Hybrid Machine Learning Algorithms for Predicting Academic Performance

Phauk Sokkhey[1]
Graduate School of Engineering and Science
University of the Ryukyus
1 Senbaru, Nishihara, Okinawa, 903-0123, Japan

Takeo Okazaki[2]
Department of Computer Science and Intelligent Systems
University of the Ryukyus
1 Senbaru, Nishihara, Okinawa, 903-0123, Japan

***Abstract*—The large volume of data and its complexity in educational institutions require the sakes from informative technologies. In order to facilitate this task, many researchers have focused on using machine learning to extract knowledge from the education database to support students and instructors in getting better performance. In prediction models, the challenging task is to choose the effective techniques which could produce satisfying predictive accuracy. Hence, in this work, we introduced a hybrid approach of principal component analysis (PCA) as conjunction with four machines learning (ML) algorithms: random forest (RF), C5.0 of decision tree (DT), and naïve Bayes (NB) of Bayes network and support vector machine (SVM), to improve the performances of classification by solving the misclassification problem. Three datasets were used to confirm the robustness of the proposed models. Through the given datasets, we evaluated the classification accuracy and root mean square error (RSME) as evaluation metrics of the proposed models. In this classification problem, 10-fold cross-validation was proposed to evaluate the predictive performance. The proposed hybrid models produced very prediction results which shown itself as the optimal prediction and classification algorithms.**

Keywords*—Student performance; machine learning algorithms; k-fold cross-validation; principal component analysis

I. Introduction

The poor performance of students in high school has become a worried-task for educators as it affects the secondary national exam and step to higher education. Mathematics is considered as the basic background for many science subjects, and give very strongly affect the national exam and for further study in higher education [1]. For example, students who are poor in mathematics are much more likely to fail in diploma national exams in Cambodia [2]. They later found themselves harder to choose a major for higher study and hard to survive in the university journey. Early prediction and classification of student performance level offers an early warning and gives a recipe for improving the poor performance of students as well as for other managerial settings. Hence, we aim to deal with the unknown behavior pattern of students which affects student performance. There are various factors affect the performance of students in mathematics; those factors consist of schooling factors, domestics or home factors, and personal or individual factors. These related factors were used as predictive features in predicting the achievement of students in mathematics.

In the age of the information revolution, analysis of the database in education environments such as learning analytics, predictive analytics, educational data mining, and machine learning techniques has become a hot area of research [3-5]. The supervised learning was used to predict, classify the students' performance and analyze their learning behaviors to follow up on their progress in classes. However, the challenging task is to find the optimal algorithm which could produce satisfying results. Machine learning algorithms such as naïve Bayes, logistic regression, artificial neural networks, decision tree, random forest, support vector machine, k-nearest neighbor, and more, were popularly used to analyze and predict academic performance [3-14]. The performance of each model is varied from dataset to dataset, which relies on the characteristics and quality of data.

In the classification problem, a reason for misclassification that declines the performance of the model is from the quality of data that disturbs the algorithms. Various literature has focused on using dimensional reduction (feature selection and feature extraction methods) to improve the prediction and classification performance. In our work, we applied principal component analysis (PCA) as a feature extraction technique to transform the original dataset into a new dataset of high quality. We also introduced 10-fold cross-validation is to evaluate the predictive performance of the models and to judge how they perform in a new dataset, the testing samples or test data.

This paper aims at proposing a novel hybrid approach of machine learning for solving the classification problem. The proposed hybrid approach is the combination of four baseline machine learning algorithms with 10-fold cross-validation and principal component analysis.

II. Related Works

Supervised learning in machine learning requires an effective prediction model for solving prediction and classification problems. As mentioned in the Introduction, the educational data mining (EDM) field has studied different machine learning techniques to determine these techniques obtaining a high accuracy to predict the future performance of students [3-5].

Table I summarized the popular and state-of-the-art classification algorithms, which were used to predict student performance in educational datasets. Several works have been investigated to find the best algorithms to predict future performance.

TABLE. I. SUMMARY OF COMMON MACHINE LEARNING CLASSIFIERS WHICH ARE USED IN PREDICTING STUDENT PERFORMANCE

Ref.	Main Results
[6]	(i) C4.5 and Randomtree were proposed. (ii) C4.5 could produce the highest accuracy.
[7]	(i) The six classifiers are decision tree (DT), random forest (RF), artificial neural network (ANN), Navie Bayes (NB), logistic regression (LR), and generalized linear model (GLM). (ii) The RF was found to be the best classifier.
[8]	(i) C4.5, NB, 3-nearest neighbor (3-NN), backpropagation (BP), sequential minimal optimization (SMO), LR were proposed, (ii) NB algorithms produced the highest classification result.
[9]	(i) Three tree-based classifiers: J48, Random Tree, and REPTree were used. (ii) J48 was found to be the best prediction model.
[10]	(i) NB, support vector machine (SVM), C4.5, CART are used to build the learning model. (ii) SVM is the best model compared to NB, C4.5, and CART.
[11]	(i) RF, multilayer perceptron (MLP), and ANN were used to classify student performance. (ii) The RF algorithms generated the highest accuracy.
[12]	(i) J48, CART, and RF classifiers were proposed with principal component analysis (PCA). (ii) PCA-RF was found to generate the highest accuracy.
[13]	(i) MLP, Radial Bias Function (RBF), SMO, J48, and NB are proposed to combine with PCA. (ii) PCA-NB generated the highest accuracy.
[14]	(i) Three Boosting algorithms (C5.0, AddaBoost M1., and AdaBoost SAMME) are proposed. (ii) The C5.0 outperformed the other two boosting models.

III. MACHINE LEARNING ALGORITHMS

We proposed hybrid models by a conjunction of machine learning algorithms with principal component analysis. We first proposed the baseline models. We then improved the performance of our proposed baseline models with k-fold cross-validation. Lastly, we proposed the hybrid machine learning model by combining it with principal component analysis as in Fig. 1.

A. The Baseline Models

There are numerous effective machine learning approaches that have been extensively applied to educational environments. For various purposes in educational settings, we need to take different machine learning techniques such as association rule mining, regression analysis, classification, and clustering [3]. Classification is a common technique in machine learning that was used in order to classify and predict the categories or predefined classes of target variables. In this work, we observed several machine learning classifiers and selected the four state-of-the-art methods which are popularly used in predicting academic performances [3-14]. The four proposed algorithms are support vector machine, naïve Bayes C5.0 of the decision tree, and random forest.

