Predicting Academic Performance of High School Students

Nguyen Dinh-Thanh[1] and Pham Thi-Ngoc-Diem[2(✉)]

[1] Song Doc High School, Ca Mau, Ca Mau Province, Vietnam
ndthanh@camau.edu.vn
[2] College of Information and Communication Technology, Can Tho University, Can Tho, Vietnam
ptndiem@ctu.edu.vn

Abstract. Students' weak learning ability is a problem that occurs in most countries around the worldwide and leads to many bad effects on students such as boredom leading to dropout, guilt with friends and with many other students. Students' poor academic results will greatly affect the teaching effectiveness and the reputation of the school. Therefore, predicting the student learning outcomes in high school can help educators to find innovative and effective solutions to support teachers, students in improving the learning and teaching quality in high schools. In this work, machine learning models were used to predict academic performance of high school students. These models were built from a dataset of 21,222 student records with 2,545 (11.99%) very good students, 7,859 (37.03%) good students, 8,099 (38.16%) average students, 2,531 (11.93%) poor students and 188 (0.89%) very poor students in high schools in Ca Mau province, Vietnam. With the use of the Synthetic Minority Oversampling TEchnique algorithm to balance the dataset before putting it into the machine learning models, the results have shown that the Random Forest, XGBoost, Light GBM models give the best results with the Accuracy of 81.69%, 80.86% and 80.82%. In addition, important features that contribute decisively in predicting academic performance were also extracted, including Grade Point Average (GPA) of semester 1 and 2, Age, Class, Academic Performance of semester 1 and 2, Father's occupation, Mother's occupation and Learning online.

Keywords: Academic performance prediction · Students' academic performance · Features extraction · Machine learning models

1 Introduction

Martinez considers academic performance is "the product given by the students and it is usually expressed through school grades" [11]. In this way, in Vietnam, academic performance of high school students is measured mainly by GPA and is divided into 5 levels of Very good, Good, Average, Poor, Very Poor, described in Table 1.

C. V. Phan and T. D. Nguyen (Eds.): ICTCC 2022, LNICST 473, pp. 123–135, 2023.
https://doi.org/10.1007/978-3-031-28790-9_8

Table 1. Academic levels and conditions

Academic performance	Condition
Very Good	The GPA of all subjects is from 8.0, in which score of each subject is from 6.5, the score of one of the three subjects of Literature, Mathematics, and Foreign Language is at least 8.0 and Physical Education is rated as Passed
Good	The GPA of all subjects is from 6.5, in which score of each subject is from 5.0, the score of one of the three subjects of Literature, Mathematics, and Foreign Language is at least 6.5 and Physical Education is rated as Passed
Average	The GPA of all subjects is from 5.0, in which the score of each subject is from 3.5, the score of one of the three subjects of Literature, Mathematics, and Foreign Language is at least 5.0 and Physical Education is rated as Passed
Poor	Students meet one of the following conditions: – The GPA of all subjects is not enough 5.0 – The score of one of three subjects Literature, Mathematics, and Foreign Language is less than 5.0 – At least one subject has a score of 2.0 to less than 3.5 – Physical Education is not passed
Very Poor	All remaining cases

In addition, there are some cases that are temporarily called downgrade in academic performance levels:

✓ Students are graded as Very Good, but due to a certain subject is graded as Average, so their academic performance is adjusted as Good.
✓ Students are graded as Very Good, but due to a certain subject is graded as Poor, so their academic performance is adjusted as Average.
✓ Students are graded as Good, but due to a certain subject is graded as Poor, so their academic performance is adjusted as Average.

Academic performance plays an important role in a student's subsequent decisions like continuing to go to school or leaving school. In Vietnam's Ca Mau province, high school dropout rate was 5.61% in 2019–2020 school year [9]. This rate is relatively high. Among the factors affecting students' dropping out of school, poor academic performance is the most important factor [2,11]. Moreover, [7,12,14,18] have determined that students' academic performance dramatically influences dropping out of school. The problem of low academic performance can cause negative effects on the family, the school and society such as the emergence of cases of violence, imbalance and in harmonic among the community members, the emergence of social classes, ... [2]. Therefore, early predicting the risk of poor academic performance in high school can help learners to improve their learning and educators to have interventions and efficient

solutions in reducing the highs school dropout rate. This is the main goal of this study.

