International Journal of Internet, Broadcasting and Communication Vol.14 No.3 285-291 (2022)
http://dx.doi.org/10.7236/IJIBC.2022.14.3.285

IJIBC 22-3-33

Applying advanced machine learning techniques in the early prediction of graduate ability of university students

Nga Pham[1,5], Pham Van Tiep[1], Tran Thu Trang[1], Hoai-Nam Nguyen[2, 3], Gyoo-Seok Choi[4], Ha-Nam Nguyen[5,6]

[1]*Faculty of Information Technology, Dainam university, Hanoi, Vietnam*
[2]*VNU University of Education, Vietnam National University in Hanoi, Vietnam*
[3] *ICT Department, Ministry of Education and Training, Vietnam*
[4] *Department of Computer Science, Chungwoon University, Incheon, Korea*
[5] *VNU University of Engineering and Technology, Vietnam National University in Hanoi, Vietnam*
[6] *Vietnam Institute for Advanced Study in Mathematic, Hanoi, Vietnam*
E-mail: {ngaptt, tieppv, trangtt}@dainam.edu.vn; nam.moet@gmail.com;
lionel@chungwoon.ac.kr; nhnam@viasm.edu.vn,

Corresponding Author: lionel@chungwoon.ac.kr(Gyoo-Seok Choi)
namnhvn@gmail.com(Ha-Nam Nguyen)

Abstract

The number of people enrolling in universities is rising due to the simplicity of applying and the benefit of earning a bachelor's degree. However, the on-time graduation rate has declined since plenty of students fail to complete their courses and take longer to get their diplomas. Even though there are various reasons leading to the aforementioned problem, it is crucial to emphasize the cause originating from the management and care of learners. In fact, understanding students' difficult situations and offering timely Number of Test data and advice would help prevent college dropouts or graduate delays. In this study, we present a machine learning-based method for early detection at-risk students, using data obtained from graduates of the Faculty of Information Technology, Dainam University, Vietnam. We experiment with several fundamental machine learning methods before implementing the parameter optimization techniques. In comparison to the other strategies, Random Forest and Grid Search (RF&GS) and Random Forest and Random Search (RF&RS) provided more accurate predictions for identifying at-risk students.

Keywords: *Graduate result; Machine learning; Graduation prediction; ; Random Forest; Support Vector Machine.*

1. Introduction

Many studies have been conducted to build an early intervention solution, with the goal of generating early

Manuscript Received: May. 10, 2022 / Revised: May. 12, 2022 / Accepted: May. 15, 2022
Corresponding Author: lionel@chungwoon.ac.kr(Gyoo-Seok Choi), namnhvn@gmail.com(Ha-Nam Nguyen)
Tel:***－****－**** +84-90-467-0471, Fax: +82-32-770-8165
Professor, Department of Computer Science, Chungwoon University, Incheon, Korea
Professor, Vietnam Institute for Advanced study in mathematics, Ha Noi, Viet Nam

predictions based on student performance [1 - 12]. Through such early detection, the academic advisor team may intervene and guidance from the beginning, thereby helping learners improve their learning and overcome difficulties more easily.

