International Journal of Internet, Broadcasting and Communication Vol.15 No.2 261-267 (2023)
http://dx.doi.org/10.7236/IJIBC.2023.15.2.261

IJIBC 23-2-30

Applying Machine Learning approaches to predict High-school Student Assessment scores based on high school transcript records

Nguyen Ba Tiến[1], Hoai-Nam Nguyen[2,3], Hoang-Ha Le[2], Tran Thu Trang[4], Chau Van Dinh[5], Ha-Nam Nguyen[5] and Gyoo Seok Choi[6]

[1] Center for Educational Testing Vietnam National University Hanoi, Vietnam
[2] VNU University of Education, Vietnam National University in Hanoi, Vietnam
[3] ICT Department, Ministry of Education and Training, Vietnam
[4] Faculty of Information Technology, Dainam university, Hanoi, Vietnam
[5] Faculty of Information Technology, Electric Power University, Hanoi Vietnam
[6] Department of Computer Science, Chungwoon University, Incheon, Korea

e-mail: tiennb@vnu.edu.vn, nam.moet@gmail.com; trangtt@dainam.edu.vn; chaudv@epu.edu.vn; hoangleha183@gmail.com; namnhvn@epu.edu.vn; chaudv@epu.edu.vn;lionel@chungwoon.ac.kr

Corresponding Author: Gyoo-Seok Choi lionel@chungwoon.ac.kr
Ha-Nam Nguyen namnhvn@epu.edu.vn

Abstract

A common approach to the problem of predicting student test scores is based on the student's previous educational history. In this study, high school transcripts of about two thousand candidates, who took the High-school Student Assessment (HSA) were collected. The data were estimated through building a regression model - Random Forest and optimizing the model's parameters based on Genetic Algorithm (GA) to predict the HSA scores. The RMSE (Root Mean Square Error) measure of the predictive models was used to evaluate the model's performance.

Keywords: *Student Assessment; Machine learning; Entrance Examination.*

1. Introduction

In recent years, education in Vietnam has experienced lots of changes in methods of selecting students to admit to higher education training programs. Besides the traditional method that used the results students got from the high school graduation subject-based exams as the only information resource for selection, individual universities have developed various alternatives to serve this purpose, of which a popular way is using the scores students got from an independent assessment developed and administered by a professional testing organization. There was such an independent assessment called High School Assessment (HSA). In 2022, there were 62,633 students taking HSA, and more than 30 universities accepted this instead of entrance examination. The assessment was developed to measure three groups of competencies: (i) creativity and problem-solving; (ii) linguistic thinking, reasoning, logic, computation and data processing; and (iii) self-study, self-discovery, and application of technology, natural and social sciences. The test consisted of three parts: quantitative thinking, qualitative thinking, and scientific, with the total raw score of 150.

Manuscript Received: April. 21, 2023 / Revised: April. 25, 2023 / Accepted: April. 27, 2023
Corresponding Author: lionel@chungwoon.ac.kr(Gyoo-Seok Choi), namnhvn@epu.edu.vn(Ha-Nam Nguyen)
Tel:*** - **** - **** +84-90-467-0471, Fax: +82-32-770-8165
Professor, Department of Computer Science, Chungwoon University, Incheon, Korea
Professor, Faculty of Information Technology, Electric Power University, Hanoi Vietnam

