019, 2020 and 2021 courses majoring in Information Technology and Informatics Pedagogy at Dong Thap University. The information that needs to be collected to build the model is: entrance grades and first semester result of each student. In addition, the article also collects more survey feedback on the accommodation, activities, and facilities that students are supported in school, which are also factors affecting the learning performance of students [Table 1]. These data were collected from the Student Affairs Department, the Training Department and the Faculty of Computer Science and Education. After standardized data processing, four models are proposed to predict student learning performance and then compare models by their accuracy. The standardized dataset is scaled 80% for model training and 20% is then used to evaluate the accuracy of the models. The performance of models was evaluated with classification accuracy (CA), precision, recall and f-score (F1) (Fig. 1).

Table 1. Data Description

Attribute	Description and Value
Entrance grade of student	0-30
Gender	[0,1]
Hometown is Dong Thap province	Hometown' student is Dong Thap province where the university is located. [0,1]
Staying in a hostel or dormitory	Student is staying in a hostel or dormitory. [0,1].
Distance between accommodation and school	Distance between accommodation and school. (1 - near, 2 - medium and 3- far).

Supporting equipment in the classrooms	Very Unsatisfied: 1. Not Satisfied: 2. Confused: 3. Satisfied: 4. Very Satisfied: 5.
Supporting equipment in lecture halls	
Books, textbooks, library references	
The usefulness of the website	
The equipment in computer practice room	
System of yards for sports	
Medical equipment and treatment	
Scenery on campus	
Wi-Fi system on campus	
Service at the school canteen	
The school's concern and help with difficult circumstances	
School feedback and inquiries	
Study counseling activities	
Expertise and professionalism of the teaching staff	
The caring attitude, friendly communication of the teaching staff	
The organization of sports activities	
Organization of cultural and artistic activities	

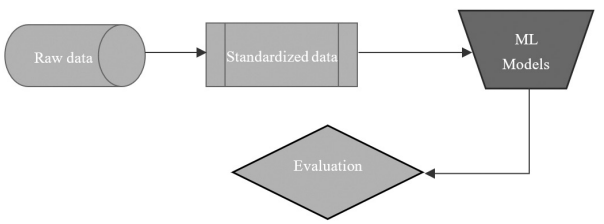

Figure 1. The general scheme of the solution

3.2. Machine learning models

3.2.1. Logistic Regression

Logistic regression models are often created with the goal of predicting the outcomes of the future based on variables. Regression model diagnostics measure how well Models describe the underlying relationships between predictors and outcomes existing within the data, either the data on which the model was built or data from a different population. The accuracy of a logistic regression model is mainly judged by considering discrimination and calibration. Discrimination is the ability of the model to correctly assign a higher risk of an outcome. Whereas calibration is the ability of the model to assign the correct average absolute level of risk (i.e., accurately estimate the probability of the outcome).

Logistic regression that predicts the value of y (y being a binary value of 0 and value 1) is called binary classification. Logistic regression is essentially a linear predictor, and the result of y not only tells what class is predicted, but also gives more information about the probability of that class.

Logistic regression aims to:

• Perform the classification problem by a linear model $z = W^TX + b$. However, because z is too large $(-\infty,+\infty)$. Therefore, the output requirements of the problem cannot be satisfied.

• An outer wrapper should be used to map z to [0,1]. This also means using the wrapper $g(z) \in [0,1]$ that follows the Bernoulli distribution.

• Probability of positive class $P(y = 1|z)$ due to $P(Y = 0|z) = 1 - P(Y = 1|z)$.

Logistic regression uses a logistic function called a sigmoid function to map predictions and their probabilities. The sigmoid function refers to an S-shaped curve that converts any real value to a range between 0 and 1. If the output of the sigmoid function (estimated probability) is greater than a predefined threshold on the graph, the model predicts that the instance belongs to that class. If the estimated probability is less than the predefined threshold, the model predicts that the instance does not belong to the class. The sigmoid function is referred to as an activation function for logistic regression and is defined as formula 1:

$$\sigma = \frac{1}{(1+e^{-t})} \quad (1)$$

where,

e = base of natural logarithms

t = numerical value one wishes to transform

Equation of Logistic regression is defined as formular 2:

$$y = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}} \quad (2)$$

here,

x = input value

y = predicted output

b0 = bias or intercept term

b1 = coefficient for input (x)

3.2.2. Support Vector Machine

Support vector machine (SVM) are mainly utilized to perform supervised classification tasks. Linear and non-linear classifications can be performed using SVM by employing a kernel function. Also, used for regression tasks and it performs classification by constructing ideal hyperplane on the trained data. Two parallel hyperplanes are constructed on the sides of the separating hyperplane. SVMs can be used for the classification of complex datasets in sophisticated applications such as handwritten digit recognition, object recognition, text classification.