Nowadays, machine learning has many applications in education such as enrollment management [1], enrollment prediction [13], predicting dropout [4], predicting learning outcomes [10,15], ... In this paper, machine learning algorithms have been applied to build models for predicting student academic performance in high school. They are Decision Tree (DT), Random Forest (RF), XGBoost (XGB), Light GBM (LGBM), Artificial Neuron Network (ANN) and Multilayer Perceptron (MLP). This is a multi-label classification problem where students are classified by the levels of Very Good, Good, Average, Poor, Very Poor.

Single machine learning models as well as machine learning models combined with Bagging were used in this study. The dataset to train and test the models was collected from 12 high schools in Ca Mau province [4]. The data collected include the student's personal information and their academic performance in the two semesters adjacent (called GPA 1 and 2, academic performance 1 and 2) to the semester in which the student academic performance is predicted. This dataset contains 21,222 students, in which the number of Very Good, Good, Average, Poor and Very Poor students are 2,545, 7,859, 8,099, 2,531 and 188 respectively. The experiment results have shown that the RF model is best with 81.69% in Accuracy, 81.47% in Precision, 81.62% in Recall and 81.53% in F1-Score.

This paper is organized as follows. Literature review is presented at Sect. 2, our method is described in Sect. 3. Our experimental results are demonstrated in Sect. 4 and finally, the conclusion and future works are drawn in Sect. 5.

2 Related Work

There are a variety of researches for student academic performance prediction by using machine learning in the recent years. [16] focused on discussing the important attributes used in predicting students' performance and prediction methods used for students' performance. In the secondary education, [19] used decision tree, random forest, and naive Bayes to predict the five-level final grade of students based on their historical data. The experiment results showed the effectiveness of machine learning techniques when predicting the performances of students on two educational datasets related to mathematics lesson and Portuguese language lesson with 33 attributes each. [17] applied three single classifiers including a MLP, J48, and PART, three ensemble algorithms encompassing Bagging, MultiBoost and Voting and nine models developed by the fusion of single and ensemble-based classifiers to predict student performance. The experiment results on 1227 records and 16 attributes showed that MultiBoost with MLP achieved 98.7% Accuracy, 98.6% Precision, Recall, and F1-Score. In high school, [12] presented a use of machine learning for the student performance prediction in technical high school using tree-based methods and obtained prediction results over 89% Accuracy, etc.

Most of the researches focused on predicting student performance in higher education [3,8]. Only some works related to high school [2,12] and secondary school [17,19] students. These studies have used machine learning algorithms as well as deep learning to build student performance prediction models. The GPA, gender, age, income, nationality, marital status, employment status, attendance are attributes used in predicting student academic performance [3]. In addition, academic performance prediction can be conducted with only one subject or based on many subjects. The prediction results depend on many different factors such as the size of the dataset, the student's features, the machine learning techniques, data processing methods, etc. Furthermore, the machine learning techniques can be used individually or in combination with others. This paper presents a comparative analysis of six machine learning algorithms for early detection at the risk of low academic performance of high school students in the next semester using a dataset collected in the two previous semesters and student features related to academic performance, personal information, and family and high schools, etc.

3 Proposed Method

The main flow of the student academic performance prediction system is shown in Fig. 1. The gathered data were divided into two datasets, one for training, and the other for testing with machine learning algorithms such as DT, RF, XGB, LGBM, ANN and MLP.