Many studies have been carried out to build models to predict test scores of students [1]-[6]. The results from these models could assist students in self-evaluating their abilities and offering more appropriate learning routes. In 2016, Mark Sweeney and his colleagues developed a method to predict student grades [1]. They tried many classical and modern methods such as Factorization Machines (FM) and Random Forest (RF) and obtained better results compared to tradition approaches. In 2019, Vijayalakshmi V. proposed a system to predict student performance using Deep Neural Network [2]. They used a variety of techniques which included Decision Tree (C5.0), Naive Bayes, Random Forest, Support Vector Machine, K-Nearest, and Deep Neural Network to conduct the experiment by using a Kalboard 360 Dataset[1]. It is an educational dataset accumulated from learning the board structure called kalboard 360. The instructive accumulation expands into 500 students with 17 features. The highlights are assembled into three essential classes: the first is Demographic features, for example, nationality and sexual orientation… The experimental results shown that the Deep Neural Network outperformed among the other six methods with an accuracy of 84%. In the 2016 study, A. A. Saa was also interested in exploiting education data and exploring the factors that were theoretically assumed to affect student achievement at higher education. They expected to develop a better model which could categorize and predict students' learning achievement based on relevant social and personal factors [3]. Z. Iqbal et al., in their 2019 study, evaluated various modern machine learning techniques to predict college student grades and found that the Restricted Boltzmann Machine (RBM) model could predict students' scores more accurately [4]. Pentel Avar presented a linear regression model that utilized data generated by the activities of students in two courses to predict their final exam scores [5]. For similar purpose, Fiseha Berhanu and Addisalem Abera used a concept of educational data mining students' performance to predict students' learning achievement, based on their academic record, using a decision tree algorithm [6].

Come up with the practical enrollment needs mentioned above, this study aimed to develop a model for predicting students' test scores based on their high school transcripts. We proposed a parameter optimization method using Genetic Algorithm (GA) for Random Forest (RF) to find the optimal parameters for HSA score prediction model. By utilizing these algorithms, suitable parameters can be selected to enhance the accuracy of the test score prediction model.

The remainder of the paper is structured as follows: a Machine Learning pipeline for predicting HSA score is described in section 2. Then, experiment results and discussions will be presented in the next section. Finally, the conclusion is presented in the last section.

2. Machine learning framework to predicts HSA score

We proposed the machine learning framework (Figure 1) with four phases:

- The first phase involves data exploit consisting of extracting, normalization, and data labeling.
- The second phase is to manage the optimal parameters set of models for fitting the training data.
- The third phase is to evaluate the models.
- The last phase is to predict the results with a new raw dataset.

[1] https://www.kaggle.com/aljarah/xAPI-Edu-Data

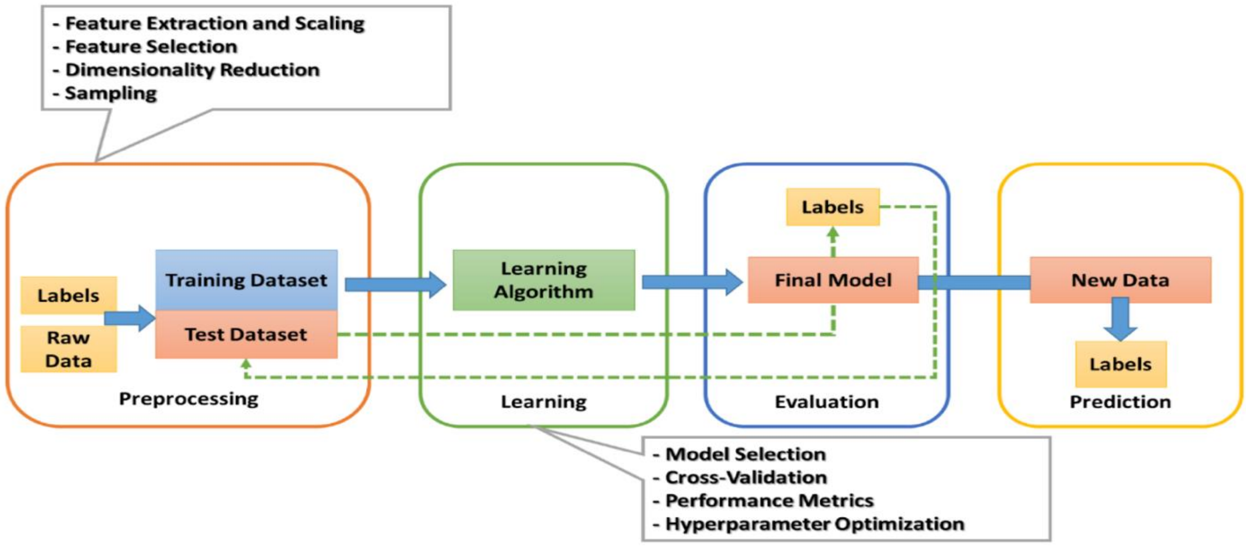

Figure 1. The optimal machine learning framework predicts HSA score.