1) Support vector machine: A Support Vector Machine (SVM) is a kind of classification algorithm obtained by the mean of a separating hyperplane [15]. The concept of SVM is to create a line or a hyperplane to separates the samples into classes. SVM is used to observe for the optimal hypersurface to separate each two different data classes. Once the data is more complex, then we create more dimensional space to have a linear separation of data.

Given a training sample $(x_i, y_i), i = 1,2,...,m$, where $x_i \in \square^n$ and $y_i \in \{-1,1\}$ are called the target classes, the classical SVM classifier is subject to solve the optimization problem:

$$\min_{w,b,\xi}\left\{\frac{1}{2}w^T w + C\left(\sum_{i=1}^{m}\xi_i\right)\right\}$$
$$\text{subject to: } y_i(w^T\phi(x_i)+b) \geq 1-\xi_i,\ \xi_i > 0, \forall i, \quad (1)$$

where $\phi(x)$ is treated for nonlinear function case mapping x into a higher dimensional space. The parameters w, b and ξ_i represent the weight, bias, and slack variable, respectively. And the optimal hyperplane is possibly to be solved using Lagrangian and then transform it into a quadratic problem of the function $W(\alpha)$ as in (2):

$$\max W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i, x_j)$$
$$\text{subject to: } \sum_{i=1}^{m}\alpha_i y_i = 0; \alpha_i \in [0, C], i = 1,2,..,m, \quad (2)$$

where $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel function and, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)$ is a set of Lagrange multipliers.

The decision function can be written as:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{m}\alpha_i y_i K(x_i, x_j) + b\right). \quad (3)$$

Different kernel functions are used to help SVM to maximize margin hyperplanes to obtain the optimal solution. The most popular used kernels are the polynomial function, sigmoid function, and radial basis function. SVM with radial bias function (RBF) kernel is one of the most commonly used kernels for the multi-classification problem since it requires fewer parameters comparing to the polynomial kernel. Consequently, RFB is an appropriate choice to be used kernel. Hence, this work applied RBF as a kernel function top to get the optimal solution.

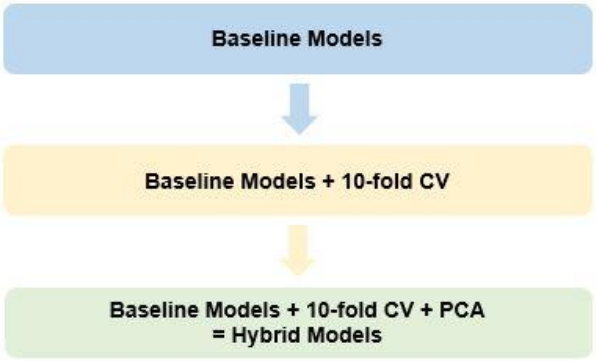

Fig. 1. Illustration of Task Procedure.

2) Naïve Bayes (NB): NB is one among the simple but effective machine learning algorithms that is preferably used in many classification problems. NB is a very attractive method for education research [16]. In the educational domain, an assumption of conditional independence is often ignored and disturbed. Considering that variables are inter-connected, the NB classifier can tolerate strong supervising dependence between independent variables. NB classifier is Bayes theorem-based method that used the idea of computing posterior probability for decision rule. NB classifier has been especially popular for educational data mining. Suppose D is a dataset of n dimensional vector $X : (x_1, x_2, x_3, ..., x_n)$ describing attributes of each student and suppose there are k classes: $C_1, C_2, ..., C_k$. NB classif-ier predicts X belong to a class C_i if and only if $P(C_i \mid X) > P(C_j \mid X)$ for all $1 \le j \le k, i \ne j$. The NB classifier is found on conditional Bayes probability as in (4):

$$P(C_i \mid X) = \frac{P(C_i) \times P(X \mid C_i)}{P(X)}. \tag{4}$$

The probability $P(X)$ is normalizing constant and $X = (x_1, x_2, ..., x_p)$ is the set of features variables with a strong assumption of independent predictors, then (4) can be rewritten as:

$$P(C_i \mid X) \propto P(C_i)\,P(X \mid C_i) = P(C_i)\prod_{j=1}^{n} P(x_j \mid C_i). \tag{5}$$

The naïve Bayes classifier holds many advantages such as it is a very simple algorithm, not contain any parameter to optimize, efficient for classification, and easy to interpret.

3) C5.0: Decision tree is a "non-parametric white-box model" which is simple and effective for classification and regression tasks while C5.0 is one of the most famous algorithms of decision tree that construct the structure in the form of tree diagram [14]. This algorithm takes care of various of the decisions automatically using fairly reasonable defaults.

C5.0 is a successor of C4.5; it builds tree structure from training set using the idea of Shannon entropy. The algorithm purifies the subset of samples via the concept of information entropy. Entropy defines the impurity of any subset of an sample set S at a specific node N is written as:

$$Entropy(S) = I(S) = -\sum_{i=1}^{c} P(c_i) \log_2 P(c_i). \tag{6}$$

The constant c is denoting the number of classes and $P(c_i)$ is the proportion of values in the class i . After obtaining the measure of purity, the algorithm needs to decide which feature to split next. The algorithm calculates homogeneity resulting from a split on each possible feature, this procedure of calculation is called information gain (IG) as shown in (7):

$$IG(S, A) = I(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} I(S_v). \tag{7}$$

One complicated matter after splitting is that a split result in more than one partition that is what we need to compute what is called split information in the following equation:

$$SplitInfo_A(S) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} \log_2 \left(\frac{|S_i|}{|S|} \right). \tag{8}$$

Then, using information gain as see formula (7), and splitting information as in (8), we then can compute the information gain ration using the following equation:

$$GainRatio(S, A) = \frac{IG(S, A)}{SplitInfo_A(S)}. \tag{9}$$

The C5.0 of a decision tree is one of the most popular machine learning algorithms that has been widely used in various applications.

