IEIE Transactions on Smart Processing and Computing, vol. 9, no. 3, June 2020
https://doi.org/10.5573/IEIESPC.2020.9.3.217

Multi-models of Educational Data Mining for Predicting Student Performance in Mathematics: A Case Study on High Schools in Cambodia

Phauk Sokkhey[1], Sin Navy[2], Ly Tong[3], and Takeo Okazaki[4]

[1] Graduate School of Engineering and Science, University of the Ryukyus 1 Senbaru, Nishihara, Okinawa 903-0213, Japan
sokkheymath15@gmail.com
[2] Ministry of Education, Youth and Sport (MoEYS), 80 Preah Norodom Blvd (41), Phnom Penh, Cambodia
navymath.sin@gmail.com
[3] Royal Academy of Cambodia, Russian Blvd, Phnom Penh, Cambodia lytongcambodia2013@gmail.com
[4] Department of Computer Science and Intelligent Systems, University of the Ryukyus 1 Senbaru, Nishihara Okinawa 903-0213, Japan okazaki@ie.u-ryukyu.ac.jp

***** Corresponding Author: Takeo Okazaki

Received October 25, 2019; Accepted February 27, 2020; Published June 30, 2020

** Regular Paper*

** Review Paper: This paper reviews the recent progress possibly including previous works in a particular research topic, and has been accepted by the editorial board through the regular reviewing process.*

***Abstract*:** Education is crucial for the development of any country. Analysis of education datasets requires effective algorithms to extract hidden information and gain the fruitful results to improve academic performance. Multiple models were used to maximize the contribution to the education environment. In this study, we used the spot-checking algorithm to compare these methods and find the most effective method. We propose three main classes of education research tools: a statistical analysis method, machine learning algorithms, and a deep learning framework. The data were obtained from many high schools in Cambodia. We introduced feature selection techniques to figure out the informative features that affect the future performance of students in mathematics. The proposed ensemble methods of tree-based classifiers provide satisfiying results, and in that, random forest algorithm generates the highest accuracy and the lowest predictive mean squared error, thus showing potential in this prediction and classification problem. The results from this work can be used as recipe and recommendation for mining various material settings in improving high school student performance in Cambodia.

***Keywords*:** Education data mining, Statistical analysis technique, Machine learning algorithms, Deep belief network, Predicting student performance

1. Introduction

The poor performance of students in high schools has become a worried-task for educators as it affects the secondary national exam and step to higher education. Early prediction and classification of student performance levels offers an early warning and gives a recipe for improving poor performance of students and other managerial settings. Ministry of education, youth and sport (MoEYS) in Cambodia is trying to improve the performance of poor performance students who are more likely to fail exam, drop out, and repeating classes and improving STEM (science, technology, engineering, and mathematics) discipline. Hence, the challenging task is to give an early predicting to prediction student performance in mathematics and investigate unknown learning patterns of students that affect student performance.

Educational data mining (EDM) provides a number of effective techniques to extract useful information for the education environment. Various analysis techniques on education research have been introduced for monitoring and anticipating academic performance to keep track of teaching, learning actions and productive results. The objectives in education research are predicting student

performance, classifying student performance levels, and analyzing student learning behaviors that affect their academic performance. The results obtained from these objectives are used at various managerial levels in the education system.

The prediction of education performance generally has two meanings. In some education research, prediction refers to methods of extracting important features and the impact of those features on the academic process. This prediction is important for the academic environment since it gives information about the underlying construct of features and its influence on academic outcomes. In the other meaning, prediction methods are used for predicting outputs such as academic scores, grades, dropout rates, and other performance measurements. In the concept of data mining, the prediction of numerical or continuous variables is called regression, and the prediction of categorical or discretized variables is called classification. Several tools have been applied to education research. The main task-based techniques are described as follows.

A. Association Rules and Causation

Association and causation are the best-known implication techniques in data mining related to tracking patterns but is more specific to dependently linked variables. In the education environment, these classical techniques are quite important in identifying the interesting relationships between variables, causal structures affecting academic performance and the underlying the structures.

B. Regression

In a statistical context, regression analysis is a combination of processes for measuring and estimating the relationship between two sets of variables called input variables and output variables. It is an explanation tool for determining the influence of input variables to output variables (outcomes).

C. Classification

A basic concept of supervised machine learning is classification. Classification is a common technique in machine learning that is used to classify and predict predefined classes or categories of target variables. In education, classification is mostly used to classify the performance of students based on their scores or grades. This classification is useful for identifying groups of poor-performance students that requires advanced assistance for improvement before final exams, as well as high-performance groups that can be awarded the scholarships.

D. Clustering

Clustering is an effective technique in data mining that discovers significant or informative clusters of objects that have similar characteristics. Clustering is an unsupervised technique that investigates objects in dataset and places them in each class due to similarity. However, classification is a supervised technique that assignes objects into predefined classes.

The preliminary part of this work was published in a previous study [1]. In the study, we used data collected from high schools in Cambodia and introduce more effective methods to boost the accuracy of our proposed algorithms. We reviews various techniques from statistical analysis technique to machine learning and deep learning. Moreover, we did a comparative study of these three main categories of tools. The primary contributions of this study are itemized in the following:

(i) Reviewing education research, its purpose, algorithms, and key findings.
(ii) Identifying the leading tools that are popularly used in education research.
(iii) Comparing prediction models of student performance in mathematics, which is the principal subject for any scientific subjects and technology.
(iv) Predicting students' outcomes so that we can find the poor-performance students for early warning and high-performance students that can be rewarded.
(v) Observing informative features that highly influence mathematics performance and academic performance.
(vi) Boosting the performance of tree-based models which as found to have high performance in paper [1] with ensemble methods in this paper.
(vii) Using the results from prediction models and ranking important features from features selection for improving student performance in high schools and various manangerial settings in MoEYS and STEM discipline in Cambodia.

2. Review of Previous Works

Several studies on EDM have been done to improve education. Numerous tools have been applied according to the objectives of the studies. The distinction of characteristics of data, the complexity of data, the level of contribution signification, and limited performance of used algorithms are implied to exist in various papers. These works falls into three main categories. The first category focuses on detecting the causal structure and relation of features. The second category is very popularly used to predict student performance, student achievement, learning outcome, and dropout rates. The last group focuses on observing important features that have a high influence on student performance and learning behaviors.

(a) Detecting causal structure and correlations of features

In classical statistical methods, the majority of tools are causal models that are used to observe the structure of input features. The most popular techniques are factor analysis, path analysis, multilevel modeling, and structural equation modeling.

(b) Predicting student performance, learning behavior, and achievement in learning outcomes

Machine learning techniques and educational data mining algorithms are popular tools for predicting students' grades, outcomes, and characteristics such as dropout rates. The aim is to predict the student outcome and final grades to improve academic performance, and to obtain hidden information for educational settings.

Table 1. Summary of previous education research.

