International Journal of Data Warehousing and Mining
Volume 18 • Issue 1

Machine Learning Based Admission Data Processing for Early Forecasting Students' Learning Outcomes

Nguyen Thi Kim Son, Faculty of Natural Science, Hanoi Metropolitan University, Vietnam & Institute of Information Technology, Vietnam Academy of Science and Technology, Vietnam*

Nguyen Van Bien, Hanoi National University of Education, Hanoi, Vietnam

Nguyen Huu Quynh, Thuyloi University, Hanoi, Vietnam

Chu Cam Tho, The Vietnam Institute of Educational Sciences, Hanoi, Vietnam

ABSTRACT

In this paper, the authors explore the factors to improve the accuracy of predicting student learning outcomes. The method can remove redundant and irrelevant factors to get a "clean" data set without having to solve the NP-Hard problem. The method can improve the graduation outcome prediction accuracy through logistic regression machine learning method for "clean" data set. They empirically evaluate the training and university admission data of Hanoi Metropolitan University from 2016 to 2020. From data processing results and the support from the machine learning techniques application program, they analyze, evaluate, and forecast students' learning outcomes based on admission data, first-year, and second-year academic performance data. They then submit proposals of training and admission policies and methods of radically and quantitatively solving problems in university admissions.

KEYWORDS

Admission Data, Linear Discriminant Analysis, Logistic Regression Machine Learning Method, Students' Learning Outcomes Forecast

INTRODUCTION

Industry 4.0 is shaped inseparably with data and data analysis, posing challenges for organizations (including universities) *that How could they be able to improve their operational capacity, effective management, and minimize the risk of failure through efficiently handling data?*

In the era of big data technology, the demand for universities to innovate their governance models and improve governance efficiency has become an urgent issue for managers. In training management, universities need to digitalize (digital transformation) management information, create large database systems to organize training management and support decision making instantly and accurately. The problem is how to use, analyze and exploit this data source effectively to adapt the management in education and improve education efficiency.

DOI: 10.4018/IJDWM.313585 *Corresponding Author

A new field that emerged to solve this problem, is data mining in education (Romero & Ventura, 2010). The field uses data mining models, machine learning techniques to extract potential knowledge in educational data. Since then, this field has been growing and obtaining many significant achievements.

Machine learning and big data mining is a rapidly developing field, which is the intersection of many related fields such as databases, statistics, machine learning, algorithms, and other related ones to extract useful knowledge from large data sets. Other names can also be used for data mining and knowledge exploration such as Knowledge discovery in databases (KDD), Knowledge extraction (KE), Data analysis or samples (Data/pattern analysis - DA/PA), or Business intelligence (BI), see Lu et. al. (2022), Schild et. al. (2022).

Over the past two decades, significant progress has been made in the field of machine learning. The field has arisen as a method of choice for developing practical software for computer vision, speech recognition, natural language processing, robotic control, and other applications. With the positive impact of the increase in the amount of educational data through digitalization, there are quite a few areas in which machine learning can positively impact education. It can be affirmed that this is an inevitable trend that demonstrates the development of education and training associated with technology.

The machine learning model of selection of the factors affecting students' output, the Naive Bayes classification model are recommended by Harvey & Kumar (2019). This model can be used to implement early intervention for students. Others model from Đambić et. al. (2016) got also positive appreciation. Another study using two data mining algorithms Naïve Bayes and Logistic Regression also gave some positive results in predicting learning outcomes and predicting forced drop-out from school (Uyen & Tam, 2019)

Entrance admission is one of the important activities of any higher education institution. Every year all higher education institutions in Vietnam develop admissions schemes for their schools. The purpose of the admissions scheme is to recruit an adequate number of students according to the allocated criteria and, more importantly, to recruit the right students who wish to study the appropriate subject.

Admissions schemes of universities are usually assigned to the training management department and training organizers to plan and implement. In some universities, there is also a dedicated center/department/unit that organizes the implementation of the university's admissions tasks. The actual process carried out in most educational institutions is based on the experience of the officer in charge and, if any, on the calculation of admission figures in previous years. This leads to undesirable admissions results.

By approaching university admissions problems from the perspective of data mining and the application of machine learning techniques, the problem of making admission plans and related problems are extraordinarily complex issues. In Vietnam and around the world, this problem is less common in education research.

In this article, we apply machine learning techniques to solve several problems in university admissions, in which data is sampled from Hanoi Metropolitan University (in Vietnam) as a case study. In the framework of this paper, based on machine learning technology, our research questions are:

1. What parameters does the dataset have to be able to use machine learning techniques to analyze and forecast students' learning outcomes (specifically in graduation type forecasts based on inputs provided)?
2. Which machine learning model should be used to solve the problem of predicting learning outcomes effectively?
3. What are the factors affecting students' output, which factors are most important in the input data that affect students' output (which subject scores in the entrance combination will be the

most important factors for achieving expected output), and which machine learning techniques should be used to solve this problem?
4. What are the proposals on training and admission policies, effective and quantitative methods of solving problems in university admissions for universities in general and Hanoi Metropolitan University in particular?