This allows students to graduate on time and avoid dropping out. Mark Sweeney et al. in their work in 2016 developed a method for predicting students' grades in the following courses based on pattern learning from existing grade data combined with additional information on students, courses, and teaching faculty [1]. They experimented with many classical and modern approaches, with the Factorization Machines (FM), Random Forests (RF), and Personalized Multi-Linear Regression model giving the lowest prediction error. In 2019, Vijayalakshmi V. suggested a system to predict student performance using a deep neural network [2]. They used a variety of techniques, including Decision Tree (C5.0), Naive Bayes, Random Forest, Number of Test data Vector Machines, K-Nearest Neighbors, and Deep Neural Network in R programming to train the model and conduct experiments using a Kaggle dataset. They evaluate the precision of each algorithm, and Deep Neural Network outperforms the other six methods with an accuracy of 84%. In research in 2016, A. A. Saa was also interested in educational data mining and exploring the factors that are theoretically assumed to influence student learning outcomes in higher education, and finding a deterministic model. classifiers and best predictors of student achievement based on relevant social and individual factors [3]. Z. Iqbal et al in their research in 2019, evaluated various modern machine learning techniques for predicting college students' grades and showed that a Restricted Boltzmann Machine (RBM) could more accurately predict student grades [4]. Beside that, many machine learning models have been tested by many authors to predict student learning outcomes in their studies: S. Huang and N. Fang developed and compare four types of mathematical models to predict student learning outcomes including multilinear regression model, multilayer cognitive network model, basis function network model. radial and support vector machine models [5]; Z. Iqbal et al use the techniques Collaborative Filtering (CF), Factor Matrix (MF) and Restricted Boltzmann Machine (RBM) to systematically analyze student learning outcomes. Accordingly, the RBM technique is said to be better than other techniques used to predict student achievement in a particular course [6]; J. Dhilipan et al used binomial logical regression models, decision trees, and Entropy and KNN classifiers to classify students. Besides, to support students to achieve higher scores, recognize final grades and improve students' academic conduct [7]; L. F. Ettinger explores academic literature published between 2010 and 2016 addressing student-led data analysis with the aim of providing early intervention to promote better academic outcomes [8]. In a recent study, S. Alturki et al., from a new perspective, provided a comprehensive foundation for understanding Educational Data Mining (EDM) [9]. In addition, the use of data mining methods in studies to predict student learning outcomes has also been pointed out by many authors in recent studies [10-12].

In this study, we utilize an academic dataset from the Faculty of Information Technology, Dainam University, Hanoi, Vietnam. We experiment with some fundamental machine learning techniques and apply the parameter optimization approach to determine the optimal strategy. Then, we propose a model to notify students about their risk levels derived from their academic performance in their first year of college.

The remainder of the paper is structured as follows. In Section 2 we describe a Machine Learning pipeline for predicting the probability of students graduating late. Then we present the experiment procedures and discuss the results in Section 3. Finally, we conclude our paper in Section 4.

2. Machine learning framework predicts the risk of late graduation of students

The system that predicts the probability of students graduating late is separated into two phases: Training and Testing. The Training phase utilizes student data that has been tagged with learning ratings. It will be preprocessed before being fed into the training model. The best model will then be found through parameter optimization. In the Testing phase, the test dataset is utilized to evaluate the accuracy of the chosen model in

the previous phase.