Tasks	References	Methods/ Technique	Dataset	Key Findings
Detecting Causal Structure, Correlation features	Mohamed et al. (2012) [2]	Structural equation modeling, factor analysis	Libyan student in Kuala Lumpur	Effects of factors on mathematics achievements
	Uysal S. (2015), [3]	Structural equation modeling	Data from PISA 2012 of OECD members	Correlations among the features and mathematics' achievement
	Tongsilp A. (2013), [4]	Path analysis	Records in private universities in Bangkok, Thailand.	Future expectation has direct effect on achievement motivation
	Kilic S. et. al. (2013) [5]	Multilevel regression modeling	TIMSS 2011 databases	Effect of economic background on students achievement in mathematics
Predicting student performance, learning behavior, and achievement in learning outcomes	Stephen et al. (2018) [6]	Multiple linear regression and PCA	Open edX and Mapple T.A records	PCA could help to improve MLR algorithms
	Kotiantis et al. (2007) [7]	C4.5, NB network, BP, 3-NN, SMO	Data obtained from written assignments	Naïve Bayes generated the highest accuracy
	Menaei-Bidgoli et al. (2013) [8]	BN, 1-NN, prazen window, MLP, GA	Logged data from the enrolled database	The proposed GA improved the prediction model
	Ermiyas et al. (2017) [9]	Multilayer perceptron, naïve Bayes, SMO	WKU registered databases	Naïve Bayes generate the highest accuracy
	Shanithi et al. (2018) [10]	Meta decision tree: AdaBoost, Bagging, Dagging and Grading	Undergraduate data recorded	The proposed meta DT give a satisfying result for prediction
	Kumar M. et al. (2017) [11]	Decision tree, naïve Bayes, random forest, Bayes network	412 post-graduate records from the University	Random forest gave the best result compare to the other three models
	Wiyono S. et al. (2019) [12]	KNN, SVM, DT	Politeknik Harapan Bersama, Indonesia	SVM generates the highest accuracy.
Extracting informative features, learning behaviors affecting students' performance	Pimpa C. (2013) [13]	Decision tree, neural network	Database from a Thailand University	Tree classifier produces higher accuracy than neural network
	Amieh E.A. et al. (2016) [14]	ANN, NB, DT, Bagging, Boosting, RF	Learning management system (LMS)	Informative features could improve the accuracy up to 25.5%
	Menaei-Bidgoli et al. (2013) [8]	BN, 1-NN, prazen window, MLP, GA	Logged data from the enrolled database	The proposed GA improved the prediction model
	Ramaswami M. et al. (2009) [15]	VotedPerceptron, naïve Bayes, OneR, PART	Higher secondary school students in State Tamil Nadu, India.	Feature reduction could improve computational time, construction cost, and predictive accuracy
	Affendey L.S. et al. (2010) [16]	Bayes based classifiers, tree-based classifiers, SMO, and LR	Dataset from bachelor students' records at PUM	NB, AODE, and RBFNetwork classifiers generated the highest accuracy.
	Arindam et al. (2018) [17]	ANN, DNN, RNN	Data from Kaggle	RNN outperforms other models
	Sintthia G. et al [18]	KNN, K-mean, and MLP	Students' record at CHRIST university	MLP outperforms K-mean and KNN.
	Zhang Y. et al. (2016) [19]	SVM, LR, and DBN	Data obtain from University of the Cordilleras, Phillipine	Deep belief network (DBN) outperformance the other two models.

(c) Observing features or behaviors affecting student performance

A typical task in education research is to extract informative features that affect academic performance keep track of learning behaviors. Features selection and information ranking algorithms are used to observe important features that affect academic performance. Another aspect is to continuously observe students' behaviors to provide timely information on their progress. The majority of these tasks were observed from e-learnings platforms such as learning management system (LMS), ittellignet totoring system (ITS), and massive open online course (MOCCs).

3. Research Methods

EDM focuses on applications of learning approaches and algorithms of data mining in the education field. EDM deals with researching, developing, and using statistical analysis techniques, machine learning, data mining, and deep learning to observe any form of education data and use the knowledge for enhancement and improvement in the education field. In this article, we categorized these techniques into three main classes with respect to the innovation, contribution, purpose, and merits.

There is no rule that a particular algorithm is best for a

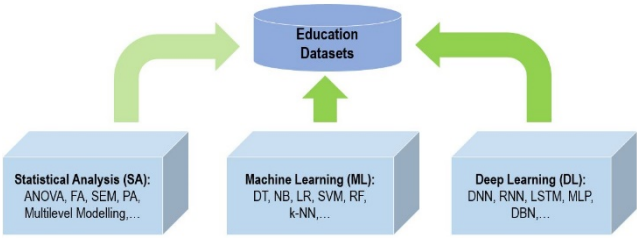


Fig. 1. Overview of research methods that have been applied in education research.

particular task; choosing one relies on experience. We evaluated a diverse set of algorithms on a dataset to see what works and drop what does not work. This process is called spot-checking algorithms.

3.1 Statistical Analysis

Statistical analysis is an integral technique in research. In an education environment, research popularly applies the analysis of variance (ANOVA) test, factor analysis, structural equation modeling, path analysis, and multilevel regression to observe the correlation and structure of indicators in academic performance. Among these techniques, structural equation modeling (SEM) is a popular tool that was generated by the geneticist Sewall Wright [20]. This technique is desired by numerous researchers since it is used to decompose multiple and interrelated dependencies in a single analysis.

SEM is constructed with two main models: a structural model and a measurement model [21, 22].

- **The structural model** (also called the inner model) presents the relationships or paths among the constructs (or latent variables). The structure shows the relation within latent variables and the regressions of latent variables on an observed variable.
- **The measurement model** is built from implicit or explicit models that relate the latent variables to its indicators.

SEM is constructed of factor analysis and multiple regression analysis. Factor analysis is a dimensional reduction technique used in SEM structure to extract factors (components) from a group of indicators that are related to each factor. We used the Kaiser-Meyer-Olkin (KMO) test to measure how suited our data are for factor analysis. We obtain $KMO = 0.87 > 0.5$, which shows that our data are acceptable for factor analysis [3]. Each factor in the measurement model can be written in the form of a linear regression (1):

$$F_i = w_{i1}x_1 + w_{i2}x_2 + \ldots + w_{ip}x_p + e_i, \quad (1)$$

where F is a factor, w is a coefficient of vector x, x is an indicator or manifest variable, and e is an error term.

In order to measure the goodness-off fitted of the causal model, a statistical measure such as a comparative fit index (CFI) and the root mean square of error approximation (RMSEA) are used. The consistency index results of the structural equation modeling are analyzed

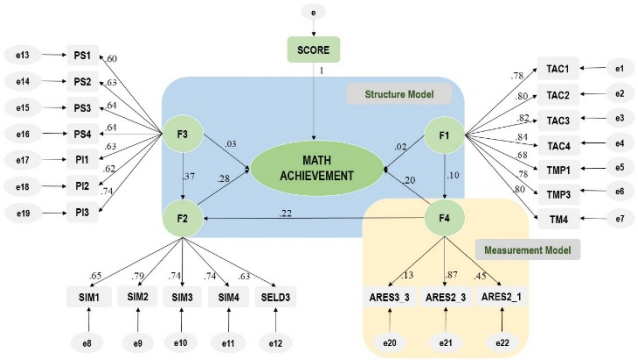


Fig. 2. SEM structure for detecting factors affecting student performance in mathematics.

Table 2. Model fitness index.