4) Random Forest (RF): As in the name indicates its meaning, the random forest is an algorithm builds the forest with a number of trees. A random forest algorithm is a tree-based tool that grows many classification trees [12]. It is a kind of ensemble classifier that combines several classification trees to create a new classifier. The concepts of bootstrap aggregation or bagging method is used to grow each tree. To classify a new example, each decision tree gives a classification for the input data which is so-called “voting for a class”. The RF algorithm chooses a class with the highest votes. The illustration of the process of random forest algorithms is shown in Fig. 2.

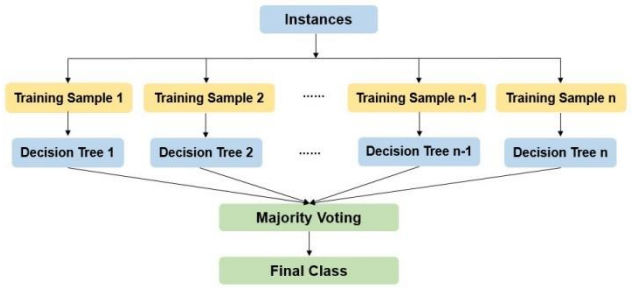

Fig. 2. Illustration of Random Forest Algorithm.

B. The k-fold Cross-Validation

Cross-validation is one of statistical technique that used to test the effectiveness of machine learning algorithms. There are various methods of cross-validation but the k-fold cross-validation is chosen since it is popular and easy to understand, also generally generates a lower bias comparing to the other cross-validation methods. The process of k-fold cross-validations is summarized as the following:

1) Shuffle the entire samples randomly
2) Split samples into k sub folds
3) In the split k sub folds:

- Take 1 fold as a holdout or test set

- Take the remaining $k-1$ folds as the training set
- Retain the evaluation score and discard the model

4) Repeat the iteration until every single fold was treated as a testing set. Finally, compute the average score of the recorded scores.

In our study, we chose the 10-fold cross-validation (will be shortly called 10-CV) to access our proposed algorithms. This process is precisely illustrated in Fig. 3.

C. The Proposed Hybrid Models

The majority task in supervised machine learning is classification. The classification problem is a hot issue in data mining and machine learning. We proposed the four most popular classifiers that hold many merits. However, the major problem for those classifiers is overfitting and noisy data which leads to misclassification and deduce the accuracy of the classification. To overcome this matter, we try to reduce irrelevant feature and non-correlated features which disturb in the classification process. In data analysis, it requires more computational resources and consumes much time when that data consists of a huge volume. Hence, the feature extraction approach to remove noises in data in order to reduce time and resource usage and regain the high quality of data. The dimensional reduction could improve accuracy and boost up the performance by combining it with classification techniques. Using more high-quality data and feature reduction is one of the effective approaches to improve the performance of machine learning models. The four proposed models: support vector machine using radial basis function kernel (SVMRBF), naïve Bayes (NB), decision tree C5.0, and random forest (RF) are the affective algorithms for the classification problem, yet there is no perfect algorithm in machine learning.

SVM is a classifier with the use of support vectors called hyperplanes to separate data into classes. Thus, for a high dimensional dataset, the input space is high and can be unclean which is mostly declining the performance of the SVM algorithm. Thus, it requires an effective feature extraction method that discards noisy, irrelevant and redundant data, and still contains the useful information of data. Removal of such features can increase the search speed and accuracy rate.

NB is a classifier that holds many advantages, yet the greatest weakness of the NB classifier is that it relies on the often-faulty assumption of equally important and independent features. If there are any features that are irrelevant to some class C_k then the whole probability goes to zeros for that class because of production in equation (5), which leads to misclassification. In order to solve this problem, feature extraction will be the best tool to reduce irrelevant features and also improve the classification performance.

In the tree-based algorithms C5.0 and RF, the major problem in the splitting process of the decision tree is overfitting. Overfitting caused by noisy data and irrelevant features that produce misclassification results. In return, overfitting lowering the accuracy of tree-based classifiers. To reduce high dimensional data which, contains noisy and irrelevant data, a commonly-used technique is to use feature extraction in order to obtain a lower-input space that contains relevant and informative input features.

In order to improve the performance of the proposed machine learning algorithms, we proposed commonly-used feature extraction approach: principal component analysis (PCA) in this study. PCA is a statistical method that transforms an original data set to a new dataset of a lower dimension. The original dataset consists of possibly correlated variables are converted into a set of linearly uncorrelated variables.

PCA is one of the most popular dimensionality reduction algorithm [17]. In the PCA procedure, the data is first transformed into standardized data with zero mean. The idea behind getting the principle components is the covariance matrix is computed in order to obtain eigenvector and eigenvalues. The eigenvector with the highest eigenvalue is treated as the principal component of new data which shows the most significant relationship of input feature. PCA is less sensitive to different datasets than other holistic methods, so it is the most widely used technique as one of the effective feature reduction methods.

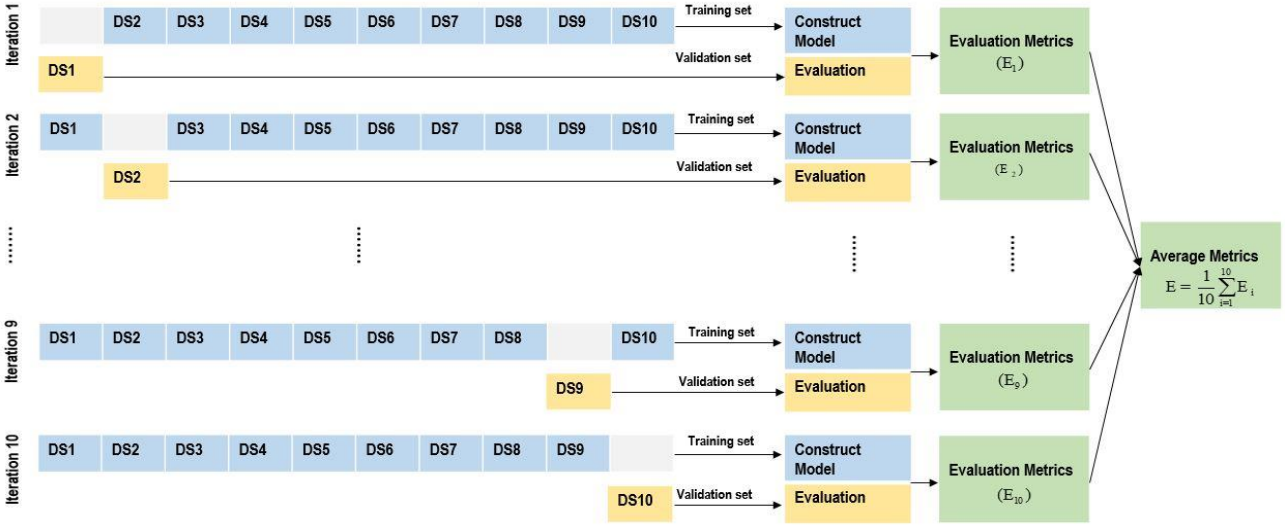

Fig. 3. Illustration of the K-Fold Cross-Validation Algorithm.