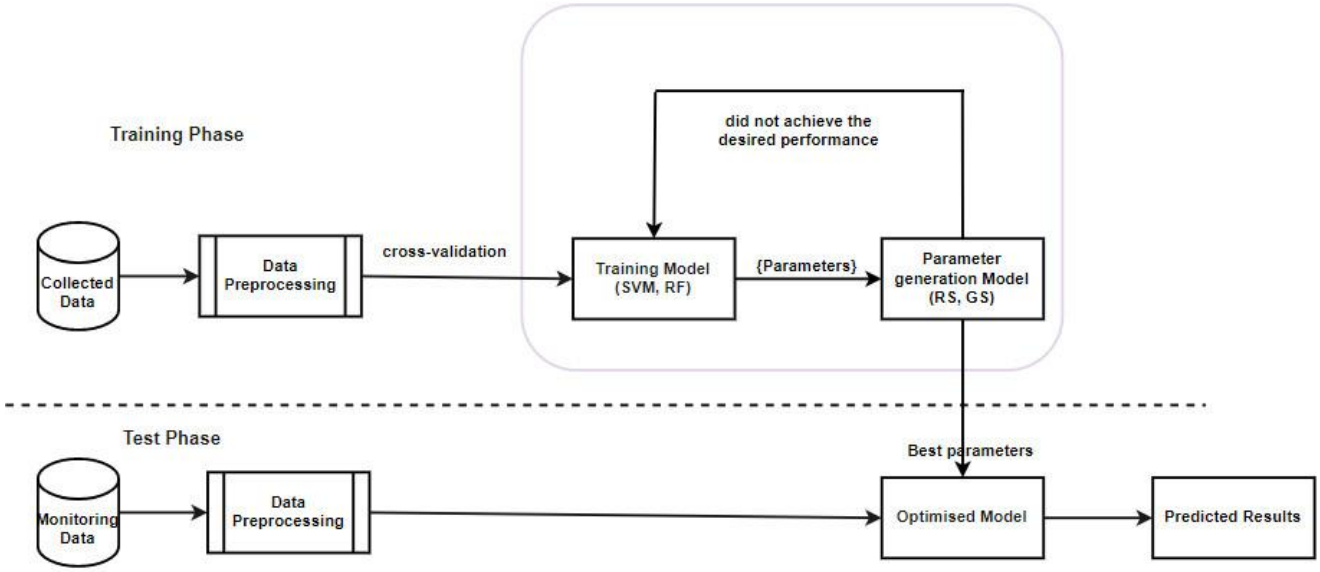

Figure 1. The Optimal machine learning framework predicts the risk of late graduation of students

To find the best model we concentrate on two factors. The first involves data processing, data normalization, and data labeling. The second is to select the optimal set of model parameters. During the Training phase, preprocessed data is utilized to evaluate parameters set generated by the optimization algorithm. The square frame in Figure 1 shows the model optimization process by comparing all metric values and choosing the hyperparameter configuration that produces the best metric value.

In this study, we manage to find the optimal parameters set of Support Vector Machine(SVM) and RF algorithms using Grid Search(GS) and Random Search(RS) techniques, respectively.

3. Experiments and results

3.1. Dataset

The dataset contains the entire score data of 71 subjects and 241 students from the Faculty of Information Technology, DU. The dataset's attributes are not the same among courses, classes, and students. To address this issue, we divided the subjects into groups and determined the average score for each group. These are foreign languages, mathematics, politics, programming, and skills subjects. In addition, we chose 15 more major subjects that had been trained in the first or second year. In the case of missing data in any subject, we supplement with the subject's average score.

With the collected data set, the GPA scores of 241 students are shown in Figure 2.