In this study, we first attempted to identify suitable machine learning models for the given dataset. Subsequently, we utilized Genetic Algorithm (GA) techniques to determine the optimal parameter set for the selected model. As it can be seen in Figure 2, a genetic algorithm mimics natural selection by evolving a population of individual solutions over time to the current problem until a termination condition is fulfilled and the best individual is returned as result of the algorithm.

One of the significant results from this research is the proposal to apply a genetic algorithm to solve the problem of selecting the best model and optimal parameter set for predicting test scores. Although randomized, the genetic algorithm utilizes historical information to guide the search towards regions of better performance within the search space. The basic techniques of the genetic algorithm are designed to simulate processes in natural systems that are necessary for evolution. A genetic algorithm has several strategy parameters, including population size, mutation operator and mutation rate, crossover operator and crossover rate, selection mechanism and selective pressure, etc. An adequate parameter setting can be crucial in obtaining high-quality solutions.

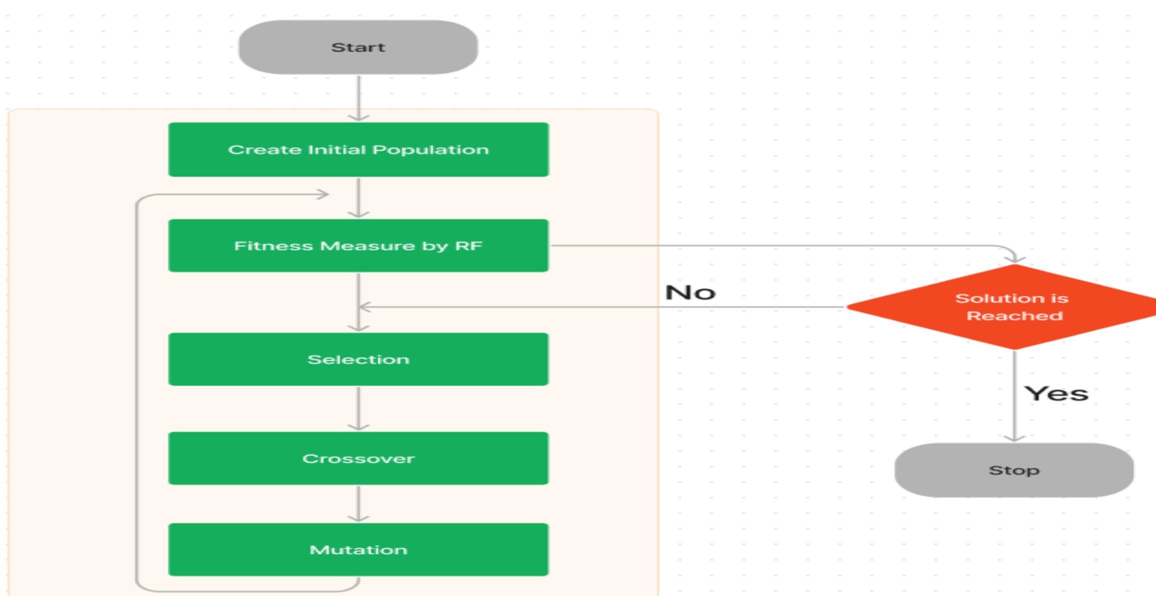

Figure 2. Algorithm for finding the best parameter set for RF based on GA

The parameters used in our implementation were:

- Population size: This parameter controls the sample size for each population. The population size used here was 100 individuals.
- Crossover operator and crossover rate: The crossover operator creates a new chromosome by combining parts of two (or more) parent chromosomes. In this paper, linear crossover, which consists of taking two chromosome (treating it as vectors) and creating a linear combination of these vectors as result, was used. The crossover rate used here was 0.8 (80%).
- Mutation operator and mutation rate: The mutation operator mutates a specific gene across the entire population, preventing the population from converging to a local minimum by keeping the solutions from becoming too similar to each other. While most of the search is performed through crossover, mutation is crucial in maintaining diversity within the population. The mutation rate used in this study was 0.03 (3%).
- Selection mechanism: selectors are responsible for selecting a given number of individuals from the population, then obtaining survivors and offspring. The selection mechanism used here was the roulette-wheel selector, which is a fitness proportional selector that applies less selective pressure over than other strategies such as tournament selector.
- Termination condition: termination condition is the criteria to determine when the Genetic Algorithm should end. The termination condition used here was that the algorithm to stop its steady state. In our case, the algorithm stopped when it reached the 60th generation without improvement.

