Chapter 13
Recommendation Model for Students Dropout at Ba Ria-Vung Tau College of Technology

Ngoc-Hoang Phan and Thi-Thu-Trang Bui

1 Introduction

In recent years, the phenomenon of students' dropout at Ba Ria-Vung Tau College of Technology is becoming more and more common. This is an issue that has received a lot of special attention from the Board of Directors as well as the teachers in the school. Through incomplete statistics from 2016 up to now, the phenomenon of students' dropout usually falls mainly after the Lunar New Year and after the first year.

From the beginning of the 2016–2017 academic year until now, Ba Ria-Vung Tau College of Technology has had nearly 458 cases of students drop out of school. In particular, the phenomenon of dropping out of school focuses mainly on students of intermedia degree. The number of dropouts for intermedia degree is 380 cases, accounting for 83%. The remaining 78 cases belonged to students at college degree, accounting for 17%.

This phenomenon can be related to many different objective and subjective factors such as boredom about learning, leading to a decline in results; being enticed by bad friends; do not have a clear study plan for themselves; study difficult subjects such as industrial electricity, information technology. The student's dropout is not only a concern of today but will still be a burden on the community and society in the future.

In order to reduce the phenomenon of dropping out of school, it is necessary to have solutions to early warn of possible cases of students drop out of school. From these early warnings, the school will have support in learning orientation counseling and recommendations to students. From there, correct the sense of learning and have

N.-H. Phan (✉) · T.-T.-T. Bui
Ba Ria-Vung Tau University, Vung Tau, Vietnam
e-mail: hoangpn@bvu.edu.vn

S. Kumar et al. (eds.), *Proceedings of International Conference on Communication and Computational Technologies*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-19-3951-8_13

a reasonable learning method for better learning results, and at the same time help learners orient themselves more accurately about the career that is right for them.

Therefore, this paper proposes an approach to apply artificial intelligence to build a model to recommend dropping out cases at Ba Ria-Vung Tau College of Technology. Thereby, managers promptly take appropriate measures and strategies to minimize the phenomenon of students drop out of school.

2 Related Studies

In 2002, in a study of Le Thanh Minh [1], it was proposed the use of association rule mining algorithm [2] and fuzzy logic set [3] to recommend necessary information for improving the quality of high school and junior high school students based on the students' graduation exam results. In the same year, author Nguyen Quoc Thong with the topic of his Master thesis [4] used association rules to predict students' scores. Predicted results will be used to recommend cases of weak students and students who need extra tutoring.

In 2006, in study [5], Superby and colleagues collected data of students including personal information, students' learning behaviors and perceptions. Based on the collected data, the authors propose a model to identify factors affecting students' learning using different algorithms such as decision trees, neural networks, random forests, and linear discriminant analysis.

In a study by Nguyen Thai Nghe in 2007 [6], it was proposed to use decision tree algorithms [7, 8] and Bayesian networks [9] to predict the learning outcomes of undergraduate and graduate students of Can Tho University (Vietnam). In another study by the same author [10], it was proposed to use matrix decomposition to predict student learning outcomes.

In 2009, author Phan Dinh the Huan studied mining methods to support the assessment and prediction of student learning outcomes based on educational data [11]. Experimental results were evaluated on the data set of learning outcomes of students at Ton Duc Thang University (Vietnam). Also in the same year, G. Dekker and colleagues used decision trees to build a model to predict the likelihood that students might drop out after the first semester [12].

In 2010, Ayesha and colleagues applied the K-means algorithm to predict student learning behavior [13]. The results obtained can help teachers make timely adjustments in the teaching process. In 2011, Nguyen Thi Van Hao built a system to predict high school graduation. In her study, based on the data of students' academic performance and behavior, the author has applied fuzzy association rule mining algorithm to predict high school graduation results [14]. Also in 2011, Bharadwaj [15], Yadav [16] applied mining algorithms to predict students' learning outcomes at the end of the semester. The data collected for the experiment includes information about students' attendance, test scores, and extracurricular activities.