The raw data is collected from a variety of high schools in Ca Mau. Each data item in this raw dataset contains many attributes of a student including the academic performance attribute. An attribute can take a numeric, string, or character value. Then, this raw dataset cannot be used to build machine learning models. The pre-processing of this dataset is necessary. The dataset obtained after pre-processing is separated into two datasets, called the D_train and D_test. During the training phase, the D_train is used to train and build the models. In the testing phase, the trained models are tested with all data point in the D_test. Every model produces a predictive value called P_predict indicating student's academic performance.

In this research, metrics of Accuracy, Precision, Recall, F1-Score are used to evaluate models while the selection of an appropriate model for predicting student academic performance in high school in Ca Mau province is based on D_test dataset and P_predict. The following sections presents the methods implemented in detail.

3.1 Data Collection

The dataset to train and test the models was collected from 12 high schools in Ca Mau province in the 2019–2020 school year [4]. This dataset is supplemented with student records in 2020–2021 school year and with many attributes. These data were processed and converted to .xlsx format. This .xlsx file called raw dataset. The details of attributes in the raw dataset are described in Table 2.

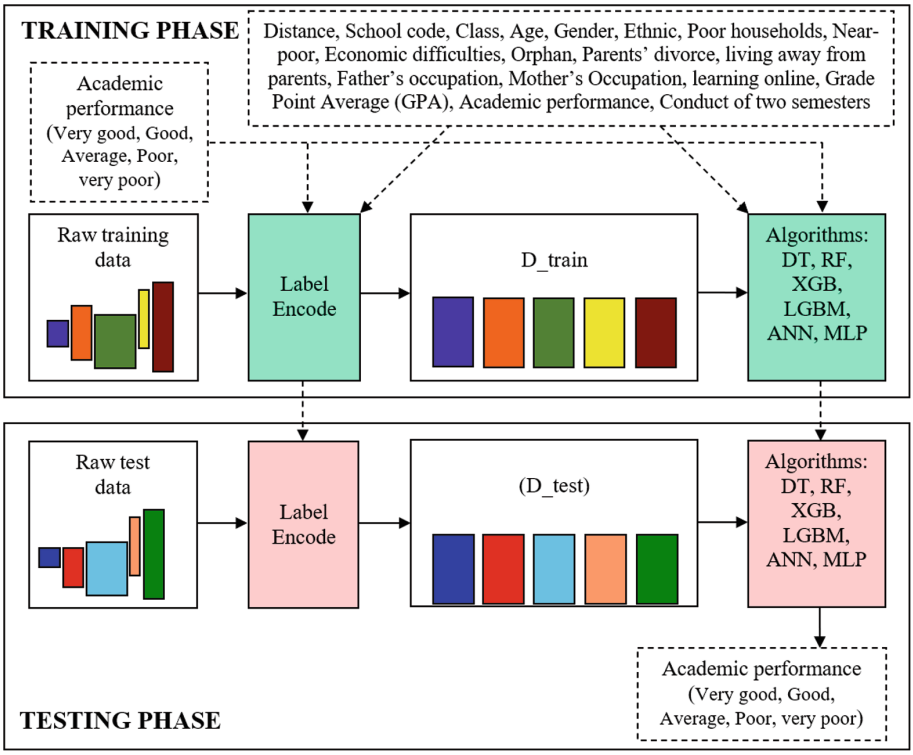

Fig. 1. Framework for student academic performance prediction

3.2 Data Pre-processing

The label encoding (LabelEncode) [5] method is applied on raw dataset before it is used to build machine learning models. An attribute in the raw dataset may have a defined set of available values. Each of these values will be encoded with a number. The resulting dataset is called the label-encoded dataset. For example, an academic performance attribute may be limited to the values Very Good, Good, Average, Poor and Very Poor in the raw dataset, then it will take one of the corresponding values of 4, 3, 2, 1 and 0 in the label-encoded dataset. The latter one includes feature vectors, each of which has the form (Distance, School Code, Class, Age, Gender, Ethnic, Poor households, Near-poor, Economic Difficulties, Orphan, Parents' divorce, Live away from parents, Father's occupation, Mother's occupation, learning online, GPA 1, Academic Performance 1, Conduct of semester 1, GPA 2, Academic Performance 2, Conduct of semester 2, and Academic Performance). More specifically, a data item in the label-encoded dataset looks like (24, 006, 10, 16, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0.5, 6.5, 2, 3, 7.5, 3, 3).