The procedure of transforming original dataset X of l dimension consisting of possibly correlated features to a new dataset Z of lower dimension $m(\mathrm{m} < l)$ consisting of linear uncorrelated features is as follows:

1) Compute mean: From the already processed data, first, find the mean of each attribute using the equation:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

2) Compute variance: In order to investigate and deviation of each feature in the dataset, we compute the variance using equation (11):

$$Var(X) = \sigma_x^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{11}$$

3) Compute covariance: Given two variables, denoted X and Y, the covariance and correlation are calculated using equation (12):

$$Cov(X,Y) = \sigma_x^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) \tag{12}$$

$Cov(X,Y)$ equals to zero means that the two attributes X and Y are independent. Using equation (11) and (12), we can obtain covariance matrix S, which the entry s_{ij}, $i \neq j$, is the covariance between the i^{th} and j^{th} variables, and diagonal s_{ii} is the variance of i^{th} variables.

4) Compute Eigenvalues and Eigenvectors: The features in the new datasets are characterized by mean of eigenvectors and eigenvalues. The obtained eigenvectors will tell the direction of new features space while the eigenvalues are its magnitude. The eigenvalues are possible to obtain by solving the equation:

$$Det(S - \lambda I) = 0, \tag{13}$$

where the covariance matrix S is symmetric, λ is the eigenvalue of the symmetric matrix S, and I is an identity matrix. The eigenvector v corresponding to each eigenvalue λ can be computed via the equation:

$$(S - \lambda I)v = 0 \tag{14}$$

We denoted $E = \{v : (S - \lambda I)v = 0\}$ as the Eigen space containing all eigenvectors.

5) Obtain orthonormal eigenvectors: By means of linear algebra concept, we can obtain the nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \lambda_m > 0$ with corresponding orthonormal eigenvectors $v_1, v_2, ..., v_m$. The eigenvectors are called the principal components of the dataset.

The proposed hybrid models by conjunction machine learning models with PCA are introduced for predicting and classifying the academic performance. The best benefits of PCA are summarized as follow:

a) Removing the high noises from samples and uncorrelated features from the collected dataset in the preprocessing step.

b) Reducing the high dimensional data to low dimensional one which remains the important characteristics of data that reduce overfitting problems.

c) Enhance the equality of features by getting rid of correlated features that effectively improve the performance of classification.

In this proposed research, we proposed the hybrid models by a conjunction of four baseline models (SVMRFB, NB, C5.0, and RF) with 10-fold cross-validation (10-CV) and principal component analysis (PCA).

IV. Datasets and Preprocessing

A. Datasets

In our study, we tried to collect all unseen features affecting student performance in mathematics subjects. Datasets contained 43 features describing the information of the learning behaviors of each student and one target variable describing the performance levels of students based on their score. The predictive features consist of the features observing from three main affected factors. These main factors contain the forty-three variables and their descriptions are shown in Table II. Table III described the predefined classes of the target variable.

To confirm the robustness and effectiveness of our proposed algorithms, we used three datasets. The first two datasets are generated datasets namely GDS1 (2000 samples) and GDS2 (4000 samples) that were constructed based on proposed structures of predictive features to the output variable as stated in [18-20]. The third dataset is the actual dataset that was collected from 22 high schools in Cambodia. The data collection was made using questionnaires form. Students were asked to provide their demographical information related to external effects such as domestic factors, individual or student factors, and school factors. The score of mathematics of students in the semester I was obtained from the administrative offices in each school. The dataset was named ADS3 that consists of 1204 samples.

TABLE. II. The Output Variable

N	Performance Levels	Score-based discretization
1	Excellent learner	90% and above
2	Good learner	75% to less than 90%
3	Average learner	60% to less than 75%
4	Slow learner	Less than 60%

TABLE. III. THE FACTORS AFFECTING STUDENT PERFORMANCE IN MATHEMATICS

N	Variables	Description	Type
Domestic Factors			
1	PEDU1	Father's educational level	Nominal
2	PEDU2	Mother's educational level	Nominal
3	POCC1	Father's occupational status	Nominal
4	POCC2	Mother's occupational status	Nominal
5	PSES	Family's socioeconomic	Ordinal
6	PI1	Parents' attention to students' attitude	Ordinal
7	PI2	Parents' time and money spending	Ordinal
8	PI3	Parents' involvement as education	Ordinal
9	PS1	Parents' feeling responsive and need	Ordinal
10	PS2	Parents' respond to children's attitude	Ordinal
11	PS3	Parents' encouragement	Ordinal
12	PS4	Parents' compliment	Ordinal
13	DE1	Domestic environment for study	Ordinal
14	DE2	Distance from home to school	Nominal
Student or Individual Factors			
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
20	SIM1	Student' s interest in math	Ordinal
21	SIM2	Student's enjoyment in math class	Ordinal
22	SIM3	Student's attention in math class	Ordinal
23	SIM4	Student's motivation to succeed in math	Ordinal
24	ANXI1	Student's anxiety in math class	Ordinal
25	ANXI2	Student's nervous in the math exam	Ordinal
26	ANXI3	Student's feeling helpless in math	Ordinal
27	POSS1	Internet's use at home	Binary
28	POSS2	Possession of computer	Binary
29	POSS3	Student's study desk at home	Binary
School Factors			
30	CENV1	Classroom environment	Ordinal
31	CU1	Content's language in math class	Nominal
32	CU2	Class session	Nominal
33	TMP1	Teacher mastering 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 involving in education's content	Ordinal
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
41	ARES1	Adequate number of math teacher	Nominal
42	ARES2	Adequate use of classroom	Nominal
43	ARES3	Adequate use of math handout	Nominal