CMINDF	GFI	AGFI	NFI	CFI	RMSEA
4.12	0.91	0.90	0.90	0.93	0.05

using another criterion such as GFI, AGFI, and CMINDF. The results of the model-fit-index are shown in the Table 2.

Table 2 shows that $1 < CMINDF < 5$, and the values of GFI, AGFI, NFI, and CFI are all greater than 0.90 [23, 24]. Our proposed SEM structure is an extension of previous structure [2, 25, 26]. The threshold of RMSEA is less than 0.08. Uysal [3] stated that an RMSE equal to 0.05 or lower shows perfect data suitability. Hair [23] proposed that a good RMSEA is less than 0.06. Steiger [24] mentioned an appropriate RMSEA is less than 0.07. Hooper [27] stated that the general reported conjunction with RMSEA in a well-fitting model has a restriction of the upper limit less than 0.08 and a lower limit close to 0. In our SEM structure, we found RMSEA is 0.05, which satisfied even the strongest threshold.

The classical SEM was introduced as an underlying model to analyze the interrelated structure between measured variables and latent constructs [21]. In our work, we proposes a novel version of SEM called predictive SEM (P-SEM). P-SEM was used for detecting the causal structure of factors or students' learning behaviors that affect their academic performance and to decompose the prediction model for predicting students' outcomes. The threshold model was introduced for the categorical response [22]. We used the merits of factor analysis and multiple regression analysis to adapt our feature-detecting model to a prediction model, P-SEM.

3.2 Machine Learning

The application of machine learning covers a wide range of research. In the last decades, machine learning has been a paradigm shift in evaluating higher education tasks and to extract the hidden useful information and desired knowledge to support the decision process for the progress of students. Large-scale and complex education databases require suitable prediction and classification algorithms to discover meaningful patterns. Machine learning is an automated process that extracts these patterns from data

[28-30].

The four main tasks of machine learning are supervised learning, unsupervised learning, reinforcement, and a recommender system. Supervised learning is a special class of machine learning tasks that creates a rule for predicting and classifying unseen data by experience from a given set of instances. The algorithms used in this task determine the learning function L using input data E (experience or training instances) and output variables O, denoted as $O = L(E)$. Hence, given any arbitrary (test) data, we can predict the output variables. The aim is to obtain a learning function L, which is called a prediction model (algorithm). Machine learning algorithms such as the decision tree (DT), naïve Bayes (NB), logistic regression (LR), support vector machine (SVM), random forest (RF), K-nearest neighbors (K-NN), and artificial neural network (ANN) methods are popularly used to predict academic performance.

A distinct algorithm handles a problem differently, and there is no one algorithm for machine learning. Different tasks require different algorithms. Hence, it is a good idea to observe different types of algorithms to see what works best. We observed many machine learning classifiers [28-32], yet we propose seven algorithms that are best known for prediction models: ANN, sequential minimal optimization (SMO) of SVM, LR, K-NN, C5.0, CART, and RF.

3.2.1 Ensemble Methods

Ensemble methods are machine learning techniques that combine several baseline models in order to preduce an optimal reodictive model. Here, ensemble methods are applied to improve the performance of our prediction models. From our previous study in paper [1], we found that the C5.0 and RF outperformed the other proposed models. In this study, we improved our tree-based models with ensemble methods. The two proposed ensemble methods are boosting ang bagging. Here, a boosted tree-based algorithm named Boosted C5.0 and a bagging tree-based algorithm named Bagged CART and RF are proposed for improvement of our proposed prediction models.

3.3 Deep Learning

Deep learning is a broader family of machine learning via the extension of ANN architecture. Recently, a deep learning framework has shown its potential and effectiveness in many complicated problems [33]. Some particular approaches of deep learning were applied to education data mining. Deep learning algorithms such as a deep neural network (DNN), multilayer percentron (MLP), recurrent neural network (RNN), long short time memory (LSTM), and deep belief network (DBN) were used to analyze logged data and time-series data from the education databases [17-19, 33, 34].

The special architecture of DBN effectively solved many problems in healthcare, finance, and natural langue processing. Various applications of DBN have been applied to a number of recent studies ranging from image classification [35-37], speech recognition and audio classification [38], and healthcare [39]. We are interested in the performance of DBN, so in this study, we proposed a DBN model as a novel trend of deep learning tool for predicting academic performance.

A restricted Boltzmann machine (RBM) is a restricted version of a Boltzmann machine used for dimensionality reduction, classification, regression, collaborative filtering, feature learning, and topic modeling. RBM is a generative stochastic framework of ANN that learns a probability distribution over its set of inputs, with the restriction that their visible units and hidden units must form a fully connected bipartite graph, and there is no connection between them in the same layer. It is a two-layer stochastic network. A combination of RMB frameworks is stacked with each other to generate a special case of a deep learning framework, called a deep belief network (DBN). The DBN architecture was conceived by Geoffrey Hilton [40]. We propose a DBN framework as a classification and prediction model to predict high school student performance in mathematics. The DBN flow contains pre-training following a greedy layer-wise learning procedure [40]. One layer is added on top of the network at each step, and only the top layer is trained as an RBM using contrastive divergence (CD) [41]. After each RBM has been trained, the weights are clamped, and a new layer is added to repeat the same procedure.

3.3.1 Training Procedure of DBN Model

A DBN training is divided into two main stages. The first stage is an unsupervised pre-training called training section that learns features only with the input values without a label. The second stage is tuning with backpropagation algorithm using a label with fine-tuning called tuning section.

1) Pre-training Section: The training section is implemented by using layer by layer training approach called "Greedy Layer" training algorithm. This section learning the parameters in each layer of the RBM. The first RBM is trained using CD algorithms to reconstruct its inputs to be accurate as possible. The hidden layer of the first RBM is treated as a visible layer for the second RBM, and the output of second layer is subsequently inserted to be aa visible layer of the next RBM, and so on. This procedure is repeated until every predefined hidden layer in the network is trained. This pre-training is implemented to obtain the pre-trained parameters before the supervised training work.

2) Fine-tuning Section: This section is supervised training stage. It uses labeled samples to supervise the DBN model in the top-down phase. The trained weights and biases of the RBM layers from unsupervised RBM learning are utilized in this section. The algorithms use parameters obtained in the training section as it will avoid falling into local optimum and overfitting which caused by random initialization of the traditional network. The supervised learning process further reduces the training error and improves the classification accuracy of the DBN classifier.

We optimized our proposed DBN by executing various

Table 3. The Hyper Parameters.

Pretraining	Values	Fine-tuning	Values
Hidden Layer	2 layers	-	-
Neuron in Layer	100-100	Momentum	0.001
No. of Epochs	30	No. of Epochs	50
Learning Rate	0.8	Learning Rate	1
Mini-batch size	30	Mini-batch size	10

models with distinct numbers of hidden layers, numbers of nodes in each hidden layer, and other hyper-parameters and we found the optimal hyper parameters in Table 3.

4. Data Collection and Preprocessing Tasks

4.1 Data Description

Analysis of academic performance is a challenging task since it relies on diverse hidden factors like domestic factors, individual factors, and schooling factors. We tried to observe all relevant features since many of effective factors are not always found in school databases [2-5, 26, 27, 37, 38]. That is why our attributes are observed from all various main factors, and we tried to observe the contribution of these attributes to the success of student performance. In our study, we aim to predict the students' performance and to analyze the related features that influence their achievement in mathematics.