Table 2. Description of the attributes in the raw dataset

Feature	Value range
Distance	The distance from the student's school to the center of Ca Mau City (unit is kilometers).
School Code	The code of high school (described by three numeric characters) of the considered student
Class	Integer: From 10 to 12
Age	Integer: From 16 to 20
Gender	Male/Female
Ethnic	Text
Poor households	x or empty, with "x", is YES, and empty is NO
Near poor	x or empty, with "x", is YES, and empty is NO
Economic Difficulties	x or empty, with "x", is YES, and empty is NO
Orphan	x or empty, with "x", is YES, and empty is NO
Parents' divorce	x or empty, with "x", is YES, and empty is NO
Living away from parents	x or empty, with "x", is YES, and empty is NO
Father's occupation	Text
Mother's occupation	Text
Learning online	Real: Rate of learning online time
Grade Point Average 1	Real: from 0.1 to 10.0
Academic Performance 1	Very Good, Good, Average, Poor, Very Poor
Conduct of semester 1	Very Good, Good, Average, Poor
Grade Point Average 2	Real: from 0.1 to 10.0
Academic Performance 2	Very Good, Good, Average, Poor, Very Poor
Conduct of semester 2	Very Good, Good, Average, Poor
Academic Performance	Very Good, Good, Average, Poor, Very Poor

3.3 Using Machine Learning Models

In this work, machine learning algorithms used for predicting academic performance are DT, RF, XGB, LGBM, ANN and MLP. The Grid-search model hyper-parameter optimization technique was chosen to find the best set of hyper-parameters for each model. The latter one with different parameters was also executed. After testing, the values of parameters are adjusted according to Table 3. With ANN, the network was chosen with 4 layers: the input layer, two hidden layers with 64 neurons each and the output layer with 5 neurons for 5 labels (Very good, Good, Average, Poor, Very Poor) to predict. With MLP, default values are used.

3.4 Processing Imbalance Dataset

The data used in this study includes 21,222 students. Figure 2 shows the number of students in each academic level. It can be concluded that the majority of students are in the Average and Good grades.

As illustrated in Fig. 2, the dataset is imbalanced between the output classes for prediction. Then, the method of balancing the dataset with SMOTE [6]

Table 3. Parameters of the models

Model	Parameters
Decision Tree	max_depth = 21
Random Forest	n_estimators=2000
XGBoost	default
Light GBM	default
Artificial Neural Network	model = tf.keras.models.Sequential() model.add(tf.keras.layers.Dense(64, activation='relu')) model.add(tf.keras.layers.Dense(64, activation='relu')) model.add(tf.keras.layers.Dense(5, activation='sigmoid')) model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
Multilayer Perceptron	hidden_layer_sizes=(100, 100, 100), max_iter=500, alpha=0.0001, solver='adam', verbose=10, random_state=21, tol=0.000000001

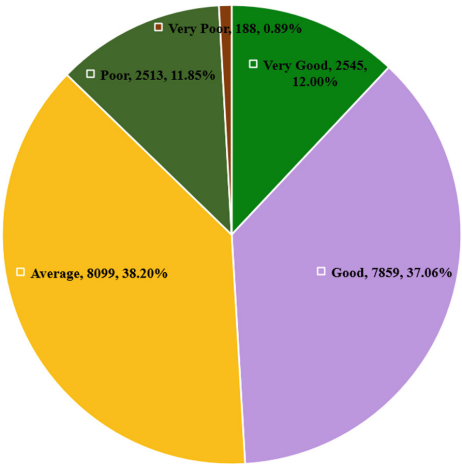

Fig. 2. The graph of the number and percentage of students at different academic levels

is used before putting this dataset into machine learning algorithms to build models. There are many methods of balancing the dataset [6], SMOTE is used because it does not change the original dataset. Moreover, the dataset is small, if Under sampling method is applied, there is not enough data to train the model. After balancing, the size of the dataset is 40,495 (8,099*5=40,495) for 5 labels.