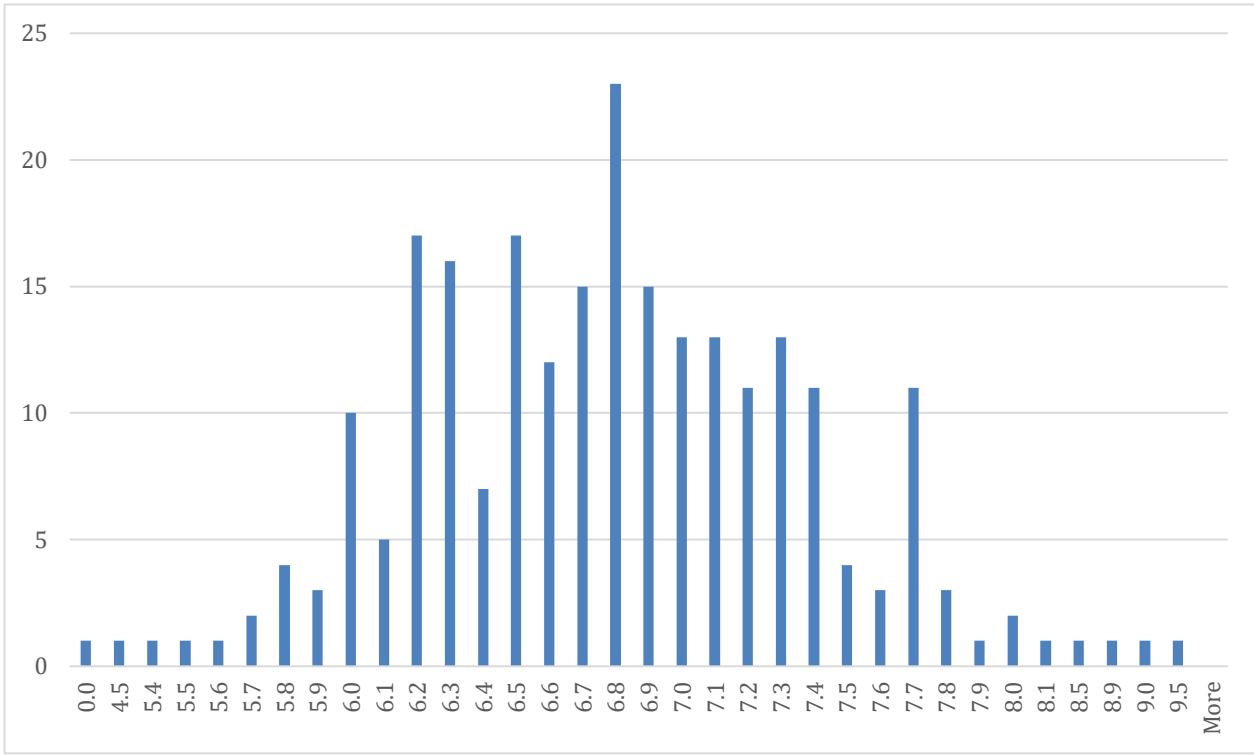

Figure 2. GPA Score Frequency Histogram

To predict students at risk of late graduation, we divided the output score data into two groups: Risky and Safety. All students with an output score smaller than 6.0 (by a 10-score scale) are classified as Risky. Safety students are those who have an output score greater than or equal to 6.0 (by a 10-score scale). According to that, there are 213 students in the Safe group and 28 students in the Risky group. Figure 3 shows the number of students divided into each group.

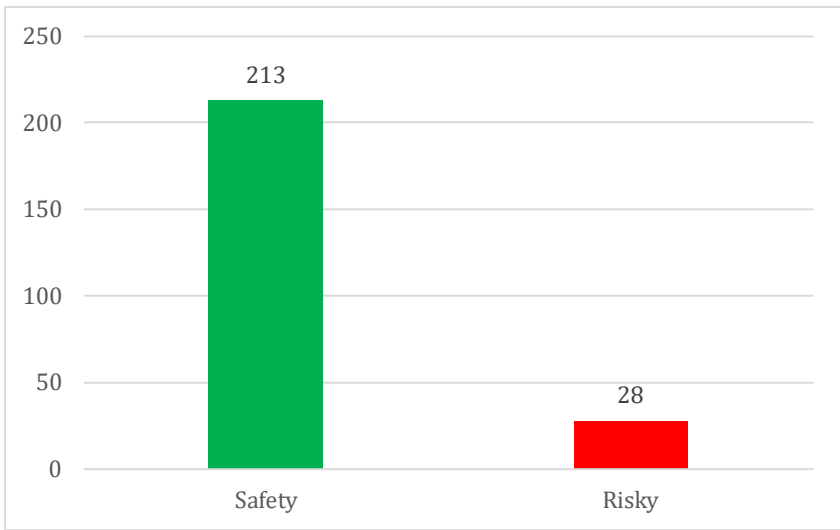

Figure 3. The ratio of safety/risky samples in the collected dataset

Figure 3 shows the unbalanced data. Due to such data characteristics, in the next section, when conducting model evaluation, we used the F1 score index.

3.2. Experimental results and discussions

We experimented with several machine learning methods such as RF and SVM. Then we utilized parameter optimization techniques like GS and RS to get the best solution. Finally, we analyze the optimized model to choose the most optimal model.

We carried out some machine learning algorithms which are RF and SVM. The optimum hyperparameter combination is then selected using hyperparameter optimization methods such as GS and RS. Finally, we examine the post-trained models to choose the best model.

According to the results shown in Figure 4, the SVM&GS, RF&GS, and RF&RS models give more accurate predictions of the Risky class than the other models. For the Safety class, the prediction results are accurate in most of the models, but in the SVM&GS model, the prediction results for the Safety class are worse than the rest of the models.