In a dataset $D \in R^{n \times d}$ suppose we have $s_1, s_2, .., s_{1n} \in R^d$ as row vectors. Let s_i be a student such that $s_i = (F_{i1}, F_{i2}, \ldots, F_{in})$ with each F_{ij} being the value of some feature (or attributes) or information describing that student. The objective of the study is to predict the student performance via the effect of features $(F_1, F_2, .., F_n)$ on student performance $\mathbf{P(s)}$ denoted with the following mapping:

$$s = (F_1, F_2, ..., F_n) \rightarrow \mathbf{P(s)}.$$

We collected a dataset of descriptive features from three main factors, domestic features, individual features, and schooling features, which include home background, academics, and attitudinal information describing each student. The data collection was done by giving out questionnaires. The questionnaires were prepared with references, assistance, and guidance from (i) reviewing literature, (ii) teachers from diverse schools, (iii) staff from the department of research (MOEYS: Ministry of Education Youth and Sport, Cambodia), and (iv) senior researchers in education. The responses were obtained from (i) students from various high schools in Cambodia, (ii) teachers in corresponding high schools, and (iii) score records from administrative offices.

Finally, after the data collection process, we obtained a dataset of 1204 records. Table 5 describes the characteristics of 43 attributes from the three main factors containing various sub-factors with shorthand variable notation listed on the left-hand side. The data type used in this study is categorical data, including binary, ordinal, and nominal data. The nominal types represent different classes of features, the ordinal type corresponds to the 5-point Likert scale of response of 1, 2, 3, 4, and 5, which represent the answers "Strongly Disagree," "Disagree," "Neural," "Agree," and "Strongly Agree," respectively. Binary responses are "Yes," and "No".

Table 4. Target variable.

Classes	Performance Levels	Score Based
1	Excellent	90 - 100%
2	Good	75 - < 90%
3	Average	60 - < 75%
4	Poor	< 60%

The output variable measured by the final scores of students in mathematics that is discretized into four performance levels shown in Table 4.

We labeled our student performance with four main classes (Poor, Average, Good, and Excellent) based on their score in mathematics in semester I. The four categories of the performance levels are presented in Table 6.

4.2 Data Preprocessing

The best algorithm is useless unless the data are ready for implementation. Coming up with features is difficult and time-consuming and requires expert knowledge. Data preprocessing is an integral step data mining as the quality of data and the useful information that can be derived from it directly affect the ability of the model. Normally, real-world data will not always be neat. It is often incomplete, inconsistent, likely to contain many errors, and not in an executable format. Data preprocessing is a proven method of resolving such issues. The contribution of this step is to obtain data ready for execution and improving the accuracy of algorithms.

- **Data Cleaning:** Even though data collecting was done with careful concentration, there are still many reasons for missing data, such as data not being continuously collected, technical problems with bio-metrics, a mistake in data entry, and much more. Ignoring this is not a good choice since it may play an important role or when data become large, it will lead to an inaccurate result. Data cleaning deals with the processes of filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data. In our datasets, the number of missing values is low, so we used the imputing methods in order to cean our data. The student questionnaire completion was done with missing some questions and inputing invalid number (outliers). In case of missing value in categorical variables, we replaced it by the modes or frequency-category value. In the output variable, there are a few missing values and outlier, we replaced it by the mean value.
- **Data Transformation:** Data transformation is the

Table 5. Factors affecting student performance.

Factors	Sub Factors	N	Attributes	Description	Type
Domestic Factors	Parents' education	1	PEDU1	Father's education level	Nominal
		2	PEDU2	Mother's education level	Nominal
	Parents' occupation	3	POCC1	Father's occupational status	Nominal
		4	POCC2	Mother's occupational status	Nominal
	Family's income	5	PSES	Family's socioeconomic status	Ordinal
	Parental involvement	6	PI1	Parents' attention to student's attitude	Ordinal
		7	PI2	Parents' time and money spending	Ordinal
		8	PI3	Parents' involvement in education	Ordinal
	Parenting styles	9	PS1	Parents' feeling responsive to needs	Ordinal
		10	PS2	Parents' response to children's attitude	Ordinal
		11	PS3	Parents' encouragement	Ordinal
		12	PS4	Parents' compliments	Ordinal
	Domestic environment	13	DE1	Domestic environment for study	Ordinal
		14	DE2	Distance from home to school	Nominal
Individual Factors	Self-discipline	15	SELD1	Number of hours for self-study	Nominal
		16	SELD2	Number of hours for private math study	Ordinal
		17	SELD3	Frequency of doing math homework	Ordinal
		18	SELD4	Frequency of absence in math class	Ordinal
		19	SELD5	Frequency of preparing for the math exam	Ordinal
	Student's interest and motivation	20	SIM1	Student' s interest in mathematics	Ordinal
		21	SIM2	Student's enjoyment of math lecture	Ordinal
		22	SIM3	Student's attention in math class	Ordinal
		23	SIM4	Student's motivation to succeed in math	Ordinal
	Student's anxiety Toward math	24	ANXI1	Student's anxiety in math class	Ordinal
		25	ANXI2	Student's nervousness in the math exam	Ordinal
		26	ANXI3	Student's feeling helpless in math	Ordinal
	Student's possession	27	POSS1	Internet use at home	Binary
		28	POSS2	Possession of computer	Binary
		29	POSS3	Student's study desk at home	Binary
School Factors	Class environment	30	CENV1	Classroom environment	Ordinal
	Curriculum	31	CU1	Content's languages in math class	Nominal
		32	CU2	Class sessions	Nominal
	Teaching methods and practice	33	TMP1	Teacher's mastery in math class	Ordinal
		34	TMP2	Teacher's absence in math class	Ordinal
		35	TMP3	Teaching methods in math class	Ordinal
		36	TMP4	Teacher's involvement in education's content	Ordinal
	Teacher's attribute and characteristics	37	TAC1	Math teacher's ability	Ordinal
		38	TAC2	Teacher's encouragement to students	Ordinal
		39	TAC3	Math teacher's connection with students	Ordinal
		40	TAC4	Math teacher's help	Ordinal
	Academic resource	41	ARES1	Availability of math teacher	Nominal
		42	ARES2	Availability of classroom	Nominal
		43	ARES3	Availability of math handout	Nominal

process of transforming or converting data from one format or structure into an executable form and not missing important information. The techniques are concerned with replacing values, label encoding, dummy encoding, and one-hot encoding. Some algorithms require normalization such as z-score and min-max transformation. In our experiment, the features of nominal and binary types were encoded into numeric type to be ready for execution. In neural network models (ANN and DBN), the input features

Table 6. Percentage of four predefined classes of student performance levels.

Classes	Poor	Average	Good	Excellent
Percentage	27.49%	15.60%	27.90%	29.01%

were normalized with max-min normalization and the target variable was encoded with one-hot encoding.

- **Data Discretization:** Data discretization is an important concept for dealing with a large range of numerical values and classifying it into ordinal or nominal values. In our study, we descritized the mathematics scores into predifined classes to categorize the performance levels of students.
- **Dimensional Reduction:** Dimensional reduction is one of the effective techniques used in a large dataset in the preprocessing step. We introduced feature selections methods in our study for two main purposes. Firstly, a features selection method is introduced in the preprocessing step for obtaining the optimal feature subset containing relevant and important features. Secondly, it was introduced for mining the highly influencing features that affected the student performance.