4 Experimental Result

4.1 Machine Learning Model Evaluation

In this study, the dataset is divided into two parts, training and testing at the rate of 30% used as a test set. Several metrics are used to compare and

evaluate the performance of the machine learning algorithms. The dataset for the experiment is imbalanced and processed into a balanced dataset. So, the evaluation measures such as Accuracy, Precision, Recall and F1-Score [5] are applied. Besides, important features are also extracted using the Gini index.

4.2 Experimental Result

The experiment was done in Python programming language (version 3.10) and the scikit-learning library [5] (version 1.0.2). This experiment uses the dataset presented in Sect. 3.1 and machine learning algorithms including DT, RF, XGB, LGBM, ANN and MLP for building models to predict student academic performance levels. Each student record consists of 21 features presented in Sect. 3.1. The experiment is run on a personal laptop configured with Chipset Intel i7 10750H 6-cores 2.6 GHz, 24 GB of memory, and a Windows 10 Home Single operating system.

The method of dividing the dataset by Hold-out [5] is used to split the dataset into a training set at the rate of 70% and testing set. Predictive results are shown in Table 4.

Table 4. Predictive results of all models

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	73.65	73.49	73.58	73.52
Random Forest	81.69	81.47	81.62	81.53
XGBoost	80.86	80.81	80.81	80.78
Light GBM	80.82	80.62	80.76	80.66
Artificial Neural Network	73.04	72.89	72.96	72.86
Multilayer Perceptron	74.32	74.16	74.27	73.92

The Accuracy measure of all models is the best (more than 73%), followed by recall, precision and F1-score measure (more than 72%). The lowest value of recall measure is 72.96% for the ANN model. The high Accuracy measure indicates that the Accuracy of predicting student academic performance is high. Some models have F1-score value that is higher 80% as RF, XGB, LGBM. This result is suitable to evaluate the models as good. A high recall value of these models also shows that the wrong predictive rate is low.

A comparison of Accuracy of all models was illustrated in Fig. 3. Each model has a column representing Accuracy as a percentage (%). Only the Accuracy measures are shown in Fig. 3 because the dataset used in the models has been balanced using SMOTE. So, the values of the Accuracy, Precision, Recall measures or F1-Score are almost the same. Three models with over 80% Accuracy are RF, XGB and LGBM, in which the RF model obtained the highest value (81.69%), while the ANN model obtained the lowest value (73.04%). If the 80%

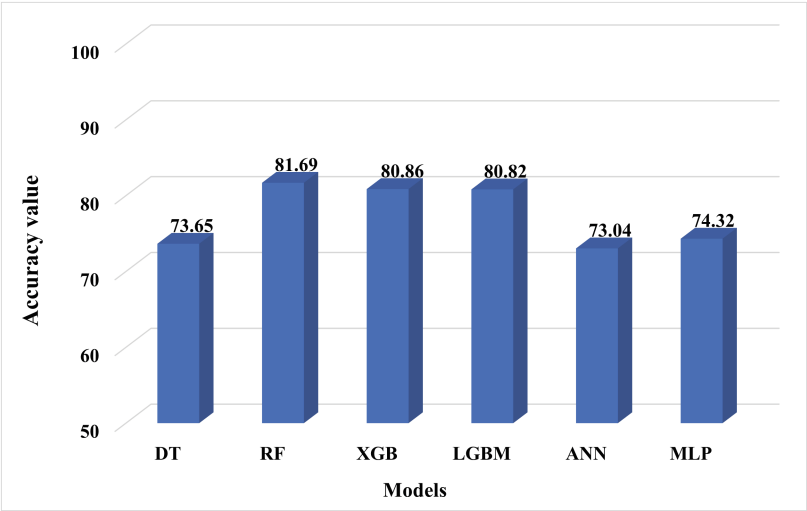

Fig. 3. Comparison of Accuracy of all models

standard is taken at the measures to select the model, the Fig. 3 shows that RF, XGB and LGBM models are qualified. The normalized confusion matrix results of three models are presented in Fig. 4, Fig. 5 and Fig. 6.