In 2012, Nguyen Dang Nhuong applied the K-means data clustering algorithm [18] to extract information from student scores of Van Lang Vocational College, Hanoi (Vietnam) [17]. Experimental results show the influence of factors such as region, family situation, ethnicity, morality… on students' learning results. From there, using the solution to classify learning results helps to quickly assess learners' perceptions, and helps to adjust teaching methods to suit learners' abilities.

Also in 2012, Bukralia [19] proposed to use machine learning techniques such as neural networks, logistic regression [20], decision trees [7, 8], machine learning. SVM support vector [21] to predict student learning outcomes under the Midwest University distance learning system. Besides, Lin propose to build a model that allows to predict which students will have difficulties in learning, so that there can be timely support solutions [22].

In 2013, Pal and colleagues proposed to use decision tree algorithm [7, 8] and Bagging technique [24] to predict student learning outcomes at Purvanchal University, India [23]. In 2014, Do Thanh Nghi and colleagues proposed to apply the random forest algorithm to help detect important subjects that affect the learning outcomes of students in information technology [25]. The collected data includes the learning results of Information Technology students from 20 to 29th courses (enrolled from 1994 to 2003). Experimental results show that 10 subjects have the most important influence on the outcomes of Information Technology students.

In 2019, Nguyen Thi Uyen and colleagues proposed a model using logistic regression to analyze the factors that have a great influence on the dropout status of students majoring in Information Technology. Experimental results show that students with low scores in the subjects of Programming Language C, Mathematics A2 (Calculus), and Ho Chi Minh Thought and with low scores in the university entrance exam tend to be forced to stop studying [26].

The above studies focus on solving problems to predict student learning outcomes and analyze the factors that influence the learning outcomes. In which, algorithms like K-means [18], decision trees [7, 8], neural networks [20], and support vector machines (SVMs) [21] are widely used. However, each study has a different input data structure, and is also different from the existing data in the system of Ba Ria-Vung Tau College of Technology. Therefore, we are interested in applying these algorithms to the school's data set. From there, based on the experimental results to find a suitable model for recommending the dropout ability of students.

3 Preprocessing Data

Currently, student information of Ba Ria-Vung Tau College of Technology can be collected including: basic personal information and information about learning results. General basic personal information of students includes full name, date of birth, place of birth, hometown, course, major, training degree, training system. For the information about learning results, in the system stores information about registered subjects and students' scores of these subjects in each semester, status subject debt, tuition debt, warning, reservation.

Student information of the 2017–2018 academic year data is used for training dataset, and the 2018–2019 academic year data is for the test dataset. In order to use model for all students of different training degree and major throughout the school, there are some data that must be preprocessed to transform into a common structure before being using.

For student achievement results, the information includes the summaries of each subject for each semester. However, for different major, the subjects in each semester will not be the same. For example, for students majoring in Automotive, in first semester, there will be specialized subjects that other majors do not have, such as Automobile Construction, Principle of Internal Combustion Engine.

Therefore, the column final scores of subjects in each semester was converted into a new data type. This new data type includes, in a semester, the total number of subjects; the number of subjects with excellent grade; the number of subjects with good grade; the number of subjects with average grade; and the number of subjects with weak grade. The final score of all subjects in the semester is used to determine excellent, good, average, and weak grade based on the current scoring regulations of our College according to the criteria for classifying learning outcomes as Table 1.

The results of converting the final scores are shown in Table 2, in which (1)—course; (2)—major; (3)—number of subjects in first semester; (4)—number of excellent subjects in first semester; (5)—number of good subjects in first semester; (6)—number of average subjects in first semester; (7)—number of weak subjects in first semester; (8)—number of absence in first semester; (9)—number of subjects in second semester; (10)—number of excellent subjects in second semester; (11)—number of good subjects in second semester; (12)—number of average subjects in second semester; (13)—number of weak subjects in second semester.