To answer the above questions, we focus on three main task groups: (1) Eliminate redundant and irrelevant attributes (i.e., keep the important ones) without having to solve the NP-Hard combinatorial optimization problem. To avoid having to solve the NP-Hard combinatorial optimization problem, we use the approach to convert the combinatorial optimization problem to the continuous optimization problem. Specifically, we convert the linear discriminant analysis (LDA) method into a data dimensionality reduction algorithm; (2) Implement logistic regression machine learning method on "clean" data set that only includes important factors to improve the accuracy of predicting graduation results; (3) We do an empirical evaluation on a data set of 2763 samples with 89 data fields of Hanoi Metropolitan University. Machine learning techniques used include logistic regression (to forecast students' graduation results) and an improved technique of Linear discriminant analysis (to forecast important factors affecting student academic performance)-Discriminative factor Selection technique.

The answers to the issues will be presented clearly in section 4 (results) and section 5 (Discussion). Earlier, in section 2, we analyzed an overview of data science studies in education, which specifically emphasized the application of machine learning techniques in the processing of data in education research. The process of building the admissions data set and the attribute schools that make up the training and testing data set for machine learning will be presented in section 3 along with a detailed description of the method of selection of machine learning techniques and how to process input data (data focused on analysis for the Primary Education major of Hanoi Metropolitan University in 5 years, from 2016 to 2020). Finally, some further research directions, which are aimed to improve the forecasting efficiency, are presented in section 6 (Conclusion).

LITERATURE REVIEW

Data mining is the process of discovering valuable information or making forecasts from data. (Bao, 2002) generalized several concepts related to the field of knowledge detection and data mining mentioned in the Introduction to knowledge discovery and data mining lecture to systematize the foundational knowledge of the field.

For machine learning, one advantage of this technology is that computers do not need to be programmed clearly. Especially, computers are fully capable of changing and improving algorithmic elements, or in other words, computers approaching artificial intelligence (AI) technology. To demonstrate the complexity and intelligence level of machine learning in solving practical problems, (Webber. & Zheng, 2020) gave the following diagram.

Machine learning techniques for data mining are applied in many fields including education science. Especially in the context of education with many changes underlying the impact of the 4.0 revolution, technology has become a part of the means of production of the educational process. On the other hand, individual learning needs are also focused. Therefore, pedagogical research is being redirected to in-depth studies of learner behaviors to establish individual learning programs; at the same time mining big data of learners to diagnose in particular, and reorient the learning process of learners, and manage/operate the educational process in general. It is also the content that machine learning can be applied in educational science research.

The research by (Kotsiantis, 2012) described the emerging field of machine learning in education. In this study, student-specific data and point data were mined as data sets for the regression machine learning method used to predict a student's future learning ability. The problem of forecasting students'

learning results is proposed for research by (Anozie & Junker, 2006). (Đambić et al., 2016) proposed solutions to reduce the number of students who were retained or expelled because of poor academic performance, improve pass-through rates for students by analyzing and using machine learning techniques to identify "at-risk" students, from which the school can bring out the necessary support to help students improve their performance practice.

Recently, (Wu et al., 2020) has studied the use of machine learning for text classification to grade students' grade in some courses, expressing the potential to use classify messages from machine learning to identify students at risk of failing the course. The results of machine learning application in education research show great potential in solving the problem of predicting student outcomes as well as the problem of admission.

There have been many studies related to data mining methods such as Decision Tree, KNN, Bayes, Combined Law, etc. to solve the problem of student capacity predicting and come up with many positive results (Xu et al., 2021), (Damuluri et al., 2020), (Thai-Nghe et al., 2011), (Nghe & Dinh, 2015). Taking advantage of the pre-eminent features of the MyMediaLite system, (Nghe, 2013) has built a method for predicting student performance. With this algorithm, it is possible to accurately indicate which students need to work hard to study in accordance with minimizing the risk of being forced to expel from school.

Some current work has focused on supervised learning algorithms such as Naïve Bayesian Algorithm, combined rule mining, artificial neural network-based algorithms (ANN), Logistic regression, CART, C4.5, J48, (Bayes Net), Simple Logistic, JRip, Random Forest, Logistic Regression analytics, ICRM2 for classification of drop-out students (Kumar et al., 2017). However, according to classification techniques, Neural Networks and Decision Trees are two methods that are widely used by researchers to predict student learning outcomes (Shahiri et al., 2015). The advantage of the neural network is that it can detect all possible interactions between predictor variables and can also perform complete detection even in complex nonlinear relationships between dependent variables and independent variables (Arsad et al., 2013), while decision trees have been used for their simplicity and ease of understanding for small or large data structure discovery and value prediction (Natek & Zwilling, 2014).

In the work of (Elbadrawy et al., 2016), two classes of forecasting modeling methods were presented. The purpose of the study was to facilitate degree planning and determine who was at risk of slipping or dropping out of class. The first class builds the model using a regression-based method, and the second class uses a matrix factorization-based method. Both methods are based on course-specific, descriptive scale-based, and personalized multi-linear resize, while matrix analysis-based methods combine with a standard matrix decomposition approach. The mentioned approach is applied to the data set created from George Mason University (GMU) transcript data, University of Minnesota (UMN) transcript data, UMN LMS data, and Stanford University MOOC data (Harvey & Kumar, 2019).

In other consideration, (Suresh & Guttag, 2019) confirming that machine learning increasingly affects people and society, and awarding of its grown potential unwanted consequences, the authors of this research states that there are seven sources of harm in machine learning: (1) Historical bias; (2) Representation Bias; (3) Measurement Bias; (4) Aggregation Bias; (5) Learning Bias; (6) Evaluation Bias; (7) Deployment Bias. This consideration is helpful for our application learning model. (Ovadya & Whittlestone, 2019) state "information hazard" might be misused from machine learning research in general: Product hazard - Research produces software that can be directly used for hard aim; Data hazard - research produces detailed information or outputs which if disseminated, then create the risk of use for harm and Attention hazard - research which directs attention towards an idea or data that increases. Machine learning and AI, in general, are quickly developed, but we always need to "keep our eyes widely open" during the research to reduce as much as possible the data bias, especially in decision making which has a huge impact on someone's life.