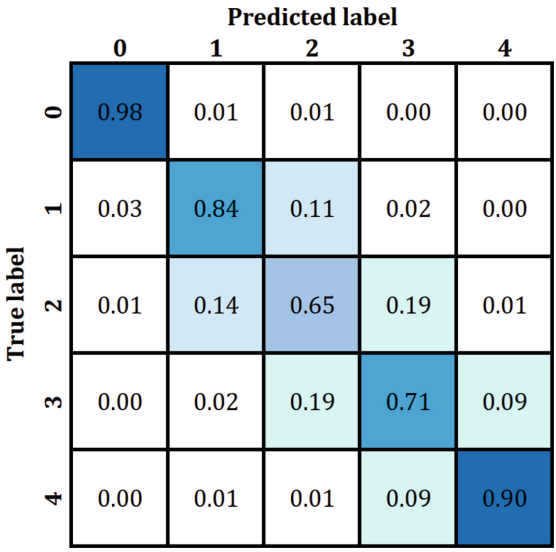

Fig. 4. Normalized confusion matrix of Random Forest model

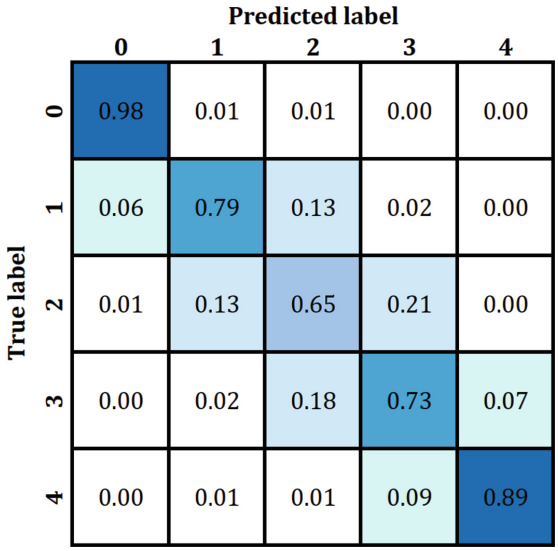

Fig. 5. Normalized confusion matrix of XGBoost model

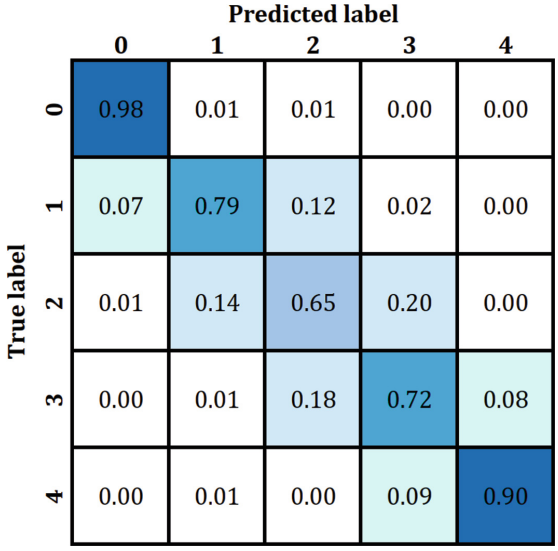

Fig. 6. Normalized confusion matrix of Light GBM model

As illustrated in Fig. 4, Fig. 5 and Fig. 6, the predictive results of Very Good (4) and Very Poor (0) levels are very high (over 90% except for the XGBoost model). The Average level is the lowest (only 65%). The objective of this study is to predict students with low academic performance (Poor and Very Poor levels) so that teachers and educators take appropriate measures to improve students

learning outcomes. The prediction Accuracy at these two levels is high (over 79% for Poor level and 98% for Very Poor level). The prediction error through the Average and Good levels is small (from 0.14% to 0.16%). That means with this prediction result, teachers can focus more on most students at risk of low academy performance to help learners to improve their learning. So, for this study, RF, XGB and LGBM models are selected in predicting student academic performance in high school.