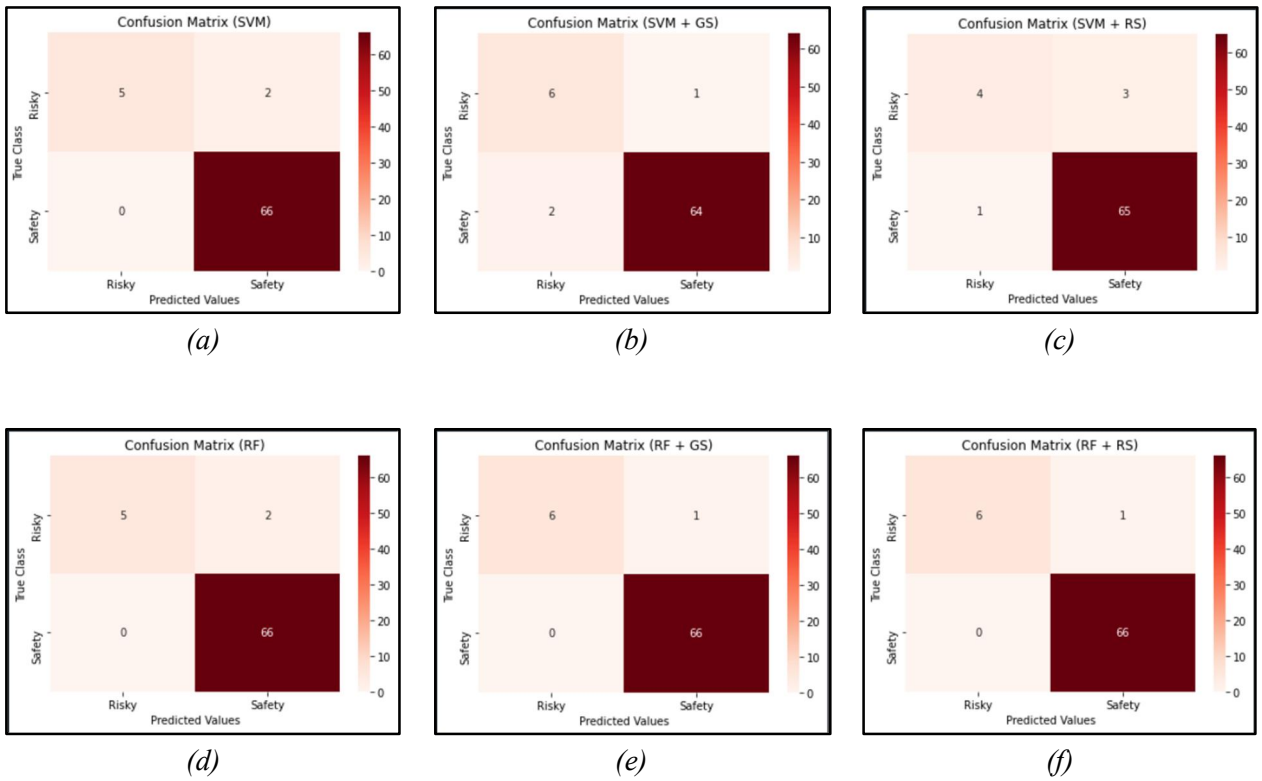

Figure 4. Confusion matric of (a) SVM algorithm, (b) GS combined with SVM, (c) RS combined with SVM, (d) RF algorithm, (e) GS combined with RF, (f) RS combined with RF

According to the results obtained in Table 1, we conclude that all models have high accuracy (all above 95%). The SVM&RS model has the lowest accuracy of 0.95. When looking at F1 scores, all samples in the Safe class have very high F1 scores (all above 0.96). However, the F1 score for the Risk class of some internal models is still low, the SVM & RS model has an F1 score of 0.67, while other models achieve quite high results such as the RF&GS model and RF&RS is the F1 score of 0.92.

Table 1. F1-score, Accuracy and Execution time

	Class	SVM	RF	SVM &GS	SVM &RS	RF &GS	RF &RS
F1-score	Safety	0.99	0.96	0.98	0.97	0.99	0.99
	Risk	0.83	0.67	0.80	0.71	0.92	0.92
Accuracy		0.97	0.95	0.96	0.95	0.99	0.99
Exe. time (s)		0.17	0.12	0.69	3.6	192.26	23.18

When comparing the execution time, it shows that the execution time in different models is quite different. The execution time of the SVM and RF classification models is very fast (both under 0.2s). With optimal models such as SVM&GS; SVM&RS; RF&GS; RF&RS execution time is also quite different. The RF&GS model shows the largest execution time (192.26s) while the SVM&GS model has the smallest execution time of only 0.69s and the RF&RS model used 23.18s. The reason for this can be seen that GS is a comprehensive algorithm that covers all combinations, so it can actually find the best in the domain. But this is also its major drawback is that it is very slow and computationally large. Testing every combination of hyperparameters requires a lot of time which is sometimes not available. Random Search is similar to grid search, but instead of using all points in the grid, it only checks a randomly selected subset of these points. The smaller this subset is, the faster but less precise the optimization will be. The larger this dataset, the more accurate the optimization but the closer it is to grid search. RS probably won't be the best score, but it can still be a good set of values that gives us a good model.