R language is one of the most popular choices for dealing with data and machine learning implementation. R is a programming language, interpreter, and platform. The experiment in our work was done using R Studio, an integrated development environment (IDE) for R.

5. Experimental Results and Analysis

We did a comparative study of multiple models of educational data mining to predict and classify high school student performance. We used two standard evaluation metrics: accuracy, and predictive mean squared error (PMSE).

5.1 Performance Evaluation Metrics

We propose two standard performance metrics to evaluate the efficiency of the proposed algorithms. In these metrics, we denote some terms like TP (true positive or simply said correctly predicted), E (error or incorrectly predicted), G^a (actual grade), and G^p (predicted grade). The first measurement metric is accuracy, which is used to measure the percentage of correct prediction.

$$\text{Accuracy} = \frac{\sum \text{TP}_i}{\sum \text{TP}_i + \sum \text{E}_{ij}} \times 100\% . \quad (2)$$

Another metric is PMSE, which is used to quantify the closeness of the predicted output to the actual output. It can be calculated as the average of the squared distance between predicted grade and the actual grade. We classified the performance levels of students in mathematics into four levels: Poor, Average, Good, and Excellent, which are represented by 1, 2, 3, and 4. Using a confusion matrix, PMSE can be easily calculated via Eq. (3):

$$\text{PMSE} = \frac{1}{M} \sum_{i=1}^{M} (G_i^a - G_i^p)^2, \quad (3)$$

where $G^a \in \{1,2,3,4\}$ is the actual predefined grades (classes) and $G^p \in \{1,2,3,4\}$ is the predicted grades.

5.2 Experimental Results

Tables 6 and 7 summarize the accuracy and PMSE results of the proposed model. A single tool may perform well in a particular problem, yet poorly in another problem. We compared three main categories of research techniques in this analysis. We observed several tools in these categories, but only the most effective were chosen. These techniques are the P-SEM of statistical analysis , seven superior classes (ANN, SMO, LR, K-NN, Boosted C5.0, Bagged CART, and RF) of machine learning algorithms, and DBN. For the instance-based K-NN classifier, we implemented the algorithm with various values of $K = 1:30$ (ranging from 1 to 30 where

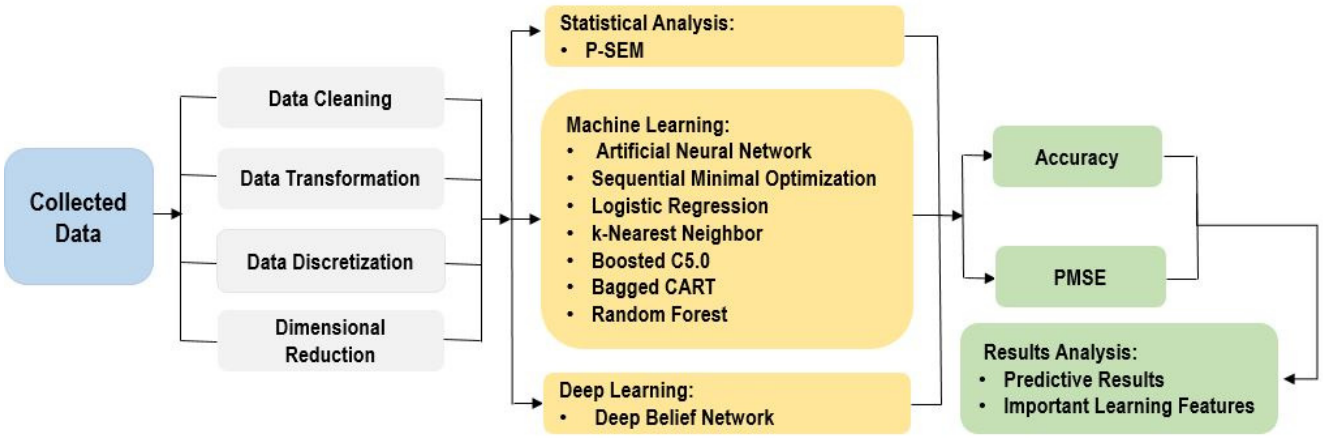


Fig. 3. The analysis procedure for predicting student performance in mathematics.

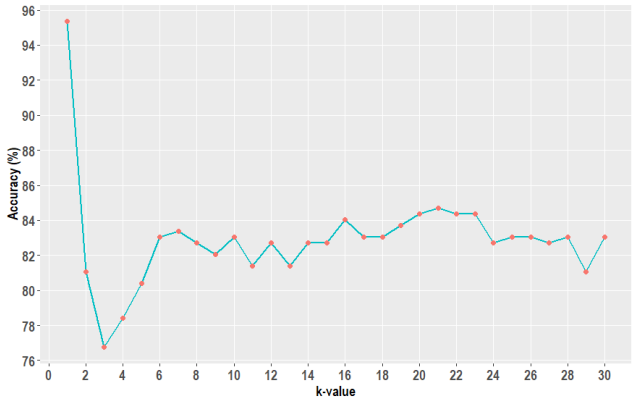


Fig. 4. Performance of K-NN with different values of parameter K.

Table 6. Accuracies of the proposed models.

Classes	Models	Lowest Acc.	Highest Acc.	Average Acc.	Std.
Statistical Analysis	P-SEM	31.50%	45.20%	38.20%	5
Machine learning	ANN	48.50%	58.12%	53.22%	4
	SMO	86.52%	94.14%	90.66%	3
	LR	52.75%	60.34%	56.37%	3
	1-NN	94.01%	96.01%	94.95%	0.8
	Boosted C5.0	95.20%	96.18%	95.67%	0.3
	Bagged CART	95.22%	96.05%	95.60%	0.3
	RF	**96.04%**	**97.10%**	**96.69%**	**0.3**
Deep Learning	DBN	72.60%	79.78%	75.76%	3

$30 = [\sqrt{\#\text{trainingset}}]$). However, the best performance of this model was obtained at $K = 1$, as shown in Fig. 4.

In Table 6, the accuracy of the nine proposed models ranges from 38.20 to 96.69%. The four leading algorithms are 1-NN (94.95%), and the three ensemble tree-based algorithms: Boosted C5.0 (95.67%), Bagged CART (95.60%), and RF (96.69%). Using ANOVA with a significance level of $p = 0.05$, we found that 1-NN is significantly different from the other models. There is no difference between the Boosted C5.0 and Bagged CART algorithms. The RF algorithm is statistically different from the other with the highest accuracy, which shows the robustness of RF and implies that it is the best predictive algorithm in this problem.

The other objective is to measure how close our prediction is to the actual grade of students. PMSE is used to measure the goodness of fit of the model. Using the confusion matrix of the algorithms and Eq. (3), we can compute PMSE , as shown in Table 7.

Fig. 6 shows the performance of the proposed models via PMSE. In contrast to accuracy, PMSE measures the average of the squared distance from the predicted output to the actual output. The smaller the PMSE, the better the model is. A zero value of PMSE shows that model is perfect. RF produced a very small PMSE, which shows that the prediction model is the best classifier and produces very little error.

