Developing Web-based Support Systems for Predicting Poor-performing Students using Educational Data Mining Techniques

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

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

***Abstract*—The primary goal of educational systems is to enrich the quality of education by maximizing the best results and minimizing the failure rate of poor-performing students. Early predicting student performance has become a challenging task for the improvement and development of academic performance. Educational data mining is an effective discipline of data mining concerned with information integrated into the education domain. The study is of this work is to propose techniques in educational data mining and integrate it into a web-based system for predicting poor-performing students. A comparative study of prediction models was conducted. Subsequently, high performing models were developed to get higher performance. The hybrid random forest named Hybrid RF produces the most successful classification. For the context of intervention and improving the learning outcomes, a novel feature selection method named MICHI, which is the combination of mutual information and chi-square algorithms based on the ranked feature scores is introduced to select a dominant set and improve performance of prediction models. By using the proposed techniques of educational data mining, and academic performance prediction system is subsequently developed for educational stockholders to get an early prediction of student learning outcomes for timely intervention. Experimental results and evaluation surveys report the effectiveness and usefulness of the developed academic prediction system. The system is used to help educational stakeholders for intervening and improving student performance.**

***Keywords*—Academic performance prediction systems; educational data mining; dominant factors; feature selection methods; prediction models; student performance**

I. Introduction

Education is considered as a key factor for the development and long-term economic growth of every country. The poor performance causes the problem of under education and shortage of skilled manpower in developing countries. That is academic performance is an important and challenging task in educational institutions. In recent years, educational institutions have tried to improve academic performance and enrich the quality of their learning process. One of the main goals in educational systems is to achieve the high performance of education to increase the best results and decrease the failure rate of poor-performing students. Due to their poor performance, it has arrived at the worrying issue in educational institutions that those students are highly possible to fail, drop out, or repeat classes [1]. To solve this problem, the prediction has recently become one of the first and foremost effective methods since at-risk students can only be accurately identified early enough through the performance of prediction [2]. Therefore, the early prediction has been considered to be a powerful method for early identification of students who are at risk and need intervention and assistance.

In the recent decade, innovation and information technology have proven its significance in many areas of applications. Educational data mining (EDM) is a research field concerned with the application of data mining, machine learning, and statistics applied to explore data in educational contexts [3]. EDM combines several interdisciplinary fields of study such as machine learning, statistics, data mining, information retrieval, psycho-pedagogy, cognitive psychology, recommender system methods, and techniques to various educational datasets to resolve educational issues [4]. In the context of educational settings, various managerial settings, planning, and scheduling required effective techniques of EDM to uncover the knowledge and information of student learning patterns to give intervention and set up a policy to improve academic performance [5][6]. Various analysis techniques have been introduced for monitoring and anticipating academic performance to keep track of teaching, learning actions, and productive results.

The EDM process comprises of five main steps, as illustrated in Fig. 1. The first step is to get the targeted data, which can be collected from school databases or surveys using questionnaires. The collected raw data is never cleaned or it may be in the undesired format, so the second step is preprocessing step where data is cleaned and transformed into an executable format. The third step is to introduce particular techniques of EDM to obtain the target of the experiment. The answers to educational questions and decision making are obtained from the interpretation of experimental results. The last step is to modify the education process accordingly or defer this step until the next investigation is conducted for a better or more accurate result. The main goal of EDM is to extract information from educational data to address important educational questions and support decision making. Several studies on academic performance have been carried out using methods from the EDM discipline. 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 existing methods require advanced techniques of EDM [5][8].

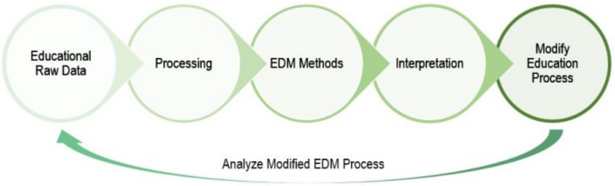


Fig. 1. Outline of EDM Process (Adapted from [7]).

Most researchers have worked on evaluating performances of students in higher education, yet the study on high school student performance evaluation is less. High school student performance is a significant indicator of developing the academic sector since it concerns the background knowledge of students for secondary education and higher education. To improve the poor performance of students in high schools, the right intervention and improvement must be made to low-performing students who are considered in the risk of failure. Poor performing students are highly possible to fail in the national exam and find themselves harder to survive in university life [9]. In the context of academic poor performance, EDM is used for timely prediction for intervention and improvement. In this study, we proposed developed models of EDM and integrate the model into a web-based system for predicting high school student performance.

In short, the modeling in this study is driven as the following research questions:

(i) Question 1: How to obtain dominant factors (highly influencing factors) that are required and sufficient for controlling student's outcomes?

(ii) Question 2: Which prediction model of EDM offers superior predictive results of student learning outcomes?

(iii) Question 3: How educators and related individuals can predict student learning outcomes (the prediction system) for giving intervention and improvement of student performance?