Table 1 Criteria of classifying learning outcomes

No	Final score (FS)	Grade
1	FS $\geq$ 8.0	Excellent
2	8.0 > FS $\geq$ 6.5	Good
3	6.5 > FS $\geq$ 5.0	Average
4	5.0 > FS	Weak

Table 2 The result of converting the final score data into the total number of subjects

No	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
0	2017–2018	CDT	7	3	3	1	0	12	6	2	3	1	0
1	2017–2018	CDT	7	4	3	0	0	14	6	1	4	1	0
2	2017–2018	CDT	7	2	5	0	0	16	6	1	2	3	0
3	2017–2018	CDT	7	1	4	2	0	21	6	2	3	1	0
4	2017–2018	CDT	7	2	3	2	0	13	6	2	1	3	0

Table 3 Results of converting total subject data to percentage data

No	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
0	0	0	0	0	0	0.86	0.14	0.0	0.83	0.17	0.0	0.0
1	0	0	0	0	0	1.00	0.00	0.0	0.83	0.17	0.0	0.0
2	0	0	0	0	0	1.00	0.00	0.0	0.50	0.50	0.0	0.0
3	0	0	0	0	0	0.71	0.29	0.0	0.83	0.17	0.0	0.0
4	0	0	1	0	0	0.71	0.29	0.0	0.50	0.50	0.0	0.0

However, it is shown that for different major, the total number of subjects in the semester is different. Similar to the training degree, the college degree will have more subjects in each semester than the intermediate degree. In addition, the number of subjects with excellent grade accounts for a very small percentage, and the value is almost zero for most of the students.

So, the academic performance data further transformed into the percentage of excellent grade subjects; the percentage of good grade subjects; the percentage of average grade subjects; and the percentage of weak grade subjects in each semester. Because the percentage of excellent grade subjects is almost zero for all students, so we combine excellent and good grade subjects into one common feature that is the percentage of excellent and good grade subjects. The transformation results provide new data that has values in the range of [0–1], which is very suitable for machine learning models.

An example of the result of converting data is shown in Table 3, in which (1)—student in province; (2)—tuition debt in second semester; (3)—number of warning in first semester; (4)—being warned in second semester; (5)—being to reservation in second semester; (6)—percentage of excellent and good subjects in first semester; (7)—percentage of average subjects in first semester; (8)—percentage of weak subjects in first semester; (9)—percentage of excellent and good subjects in second semester; (10)—percentage of average subjects in second semester; (11)—percentage of weak subjects in second semester; (12)—percentage of weak subjects in whole academic year;

In addition, several features that may have affect to the ability of students' dropout was selected such as learning away from home; having problems with tuition debt difficulties; being warned; or being to reservations. In Table 3, the 12 information about students is presented and will be used as features to train and test algorithms including K-means [18], decision trees [7, 8], neural networks [20], and support vector machine [21] to find the appropriate model for recommending the dropout ability of students.

4 Experiment Results

4.1 Training Models

In this paper, the training and testing process was conducted on the Google Colab using Python programming language, data processing support libraries such as numpy, pandas, and support library for building models like sklearn, keras, tensorflow.

For the kMeans algorithm, the kMeans model in the sklearn.cluster library is used. In which, the parameter n_clusters is set to 2, corresponding to 2 outputs of the model. The output will determine the warning to dropout case or not to dropout case.

For decision tree algorithm, DecisionTreeClassifier model in sklearn.tree library is used with default parameters.

For the neural network model, the Sequential model with Dense classes from keras library through tensorflow is used. The neural network model has 12 inputs, 18 neurons in the hidden layer, and 2 output neurons. The number of input neurons corresponding to 12 features was extracted from the data set of students of Ba Ria-Vung Tau College of Technology shown in Table 3. The number of output neurons is 2 according to the ability to warn students to dropout case or not to dropout case. Besides, for training neural network model, Relu activation function is used for hidden layer neurons, and softmax function is used for output neurons. The neural network is trained using the loss function categorical_crossentropy, and the optimization function adam.

For the support vector machine algorithm, the SVC model of the sklearn.svm library is used. The support vector machine model is established with a core function of Radial Basic Function (rbf), also known as a Gaussian kernel. This is the most used function in practice and is also the default choice in the sklearn library.

These models are trained using the dataset collected from the 2017–2018 academic year. Training results are evaluated using metrics such as accuracy, precision, recall, and f1-score. The results of model training are shown in Table 4.

In addition, the results of models training are also represented through the confusion matrix. This helps specify how each class is classified, which class is most correctly classified, and which class is often misclassified into another. The results represented by the confusion matrix are shown in Fig. 1.