B. Preprocessing Tasks

Data preprocessing is an integral step in data mining that is used to transform the raw dataset into a clean and executable format to be ready for implementation. The preprocessing step is not only used to ensure the readiness of data suitable and ready for modeling but also to improve the performance of the models. The preprocessing tasks in this study contain some operations such as data cleaning or cleansing, data transformation, and data discretization. During data collection, the questionnaire completion was done with missing some questions and inputting invalid value (outliers). In our datasets, the number of missing values is low, so we used the imputing method in order to clean our data. We replaced the missing value in our categorical variables by its modes or high frequency-category values. In the output variable, there is a few missing value and outliers, then we replaced it by the mean value. For simplicity, we transformed some numerical features into ordinal types. In our study, we also discretized the output variables into four performance levels as shown in Table I.

V. EVALUATION METRICS

The performance of each proposed model in analyzing and predicting student performance can be evaluated from the analysis of the graphical confusion matrix. Without loss of generality, our output variable can be categorized into four ordinal categories as mention in Table I. Table IV shows the graphical confusion matrix which represents four classes of student performance level in mathematics subject. Class 1 presents the highest class, Class 2 denotes the second upper class, Class 3 describes the third class lower, and Class 4 denotes the lowest (poor) group of students. The below parameters are calculated.

A. Classification Accuracy

Accuracy is used to quantify the percentage of correctly predicted. Here, we want to evaluate the potential of our prediction model by measuring the percentage of correctly predicted the level of student performance as in (15):

$$Accuracy = \frac{\sum a_{ii}}{\sum a_{ij}} \times 100\% \tag{15}$$

B. Root Mean Square Error (RMSE)

We aim not only to predict the ability of students' performance levels but also to estimate how much our prediction is close to their performance level. We encoded these ordinal performance levels {slow, average, good, excellent} as {1,2,3,4}, respectively. The RMSE can be computed as:

$$RMSE = \sqrt{\sum_{i=1}^{M} \frac{(Pl_i^a - Pl_i^p)^2}{M}} \tag{16}$$

where $Pl^a \in \{1,2,3,4\}$ is the actual performance level and $Pl^p \in \{1,2,3,4\}$ is the predicted performance level. Contrasting with accuracy, the smaller the RMSE, the better the model is. RMSE equal to 0 shows the prediction model is perfect.

TABLE. IV. GRAPHICAL CONFUSION METRIC

		Predicted Classes			
		Slow	Average	Good	Excellent
Actual Classes	Slow	a_{11}	a_{12}	a_{13}	a_{14}
	Average	a_{21}	a_{22}	a_{23}	a_{24}
	Good	a_{31}	a_{32}	a_{33}	a_{34}
	Excellent	a_{41}	a_{42}	a_{43}	a_{44}

VI. EXPERIMENTAL RESULTS

In our experiments, we proceed in three phases. Phase 1 is to implement for the result of the baseline models. Phase 2 is to improve the baseline models by 10-fold cross-validation (10-CV). Phase 3 is to execute a hybrid model which is the combination of the baseline models with 10-CV and PCA.

A. Result of Baseline Models

We proposed four most popular machine learning techniques, random forest (RF), C5.0 of the decision tree, support vector machine using radial basis function kernel (SVMRBF), and naïve Bayes (NB) of the Bayesian network. The two performance metrics, classification accuracy, and RMSE are shown in the tables.

From Table V, VI, and VII, NB was found to be the poorest model, while the RF technique generates the highest performance with respect to both classification accuracy and RMSE, which shown itself as the potential model.

B. Results of Baseline Models with k-fold Cross-Validation

The k-fold cross-validation is a technique that is popularly used in prediction and classification models to split the dataset into $k-1$ sub folds for training and 1 fold for testing sets, then rotate the folds. In this experiment, we used 10-fold cross-validation, since it performs best at this split. 90% of the data was used in the training section, and 10% was used for testing purposes as shown in Fig. 3. Lastly, when all interactions were done, an average of all evaluation metrics is computed.

From Table VIII, the accuracy of SVMRBF was improved by 2%. The performance of the poor NB classifier was then much improved by to 68.03%. The 10-CV technique improved C5.0 and RF with an accuracy increase of 27% and 15%, respectively.

From Table IX, by shuffling the dataset GDS2 with 10-CV, the accuracy of SVMRBF algorithm was improved from 75.52% to 91.15%, which is a very good improvement. NB increased by an accuracy of 9%. The tree-based classifiers C5.0 and RF were improved by the accuracy of 9% and 6%, respectively.

TABLE. V. PERFORMANCE OF BASELINE MODELS TO GDS1

Bassline Models	Accuracy	RMSE
SVMRBF	75.01%	0.516
NB	35.79%	1.191
C5.0	78.42%	0.487
RF	80.06%	0.431

TABLE. VI. PERFORMANCE OF BASELINE MODELS TO GDS2

Bassline Models	Accuracy	RMSE
SVMRBF	75.52%	0.489
NB	67.68%	0.664
C5.0	86.18%	0.372
RF	90.37%	0.321

TABLE. VII. PERFORMANCE OF BASELINE MODELS TO ADS3

Bassline Models	Accuracy	RMSE
SVMRBF	86.44%	0.823
NB	65.02%	1.016
C5.0	76.55%	0.845
RF	89.23%	0.516

TABLE. VIII. PERFORMANCE OF BASELINE MODELS AND BASELINE MODELS+10-CV TO GDS1

Models	Accuracy	RMSE
SVMRBF	75.01%	0.516
SVMRBF + 10-CV	77.08%	0.456
NB	35.79%	1.191
NB+ 10-CV	68.03%	0.654
C5.0	78.42%	0.487
C5.0+ 10-CV	95.24%	0.185
RF	80.06%	0.431
RF+ 10-CV	96.48%	0.143

TABLE. IX. PERFORMANCE OF BASELINE MODELS AND BASELINE MODELS+10-CV TO GDS2

Models	Accuracy	RMSE
SVMRBF	75.52%	0.489
SVMRBF + 10-CV	94.15%	0.274
NB	67.68%	0.664
NB+ 10-CV	76.47%	0.498
C5.0	86.18%	0.372
C5.0+ 10-CV	95.69%	0.174
RF	90.37%	0.321
RF+ 10-CV	96.58%	0.139