II. Literature Review

A. Feature Selection Methods and Prediction Models in EDM

Estrera et al. [10] predicted student performance for academic ranking in a university in the Philippine using the information records from high schools. Three prediction models of data mining were used in this prediction. The models used in the analysis are decision tree (DT), naïve Bayes (NB), and k-nearest neighbor (KNN). The data used in the prediction was obtained from the information provided by the admitted students to a university using a survey questionnaire conducted on them. To get a better prediction and better understanding of learning behaviors for assessment of students' success, the authors proposed feature selection methods: Chi-square (CHI), information gain (IG), and gain ratio (GR) in this study. As a result, the DT algorithm generates the highest accuracy of 90.67%.

Dimic et al. [11] studied the behavior patterns of students in the blended learning environment. Dataset used in the study was created by integrating data from multiple sources into a form applicable for data mining technique application. Dataset of 225 instances was obtained. The experiment has focused on data preprocessing steps in data mining. Feature selection methods such as Information gain (IG), Symmetrical uncertainty (SU), Relief (REF), correlation-based feature selection (CB), wrapper method, classifier subset evaluator methods were used to extract the most important features. The dependencies of features were computed using information measure (MI). The prediction models: naïve Bayes (NB), aggregating one-dependence estimators (AODE), decision tree (DT) and support vector machines (SVM) were used as prediction models using different feature subsets from each feature selection method. The results indicated that the REF, wrapper method, and MI acted as the most successful features selection methods in selecting the optimal feature sets. The presented research concluded that selecting the subsets of lower cardinality of students' learning activities gives a significant improvement in predictive accuracy in a blended learning environment.

Zaffar and Savita [12] investigated the analysis of feature selection methods in improving the performance of prediction models for predicting student academic performance. The study utilized six feature selection methods: correlation feature selection (CFS), Chi-squared, Filtered, information gain (IG), principal component (PC), and Relief. Fifteen classifiers were used: Bayesian Network (BN), naïve Bayes (NB), Naïve Bayes Updateable (NBU), Multilayer Perceptron (MLP), Simple Logistic (SL), Sequential Minimal Optimization (SMO), Decision Table (DT), OneR, PART, JRip, Decision Stump (DS), J48, Random Forest (RF), Random Tree (RT), and REP Tree (RepT). The experiments indicated that there is a significant improvement of 10 to 20% accuracy when using different feature selection sets.

Saa et al. [13] used information systems record as features that contain their high school records, and the university records to predict the student performance in higher education. The dataset used in the study was collected from a private university. The dataset consists of 34 features and 56,000 samples contained students' personal information. He introduced decision tree (DT), artificial neural network (ANN), random forest (RF), naïve Bayes (NB), logistic regression (LR), and generalized linear model (GLM) as prediction models. The study was to use student information record systems to predict the performance levels of students and identify the weakness and factors that affect student learning outcomes. Hence, information gain (IG) was used for selecting highly influencing factors. The experimental results suggested that the RF algorithm was the most appropriate prediction model in the prediction problem and the important factors affecting student performance were identified.

B. Early Prediction System using EDM

Early prediction or warning systems for predicting student performance is regarded as the improvement or next step in academic performance. It is referred to as prediction methods capable of discovering important and useful information about student learning patterns and risks of students such as retention, drop-out, and students' outcomes in an early stage. The purpose of using an educational early warning system is to give an earlier prediction of academic performance using features that influence students' success. Performance exhibited by students in their learning could be predicted in advance and possible failure can be prevented by the timely intervention [1].

Hu et al. [14] investigated students' interaction data in an online undergraduate course by using EDM techniques to develop an academic prediction system that could predict the students' learning outcomes exhibited by students in the course recorded in a learning management system (LMS). Various prediction models were used to predict the performance of the pass/fail of students. The optimal classification algorithm in the prediction system is the Classification and Regression Tree (CART) supplemented by AdaBoost. The experiment produced a classification accuracy of 90%. The study concluded that the early warning prediction system successfully predicted students' learning performance in an online course.

Akcapina et al. [15] proposed learning analytics to develop an early warning system for predicting at-risk students registered in an online course in a university. The study was carried out using a dataset of 76 second-year students registered in the Computer Hardw Course. The prediction model used in the system is a k-nearest neighbor (KNN). The experiment was examined regarding data obtained in Week 3, 6, 9, 12, and 14 to predict if at the end of the term student will be unsuccessful or successful. In the data of the first 3 weeks, the prediction rate of predicting unsuccessful student is 74%, while in the week 14th, the prediction rate increased to an accuracy of 89%.

Lee and Chung [16] developed a dropout early warning system based on machine learning to improve the performance of dropout prediction. The study dealt with the problem with the class imbalance between non-dropout and dropout groups of students. The two baseline prediction models used in this early warning system are random forest (RF) and boosted decision tree (DT). The RF and boosted DT are combined with the synthetic minority oversampling technique (SMOTE). The data used in this study is 165715 records of high school students taken from the National Education Information System (NEIS) of South Korea. The combination of a boosted decision tree that combined with SMOTE produced the best results and improved the performance of the dropout early prediction system.

C. The Current Study

Even if there were some existing works have proposed early perdition systems using popular algorithms in EDM in higher education or online courses, but still, a lot of attention is needed to build an academic performance prediction system with the analysis and help of developed prediction methods in EDM to get high and superior classification results. This work proposes a study of developing an academic performance prediction system (APPS) to predict student performance in high schools. The study compost of the selection of informative data, a proposed feature selection method, developed EDM models, and a web-based support system (the APPS) for timely-intervention to poor-performing students. The framework of the study is illustrated in Fig. 2.