From the above overview, we can see that the educational problems solved through data mining are quite varied. The problem of forecasting student learning outcomes has been studied by many authors and the data mining techniques used are diverse. However, the data used in these studies are mainly based on student academic performance transcripts. The problem of analyzing admission data of university students and analyzing the influence of admission data in the problem of forecasting student graduation results is a new problem that has not been researched yet. Moreover, the use of the Linear discriminant analysis technique to predict important factors in admission data affecting student learning outcomes is a typical characteristic in this article, especially in the context of administrative processing data in Vietnam. This is a valuable and unfamiliar issue.

THE PROPOSED METHOD

Model of the Proposed Method

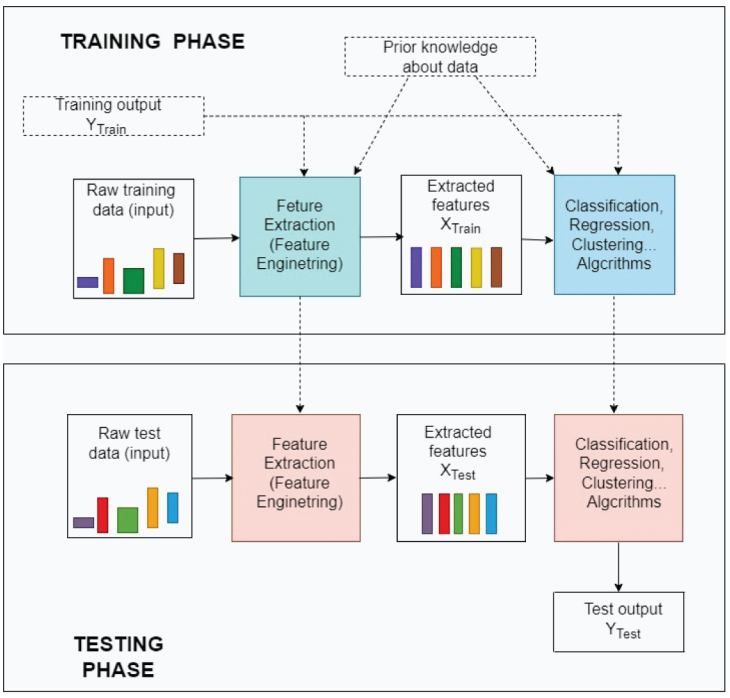
The majority of machine learning problems can be shown in Figure 1.

In this model, there are two major phases: the training phase and the testing phase. With supervised learning problems, we have input and output, while with unsupervised learning problems, we only have input. Literally, training is the data set put into training after testing to re-test the results through the problem of re-analysis, layering problems, etc.

Each problem, each forecasting model has its own pros and cons. From the factors of the built-in training data set, the researcher must build the appropriate models and algorithms to solve the questions raised in each specific case.

The data used for this article is admissions data and training data from Hanoi Metropolitan University, code HNM, which is a public university under the People's Committee of Hanoi. Hanoi Metropolitan University is a higher education institution in the national education system, organizing

Figure 1. A general model for machine learning problems (Tiep, 2018)

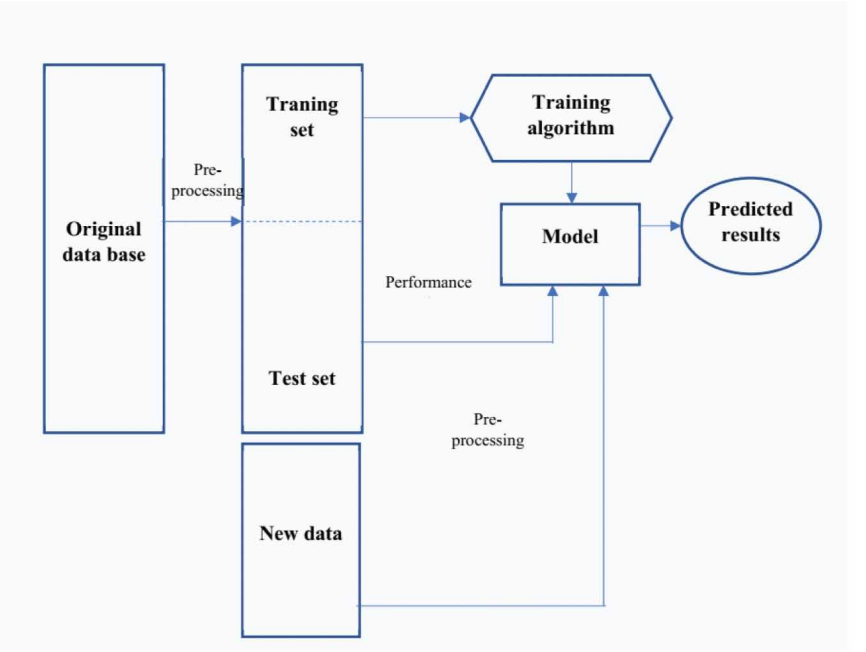


the training of multidisciplinary human resources, multidisciplinary fields of college, university, and post-university level; organize vocational education activities according to social needs and in accordance with the provisions of law. Primary Education is one of the traditional majors of the school, starting in 1959, the Primary Education department of Hanoi Metropolitan University has provided numerous primary teachers for primary schools in Hanoi city and nearby areas. In this article, we look at the issue of university admissions, formal systems, Primary education major of Hanoi Metropolitan University.