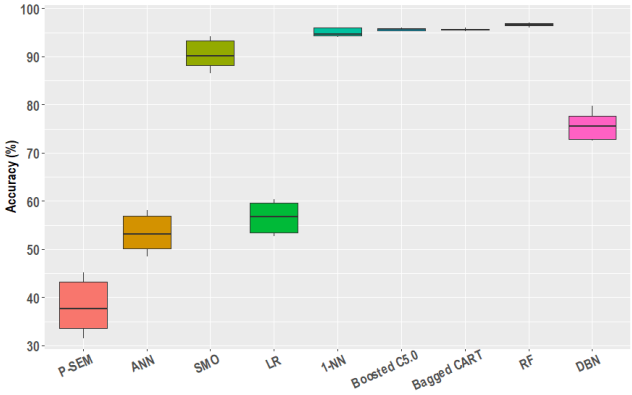


Fig. 5. Boxplot indicating the accuracies of the proposed models.

Table 7. PMSE of the proposed models.

Classes	Models	PMSE	Std.
Statistical Analysis	P-SEM	1.385	0.055
Machine learning	NN	1.162	0.041
	SMO	0.504	0.062
	LR	1.015	0.083
	1-NN	0.068	0.005
	Boosted C5.0	0.015	0.004
	Bagged CART	0.015	0.003
	RF	**0.013**	**0.003**
Deep Learning	DBN	0.714	0.042

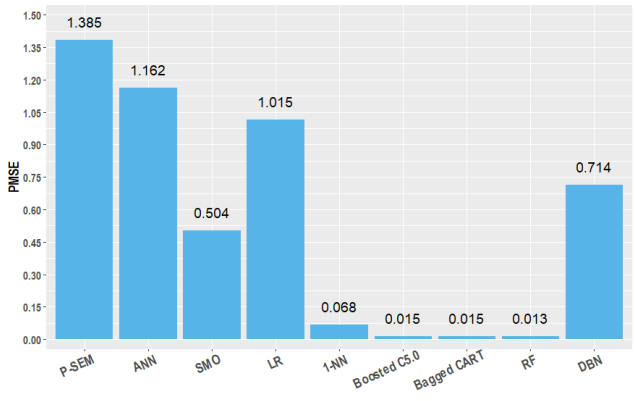


Fig. 6. Comparison of PMSE of the proposed models.

6. Feature Selection

In dealing with data, the quality and information of data are very important. Feature selection is an integral technique in statistics, pattern recognition, machine learning, and data mining. Feature selection solves two main important tasks. The first main problem in supervised learning is overfitting. Overfitting lowers the accuracy of algorithms due to redundant and irrelevant features. Irrelevant or less informative variables provide no more information and disturb the performance of the model. They are also called noisy features. To solve this problem, we qualify the data by choosing the optimal subset of data containing the most relevant features using the feature selection technique. The second task is to rank the most important features. A crucial concept in the education field is extracting useful information and learning the behaviors of students in the learning process [39, 40]. That is why the last objective of our study is to introduce techniques to identify the highly influenced features of student performance in mathematics.

In order to effectively and accurately select the informative features, we proposed several learning feature selection algorithms. Feature selection can be classified into three categories: embedded approaches, wrapper approaches, and filter approaches [39]. We introduce four feature selection methods: Information Gain Feature Evaluation (IGFE), Chi-Square Based Feature Evaluation (CSBFE), Symmetrical Uncertainty Feature Evaluation (SUFE), and Relief Feature Evaluation (RFE) [40, 41].

a) Information Gain Feature Evaluation (IGFE)

Information Gain (IG) is an important measurement used for ranking informative features. It explains how well a given attribute separates instances with respect to their target classes. IG uses a concept based on Shannon's informative entropy to split the instances. The Shannon informative entropy determines the impurity of an example dataset D with a different class c_i.

$$I(D) = -\sum_{i=1}^{c} P(c_i)\log_2 P(c_i), \tag{4}$$

where c is the number of target classes and $P(c_i)$ is the proportion of examples in D with respect to class c_i. For $D = \{D_1, D_2, ..., D_m\}$, the m partitions of D, the infrmation of attribute A in example dataset D can be calculated with a summation rule (5):

$$I(D \mid A) = \sum_{i=1}^{m} \frac{|D_i|}{|D|} \times I(D_i). \tag{5}$$

The information of attribute A in dataset D is denoted as:

$$IG(D, A) = I(D) - I(D \mid A). \tag{6}$$

b) Chi-Squared-Based Feature Evaluation (CSBFE)

The chi-squared method is a statistical technique to determine the dependency on each input variable to the output variable. The test uses the idea of chi-squared scores of the classes to obtain a ranking list of all attributes. The list of order informative features can be obtained through the following equation:

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c} \frac{(n_{ij} - \varepsilon_{ij})^2}{\varepsilon_{ij}}, \tag{7}$$

where l is the number of classes or determined intervals of a particular feature, c is the number of classes in a target variable, n_{ij} is the observed (actual) frequency of a sample of the i^{th} interval and j^{th} class, and ε_{ij} is the expected frequency of n_{ij}.

c) Symmetrical Uncertainty Feature Evaluation (SUFE)

Symmetric Uncertainty (SU) is one of the leading feature selection techniques. SU determines the correlation between a feature and target variable using entropy and information gain theory. The calculating can be done via the following equation:

$$SU(A, D) = \frac{IG(D, A)}{I(D) + I(A)}, \tag{8}$$

where $I(D)$ and $I(A)$ are entropies based on probability of a class associated with the example set D and attribute A, respectively, and $IG(D, A)$ is the information gain shown in Eq. (6).

d) Relief Feature Evaluation (RFE)

The relief algorithm is used to select samples at random, compute the difference between nearest neighbors, and modify a feature weighting vector to give higher weight to features that separate the instance from neighbors of different classes. The relief tries to compute the probability to assign the weight for each feature A:

$$\begin{aligned} w_f = {} & P(\text{different value of } A \mid \text{different class}) \\ & - P(\text{different value of } A \mid \text{same class}) \end{aligned} \tag{9}$$

Table 8. Ranking top-10 important features.

Rank	Features	Description
1	SIM1	Student's interest in math
2	SELD1	Number of hours for self-study
3	SELD3	Frequency of math doing homework
4	PEDU2	Mother's education level
5	ARES3	Availability of math' handouts
6	POCC1	Father's occupational status
7	ANXI1	Level anxiety in math class
8	AXNI3	Level of feeling helpless in class
9	SIM2	Level of enjoying with math lecture
10	ANXI2	Level of nervous in the math exam

The ranks were manually generated for every attribute by each technique (IGFE, CSBFE, SUFE, and RFE). We calculated the average rank of each attribute and provided a the top-10 list in Table 8.

The level of interest of the students in the subject of mathematics is the most important, followed by the number of hours of self-study and frequency of doing homework. Moreover, the mother's education level also affects the students' performance, which maybe due to the involvement of the mother in following up on their child's study. There is still lacking in mathematics' handout due to limited sources of handout of each school, restriction in language usage, and methods of obtaining a handout. Anxiety in mathematics class, the feeling of helplessness, and nervousness on exam are still problems for children; this suggests more assistance from teachers is needed. From the early prediction, we can identify the group of students poor performance. With the results from feature selection, we can identify the behaviors or causes that affect student performance, which will be used as for improvement and giving assistance to at-risk groups.