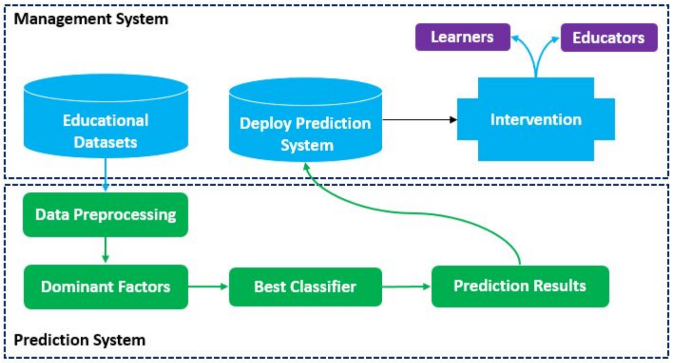


Fig. 2. The Framework of the Development of the APPS for Educational Settings.

III. Methodology

A. Participants and Data

To give intervention to high school students, informative data describing student learning patterns and highly influencing factors is required. However, in most developing countries, there is a shortage of educational data in high schools. Even if there exists, most of the existing data are only students' personal information which is not so useful for intervention purposes. Hence, this study carefully designs a questionnaire form concerning related important factors affecting student performance. The questionnaires for data collection were prepared with references, assistance, and guidance from (i) review literature, (ii) teachers from diverse educational institutions, (iii) staff from the department of research (MoEYS: Ministry of Education Youth and Sport, Cambodia), and (iv) senior researchers in the education field.

The target of the study is to improve the poor performance of students in high schools. The data used in this study was obtained from many high schools in Cambodia. Educators and related individuals can access a given online repository where survey questionnaires were designed and subsequently the survey for data collection can be conducted using Google document at any time they need. However, the time of conducting the survey is a critical factor. Oftentimes, the good time for intervention is before the final exam of the academic year. The data used in this study was collected by sharing questionnaires to high school students in the semester I. The reason is that this is a good time when the students already started their classes. They have managed the overall pictures of their learning habits, learning outcomes, and observed factors that have significant impacts on their learning outcomes during the semester. Collecting data in this period can help in predicting students' final grades and performance levels, so that the intervention can be implemented at the beginning of Semester II, especially before the final national exam. The questionnaire comprises 50 questions covering their personal information (6 questions), domestic factors (17 questions), students or individual factors (15 questions), school factors (14 questions), and score record (1 question) as shown in Table I. After students respond to the questionnaire, data can be collected and automatically stored in a repository where users can easily download the data with the right format and upload it in the prediction system to get prediction results. However, personal information is hidden since it contains some information that needs to be protected and some information that contains students' identity so that it cannot be used for intervention. The data used in the prediction consists of 43 predictors/features and one output variable. The output or target is the performing levels of students based on their score record.

B. Data Preprocessing

Data preprocessing is a boring but important phase that concern various data operations. Each operation aims to help EDM build a better predictive model. It is quite an important task to consider before putting into prediction. The proposed algorithms require data cleaning, data transformation, and data discretization to transform the data into an executable format and improve the performance of the models. The data preprocessing tasks and the experiment in our work were done using R Studio, an integrated development environment (IDE) for R programming language.

C. Evaluation Metrics

The classification performance is evaluated based on evaluation metrics of prediction tasks. We use two standard evaluation metrics to evaluate the performance of our proposed models. The two metrics are Accuracy (ACC) and Root Mean Square Error (RMSE).

TABLE I. FEATURES AFFECTING STUDENTS PERFORMANCE

Factors	ID	Predictors (number of questions)	Data types
		Personal information (6)	
Domestic	PEDU	Parents' educational levels (2)	Nominal
	POCC	Parents' occupational status (2)	Nominal
	PSES	Parents' socioeconomic levels (3)	Ordinal
	PI	Parents' involvement (4)	Ordinal
	PS	Parenting styles (4)	Ordinal
	DE	Domestic environment (2)	Ordinal
Student	SELD	Self-disciplines (5)	Ordinal
	SIM	Students' interest and motivation (4)	Ordinal
	ANXI	Students' anxiety toward their classes and exams (3)	Ordinal
	POSS	Students' possession materials (3)	Nominal
School	CENV	Class environment (1)	Ordinal
	CU	Curriculum (2)	Nominal
	TMP	Teaching methods and practices (4)	Ordinal
	TAC	Teachers' attribute & characteristics (4)	Ordinal
	RES	Academic resource (3)	Nominal

First, we want to compute the rate of our correct prediction. Hence, the first and foremost used metric in classification, called accuracy is used. From Table II, TP denoted the number of correct predictions, and E the denotes incorrect predictions. Accuracy of classification can be computed by using (1).

$$\text{ACC} = \frac{\text{Correctly predicted values}}{\text{Total values}} = \frac{\sum \text{TP}_\text{i}}{\sum \text{TP}_\text{i} + \sum \text{E}_\text{ij}} \qquad (1)$$