Table 4 Model trainning results

	kMeans	Decision tree	Neural network	SVM
Accuracy	0.5	0.99	0.96	0.95
Precision	0.16	1.0	0.9	0.92
Recall	0.48	0.97	0.85	0.75
F1-score	0.24	0.98	0.88	0.83

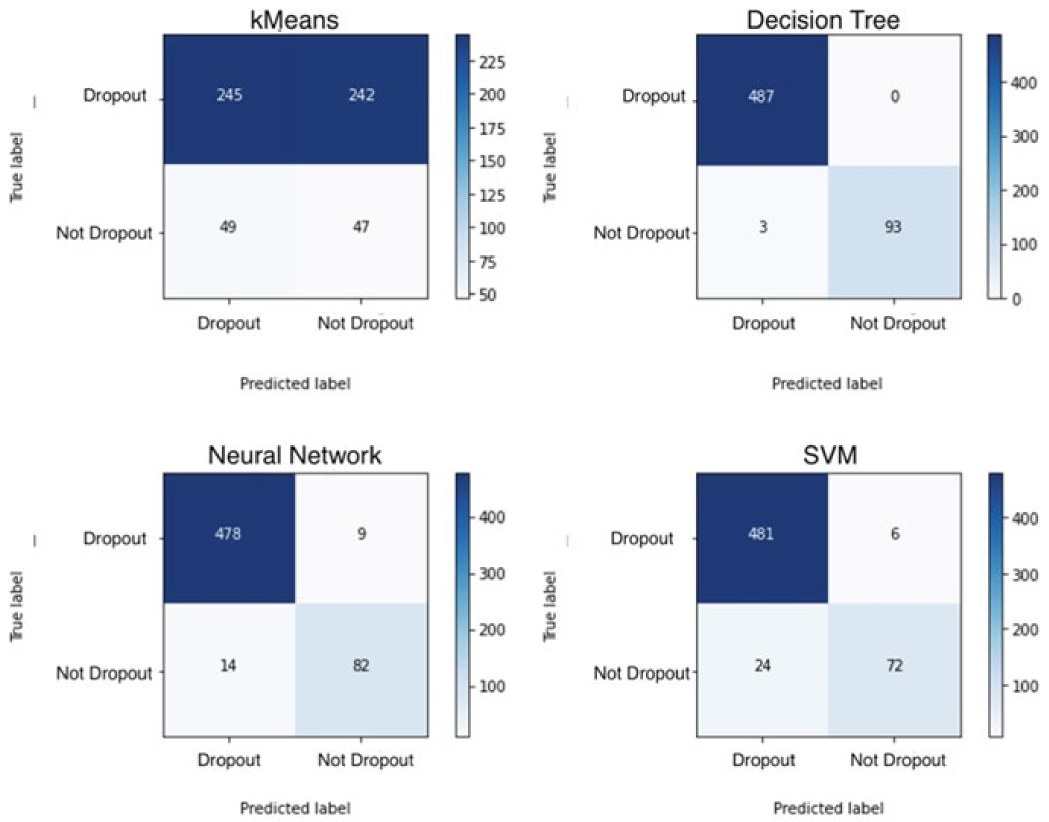

Fig. 1 Confusion matrix of models training

As Fig. 1 shows, the kMeans model gives the lowest training results. The decision tree model gives the best results, but this also has the potential to lead to overfitting of the training data. The two neural network and support vector machine (SVM) models give good results and are quite similar.

4.2 *Testing Models*

These trained models are tested on the 2018–2019 academic year dataset. Similar to training phase, test results are evaluated using metrics such as accuracy, precision, recall, and f1-score. The test results are shown in Table 5.

Experimental results show that the accuracy of these models is over 90% and is almost equivalent. However, by comparing using precision metric, the kMeans algorithm gives very low precision, almost zero. Which means the percentage of student dropout case were correctly classified is very low. Meanwhile, the precision increases for the decision tree algorithms—38%, the neural network—63%. The support vector machine algorithm achieves up to 100% precision.