So based on the results given in Table 1, RF + RS was chosen because it has the best results and relatively fast running time. However, with the current small data set, all models have small execution times, and the results are almost instantaneous.

4. Conclusion

In this work, we applied several machine learning approaches to predict student learning outcomes. After conducting tests on the models, the RF and SMV models gave the best output results, with an accuracy of over 95%. We then applied the optimization techniques GS and RS to get better results. Specifically, with SVM&RS models; SVM&GS; RF&RS, and RF&GS accuracy are both above 95%, and even with RF&RS and RF&GS models, accuracy reaches 99%. However, the above results are performed on a limited data set and a small amount of data. Therefore, the obtained results are still limited. In addition, practice shows that learning outcomes do not always fully reflect the impact on student learning outcomes. This will be considered by us in future research. In future studies on this issue, we will add to the data attributes that have been shown to have an impact on student achievement such as family origin, gender, enrollment region, ...

References

[1] Callender, Claire and Feldman, Rayah, “Part-time undergraduates in higher education: a literature review | VOCEDplus, the international tertiary education and research database.” https://www.voced.edu.au/content/ngv%3A65352 (accessed Oct. 19, 2021).

[2] Z. Iqbal, A. Qayyum, S. Latif, and J. Qadir, “Early Student Grade Prediction: An Empirical Study,” Feb. 2019, pp. 1–7. doi: 10.23919/ICACS.2019.8689136.

[3] M. Sweeney, J. Lester, H. Rangwala, and A. Johri, “Next-Term Student Performance Prediction: A Recommender Systems Approach,” 2016. doi: 10.5281/ZENODO.3554603.

[4] Vijayalakshmi V. and K. Venkatachalapathy, "Comparison of Predicting Student's Performance using Machine Learning Algorithms," 2019, doi: 10.5815/ijisa.2019.12.04.

[5] U. Awan and I. Knight, "Domestic sector energy demand and prediction models for Punjab Pakistan," *Journal of Building Engineering*, vol. 32, p. 101790, Nov. 2020, doi: 10.1016/j.jobe.2020.101790.

[6] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, "Machine Learning Based Student Grade Prediction: A Case Study," *arXiv:1708.08744 [cs]*, Aug. 2017, Accessed: Oct. 19, 2021. [Online]. Available: http://arxiv.org/abs/1708.08744

[7] J. Dhilipan, N. Vijayalakshmi, S. Suriya, and A. Christopher, "Prediction of Students Performance using Machine learning," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1055, no. 1, p. 012122, Feb. 2021, doi: 10.1088/1757-899X/1055/1/012122.

[8] L. F. Ettinger, "Using Learning Analytics to Predict Academic Outcomes of First-year Students in Higher Education," p. 53.

[9] S. Alturki, I. Hulpuș, and H. Stuckenschmidt, "Predicting Academic Outcomes: A Survey from 2007 Till 2018," *Technology, Knowledge and Learning*, Sep. 2020, Accessed: Feb. 15, 2022. [Online]. Available: https://www.academia.edu/54133587/Predicting_Academic_Outcomes_A_Survey_from_2007_Till_2018

[10] R. O. Aluko, E. I. Daniel, O. Shamsideen Oshodi, C. O. Aigbavboa, and A. O. Abisuga, "Towards reliable prediction of academic performance of architecture students using data mining techniques," *Journal of Engineering, Design and Technology*, vol. 16, no. 3, pp. 385–397, Jan. 2018, doi: 10.1108/JEDT-08-2017-0081.

[11] S. Alturki and N. Alturki, "Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions," *JITE:IIP*, vol. 20, pp. 121–137, 2021, doi: 10.28945/4835.

[12] L. F. Ettinger, "Using Learning Analytics to Predict Academic Outcomes of First-year Students in Higher Education," p. 53.