TABLE II. CONFUSION MATRIX OF THE PREDICTION

		Predicted Classes			
		HR	***MR***	***LR***	***NR***
Actual Classes	***HR***	TP_1	E_{12}	E_{13}	E_{14}
	MR	E_{21}	TP_2	E_{23}	E_{24}
	LR	E_{31}	E_{32}	TP_3	E_{34}
	NR	E_{41}	E_{42}	E_{43}	TP_4

Our target is the predefined classes of student performance based on students' learning outcomes (scores). On the other hand, it is hard to predict their real ability or performance levels. Therefore, it is also important to measure how close our prediction to the true value. Hence, another metric, root mean squared error (RMSE) is proposed. The groups of poor-performing students are classified into four levels: high risk (HR), medium risk (MR), low risk (LR), and no risk (NR), which are represented by 1, 2, 3, and 4. Using a confusion matrix in Table II, RMSE can be calculated using (2).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^a - y_i^p)^2} \qquad (2)$$

where $y^a \in \{1,2,3,4\}$ is the actual performance level and $y^p \in \{1,2,3,4\}$ is the predicted performance level. In contrast to the ACC, the smaller the RMSE, the better the model is.

D. Feature Selection Methods and Dominant Set

The performance of children, adult or adolescent can be affected by many influencing factors, especially external factors, motivation, and longitudinal factors. The main factor for students to be successful in their academic lives is not always about cleverness or IQ, but about discipline, motivation, and passion, which affect by environments around (themselves, parents, educators, and friends). The predictors are obtained for possibly effective factors that are categorized into three main factors: home or domestic factors, student or individual factors, and school factors [17]. As dimensionality of domain expands, the number of features affects student performance increases. However, it is not necessary to input all the features in the prediction system or not convenient to consider all factors for intervention. Hence, we wish to obtain the optimal set of factors with less dimension are needed and sufficient to control the success of students and improve the performance of classification models. We call that optimal set as the dominant set. Dominant factors play two important roles. First and foremost, determining dominant factors is used to enrich the quality of data, reduce computational costs, and improve prediction or classification performance. Secondly, the dominant factors describe the learning behaviors and factors that affect students' achievement and well monitor or assess the target in the academic system.

This study proposed feature selection methods to gain informative features. By gaining the informative features or dominant set, it can improve the performance of prediction models and use it as a recommendation to learn behaviors of students for intervention. Feature selection (FS) is a popular technique in data mining that is used to accomplish this purpose. There are three main approaches to feature selection, filter methods, wrapper methods, and embed/hybrid methods [18]. Wrapper and embed/hybrid methods are mostly computationally expensive to run for optimal feature subsets [19]. Filter-based selection methods are simple but effective in selecting important features and enhance the quality of prediction and classification performance. Filter feature selection is independent of classifiers and more scalable than wrapper methods [20]. We observe the performance of each feature selection method and selected the method that boosts the performance of classification accuracy. Three existing FS methods and a proposed FS method are studied and compared.

1) Information Gain (IG): IG is a commonly used feature selection method aiming at reducing dimensions of big data and improving the performance of prediction models [11][13]. IG measures the relevance of a feature by separating the training samples of input features to its target class. The algorithms use the concept of Shannon's entropy in information theory to rank the importance of input features [20].

2) Chi-square (CHI): CHI is a widely used algorithm especially for testing the independence of two discrete variables [10] & [12]. It is one of the famous variable tests in statistics and a popularly used feature selection method machine learning. The algorithm uses the concept of the chi-square score of the classes to get the rank list of all attributes [21].

3) Mutual Information (MI): MI is a method in the theory of information which is used to calculate or measure the dependency between random variables [22]. It is a symmetric measurement that can recognize non-linear relationships between variables. This property has made MI as a famous method for feature selection since other widely used criterion or method can only handle linear dependencies.

4) The Proposed FS Method (MICHI): Most of the filter feature selection algorithms such as information gain (IG), symmetric uncertainty (SU), and mutual information (MI) are mutual information-based methods [22][23]. These algorithms utilize the concept of mutual information (MI) and information theory. Chi-square is one of the robust feature selection methods that it is efficient for any dataset with categorical input features [21]. MI and CHI are the two popular and effective FS algorithms; however, we believe that the combined-FS algorithm is better than trusting on a single algorithm. The MICHI: MICHI is a proposed novel feature selection method which is the combination of CHI and MI algorithm based on the ranked feature scores.

In this study, we proposed a MICHI feature selection method as a combination of MI and CHI algorithms based on the ranked vector score. Since, different feature selection methods generate their feature score differently, before combining them, we first normalize the scores of both MI and CHI scores into the same format scale. The normalization can be done as in (3).

$$\overline{\mathrm{MI}} = \frac{\mathrm{MI}_i - \mathrm{MI}_{min}}{\mathrm{MI}_{max} - \mathrm{MI}_{min}} \tag{3}$$

Similarly, the score of CHI can be normalized as in (4).

$$\overline{\mathrm{CHI}} = \frac{\mathrm{CHI}_i - \mathrm{CHI}_{min}}{\mathrm{CHI}_{max} - \mathrm{CHI}_{min}} \tag{4}$$