7. Discussion and Conclusion

EDM approaches contribute a high level of knowledge extraction to the education environment. In this study, we wished to predict high school student performance with an analysis of various algorithms EDM. Data were collected from various high schools in Cambodia. We reviewed education research, its purposes, algorithms, key findings, and the popular algorithms applied in the education environment.

To maximize the information and benefits for education institutes, we introduced multiple models in this study. We produced multiple classifiers to find the model that can best classify the performance levels of students in mathematics. Hence, three different main categories of research methods were studied and compared. We applied the spot-checking algorithm process to evaluate diverse algorithms on our dataset and see what works and drop what does not work. We also optimized the algorithms. We boosted the tree-based algorithms with ensemble methods. We found that 1-NN, Boosted C5.0, Bagged CART, and RF generated accuracies of 94.95%, 95.67%, 95.60%, and 96.69%, respectively. The RF algorithms generated the highest accuracy with the lowest PMSE as the best algorithm.

We observed the main effective feature evaluation approaches to obtain a better understanding of important factors affecting student performance. We performed a series of approaches to extract the importance of features, and results showed that the top 10 features that had a gread impact on student performance in mathematics were students' interest in mathematics, amount of self-study, frequency of doing homework, mother's education level, math handouts, father's occupational status, level of anxiety in math class, level of feeling helpless in class, level of enjoyment in math class, and level of nervousness in math exam. If we can effectively predict and classify the performance level of students and understand the learning behaviors and affect associated features, then we can provide the right action for improvement and enhancement. The results from this research could provide a warning for poor-performance students and effective learning behaviors that could be used for students, parents, teachers, and related educators to keep track of student performance.

In conclusion, this paper presented a comparative study of EDM models in predicting high school student performance in mathematics in Cambodia. Additonally to previous study [1], we improved our prediction models by using feature selecton methods to obtain optimal feature sets and then applying ensemble methods for our tree-based algorithms. Using our actual dataset of 1204 samples, we found that RF algorithm generated the highest accuracy and the smallest PMSE with accurate result. From ranking informative features of feature selection, we obtain highly influencing factors that affect student performance in mathematics. The results from this study will be used as (i) the early warning to poor performance group of students in high schools (ii) recommendation and action strategies regarding with mining highly influencing factors and learning behaviors, and (iii) managerial settings, scheduling and planing in educational institutions and STEM discipline in Cambodia.

References

[1] Sokkhey P. and Okazaki T., "Comparative Study of Prediction Models on High School Student Performance in Mathematics", *Journal of IEIE Transaction on Smart Processing and Computing,* Vol. 8, No. 5, pp. 394-404 Oct. 2019. Article (CrossRef Link)

[2] Mohamed Z.G. A., Mustafa B. M., Lazim A., and Hamdan A. M., "The Factors Influence Students' Achievement in Mathematics: A Case for Libyan's Students". *Australian Journal of Basic and Applied Science*, Vol. 17, Issue 9, pp. 1224-1230, 2012. Article (CrossRef Link)

[3] Uysal S., "Factors Affecting the Mathematics achievement of Turkish students in PISA 2012", *Academic Journals*, Vol. 10, Issue 12, pp. 1670-1678, Jun. 2015. Article (CrossRef Link)

[4] Asanee Tongsilp, "A Path Analysis of Relationships

between Factors with Achievement Motivation of Students of Private Universities in Bangkok, Thailand", *Procedia-Social and Behavioral Sciences*, Vol. 88, pp. 229-238, 2013. Article (CrossRef Link)

[5] Kilic S. and Askin O.E., "Parental Influence on Students' Mathematics Achievement: the Comparative Study of Turkey and best performer countries in TIMSS 2011", *Procedia-Social and Behavioral Sciences*, Vol. 106, pp. 2000-2007, 2013. Article (CrossRef Link)

[6] Stephen J.H. Yang et al., "Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis", *Journal of Information Processing*, Vol. 26, pp. 170-176, Feb. 2018. Article (CrossRef Link)

[7] Kotsiantis S., Piarrekeas C., and Pintelas P., "Predicting Students' Performance in Distance Learning using Machine Learning Techniques", *Applied Artificial Intelligent*, Vol. 18, pp. 411-426, 2007. Article (CrossRef Link)

[8] Minaei-Bidgoli B., Kashy D.A., Kortemeyer G., and Punch W.F., "Predicting Student Performance: An Application of Data Mining Methods with the Education Web-Based System LON-CAPA", *33rd ASEE/IEEE Frontiers in Education Conference*, Vol. 1, 2003. Article (CrossRef Link)

[9] Ermiyas B., and Feidu A.G., "Student Performance Prediction Model using Machine Learning Approach: The Case of Wolkite University", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 7, Issue 12, Feb. 2017. Article (CrossRef Link)

[10] Shanthini A. Vinidhini G., and Chandrasekaran R.M., "Predicting Students' Academic Performance in the University Using Meta Decision Tree Classifiers", *Journal of Computer Science*, Vol. 14, Issue 5, pp. 654-662, Feb. 2018. Article (CrossRefLink)

[11] Kurma M., and Sigh A.J.., "Evaluation of Data Mining Techniques for Predicting Students' Performance", *I.J. Modern Education and Computer Science*, Vol. 8, pp. 25-31, 2017. Article (CrossRefLink).

[12] Wiyono S., and Abidin T., "Comparative Study of Machine Learning KNN, SVM, and Decision Tree Algorithm to Predict Student's Performance", *International Journal of Research-GRAMTJAALAYAH*, Vol. 1, Issue 1, Jan. 2019. Article (CrossRefLink)

[13] Pimpa Cheewaprakokit, "Study of Factors Analysis Affecting Academics Achievement of Undergraduate Students in International Program", Proceedings of International MultiConference of Engineers and Computer Sciences, Vol 1, Mar. 2013. Article (CrossRefLink)

[14] Amrieh E.A., Hamtini T., and Aljarah I., "Mining Education Data to Predict Student's Academic Performance Using Emsemble Methods", *International Journal of Database Theory and Application*, Vol. 9, No. 8, pp. 119-136, 2016. Article (CrossRefLink)

[15] Ramaswami M., and Bhaskaran R., "A Study on Feature Selection Techniques in Education Data Mining", *Journal of Computing*, Vol. 1, Issue 1., Dec. 2009. Article (CrossRefLink)

[16] Afendey L. S., Paris I.H.M., Mustapha N., and Salaiman M.N., "Ranking Influencing Factors in Predicting Students Academic Performance", *Information Technology Journal*, Vol. 9, pp. 832-837, Apr. 2010. Article (CrossRefLink)

[17] Arindam M., and Joydeep M., "An Approach to Predict Student Performance Using Recurrent Neural Network (RNN)", *International Journal of cComputer Applications*, Vol. 181, No. 6, Jul. 2018. Article (CrossRefLink)

[18] Sinthia G., and Balamurgan M., "Analyzing Student's Academic Performance Using Multilayer Perceptron Model", *International Journal of Recent Technology and Engineering*, Vol. 7, Issue 5S3, 2019. Article (CrossRefLink)