TABLE. X. PERFORMANCE OF BASELINE MODELS AND BASELINE MODELS+10-CV TO ADS3

Models	Accuracy	RMSE
SVMRBF	86.44%	0.823
SVMRBF + 10-CV	90.66%	0.678
NB	65.02%	1.016
NB+ 10-CV	92.44%	0.145
C5.0	76.55%	0.845
C5.0+ 10-CV	94.82%	0.114
RF	89.23%	0.561
RF+ 10-CV	98.22%	0.113

From Table X, the NB accuracies improved rapidly from 65.44% to 90.66%. SVMRBF could yields around 4% better than the previous baseline SVMRBF. C5.0 and RF are tree-based classifiers that could produce a high risk of over-fitting. With a 10-CV, we can not only obtain better performance but also avoid overfitting problems too. By mean of 10-CV, accuracies of C5.0 and RF were improved to 94.82% and 98.22% which improved 18% and 9%, respectively.

C. Results of Proposed Hybrid Models

Our proposed hybrid models were constructed by combing the baseline models with a feature reduction approach, PCA. Feature extraction is one of the powerful methods in classification models that are used for the purpose of removing irrelevant or non-related features. Dimensionality reduction via PCA [13] can definitely serve as regularization in order to prevent overfitting and improve the model accuracies. Often, people end up making a mistake in thinking that PCA selects some features out of the dataset and discards others. The algorithm actually constructs a new dataset of properties based on a combination of the old ones.

In this section, we proposed the hybrid models as the combination of 10-CV in the previous section to PCA in order to avoid overfitting and more improvement in predicting performance. Tables XI, XII, and XIII describe the results of the proposed models to the three datasets, GDS1, GDS2, and ADS3, respectively.

We visualized the performance of the proposed models to the three datasets GDS1, GDS2 and ADS3 in Fig. 4, 5, and 6, respectively. In Fig. 4, the accuracy based in dataset GDS1, our proposed hybrid models boost the accuracy of SVMRBF from 75.01% to 83.88%, NB from 35.79% to 86.27%, C5.0 from 78.42% to 98.32%, and RF from 80.06% to 98.92%.

In Fig. 5, the hybrid models improved SVMRBF, NB, C5.0, and RF with accuracies of 20%, 23%, 12%, and 9%, respectively. In Fig. 6, the proposed hybrid SVMRBF could improve the classification accuracy from 86.44% to 97.01%. Classification through NB could yields 30% better than baseline NB. The accuracies of C5.0 and RF were improved to 99.25% and 99.72% correctly classified.

TABLE. XI. PERFORMANCE OF BASELINE MODELS, BASELINE MODELS +10-CV, AND HYBRID MODELS TO GDS1

Models	Average Accuracy	Lowest Accuracy	Highest Accuracy	Std.	Average RMSE	Lowest RMSE	Highest RMSE	Std.
SVMRBF	75.01%	70.27%	77.01%	1.421	0.516	0.460	0.691	0.059
SVMRBF+10-CV	7.08%	75.67%	78.89%	1.124	0.496	0.456	0.524	0.024
Hybrid SVMRBF	**83.88%**	**82.01%**	**85.05%**	**1.123**	**0.414**	**0.396**	**0.437**	**0.016**
NB	35.79%	32.41%	37.27%	1.861	1.191	1.045	1.411	0.127
NB+10-CV	68.03%	66.61%	69.82%	1.363	0.645	0.577	0.768	0.070
Hybrid NB	**86.27%**	**83.40%**	**90.35%**	**2.695**	**0.521**	**0.456**	**0.608**	**0.060**
C5.0	78.42%	75.41%	82.72%	2.429	0.487	0.449	0.543	0.038
C5.0+10-CV	95.24%	93.18%	96.28%	0.806	0.185	0.158	0.242	0.026
Hybrid C5.0	**98.32%**	**97.18%**	**99.28%**	**0.564**	**0.067**	**0.043**	**0.145**	**0.027**
RF	80.06%	77.25%	83.21%	1.860	0.431	0.371	0.495	0.037
RF+10-CV	96.48%	95.21%	97.52%	0.764	0.143	0.122	0.189	0.015
Hybrid RF	**98.92%**	**97.06%**	**99.78%**	**0.817**	**0.056**	**0.031**	**0.126**	**0.026**

TABLE. XII. PERFORMANCE OF BASELINE MODELS, BASELINE MODELS+10-CV, AND HYBRID MODELS TO GDS2

Models	Average Accuracy	Lowest Accuracy	Highest Accuracy	Std.	Average RMSE	Lowest RMSE	Highest RMSE	Std.
SVMRBF	75.52%	70.00%	77.80%	1.606	0.489	0.449	0.524	0.023
SVMRBF+10-CV	94.15%	92.89%	95.51%	0.909	0.274	0.234	0.347	0.050
Hybrid SVMRBF	**96.32%**	**95.52%**	**96.89%**	**0.591**	**0.182**	**0.173**	**0.202**	**0.011**
NB	67.68%	65.01%	69.82%	1.614	0.664	0.584	0.768	0.059
NB+10-CV	76.47%	73.52%	78.21%	1.624	0.498	0.466	0.550	0.031
Hybrid NB	**91.42%**	**88.71%**	**95.05%**	**1.593**	**0.321**	**0.288**	**0.383**	**0.033**
C5.0	86.18%	84.45%	88.41%	1.454	0.372	0.319	0.533	0.059
C5.0+10-CV	95.69%	93.28%	97.28%	1.026	0.174	0.141	0.197	0.017
Hybrid C5.0	**98.62%**	**98.18%**	**99.54%**	**0.475**	**0.067**	**0.043**	**0.145**	**0.028**
RF	90.37%	89.01%	91.80%	1.021	0.321	0.286	0.345	0.018
RF+10-CV	96.58%	95.21%	98.50%	0.928	0.139	0.114	0.189	0.016
Hybrid RF	**99.08%**	**97.60%**	**99.80%**	**0.732**	**0.057**	**0.031**	**0.126**	**0.027**