Next, we can get the vector score of the MICHI algorithm which rearranges the order of importance of features base the combined scores as in (5).

$$\mathrm{MICHI} = \begin{pmatrix} \overline{\mathrm{MI}} \\ \overline{\mathrm{CHI}} \end{pmatrix} \tag{5}$$

The score contains the information of both MI and CHI scores. Recall that to get the magnitude of a vector is given by the Euclidean norm of the vector. Hence, the score of the magnitude of the score vector can be computed using (6).

$$|\mathrm{MICHI}_i| = \sqrt{\left(\overline{\mathrm{MI}_i}\right)^2 + \left(\overline{\mathrm{CHI}_i}\right)^2} \tag{6}$$

This means that the score of a feature in the MICHI algorithm can be computed as the norm of its score generated by the MI algorithm and score generated by the CHI algorithm.

E. Classification Algorithms

Several EDM techniques from many works of literature [2]- [13] were considered. 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. Three classes of EDM techniques consist of statistical analysis techniques (predictive structural equation modeling), machine learning techniques (random forest, logistic regression, C5.0 of the decision tree, sequential minimal optimization, and multilayer perceptron), and a deep learning framework called deep belief network were executed and compared [24]. The random forest was found to be the best prediction model. The improvement of the previously proposed prediction models and additional models for predicting student performance was further carried out [25]. K-nearest neighbor (KNN), ensemble decision tree (Boosted C5.0 and Bagged CART), and random forest (RF) outperformed the rest prediction models. The developed prediction models are proposed in earlier works [26][27]. In this study, we use KNN, and three developed classifiers as our prediction models.

1) K-nearest neighbor (KNN): KNN is one of the effective non-parametric EDM models used for classification tasks. The KNN is an effective classifier and produces higher classification results [25]. Like many other classifiers, the k-NN classifier is noise-sensitive. Its accuracy highly depends on the quality of the training data. Noise and mislabeled data, as well as outliers and overlaps between data regions of different classes, lead to less accurate classification. It performs much better with the dominant set of important features.

2) Hybrid C5.0 and Hybrid RF: In the earlier work [26], we have proposed hybrid machine learning models which are the combination of baseline classifiers (support vector machine (SVM), naïve Bayes (NB), C5.0, and random forest (RF)) with principal component analysis (PCA) and validated by 10-fold cross-validation. The Hybrid C5.0 (C5.0+PCA+10-CV) and Hybrid RF (RF+PCA+10-CV) were found to be the best classifiers in our classification problem.

3) Improved Deep Belief Networks (IDBN): The IDBN is the optimization approach of deep belief network (DBN) model. We proposed an optimization approach composed of (i) feature selection method, (ii) optimization of hyper-parameter, and (ii) regularization method. The developed model has introduced in our earlier work [27]. The improved model was found to produce the most classification results when using larger datasets.

IV. Experimental Results of Prediction Models

A. Experimental Results

This section gives comparative results of the four proposed classifiers on the feature set of each FS method. Table III indicates the experimental results of the proposed classifiers with the original dataset. Table IV to Table VII shows the performance of the classifiers on subsets selected by IG, CHI, MI, and the proposed MICHI. The dominant set of each FS algorithm is found and the experimental results on each set are studied and compared.

Table III indicates the experiment results of the four classifiers with the original dataset concerning the two metrics: ACC and RMSE. The two tree-based models, Hybrid C5.0 and Hybrid RF generate the highest ACC and lowest RMSE.

The results presented in Table IV demonstrate the performance of the four classifiers using datasets from the IG feature selection method. The performance of Hybrid C5.0 and Hybrid RF are comparatively better than the other models. Hybrid RF generates the highest ACC and lowest RMSE with both selected sets and dominant sets.

From Table V, the performance of KNN is significantly improved when using the dominant sets containing the best 5 features set selected by the CHI algorithm. Hybrid C5.0 and Hybrid RF generated the most successful classification performance in this classification problem. The dominant sets improved the performance of both hybrid models.

The results of ACC and RMSE of the four classifiers using subsets from the MI algorithm are shown in Table VI. Hybrid C5.0 and Hybrid RF outperform the other models when using the selected set. However, the performance of KNN, Hybrid C5.0, and Hybrid RF are comparatively improved when considering the dominant sets.

Table VII demonstrates the performance of the proposed classifiers with the input feature subsets from the proposed MICHI algorithm. The performance is significantly improved when using the dominant sets. Hybrid RF produced the most successful classification result.