Every year the university's training management department has statistical data on the results of the subjects of students in the university. The Department of Academic and Student Affairs issued with the student transcript form, assessment and accountability, information about the educator, and the school's finance with the Hanoi People's Committee and the Ministry of Education and Training. The raw data is provided by the Department of Academic and Student Affairs. All data, including admission scores, academic score from first year to fourth year, graduation scores, and the financial situation can be extracted from training management software for data analysis. The data set includes 2763 observational samples of Primary education students from D2016 to D2020 and 33 attribute variables. Each observation sample is shown in 1 row.

In the model depicted in Figure 1, the selection of important factors (also known as important features or important attributes) is included in the "Feature extraction" stage, that is, first we proceed to select the important factors and then use the Logistic Regression machine learning method. The reason for this is to reduce the data dimensionality and remove redundant and irrelevant attributes in the data set and thus increasing the predictive accuracy of the Logistic Regression model. It should also be noted here that the selection of important factors also belongs to the data preprocessing stage shown in Figure 2.

Figure 2. Predictive model of student learning classification



Determining the important attribute is an NP-Hard optimization problem on discrete space. Therefore, finding a factor selection algorithm with lower complexity is an important issue when solving this second subproblem. An appropriate way to solve the second subproblem with acceptable computational complexity is to reduce the discrete space optimization problem to the continuous space optimization problem as shown in (Tao et al., 2016). The technique used is the Discriminative Factor Selection technique. The factor selection method (combining the popular transformation-based dimensionality reduction method linear discriminant analysis and sparsity regularization in the study of (Tao et al., 2016) will be used to deal with this problem.

Machine Learning Techniques Used in Forecasting

The data set of student learning outcomes obtained from the Hanoi Metropolitan University is not linearly separable, so some machine learning methods such as linear regression, PLA, etc. are not applicable.

Logistic regression is used to solve the classification problem. It provides the probability that an event will happen or not (according to 0 and 1) based on the value of the input variables. For example, predicting whether an e-mail is classified as spam or whether a student will graduate are cases that can be considered as the binomial outcome of Logistic Regression. There can also be the multinomial outcomes of Logistic Regression, e.g., prediction of student academic performance as good, good, average, or poor... Therefore, logistic regression refers to the definite prediction of the target variable. While linear regression deals with predicting values of continuous variables. Logistic Regression has the following advantages: simple implementation, computational efficiency, efficiency from a training perspective, and ease of regularization. No normalization is required for the input features. This algorithm is mainly used to solve problems on an industrial scale. Also, logistic regression is not affected by a slight noise in the data. Logistic Regression has the following disadvantages: inability to solve non-linear problems because its decision surface is linear, prone to overfitting problems, and will not perform well unless all variables are independent is determined.

Logistic Regression (Keselj, 2009) works by extracting a set of weighted features from the input set, taking the log, and combining them linearly, that is, each component is multiplied by a weight and then added up. Logistic regression is a type of regression that predicts the probability of an event occurring by fitting data to a logistic function. As with many forms of regression analysis, logistic regression uses several predictor variables that can be numerical or categorical.

The logistic regression hypothesis is defined as:

$$h_\theta\left(x\right) = g\left(\theta^T x\right) \tag{1}$$

where *g* is a sigmoid function defined as:

$$g\left(z\right) = \frac{1}{1+e^{-z}} \tag{2}$$

The cost function for logistic regression is given as follows:

$$j\left(\theta\right) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right) - \left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right] \tag{3}$$

To find the minimum of this cost function, in machine learning we will use a built-in function called fmin_bfgs[2], which finds the best parameters θ for the logistic regression

cost function for a fixed data set (including x and y values). The parameters are the initial values of the parameters to be optimized and a function that, when fed to the training set and a special θ, computes the logistic regression cost and the gradient associated with θ for data set with values x and y. The final θ value will be used to draw the decision boundary of the training data.

The above learning outcome classification problem is a multi-class problem that leads to logistic regression with the sigmoid function which will be ineffective. Therefore, we propose to apply logistic regression with softmax function for the first subproblem. With this algorithm, the output can be expressed as a probability. However, the performance of the Logistic Regression machine learning method is affected by the input data set, which has large dimensions and redundant or irrelevant attributes. We therefore have to solve the second subproblem of determining the important attributes (namely, determining which subject is the most important among the entrance exams of the input dataset).

The technique for determining the important attribute is as shown below.

The important factor selection algorithm is the key factor to improve the accuracy. Therefore, we present the important factor selection algorithm as below.

Assume that we have a set of n samples (i.e., n students) $X = \left[x_1, x_2, \ldots, x_n\right]^T \in \mathbb{R}^{n \times d}$. These samples belong to c classes. Factor selection is done by finding the projection matrix W of the objective function (Tao et al., 2016):

$$\min_W - \frac{W^T S_{\text{int}er} W}{W^T S_{\text{int}ra} W} + \alpha W_{\infty,1} \tag{4}$$

where, W is the required projection matrix. S_{intra} and S_{inter} are the within-class scatter matrix and the between-class scatter matrix respectively. $W_{\infty,1}$ is $\ell_{\infty,1}$-norm regularization of W matrix.