3. Experiments and results

3.1. Dataset

The dataset used in this experiment was collected from 1954 candidates who participated in an independent competency assessment (High School Assessment – HSA) conducted for the purpose of selecting high school graduates for admission to universities in 2022. The dataset consists of 35 attributions, of which all reflected students' personal and learning achievement attributes, as follows:

- The first attribution for gender
- The next 33 attributions are presented the academic records of 3 years high school achievement of students consisted of annual GPA, annual final grades in Mathematics, Vietnamese and Literature, Physics, Chemistry, Biology, History, Geography, Citizenship Education, and Foreign Languages (9 subjects), and personal annual academic qualification (i.e. excellent, very good, good, passable, bad…).
- The last attribution for HSA value

The HSA also covers these 9 subjects. The dataset is ensured of being comprehensive, accurate and possessing no missing data. Thanks to the provision of data directly from the management system.

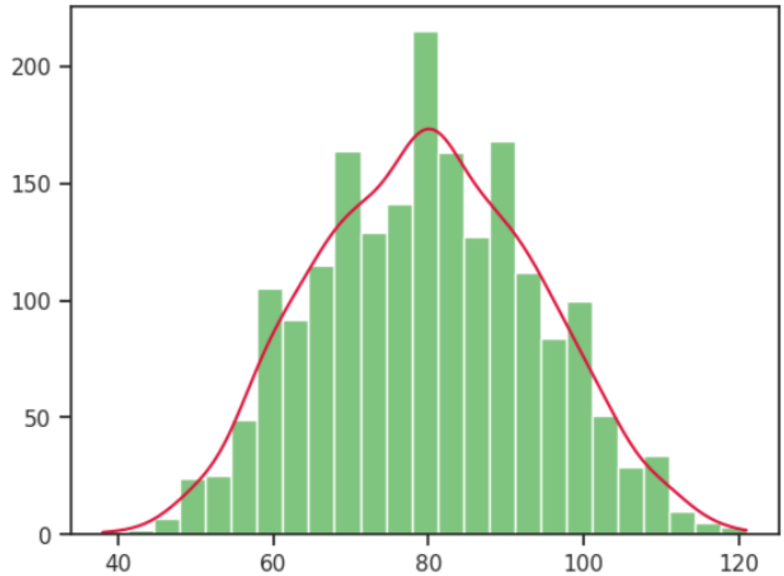

Figure 3. HSA Score Frequency Histogram with around two thousand candidates

Figure 3 represents the test has a proper discrimination with the raw scores range from 40 to 120, with a focus ranges from 70-90 and fewer scores falling below 60 and over 100.

3.2. Experimental results and discussions

We tried with several machine learning methods, including Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), Artificial Neural Networks (ANN), and Gradient Boosting (GB).

Table 1. The result of ML methods with default parameters to deal with dataset

	SVM	LR	RF	ANN	GB
RMSE	13.76	12.36	9.85	12.21	10.61

The Table 1 showed that the RF model returned the best performance. While the GB model's results are the second and the results of SVM model is the worst among 5 models. Then, we applied some parameters optimal methods like Grid Search, Random Search and Genetic Algorithm to find the optimal hyper-parameter set of RF to deal with this data set.

Table 2. Results of parameters tuning based on GA, GS, RS

	RF + Random Search	RF + Grid Search	RF + Genetic Alg.
RMSE	9.81	9.8	9.66
Exe. time (s)	145	1152	6181

The results in Table 2 showed the RMSE and execute time of the combination of Random search (RS), Grid search (GS), and Genetic algorithm (GA). The RF + GA model was found to be the highest accurate, although it took the longest time to run. It spent very long time for obtaining the optimal RMSE value of 9.66.