TABLE III. Results of Proposed Models on Original Datasets

Proposed Models	**KNN**	**Hybrid C5.0**	**Hybrid RF**	**IDBN**
ACC (%)	95.95	99.25	99.72	83.14
RMSE	0.261	0.073	0.041	0.759

TABLE IV. The Experimental Results using Subset from IG (29 Features)

Models	**Selected set**		**Dominant set**		
	ACC	***RMSE***	***N***	***ACC***	***RMSE***
KNN	95.34	0.257	5	97.34	0.153
Hybrid C5.0	99.85	0.040	29	99.85	0.040
Hybrid RF	99.87	0.038	29	99.87	0.038
IDBN	85.63	0.571	29	85.63	0.571

TABLE V. The Experimental Results using Subset from IG (29 Features)

Models	**Selected set**		**Dominant set**		
	ACC	***RMSE***	***N***	***ACC***	***RMSE***
KNN	95.34	0.257	5	99.17	0.087
Hybrid C5.0	99.85	0.040	29	99.86	0.026
Hybrid RF	99.87	0.038	29	99.95	0.015
IDBN	85.63	0.571	29	85.45	0.608

TABLE VI. The Experimental Results using Subset from MI (30 Features)

Models	**Selected set**		**Dominant set**		
	ACC	***RMSE***	***N***	***ACC***	***RMSE***
KNN	95.34	0.257	5	99.77	0.047
Hybrid C5.0	99.85	0.040	29	99.85	0.040
Hybrid RF	99.87	0.038	29	99.89	0.035
IDBN	85.63	0.571	29	87.03	0.525

TABLE VII. The Experimental Results using Subset from MICHI (29 Features)

Models	**Selected set**		**Dominant set**		
	ACC	***RMSE***	***N***	***ACC***	***RMSE***
KNN	95.34	0.257	5	99.85	0.011
Hybrid C5.0	99.85	0.040	29	99.89	0.035
Hybrid RF	99.95	0.011	*	99.98	0.008
IDBN	85.63	0.571	29	87.01	0.542

(*: from a set of five to 29 features, the values of ACC and RMSE are not statistically different)

B. Summary and Discussion

This study aims to obtain both an optimal prediction model and dominant set as useful information for educational stakeholders. The optimal method is combined with a dominant set to get accurate and informative results. Hybrid RF generates the highest ACC and the smallest RMSE. The result indicates that the proposed Hybrid RF with the dominant set selected by the MICHI algorithm performs the most successful classification result with an accuracy of 99.98% and RMSE of 0.008. Additionally, the experimental results indicate that the proposed feature selection method MICHI extracts the best dominant set. MICHI is a proposed novel feature selection method which is the combination of CHI and MI algorithms based on the ranked feature scores. The feature set selected by the MICHI algorithm is rank manually of the most important factors affecting student performance. The set does not only improve the performance of the prediction model but also describe the factors and student learning behaviors that require assistance and intention for improvement. Early predictions combined with counseling and intervention is known as an effective solution for improving the given problem. Therefore, the Hybrid RF model and the dominant set from the MICHI algorithm are combined and integrated into the APPS.

V. Design of the Apps

A. The Design of the APPS

This section presents the design of the APSS. The design consists of the web-based application linked to the server via the internet. The clients can assess the application and upload their data to obtain the results. The web-based application was created using Shiny (an open-source R package in R language). Since our experiment is done in R, hence Shiny App is the best choice. The shiny application is built using R language with its extension of the simplest code from HTML. CSS, and JavaScript. However, it is good that Shiny helps us to turn our analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge.

The architecture of the academic prediction system is shown in Fig. 3. The recorded data that input in UI (user interface) is transferred to the server where the prediction model is executed to figure out what is the performance level of students. The results are then sent back to the client or user on the UI screen. The users are educational stakeholders such as teachers, administration office, or related individuals who have a dataset containing their student information. They can input the collected dataset and obtain the results. Fig. 4 presents the introduction of the system.

The prototype illustrated in Fig. 5 presents the operation follow of tasks in using the APPS. The flow chart aims to introduce to users how to use this system and the steps in using it to obtain the prediction results.

The system gives a link to an online repository for data collection by conducting a survey questionnaire designed using Google document so that users (educators, teachers, and schools) can decide when is the suitable time to make an early prediction for intervention. Once the data collection is done, users can upload the data in the system (*File Upload* button). The results of prediction will be released in the system; in which users can download the prediction results, identify the poor performing group of students for intervention, and other useful information for future use.

Fig. 6 demonstrates the information of students (the dominant factors) and the rank of highly influencing factors affecting student performance. The description of student data can be stored in the interface and viewed to understand students' information. The results of the prediction are shown in Fig. 7 where at the bottom, there is a button that users can download the prediction results as a dataset in CSV format for the use of any settings in the intervention.

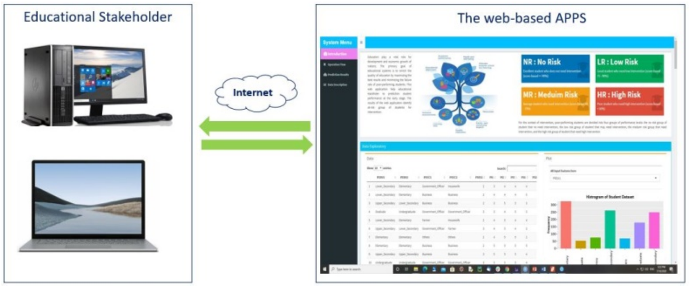
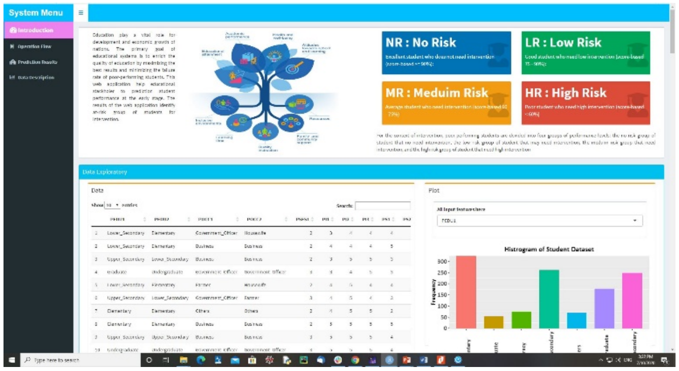


Fig. 3. Academic Performance Prediction System Architecture.