We solved using a Quasi-Newton method as implemented in Mark Schmidt's optimization toolbox[1] to solve the problem (4)

The projection matrix W is factor selectable because if all the components of the j[th] row of W are zero, the factors (i.e., columns of the matrix X) will not contribute to the data representation XW and therefore it will be removed. Thus, for a factor to be selected by the algorithm, it will have to have at least one element of the j[th] row of W that must be non-zero. Problem (4), forcing many rows of W to be zero, helps us to select a few factors.

The problem of determining important factors and prediction is done according to the following steps:

Step 1: Collect student's academic results information.
Step 2: Perform data preprocessing.
Step 3: Separate the collected data into two sets. One is training set, and the other is a test set.
Step 4: Model training.
  Step 4.1: Training the forecast model according to logistic regression with a softmax function.
  Step 4.2: Select important factors according to formula (4).
Step 5: Use a forecasting model to forecast student's academic performance and identify important attributes.
Step 6: Evaluate the results and accuracy of the forecasting model.

RESULTS

Data

The input data includes four admission data attribute fields (national high school exam scores in Math, Literature, English, and total test scores) and 85 component subject scores (including different electives for each student).

Unnecessary data and overlooked variables, which were not evaluated in this study, were removed. Some data on physical or gifted subjects and variables related to students' financial status were eliminated. The factors with minute data, or with a lot of empty data, are also removed. Electives largely fall into this empty data array. We also focused on specific checkpoint data and removed letter grades from the test score data. During this process, we did some cross-checked for harmful sources to reduce the data bias. We selected individual variables for each student to test their correlation with the variable of interest. Therefore, from 2763 samples with 89 attribute variables, the data set was cleaned to include only the data of 933 observed samples, and the variables were limited to 39 variables.

The data is divided into training and testing subsets. The training set consists of 90% of the original data set and the test set contains 10%. The user interface of the model training can be shown in Figure 3

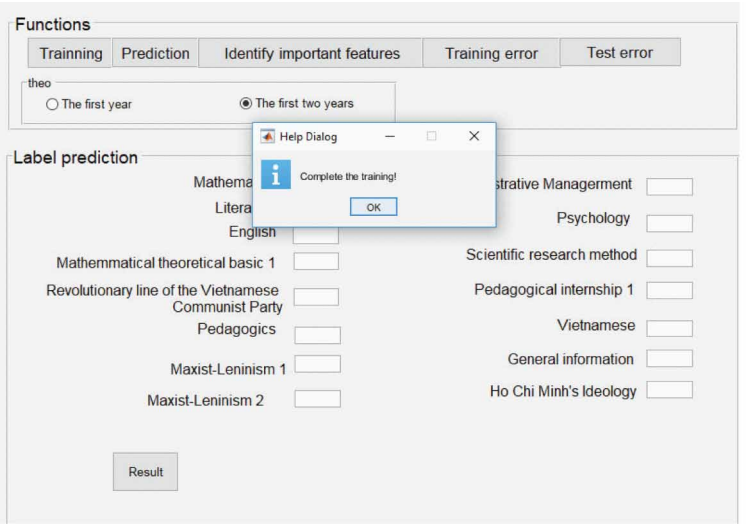
Data is shown for two forecast scenarios:

Case 1: Based on the National University Entrance Exam Score and the first-year course grade in university.

Case 2: Based on the National University Entrance Exam Score and the first and second-year course grade in university.

Both models were used to make predictions about student learning outcomes, and the models were analyzed for accuracy.

Figure 3. Model training



Forecasting Students' Learning Outcomes

Logistic regression was the first classification technique used to test this data set. Math, Literature, and English scores were selected as considered variables for all classification techniques. Course grades for year 1, year 2 are all highly correlated. Any of these variables could be leveraged for assessment. In this section, we run two methods to compare the results. The first method is Logistic Regression and the second method is our proposed method. The Logistic Regression method differs from our proposed method in that: Logistic Regression method is trained and predicted on the original data set (i.e., the data set has not reduced dimension and exists in redundant or irrelevant), while our proposed method is trained and predicted on data set has removed redundant and irrelevant attributes.

The forecast results are shown in Table 1.

Logistic regression analysis is used to check the graduation code of each student, based on 39 input data schools, the number of national university entrance examination scores, the course grades of 35 subject's exams in the first two years in the university. The problem is approached with 2 levels of input data. Level 1 is based on 22 factors, including university entrance and 18 subjects' scores of the first year of university. Level 2 is an analysis based on the total of 39 grade scores from 35 university freshmen and sophomore subjects, along with national university entrance exam scores (4 factors). Level 1, the root mean squared error (RMSE) of our proposed method is 1.8.

The forecast results using the model at level 2, based on a total of 39 grades in 35 first- and second-year university subjects, along with scores on the national entrance examination. Level 2, the root mean squared error (RMSE) of our proposed method is 1.5. The error is acceptable and in agreement with the statistical data.

From Table 1 we can see, the RMSE error of our proposed method is 1.8, which is lower than that of Logistic Regression which is 0.3. Same for level 2, RMSE error of our proposed method is 1.5, it is lower than that of Logistic Regression which is 0.4. From this result, we can see the effectiveness of the important attribute selection in our proposed method. Also from this result, we can see that our input dataset includes some redundant and unrelated attributes.