Furthermore, the Accuracy measure in Average level is not good (65%). This can be explained as follows. The score range between Average and Good levels as well as between Average and Poor levels is quite unclear. Because students can be downgraded from Good level to Average level, so the error of predicting a student from Average to Good level is still high (approximately 0.21% in XGBoost model).

4.3 Important Features

The technique of extracting the features aims to find important features and to reduce the number of features in a dataset. During the experimental process, the features extraction was also performed. The RF model was chosen to extract essential features because this model is the best among three models RF, XGB and LGBM. The weight by Gini index operator is applied to calculate the weights of attributes used to build the RF model.

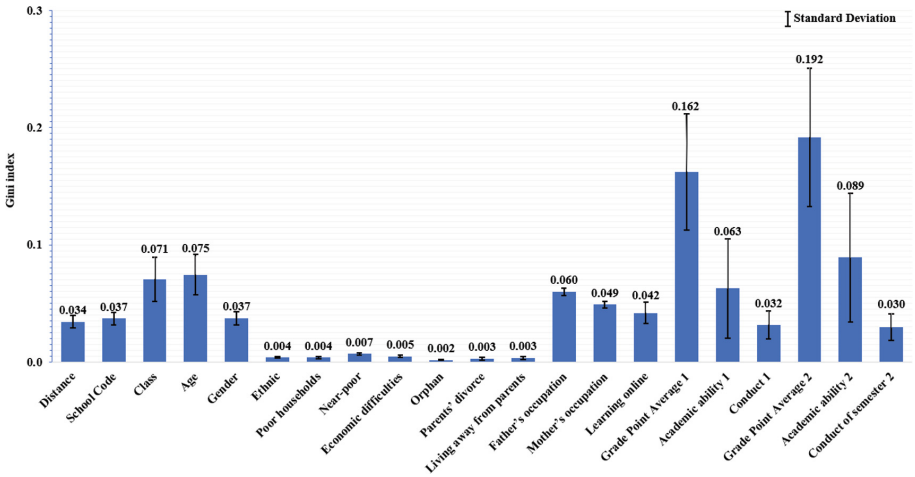

Fig. 7. The result of features extraction in model Random Forest

Figure 7 has illustrated the weights of features in the dataset. The GPA 1 and GPA 2 features are the most important with a Gini coefficient of 0.162 and 0.192, while the Orphan feature is the least important because Gini score is

the smallest (0.002). As shown in Fig. 7, GPA (GPA1 and GPA 2), Age, Class, Academic performance (1 and 2), father's occupation and mother's occupation, learning online are decisive features that play a significant role in determining student academic performance from the dataset.

5 Conclusion and Future Work

In this research, an experiment was done with six machine learning models to predict student academic performance. As a result, three models RF, XGB and LGBM have been chosen. These models obtained performance measures over 80%. More specifically, the RF model is the best among the three models. This model achieved Accuracy, Precision, Recall, F1-Score more than 81% and a higher Accuracy than the two XGB and LGBM models. Moreover, the essential features that significantly influence student academic performance have been also extracted. They are GPA (1 and 2), Age, Class, Academic Ability (1 and 2), father's occupation, mother's occupation and learning online. These futures play an important role in teachers' decision making as well as considering factors affecting student learning.

Predicting student learning results as soon as possible is useful for school administrators as well as educational managers and teachers. Using predicted results, they can recommend remedial strategies and suitable solutions to improve the quality of learning and teaching in their high schools.

In the future, a learning outcomes prediction in each subject will be studied based on using more machine learning models and a bigger dataset to achieve higher measures. The results of this study can also be improved to predict the academic performance of high school students in the Mekong Delta provinces of Vietnam.