Education play a vital role for development and economic growth of nations. The primary goal of educational systems is to enrich the quality of education by maximizing the best results and minimizing the failure rate of poor-performing students. This web application help educational stackholer to prediction student performance at the early stage. The results of the web application identify at-risk group of students for intervention.

NR : No Risk
Excellent student who does not need intervention (score-based >= 90%)

LR : Low Risk
Good student who need low intervention (score-based 75 - 90%)

MR : Meduim Risk
Average student who need intervention (score-based 60 - 75%)

HR : High Risk
Poor student who need high intervention (score-based < 60%)

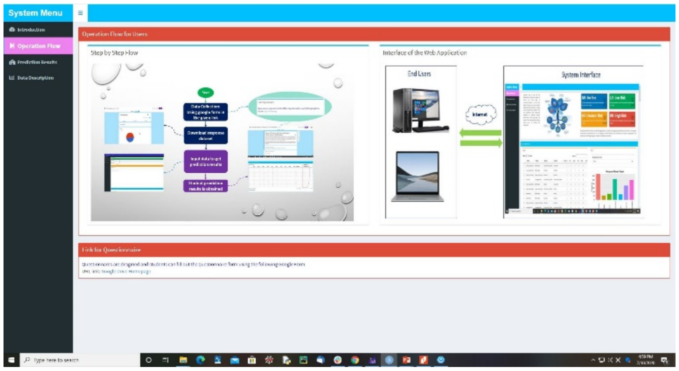
For the concept of intervention, poor performing students are divided into four groups of performance levels: the no risk group of student that no need intervention, the low risk group of student that may need intervention, the medium risk group that need intervention, and the high risk group of student that need high intervention

Data Exploratory

Show 10 entries Search:

All input features here: FEDU1

Fig. 4. A Prototype of the Introduce Interface of the APPS.


System Menu
Introduction
Operation Flow
Prediction Results
Data Description
Start
Data Collection using google form in the given link
Download response dataset
Input data to get prediction results
Student prediction results is obtained
End Users
Internet
System Interface
Type here to search

Operation Flow for Users

Step by Step Flow

Interface of the Web Application

Link for Questionnaire

Questionnaires are designed and students can fill out the questionnaire form using the following Google Form URL link: Google Drive Homepage

Fig. 5. Operation Flowchart to Instruct users of the Overall Process to Get Prediction Results.

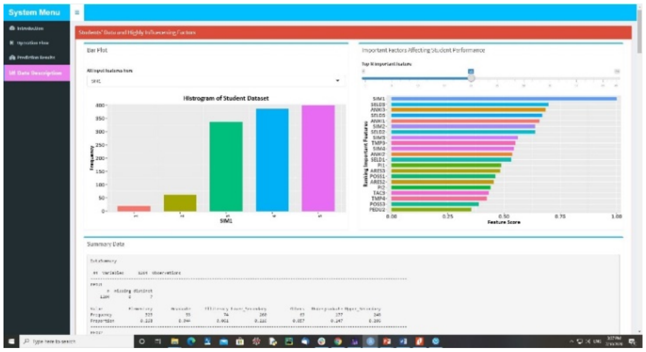


Fig. 6. Summary of the Dominant Factors and Ranking the Highly Influencing Factors.

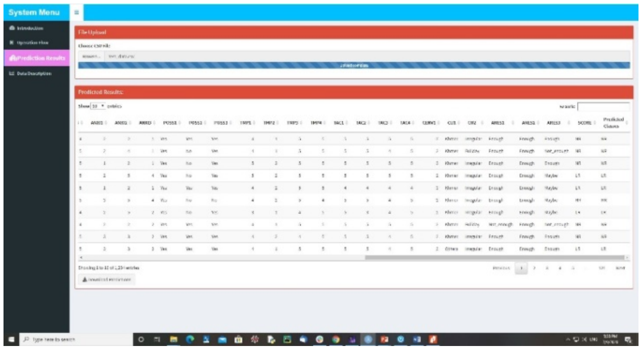


Fig. 7. Prediction Results Identifying the at-Risk Levels of Poor-Performing Students.

B. The System Deployment

Once the development of the Shiny app (the APPS) is done, it is shared or distributed with users. There is two basic option that can share. The first one is the Shiny app is shared as R scripts, users can use and edit from runGitHub. Users can use these scripts to launch the app from their R session. Users with no knowledge of programming or no care of how it works, the second type is the most comfortable way. Users can use the app from a web page or browser, which is from Shinyapps.io. This is definitely the most user-friendly way to share a Shiny app to users. They can navigate to our app through the internet with a web browser.