One of the new factors of this study is that we can build a new dataset and run a predictive test program. When we have the data of the national entrance examination and the grade of the first year or the first two years of students, based on the construction program, we can predict the expected graduation classification type of student. With 180 new data samples, the program gives us the results of predicting the graduation type of those 180 students. The results are shown in Figure 4.

Forecasting the Main Factors Affecting the Output Results

The most important factor identification is built on the Discriminative Factor Selection technique (see Tao et al., 2016) with many data fields and built-in sub libraries.

The most important factor affecting a student's output result is highly correlated with that student's output, but not with the size of the attributes or the size of the dataset. Data with a higher total admission score will correlate better with the type of graduated classification. Art, physical education, political science, and general subjects were found to have a slight correlation with the student's total graduation score. The training and testing metrics show relationships between the Literature mark and the rest of the variables, with the student's total graduation score and graduated classification type.

Table 1. The prediction error of the two methods

Ordinal number	Method	RMSE for Level 1	RMSE for Level 2
1	Logistic Regression	2.1	1.9
2	Our proposed method	1.8	1.5

Figure 4. Test new data

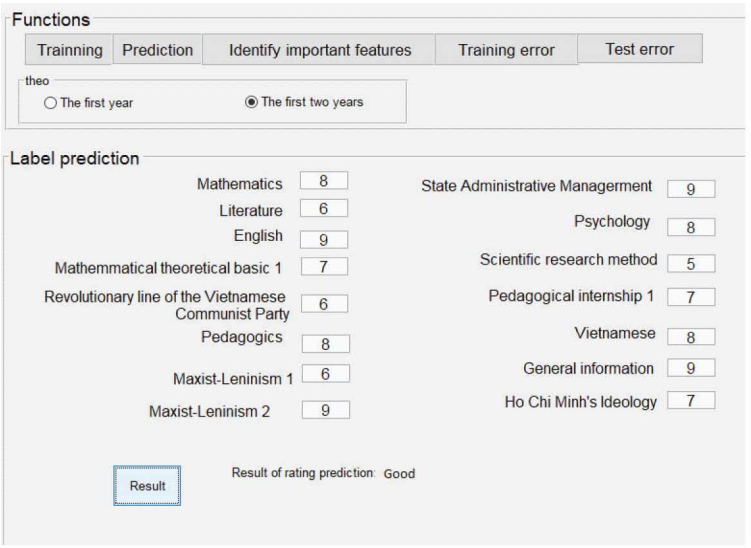
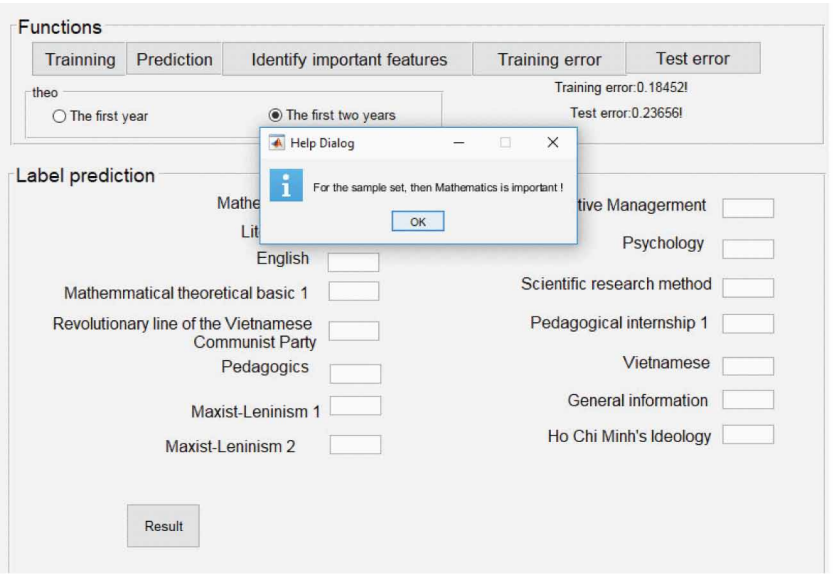


Figure 5 depicts the training results for the training set consisting of data samples of students majoring in Primary Education, Hanoi Metropolitan University. According to the results of this training, Mathematic will be the most important factor affecting the output of students in the entrance exam subjects combinations for the Primary education major of Hanoi Metropolitan University. The input data considered includes admission information and the students' performance of the first two years in university.

Figure 5. The most important factors



DISCUSSION

There are many different classifiers used to evaluate the accuracy of a model for data sets on admission and students' learning progression. Collecting, synthesizing, and analyzing data from education stakeholders is time-consuming, expensive, and not always with a high degree of accuracy. Especially in the current situation of education, digital transformation, and digital data storage in the education sector in Vietnam have not been conducted methodically and thoroughly, including Hanoi Metropolitan University. The management of admission information, including national university entrance exam scores, is managed by the Ministry of Education and Training. Meanwhile, universities only manage course grade data during training and GPA. In fact, although universities are interested in building student data sets, the built data is managed by many different departments (training department, student affairs department, training faculties, etc). The data is therefore difficult to synchronize, and the connection to exploit and use the data of the stakeholders is also difficult.

On the other hand, other student data such as culture, family tradition, financial status, personal aspirations, career orientation, high school results, study plan, training program, the teaching staff, the facilities of the educational institution, the participation in social organizations, mass organizations, anthropological factors, cultural factors, economic factors, psychology, etc. need to be studied and collected. There also needs to ensure that the data set includes many fields of information, many representative parameters, mutual influence directly affecting the training process and the learning outcome of students. These factors need to be analyzed, processed, and improved to make education data more meaningful to students, lecturer, and other stakeholders.