[19] Zhang Y., Shah R., Chi M., "Deep Learning+Student Modeling+Clustering: A Recipe for Effective Automatic Short Answer Grading", *Proceeding of the 9th International Conference on Education Data Mining,* 2016. Article (CrossRefLink)

[20] Wright S., "Correlation and Causation", *Journal of Agriculture Research*, Vol. 20, No. 7, pp. 557-586, 1921. Article (CrossRefLink)

[21] Randall E.S., and Richard G.L., *A Beginner's Guide to Structural Equation Modeling (3rd ed.)*, Routledge, 2010. Article (CrossRefLink)

[22] Anders S., and Sophia R.H., "Structural Equation Modeling: Categorical Variables", *Entry for the Encyclopedia of Statistics in Behavioral Science*, Wiley, 2005. Article (CrossRefLink)

[23] Hair J.F., Black B., Balin B., Anderson A.E., Tatham R.L., *Multivariate Data Analysis (7th ed.)*, New Jersey:Prentice-Hall, 2010. Article (CrossRef Link)

[24] Steiger, J.H, "Structural Model evaluation and modification", *Multivariate Behavioral Research*, 25, 1990.

[25] Mohamed Z.G.A., Mustafa B.M., Lazim A., and Hamdan A.M., "Path Analysis of the Factor Influencing Students' Achievement in Mathematics", *Australian Journal of Basic and Applied Science,* Vol. 7, Issue 4, pp. 437-442, 2013. Article (CrossRefLink)

[26] Mohamed Z.G.A., Mustafa B.M., Lazim A., and Hamdan A.M., "Two Models of Describing Students' Achievement in Mathematics: A Comparative Study", *Australian Journal of Basic and Applied Science,* Vol. 7, Issue 6, pp. 184-198, 2013. Article (CrossRefLink)

[27] Hooper D., Coughlan J., Mullen M., "Structural Equation Modelling: Guiding for Determining Model Fit", *Electronic Journal of Business Research Methods,* Vol. 6, Issue 1, pp. 53-60, 2008. Article (CrossRef Link)

[28] John D.K., Brain M.N., and Aoife D., *Fundamentals of Machine Learning for Predictive Data Analytics (1st ed.)*, Cambridge: The MIT Press, 2015. Article (CrossRef Link)

[29] Scott V. Burger, *Introduction to Machine Learning with R: Rigorous Mathematical Analysis (1st ed.)*, O' Reilly Media, Mar. 2018. Article (CrossRefLink)

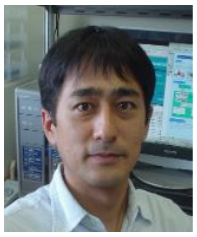
[30] Mohri M., Afshin R., and Ameet T., *Foundations of Machine Learning (2nd ed.)*, The MIT Press, 2018. Article (CrossRefLink)

[31] YU-Wei, and Chu D., Batia, and AshishSignh, *Machine Learning with R Cookbook (2nd ed.),* PACKT Publishing, 2017. Article (CrossRefLink)
[32] Karthik R., and Abhishek S., *Machine Learning Using R: With Time Series and Industry-Based Use Cases in R (2nd ed.)*, Apress, 2019. Article (CrossRefLink)
[33] Goodfellow I., Bengio Y., and Courville A., *Deep Learning,* The *MIT Press*, 2016. Article (CrossRefLink)
[34] Hinton G.E., Osindero S., and Teh Y.W., "A Fast Learning Algorithm for Deep Belief Networks", *Neural Computer,* Vol. 18, Issue 7, pp. 1527-54, 2006. Article (CrossRefLink)
[35] Hinton G.E., "A Practical Quide to Training Restricted Boltzmann Machines", *Momentum*, Vol. 9, Issue 1, 2010. Article (CrossRefLink)
[36] Bengio Y., Lamblin P., Popovici P., and Larochelle H., "Greedy Layer-Wise Training of Deep Networks", *Advances in Neural Information Processing Systems 19,* MIT Press, Cambridge, MA, 2007. Article (CrossRefLink)
[37] Caro D.H., McDonald J.T., and Wills J.D., "Socioeconomic Status and Academic Achievement Trajectories From Childhood to Adolescence", *Canadian Journal of Education*, Vol. 32, No. 3, pp. 558-590, 2009. Article (CrossRefLink)
[38] Serpil K. and Oykum E.A., "Parental Influence on Students' Mathematics Achievement: the Comparative Study of Turkey and Best Performer Countries in TMISS", *Procedia-social and behavioral science*, Vol.106, pp. 2000-2007, Dec. 2013. Article (CrossRefLink)
[39] Das Sanmay, "Filters, Wrappers and a Boosting-based Hybrid for Feature Selection", *In proceedings of the International Conference on Machine Learning 2001*, pp. 74-81, 2001. Article (CrossRefLink)
[40] Phyu T.Z., and Oo N.N, "Performance Comparison of Feature Selection Methods", *MATEC Web of Conferences42*, 2016. Article (CrossRefLink)
[41] Lei Ma, et al., "Evaluation of Feature Selection Methods for Object-based Land Cover Mapping of Unmanned Aerial Vehicle Imagery Using Random Forest and Support Vector Machine", *International Journal of Geo-Information*, Vol 6., Issue 56, 2017. Article (CrossRefLink)

Phauk Sokkhey was born in Kompong Thom province, Cambodia. He received his bachelor degree in Mathematics from Royal University of Phnom Penh (RUPP), Cambodia, in 2010, and later received his Master degree in Applied Mathematics from Suranaree University of Technology (SUT), Thailand, in 2013. He was a lecturer of mathematics at the department of foundation year at the Institute of Technology of Cambodia (ITC). Sokkhey is currently a Ph.D. student at the University of the Ryukyus, Japan. His current research are statistical analysis, machine learning, data science, and education data mining.

Sin Navy received her B.Sc degree in Matheamtics in 2009 from Royal University of Phnom Penh, Cambodia. She was a mathematics teacher in high school for three years. She later earned her Master's degree in Statistics from Institut Teknologi Sepuluh Nopember, Indonesia in 20015. Navy recently works as research assistance at Ministry of Education of You, and Sport of Cambodia. Her research interest are statistical analysis, qualitiative and quantitative researches on mining academic performance in educational system.

LY Tong was born in Batambang province, Cambodia. He obtained Bachelor's Degree in Mathematics from the Royal University of Phnom Penh (RUPP) in 2006 and later finished a Master's Degree in Mathematics from RUPP in 2010, Cambodia. In 2015, he obtained another Master's Degree in education, specializing in Mathematics Education from Hiroshima University, Japan. Currently, he is working as a researcher at the Royal Academy of Cambodia. His research interest are statistical analysis, data science, students' motivation, engagement and misconceptions in learning mathematics and other subjects.



Takeo Okazaki took B.Sc., M.Sc. from Kyushu University in 1987 and 1989, respectively. He had been a research assistant at Kyushu University from 1989 to 1995. He earned his Ph.D. from the University of the Ryukyus in 2014. He is currently a professor at the University of the Ryukyus. His research interests are statistical data normalization for analysis, statistical analysis, data analysis, genome informatics, tourism informatics, geographic information systems, and data science.