TABLE. XIII. PERFORMANCE OF BASELINE MODELS, BASELINE MODELS+10-CV, AND HYBRID MODELS TO ADS3

Models	Average Accuracy	Lowest Accuracy	Highest Accuracy	Std.	Average RMSE	Lowest RMSE	Highest RMSE	Std.
SVMRBF	86.44%	81.06%	90.69%	2.56	0.823	0.691	1.016	0.089
SVMRBF+10-CV	90.66%	86.66%	95.00%	2.86	0.678	0.364	0.813	0.131
Hybrid SVMRBF	**97.01%**	**95.34%**	**98.67%**	**1.112**	**0.178**	**0.114**	**0.230**	**0.042**
NB	65.02%	60.24%	69.21%	2.961	1.016	0.834	1.331	0.164
NB+10-CV	94.82%	91.18%	96.27%	1.165	0.154	0.133	0.232	0.035
Hybrid NB	**98.94%**	**98.01%**	**99.69%**	**0.607**	**0.145**	**0.042**	**0.230**	**0.049**
C5.0	76.55%	70.43%	80.73%	2.851	0.845	0.703	0.991	0.097
C5.0+10-CV	97.54%	94.79%	99.50%	1.923	0.114	0.070	0.160	0.034
Hybrid C5.0	**99.25%**	**98.21%**	**100%**	**0.606**	**0.073**	**0.000**	**0.145**	**0.045**
RF	89.23%	86.71%	92.35%	1.566	0.561	0.411	0.667	0.066
RF+10-CV	98.22%	95.69%	99.52%	1.353	0.113	0.070	0.160	0.034
Hybrid RF	**99.72%**	**99.01%**	**100%**	**0.357**	**0.041**	**0.000**	**0.077**	**0.029**

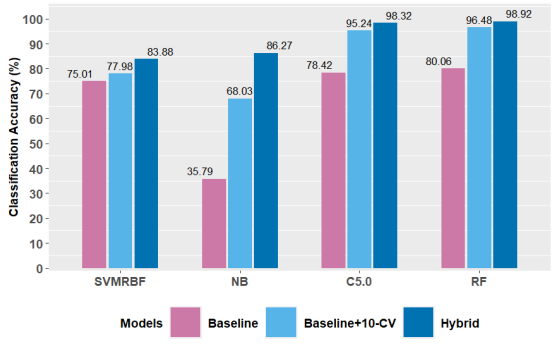

Fig. 4. Performance-based on the accuracy of GDS1

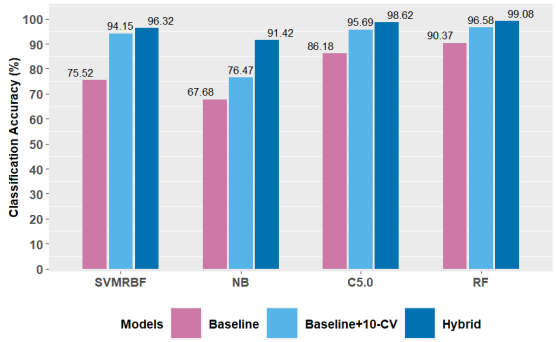

Fig. 5. Performance-based on the accuracy of GDS2

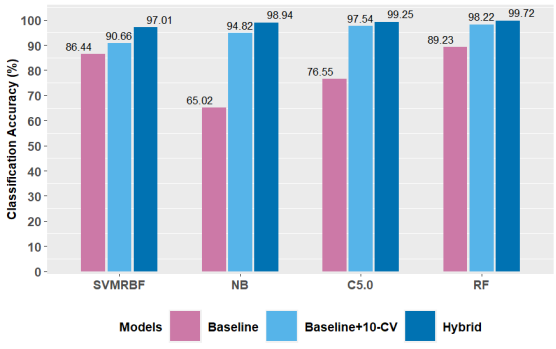

Fig. 6. Performance-based on the accuracy of ADS3

Fig. 7, 8, and 9 demonstrated the performance based on the accuracy of each model via each phase. We found the improvement by using 10-CV combined with PCA gives the best result in predicting student performance. The figures show the performance of the RMSE of the models in each step. The proposed hybrid models could generate a very small RMSE. The hybrid RF algorithm produced the smallest value of RMSE which shows itself as the best predictive model in this prediction problem.

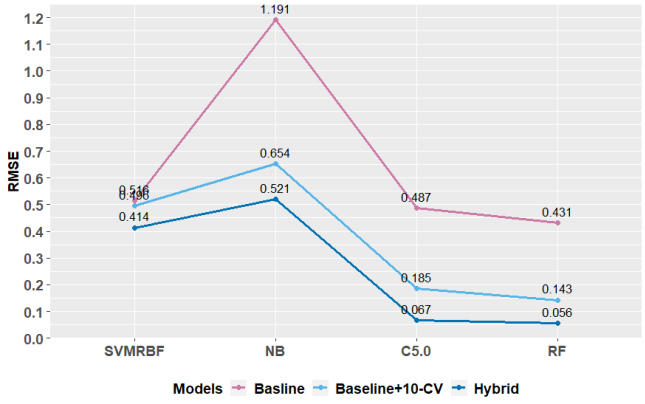

Fig. 7. Performance-based on the RMSE of GDS1.

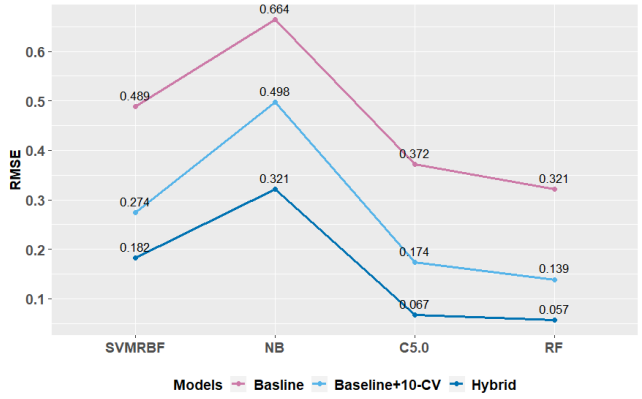

Fig. 8. Performance-based on the RMSE of GDS2.