The student evaluation, according to psychology, may be affected by the behaviors potentially. Xu et al. divided students into three categories based on the detailed records of learning activities on MOOCs platforms (Xu & Yang, 2016). Then, the authors built a predictor based on SVM to predict certification obtaining. Hana et al. (Bydžovská, 2016) compared the traditional machine learning algorithms for student learning prediction, including SVM, Linear Regression, and Random Forest, where SVM is the best on both study-related data and social behavior data. However, SVM suffers from computation cost in big data due to its optimization limitation.

3.2.3. Decision trees

Decision trees (DTs) are a non-parametric supervised learning method used for classification and regression. It learns the splitting rule to divide the data according to their features and obtains the labels by voting at leaf nodes. Decision trees could deliver interpretable results and thus obtain much attention for academic prediction (Al-Barrak & Al-Razgan, 2016; Saa, 2016; Alsalman, 2019).

Based on different feature selection methods and pruning rules, the DT model has three main algorithms, including: ID3, CART, and C4.5. Among these DT algorithms and other machine learning algorithms, the DT showed a higher precision on their used data set.

3.2.4. Random forest

A random forest classifier is more precise than a single classification tree in the sense that it has lower mean-squared prediction error. By bagging a classifier, the bias will remain the same but the variance will decrease. One way to further decrease the variance of the random forest is by construction trees that are as uncorrelated as possible. Breiman introduced in 2001 random forests with random inputs. In these forests, instead of finding the best variable and partitioning among all the variables, the algorithm will now randomly select $p < m$ random covariates and will find the best condition among those p covariates.

The novelty of random forest model is in the tree-growing procedure. Instead of finding the best condition among all the predictors, the algorithm will now randomly select a subset of predictors and will find the best condition among these, this modification greatly improved the accuracy of random forests. Random forests were used to build two classifiers: one that predicts if a student will complete their undergraduate program, the other that predicts the major of a student who completed a program in a major university in Canada over 10 years (Beaulac & Rosenthal, 2019).

3.2.5. Metrics

Accuracy, Precision, Recall and F1-Score are three metrics that are used to measure the performance of a machine learning algorithm.

Precision is the ratio of true positives over the sum of false positives and true negatives. It is also known as a positive predictive value. It is the possibility of classifying actual data belonging to positive class into positive class instead of other class. It is calculated as: Precision = TP/(TP+FP).

Recall is the ratio of correctly predicted outcomes to all predictions. It is also known as sensitivity or specificity. It is the possibility of classifying actual data belonging to positive class into positive class instead of other class. It is calculated as: Recall = TP/(TP+FN).

Accuracy is a metric which describes or classifies data into proper class label or class. The extent to which classification of data is done correctly is observed by calculating accuracy. Accuracy is the ratio of correct predictions out of all predictions made by an algorithm. It can be calculated by dividing precision by recall or as 1 minus false negative rate (FNR) divided by false positive rate (FPR)

Accuracy = (TP + TN)/(TP + FP + FN + TN)

The F1-score combines these three metrics into one single metric that ranges from 0 to 1 and it considers both Precision and Recall. This metric computes how many times a model made a correct prediction across the entire dataset.

F1-Score = 2 * Precision * Recall / (Precision + Recall)

4. Results and discussion

Student's academic data was classified using ML classifiers, training, and testing the data was carried on with the help of Scikit-learn library. Classification results were analyzed in terms of Accuracy, Precision, Recall, and F1 Score.

The results predicted by each of these algorithms, when compared with the actual result of that semester, will give the accuracy of the predictions. The predictive accuracy of each of these algorithms is computed and is tabulated below [Table 2].

Table 2. Predictive accuracy of the algorithms

No.	Machine learning model	Accuracy (%)
1	Regression Logistic (RL)	97.36
2	Random Forest (RF)	94.73
3	Support Vector Machines (SVM)	93.56
4	Decision Trees (DTs)	86.84

It was observed that RL is outstanding by yielding 97.36% classification accuracy compared to other techniques like RF, SVM and DTs. Whereas RF yields the promising classification results of 94.73%, followed by SVM-93.56% and DTs-86.84%.

Besides, RL also shows good performance through the measurement Precision, Recall and F1-Score. Specifically, RL accounts for 96.29%, 92.85% and 94.53% for Precision, Recall and F1 Score respectively. SVM followed with 94.65%, 92.34% and 93.48% for Precision, Recall and F1 Score respectively. RF stays in 3rd out of 4 models. Precision, Recall and F1 Score of RF are 93.23%, 90.43%, 91.80 respectively. Last one is DTs with 90.91% for Precision, 71.43% for Recall and 80% for F1 Score [Table 3].