When comparing the execution time, it shows that the execution time in different approaches is different. The RF + GA model shows the longest execution time (6181s) while the RF + RS model has the shortest execution time of only 145s and the RF + GS model took about 1152s. The reason why tuning hyperparameters in a GA takes a long time is that the search space for the hyperparameters can be very large, therefore the effects of changing one hyperparameter are more likely to alter the effects of another hyperparameter. This means that finding the best combination of hyperparameters requires a large number of experiments, which can be time-consuming. 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 is, the more accurate the optimization can get but the closer it is to Grid search. Although RS probably won't be the best score, it can still be a good set of values that gives us a good model.

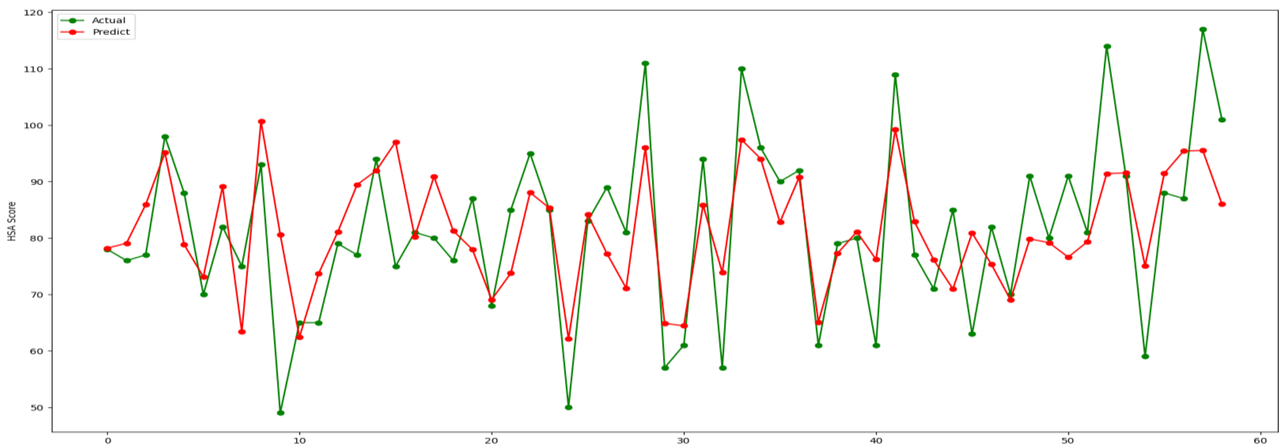

Figure 6. Comparison of the predicted results and actual results with optimal parameters using RF + GA

Based on the results given in Table 2, RF + GA was chosen because of being shown best results. Because of the small data set, all models have small execution times, and the results are almost instantaneous.

4. Conclusion

There are many approaches, frameworks as well as platforms to predict HAS score based on Machine Learning techniques. In this paper, we proposed a novel method to find the optimal hyper-parameters of Machine Learning algorithms to predict HAS score based on tuning methods such as Grid Search, Random Search and Genetic Algorithm. Firstly, five conversion models consisting of SVM, LR, RF, ANN and GB are implemented. The obtained results shown that the Random Forest algorithm gave relatively good results with a RMSE of 9.66. Then we applied the tuning techniques consist of Grid Search, Random Search to manage the optimal results. Then, we proposed a novel method using Genetic Algorithm to manage the optimal parameters of the RF model to enhance the accuracy of predicting HSA results based on students' high school. Experimental results on the aforementioned dataset have demonstrated significant outcomes, with RMSE is 9.66, much lower than that of other studies.

In the future research, we plan to incorporate additional attributes including family background, gender, and enrollment region, which have been demonstrated to impact student achievement.

References

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

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

[3] A. A. Saa, "Educational Data Mining & Students' Performance Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, p. 9, 2016.

[4] 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.

[5] S. Huang and N. Fang, “Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models,” *Computers & Education*, vol. 61, no. 1, pp. 133–145, Feb. 2013.

[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.

[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] D. Kabakchieva, “Predicting Student Performance by Using Data Mining Methods for Classification,” *Cybernetics and Information Technologies*, vol. 13, Match 2013, doi: 10.2478/cait-2013-0006.