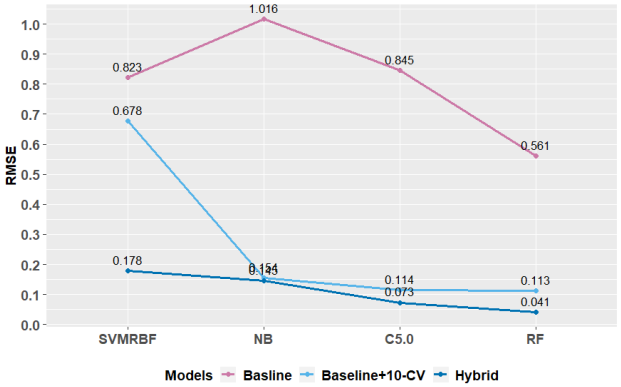

Fig. 9. Performance-based on the RMSE of ADS3.

From the results, by using 10-CV, we can improve the performance of our baseline models. Additionally, we observed that the proposed novel hybrid models could boost up the classification performance to the superior results. This proposed hybrid models can be regarded as an optimal prediction models for solving prediction and classification problems.

VII. Conclusion

This paper introduced the four popular classifiers of machine learnings to predict student performance. The four proposed algorithms are SVMRBF, NB, C5.0, RF. The procedure was made with three phases. Firstly, we observed the performance of those baseline methods. Secondly, we improved the performance with 10-CV. Lastly, we combined the PCA to baseline models, and 10-CV method to improve the classification performance. Based on classification accuracy and RMSE as measurement parameters, it shows that the proposed hybrid models by conjunction of the proposed models with PCA and 10-CV produced very satisfying results. In conclusion, by combining the baseline models with principal component analysis, and evaluated by k-fold cross-validation, the proposed hybrid models produced a high performance which shows itself as a potential algorithm for solving prediction and classification problem.

References

[1] Herbert K., "The New Book of Popular Science", World Applied Sciences Journal, Daribury, Connecticut: Grolier Inc., 1978.

[2] Ministry of Education, Youth and Sport, “Education in Cambodia: Finding from Cambodia's Experience in PISA for Development”, Phnom Penh: Author, 2018.

[3] S. Slater, S. Joksimovic, V. Kovanovic, R.s Baker, and D. Gasevic, "Tools for Educational Data Mining: A Review", Journal of Educational and Behavioral Statistics, Vol. 42, No. 1, 2016, pp. 88-106.

[4] Pooja Thakar, Anil Mehta, and Manisha, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", International Journal of Computer Application, Vol. 100, No.12, January 2015, pp. 60-68.

[5] C. Romero and Ventura., “Educational Data Mining: A Review of the State of Art”, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40, No. 6, 2010, pp. 601-618.

[6] Akinrotimi A.O, and Aremu D.R, "Student Performance Prediction Using Randomtree and C4.5," Journal of Digital Innovation and Contemporary Research, Engineering and Technology, Vol. 6, No. 3, 2018, pp. 23-34.

[7] Amjad A. S., Mostafa Al-Emran, and Khaled S., “Mining Student Information System Records to Predict Students’ Academic Performance”, Springer Nature Switzerland AG 2020, AMLTA 2019, AISC 921, 2019, pp. 229-239.

[8] Kotsiantis S., Piarrekeas C., and Pintelas P., "Predicting Students' performance in Distance Learning using Machine Learning Techniques", Applied Artificial Intelligence, Vol. 18, 2007, pp. 411-426.

[9] Hamoud A. K., Hashim A. S., and Awadh W. A., "Predicting Student Performance in Higher Education Institutes Using Decision Tree Analysis", International Journal of Interactive Multimedia and Artificial Intelligent, Vol. 5, No. 2, February 2018, pp. 26-31.

[10] Daud A., Aljonhani N.R., Abbasi R.A., Lytras M.D., Abbas F., Alowibdi J.S., “Prediction student performance using advanced learning analytics”, Proceedings of 26th International Conference on World Wide Web, Companion, Perth, Australia, April 2016, pp. 416-421.

[11] M.S. Mythili,, A.R.M. Shanavas, “An Analysis of Students' Performance using Classification Algorithms”, IOSR Journal of computer Engineering (IOSR-JCE), Vol 16, No. 1, January 2014, pp. 63-69.

[12] Aung Nway Oo, “Comparative Study of Principle Component Analysis based on Decision Tree Algorithm”, International Journal of Advances in Scientiofic Research and Engineering, Vol. 4, No. 6, June 2018, pp. 122-126.

[13] Karthikeyan T., Thangaraju P., “PCA-NB Algorithms to Enhance the Predictive Accuracy”, International Journal of Engineering and Technology, Vol. 6, No. 1, 2014, pp. 381-387.

[14] Farid J., Ahmad A.S., “Building student’s performance cesissiion tree classifier using boosting algorithm”, Indonesian Journal of Electrical Engineering and Computer Science, Vol. 14, No. 3,2019, pp. 1298-1304.

[15] Babak M.A., Seyed K.S., Maryam M.M., “Support vector machine-based arrhythmia classification using reduced features of heart rate variablitiy singanal”, Ariticial Intelligence in Mechine (Elsevier), vol. 44, 2008, pp. 51-64.

[16] Humera S., Raniah Z., Kavitha G., “Prediction of Student Performance in Semester Exam Using a Naïve Bayes Classifier”, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, No. 10, October 201 5, pp. 9823-9829.

[17] Jollifie I.T, "Principle components analysis and factor analysis", Springer 1986.

[18] Phauk Sokkhey and Takeo Okazaki., "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, October 2019, pp. 394-404.

[19] 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, N0. 9, 2012, pp. 1224-1230.

[20] Uysal S., "Factors affecting the Mathematics achievement of Turkish students in PISA 2012", Academic Journals, Vol. 10, June 2015, pp. 1670-1678.

Authors’ Profile

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 Institute of Technology of Cambodia (ITC). Sokkhey is currently a PhD candidate at the University of the Ryukyus, Japan. His currently research are statistical causal relationship analysis, machine learning, educational data mining, and data science.

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 University of the Ryukyus in 2014. He is currently a professor at the university of the Ryukyus. His research interests are statistical analysis, data analysis, genome informatics, tourism informatics, geographic information systems, and data science. He is a member of JSCS, IEICE, JSS, GISA, and BSJ Japan.