Therefore, the Ministry of Education and Training, the Department of Education and Training in the localities and higher education institutions need to pay serious attention to building large, synchronous, interconnected data sets between management units and educational institutions in order to effectively exploit data, which was used in the management tasks and improve the quality of education.

Predictive outcomes could be better if it provided educators with information that could be used to help improve student learning outcomes. Several mistakes such as missing data, empty data fields, errors in model could be problematic for data analysis.

Determining the right machine learning technique to accurately classify, accurately predict (with acceptable error) students' learning outcomes is also one of the core factors to the success of the research.

The forecast results for the degree classification of students, in both cases, show an increasing accuracy relationship with a gradually decreasing standard deviation in the forecast results, which RMSE diminishes from 1.8 to 1.5. This may explain why universities should actively shift to two-phase teaching. Based on the scale of general training and admission scores, first and second years of university, we can forecast which students' strengths are suitable for which majors of the Hanoi Metropolitan University in particular, and arbitrary universities in general (when appropriate data are available).

At the same time, this result also honestly reflects that the more attributes the data set has (the higher the complexity of the data set), the more accurate the forecast results.

Determining the predictive results for the main influencing factor on student graduation outcomes in the subjects of the admission combination will assist decision-makers in choosing the right combination of admissions for the orientation of the career standard outcomes. At the same time, the coefficient for each subject in the appropriate admission combination should be determined to select the students who have the most suitable academic ability for the requirements of professional development.

On the other hand, calculating the main factors affecting student outcomes also allows higher education and training institutions to promote the quality of their work as academic consultants and guide the selection of majors and subjects matching for each student.

In this experiment, we observe that the training error and the prediction error still need to be reduced. The cause of this error is that the number of training samples in the experiment is still low compared to the dimensionality of the data, which causes the machine learning model to be underfitting. To overcome this limitation, in the near future, we will study to reduce the data dimension and add training samples.

Moreover, the research has just only focused on one case study with simple raw data (the scores). For validity, it needs rechecking with more detailed data such as socio-economic condition, humanity, learning environment, etc. With these limitations, the results of the research stay only one of evidence for the committee to make decisions.

CONCLUSION

The logistic regression model can be used to make forecasts of other performance metrics, using the same data set. Educators can use the model to select variables they want to assess the correlation between variables used to forecast student academic performance in tests. This forecasting model can then be used to make early interventions for students, which can help improve the student's future success. Timely feedback is important for educators to be able to come up with early intervention strategies. This model can be implemented quickly and easily after the data set was built.

The logistic regression model can be improved to give better forecasts of student academic performance. It will be interesting to compare the performance of classification models on other data sets, or even improve the model to increase the predicted accuracy of the current model.

From the results of the initial analysis, we found that the admissions problem can be radically and quantitatively solved. If data such as 3-year high school scores, family economic status, career interests, and hobbies, self-study skills, etc. were fully collected, the forecast of student academic performance is entirely possible. Since then, the admissions problem has not only been in the scope of the admissions scheme but also expanded into the field of admissions counseling.

Within the scope of this paper, we make statements based on the model, which was trained on a specific dataset of Hanoi Metropolitan University. In order to avoid biases of expert perception of the social sciences, we will continue to work on combining different machine learning models that can include expert perceptions.

Future research may also include an assessment of student behavioral features, human factors, personal and historical factors of the student's learning process, as well as academic attitudes and other socio-economic factors as they relate to student academic performance based on evaluating methodology. Additional analysis can be conducted using different classifications on the same data set, including multi-level awareness and artificial neural networks.

FUNDING AGENCY

Publisher has waived the Open Access publishing fee.

REFERENCES

Anozie, N. O., & Junker, B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. *AAAI Workshop - Technical Report*, *WS-06-05*, 1–6.

Arsad, P. M., Buniyamin, N., & Manan, J. L. A. (2013). A neural network students' performance prediction model. *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications, ICSIMA 2013*. doi:10.1109/ICSIMA.2013.6717966

Bao, H. T. (2002). Introduction to Knowledge Discovery and Data Mining (in Vietnamese). In Data Mining and Knowledge Discovery Handbook. doi:10.1007/978-0-387-09823-4_1

Đambić, G., Krajcar, M., & Bele, D. (2016). Machine learning model for early detection of higher education students that need additional attention in introductory programming courses. *International Journal of Digital Technology & Economy*, *1*(1), 1–11.

Damuluri, S., Islam, K., Ahmadi, P., & Qureshi, N. (2020). Analyzing navigational data and predicting student grades using support vector machine. *Emerging Science Journal*, *4*(4), 243–252. doi:10.28991/esj-2020-01227

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting Student Performance Using Personalized Analytics. *Computer*, *49*(4), 61–69. doi:10.1109/MC.2016.119

Harvey, J. L., & Kumar, S. A. P. (2019). A Practical Model for Educators to Predict Student Performance in K-12 Education using Machine Learning. *2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019*, 3004–3011. doi:10.1109/SSCI44817.2019.9003147

Keselj, V. (2009). *Speech and Language Processing* (D. Jurafsky & J. H. Martin, Eds.). Pearson Prentice Hall.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, *37*(4), 331–344. doi:10.1007/s10462-011-9234-x