References

1. Aksenova, S.S., Zhang, D., Lu, M.: Enrollment prediction through data mining. In: 2006 IEEE International Conference on Information Reuse Integration, pp. 510–515 (2006). https://doi.org/10.1109/IRI.2006.252466
2. Al Zoubi, S., Younes, M.: Low academic achievement: causes and results. Theory Pract. Lang. Stud. **5**, 2262–2268 (2015). https://doi.org/10.17507/tpls.0511.09
3. Alturki, S., Hulpus, I., Stuckenschmidt, H.: Predicting academic outcomes: a survey from 2007 till 2018. Technol. Knowl. Learn. **27** (2022). https://doi.org/10.1007/s10758-020-09476-0
4. Dinh-Thanh, N., Thanh-Hai, N., Thi-Ngoc-Diem, P.: Forecasting and analyzing the risk of dropping out of high school students in Ca Mau Province. In: Dang, T.K., Küng, J., Chung, T.M., Takizawa, M. (eds.) FDSE 2021. CCIS, vol. 1500, pp. 224–237. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-8062-5_15
5. Fabian, P., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**(85), 2825–2830 (2011). http://jmlr.org/papers/v12/pedregosa11a.html
6. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98074-4

7. Goulet, M., Clément, M.-E., Helie, S., Villatte, A.: Longitudinal association between risk profiles, school dropout risk, and substance abuse in adolescence. Child Youth Care Forum **49**(5), 687–706 (2020). https://doi.org/10.1007/s10566-020-09550-9
8. Hellas, A., et al.: Predicting academic performance: a systematic literature review. In: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, pp. 175–199 (2018). https://doi.org/10.1145/3293881.3295783
9. Hoang Du, L.: Report No. 1495/BC-SGDDT dated July 28, 2020, on assessing the performance of tasks for the 2019–2020 school year (2020)
10. Huynh-Ly, T.N., Thai-Nghe, N.: A system for predicting student's course result using a free recommender system library - MyMediaLite. In: Information Technology Conference (2013)
11. Lamas, H.: School performance. Propóitos y Representaciones **3**, 351–386 (2015). https://doi.org/10.20511/pyr2015.v3n1.74
12. de Melo Junior, G., Oliveira, S., Ferreira, C., Filho, E., Calixto, W., Furriel, G.: Evaluation techniques of machine learning in task of reprovation prediction of technical high school students. In: 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), pp. 1–7 (2017). https://doi.org/10.1109/CHILECON.2017.8229739
13. Nandeshwar, A., Chaudhari, S.: Enrollment prediction models using data mining. In: 2006 IEEE International Conference on Information Reuse and Integration (2009)
14. Ogresta, J., Rezo, I., Kožljan, P., Pare, M.H., Ajduković, M.: Why do we drop out? Typology of dropping out of high school. Youth Soc. **53**, 934–954 (2020). https://doi.org/10.1177/0044118X20918435
15. Phuoc Hai, N., Tian-Wei, S.: Predicting the student learning outcomes based on the combination of Taylor approximation method and grey models. VNU J. Sci. Educ. Res. **31**, 70–83 (2015)
16. Shahiri, A., Husain, W., Abdul Rashid, N.: A review on predicting student's performance using data mining techniques. Procedia Comput. Sci. **72**, 414–422 (2015). https://doi.org/10.1016/j.procs.2015.12.157
17. Siddique, A., Jan, A., Majeed, F., Qahmash, A., Quadri, N.N., Wahab, M.: Predicting academic performance using an efficient model based on fusion of classifiers. Appl. Sci. **11**, 11845 (2021). https://doi.org/10.3390/app112411845
18. Stevenson, N., Swain-Bradway, J., LeBeau, B.: Examining high school student engagement and critical factors in dropout prevention. Assess. Effective Interv. **46**(2), 155–164 (2021). https://doi.org/10.1177/1534508419859655
19. Ünal, F.: Data mining for student performance prediction in education. Data Mining - Methods, Applications and Systems, pp. 1–9 (2020). https://doi.org/10.5772/intechopen.91449