Table 5 Model testing results

	kMeans	Decision tree	Neural network	SVM
Accuracy	0.95	0.92	0.95	0.96
Precision	0.04	0.38	0.63	1.0
Recall	0.58	0.57	0.5	0.38
F1-score	0.07	0.47	0.56	0.55

On the other hand, by comparing using recall metric, the support vector machine algorithm gives a low result of 38%. Meanwhile, the recall increases for the remaining algorithms including kMeans, decision trees and neural networks. The recall are almost similar for these algorithms, ranging from 50 to 58%. This proves that the omission of really positive points is acceptable.

For the f1-score metric, which is determined by the harmonic mean of the two indexes of precision and recall, the neural network algorithm and the support vector machine algorithm give the best results (55% and 56%). In terms of results, the higher the f1-score, the better the classifier. In this case, the neural network algorithm and the support vector machine algorithm give almost the same f1-score result. However, the precision and recall indexes of the neural network have similarity (63% and 50%), while the support vector machine have a large difference between the precision and recall indexes (100% and 38%). Therefore, according to all these metrics, the neural network algorithm gives the best results and is proposed to be used as a model for recommending the possibility of students dropout at the Ba Ria-Vung Tau College of Technology.

All the experimental results on the test dataset can be represented using the confusion matrix. The classification results of these models for each warning dropout or no warning dropout are shown in the confusion matrix in Fig. 2.

5 Conclusions

This paper proposes an approach using machine learning algorithms to build a model to recommend the dropout ability of students at Ba Ria-Vung Tau College of Technology. The experimental results, which using metrics such as accuracy, precision, recall, and f1-score, show that the neural network is the model that gives the best and most suitable results. Therefore, the neural network is proposed to be selected as a model for recommending the possibility of student dropout at Ba Ria-Vung Tau College of Technology.

In the future, we plan to expand research and develop models specific to each major and training level. In addition, we need to collect more other data that may be has affect to the ability of student drop out to improve data quality, increase accuracy.