Kumar, M., Singh, A. J., & Handa, D. (2017). Literature Survey on Educational Dropout Prediction. *International Journal of Education and Management Engineering*, *7*(2), 8–19. doi:10.5815/ijeme.2017.02.02

Lu, X., Gu, D., Zhang, H., Song, Z., Cai, Q., Zhao, H., & Wu, H. (2022). Semi-Supervised Sentiment Classification on E-Commerce Reviews Using Tripartite Graph and Clustering. *International Journal of Data Warehousing and Mining*, *18*(1), 1–20. doi:10.4018/IJDWM.307904

Natek, S., & Zwilling, M. (2014). Student data mining solution-knowledge management system related to higher education institutions. *Expert Systems with Applications*, *41*(14), 6400–6407. doi:10.1016/j.eswa.2014.04.024

Nghe, N. T. (2013). *A system for predicting students's course result using a free recommender system library - MyMediaLite*. Information Technology Conference 2013 At Can Tho University.

Nghe, N. T. & Dinh, T. Q. (2015). A consultancy support system for university entrance test. *Journal of Sciences - Cantho University*, 152–159.

Ovadya, A., & Whittlestone, J. (2019). *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning.* https://arxiv.org/abs/1907.11274

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, *40*(6), 601–618. doi:10.1109/TSMCC.2010.2053532

Schild, E., Durantin, G., Lamirel, J., & Miconi, F. (2022). Iterative and Semi-Supervised Design of Chatbots Using Interactive Clustering. *International Journal of Data Warehousing and Mining*, *18*(2), 1–19. doi:10.4018/IJDWM.298007

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. doi:10.1016/j.procs.2015.12.157

Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Proceedings of CHI '21: ACM CHI Conference on Human Factors in Computing Systems (CHI '21)*, *1*(1), 1–13. doi:10.1145/1122445

Tao, H., Hou, C., Nie, F., Jiao, Y., & Yi, D. (2016). Effective Discriminative Feature Selection With Nontrivial Solution. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(4), 796–808. doi:10.1109/TNNLS.2015.2424721 PMID:25993706

Thai-Nghe, N., Horváth, T., & Schmidt-Thieme, L. (2011). Factorization models for forecasting student performance. *EDM 2011 - Proceedings of the 4th International Conference on Educational Data Mining*, 11–20.

Tiep, V. H. (2018). *Introduction on Machine Learning*. https://machinelearningcoban.com/ebook/

Uyen, N. T. & Tam, N. M. (2019). Predicting student's academic performance by applying data mining technique. *Journal of Sciences - Vinh University, 48*, 68–73.

Webber, K. L., & Zheng, H. Y. (n.d.). Big data on campus: Data Analytics and Decision Making in Higher Education. Johns Hopkins University Press.

Wu, J. Y., Hsiao, Y. C., & Nian, M. W. (2020). Using supervised machine learning on large-scale online forums to classify course-related Facebook messages in predicting learning achievement within the personal learning environment. *Interactive Learning Environments*, *28*(1), 65–80. doi:10.1080/10494820.2018.1515085

Xu, Z., Yuan, H., & Liu, Q. (2021). Student Performance Prediction Based on Blended Learning. *IEEE Transactions on Education*, *64*(1), 66–73. doi:10.1109/TE.2020.3008751

ENDNOTE

[1] https://www.cs.ubc.ca/~schmidtm/Software/minFunc

Nguyen Thi Kim Son obtained Ph.D. degree in Mathematics at Hanoi National University of Education in 2010. She is currently working as a lecturer at the Faculty of Natural Science, Hanoi Metropolitan University, Hanoi, Vietnam. She has 21 years of teaching experience in the Faculty of Mathematics and Informatics, Faculty of Natural Science. Her areas of interest include mathematical modeling, machine learning, deep learning with applications in educational research, assessment & evaluation.

Nguyen Van Bien is a lecturer of Physics education at the Hanoi National University of Education. He received his undergraduate degree from the Hanoi National University of Education and his PhD in physics education from the University of Koblenz Landau (Germany) in 2007. He has been at Hanoi National University of Education since 2001 and has just served as Vice Dean of the Faculty of Physics since 2012. He is now a national consultant for STEM Education of SESDP2 (Second Secondary Education Sector Development Program). His current research effort is devoted entirely to physics education at the high school level and the college level. He has given courses such as "Physics high school curriculum analysis", "Integrated science education", "Assessment in physics education" and "ICT in physics education" for undergraduate and postgraduate students in Physics education. He has published articles, books on constructing physics apparatus, active learning in high school physics and STEM education.

Nguyen Huu Quynh is currently an Associate Professor in the Faculty of Information Technology at the Thuyloi University. He received the B. S. degree in informatics, M.S. and Ph. D. degrees in computer science, all from Hanoi University of Viet Nam, in 1998, 2004 and 2010, respectively. His current research interests include machine learning, content based image retrieval, artificial intelligence. Associate Professor Quynh has published over 80 international journal articles and international conference papers, which includes dozens of SCIE publications.

Chu Cam Tho is an Associated Professor on Education and Head of Research Division on Educational Assessment (RDEA) at The Viet Nam Institute of Educational Sciences (VNIES). Research interests: Psychological and educational basis of the process of formation and development of learners' thinking, through Mathematics Education in particular; Solutions in the curriculum, content, teaching methods to develop thinking for learners through Maths from preschool to high school; School models and management, assessment & evaluation of general education schools; Professional development & Training for Competency of Developing Educational Programs for Teachers; Educational Research and Assessment & Evaluation.