Table 3. Precision, Recall and F1-Score of the algorithms

No.	ML model	Precision (%)	Recall (%)	F1 Score (%)
1	RL	96.29	92.85	94.53
2	RF	94.65	92.34	93.48
3	SVM	93.23	90.43	91.80
4	DTs	90.91	71.43	80

5. Conclusions

The problem of predicting the situation of being forced to stop studying is quite urgent. At Dong Thap University, this has not been done until the student is forced to drop out through the calculation of cumulative points each term. In this paper, we propose ML models to predict learning alert status like Logistic Regression, Random Forest, Support Vector Machines and Decision Trees. By this study, the affecting models can be applied for the student's academic alert status prediction so that the school can take measures to support students in learning in the next semesters. The experiment with student data in Information Technology and Computer Science Pedagogy has proven the feasibility of the method. In the future, the dataset will be enhanced and expanded to experiment with other advanced models for student data from other disciplines.

Acknowledgements: This research was supported by the project of Dong Thap University, with the project coded SPD2021.01.27.

References

Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: a case study. *International journal of information and education technology*, *6*(7), 528. doi: 10.7763/IJIET.2016.V6.745.

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, *37*, 13-49.

Alsalman, Y. S., Halemah, N. K. A., AlNagi, E. S., & Salameh, W. (2019, June). Using decision tree and artificial neural network to predict students academic performance. In *2019 10th international conference on information and communication systems (ICICS)* (pp. 104-109). IEEE.

Anand, M. (2019). Advances in edm: a state of the art. *Software Engineering: Proceedings of CSI 2015*, 193-201.

Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, *60*, 1048-1064.

Bydžovská, H. (2016). A Comparative Analysis of Techniques for Predicting Student Performance. *International Educational Data Mining Society*.

Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human behavior*, *107*, 105584.

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, *96*, 346-353.

Dewan, M. A. A., Lin, F., & Wen, D. (2015, August). Predicting dropout-prone students in e-learning education system. In *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)* (pp. 1735-1740). IEEE.

Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., & Ali, S. (2019). Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review*, *52*, 381-407.

Iqbal, Z., Qadir, J., Mian, A. N., & Kamiran, F. (2017). Machine learning based student grade prediction: A case study. *arXiv preprint arXiv:1708.08744*. DOI: arxiv.org/abs/1708.08744.

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, *1*(9), 982.

Liu, Q., Tong, S., Liu, C., Zhao, H., Chen, E., Ma, H., & Wang, S. (2019, July). Exploiting cognitive structure for adaptive learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 627-635).

Nguyen, P. H., Sheu, T. W., & Nagai, M. (2015). Predicting the student learning outcomes based on the combination of Taylor approximation method and Grey models. *J. Sci. VNU J. Sci. Educ. Res*, *31*(2), 70-83.

Reeves, B. (2018). Development of rubrics to support teacher judgement of student proficiency in ethical Decision-Making (Masters Research thesis).

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *10*(3), e1355.

Saa, A. A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5). doi: 10.14569/IJACSA.2016.070531.

Sultana, S., Khan, S., & Abbas, M. A. (2017). Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. *International Journal of Electrical Engineering Education*, *54*(2), 105-118.

Sheikh, W., Parulkar, A., Ahmed, M. B., Osler, B., Liu, B., Sharma, E., ... & Chu, A. (2020). Recursive feature elimination (rfe) selects for important features in predicting post-operative cardiac arrest (poca) in patients undergoing coronary artery bypass grafting (cabg): insights from the sts registry. *Journal of the American College of Cardiology*, *75*(11_Supplement_1), 1533-1533.

Tanuar, E., Heryadi, Y., Abbas, B. S., & Gaol, F. L. (2018, September). Using machine learning techniques to earlier predict student's performance. In *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)* (pp. 85-89). IEEE.

Tokan, M. K., & Imakulata, M. M. (2019). The effect of motivation and learning behaviour on student achievement. *South African Journal of Education*, *39*(1).

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2019). Implementing AutoML in educational data mining for prediction tasks. *Applied Sciences*, *10*(1), 90.

Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., ... & Wang, S. (2020, April). Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 04, pp. 6153-6161).

Xu, B., & Yang, D. (2016). Motivation classification and grade prediction for MOOCs learners. *Computational intelligence and neuroscience*, *2016*, 4-4. doi: 10.1155/2016/2174613