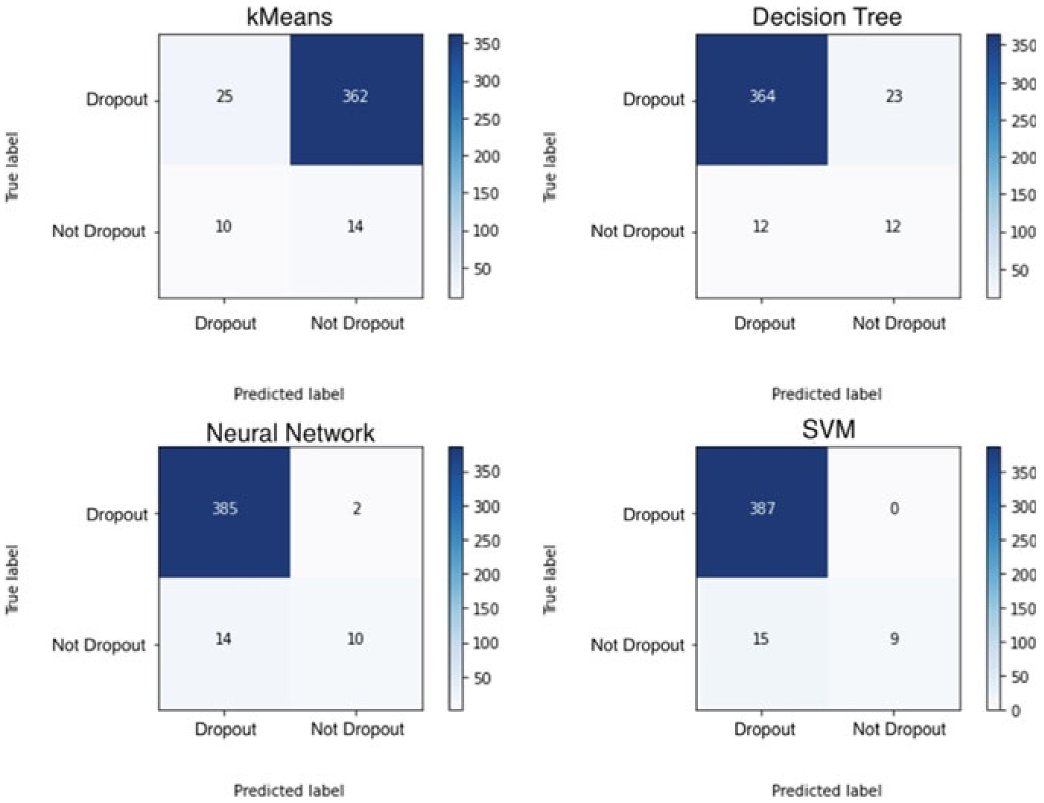

Fig. 2 Confusion matrix of models test

References

1. Minh LT (2002) Mining graduation exam scores for student classification assessment. Master Thesis, Ho Chi Minh City University of Natural Sciences, Vietnam
2. Agrawal R, Imielinski T, Swami A (1993) Mining associations between sets of items in massive databases. In: Proceeding of ACM-SIGMOD international conference on management of data, pp 207–216, Washington, USA
3. Zadeh LA (1965) Fuzzy sets. J Inf Control 8(3):338–353
4. Thong NQ (2002) Developing some data mining applications in education and training, Master Thesis, Ho Chi Minh City University of Natural Sciences, Vietnam (2002).
5. Superby JF, Vandamme JP, Meskens N (2006) Determination of factors influencing the achievement of the first-year university students using data mining methods. In: Workshop on Education
6. Thai-Nghe N (2007) An analysis of techniques in predicting learning outcomes. In: Proceedings of the 10th Vietnam national conference on information technology, pp 19–31
7. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman & Hall
8. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, CA
9. Pearl J (1985) Bayesian networks: a model of self-activated memory for evidential reasoning. In: Proceedings of Cognitive Science Society, pp 329–334, UC Irvine
10. Thai-Nghe N, Drumond L, Horváth T, Schmidt-Thieme L (2011) Multi-relational factorization models for predicting student performance. In: Proceedings of the KDD 2011 workshop on knowledge discovery in educational data

11. Huan PDT (2009) Research and application of combined law mining method on educational data. Master Thesis, Ho Chi Minh City University of Natural Sciences, Vietnam
12. Dekker G, Pechenizkiy M, Vleeshouwers J (2009) Predicting students drop out: a case study. In: Proceedings of the 2nd international conference on educational data mining, pp 41–50
13. Ayesha S, Mustafa T, Sattar AR, Inayat Khan M (2010) Data mining model for higher education system. Eur J Sci Res 43(1):24–29
14. Hao NTV (2011) Building a system to predict high school graduation results. Master Thesis, Lac Hong University, Dong Nai, Vietnam
15. Bharadwaj BK, Pal S (2011) Mining educational data to analyze student's performance. Int J Adv Comput Sci Appl (IJACSA) 2(6):63–69
16. Yadav SK, Bharadwaj BK, Pal S (2011) Data mining applications: a comparative study for predicting student's performance. Int J Innov Technol Creative Eng (IJITCE) 1(12):13–19
17. Nhuong ND (2012) Data mining on learning outcomes of students at Van Lang Vocational College, Hanoi. Master Thesis, University of Technology, Vietnam National University, Vietnam
18. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, pp 281–297
19. Bukralia R, Deokar A-V, Sarnikar S, Hawkes M (2012) Using machine learning techniques in student dropout prediction. In: Burley H (ed) Cases on institutional research system. IGI Global, pp 117–131
20. Hastie T, Friedman J-H, Tibshirani R (2001) The elements of statistical learning: data mining, inference, and prediction. Springer
21. Vapnik V (1995) The nature of statistical learning theory. Springer
22. Lin SH (2012) Data mining for student retention management. ACM J Comput Sci Colleges 27(4):92–99
23. Pal A-K, Pal S (2013) Analysis and mining of educational data for predicting the performance of students. Int J Electron Commun Comput Eng 4(5):2278–4209
24. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
25. Nghi DT, Khang PN, Trung NM, Hung TT (2014) Discovering important subjects that affect the learning outcomes of students in information technology. J Sci Can Tho Univ 33:49-57
26. Uyen NT, Tam NM (2019) Applying data mining algorithms in predicting student learning outcomes. J Sci Vinh Univ 48(3A):68–73