C. The System Evaluation

In evaluating the usability of the web application of the APPS, we designed a subjective questionnaire. The evaluation is carried out based on ten characteristics of the system: useful, motivating, user-friendly, relevant, reliable, efficient, organized, time cost, adaptable, and sophisticated. The questionnaire was designed with a 5-point Likert scale ranging from 1 to 5 (1: Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly agree). The participant of 57 students and 10 teachers are invited to join the presentation and test the system. The response regarding the survey of usability of the web application is shown in Table VIII. When writing 11 (1) mean 11 students and 1 teacher agree with the given statement.

The student participants are 35 males and 22 female students, and teacher participants are 7 male teachers and 3 female teachers. Most students and teachers agreed that the system is useful and effective for predicting student performance and identify the poor performing student for intervention. The survey result shows that 82.08% supported that the system is useful (55.22% agree, 26.86% strongly agree), 83.58% thought that the system is motivating (62.68% agree, 20.89% strongly agree), 91.04% felt that the interface is friendly (82.08% agree, 8.95% strongly agree), 85.07% believed the information was relevant (58.20% agree, 14.92% strongly agree), and 73.13% thought that the system was reliable (62.68% agree, 20.89% strongly agree), 82.08% reported the efficient of the system (55.12% agree, 26.86% strongly agree), 74.62% claimed that the system was well-organized (58.78% agree, 20.89% strongly agree), 91.04% realized that the system speed (time cost) was fast (64.17% agree, 26.86% strongly agree), 88.05% of participants perceived that the system was adaptable (62.68% agree, 20.89% strongly agree), and 92.58% felt that the system was sophisticated (77.61% agree, 14.92% strongly agree). The analysis of the evaluation is shown in Fig. 8. The evaluation survey indicates the effectiveness and usefulness of the developed academic prediction system.

TABLE VIII. THE SURVEY RESULTS FOR EVALUATING THE APPS

Statement	Description	5-point Likert Scale				
		1	2	2	4	5
Useful	The system has helped student/instructor	0 (0)	0 (0)	11 (1)	31 (6)	15 (3)
Motivating	it is interesting to see that the system can give a feedback response for educators of the challenges the students face that affect their learning outcomes	0 (0)	0 (0)	12 (1)	37 (5)	10 (4)
User-friendly	The interface is easy to use	0 (0)	0 (0)	4 (2)	52 (3)	1 (5)
Relevant	It's easy to find the information I need	0 (0)	0 (0)	17 (3)	35 (5)	15 (2)
Reliable	I feel comfortable using the system	0 (0)	0 (0)	15 (3)	35 (4)	7 (3)
Efficient	It produces results immediately after feeding in the information, and results are given correctly, easily and fast.	0 (0)	0 (0)	11 (1)	31 (6)	15 (3)
Organized	It's easy to learn its use, the interface is simple and well structure.	0 (0)	0 (0)	14 (3)	31 (5)	12 (2)
Time cost	The data can be obtained anytime and fast with the questionnaire in Google form and results of prediction can obtain immediately after data collection	0 (0)	0 (0)	5 (1)	39 (4)	13 (5)
Adaptable	Student's weakness is known so that the right intervention can be put in place	0 (0)	0 (0)	6 (2)	44 (4)	7 (4)
Sophisticated	This is innovative technology in educational system	0 (0)	0 (0)	5 (0)	46 (6)	6 (4)

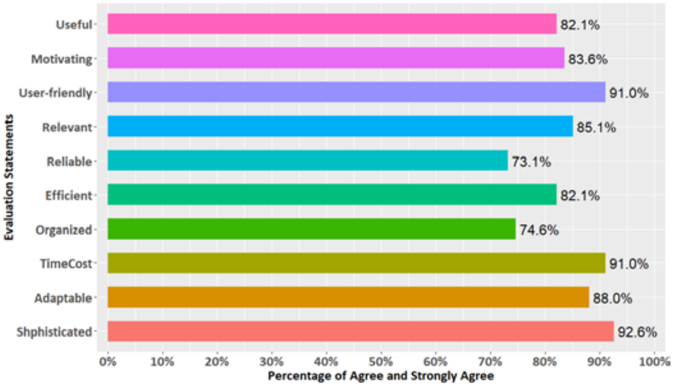


Fig. 8. User Feedback Rating the Characteristics of the APPS.

VI. Conclusion

The purpose of this study is to use EDM techniques to give an early-stage prediction for intervention and improving student performance based on a developed academic performance prediction system (APPS). The system gives faster and more comfortable ways of users to get in-time data of students for early predicting student performance levels and learning patterns and improving academic outcomes. The APPS composes of a developed prediction model and an effective feature selection method for determining the dominant factors for the success of student performance. We proposed a comparative study of EDM of prediction and classification task, the outperformed prediction models are then developed and optimized to get the most successful classification results. The comparative experiment of four classifiers (KNN and Hybrid C5.0, Hybrid RF, and IDBN) is carried out using feature sets from four FS methods (IG, CHI, MI, and the Proposed MICHI method). The analysis of dominant factors is cooperated and combined with the best classifier. The dominant set obtained from the MICHI algorithm significantly improves the performance of prediction models and used as a set of highly influencing factors that need to be considered for intervention and improvement of student performance. The experimental outcomes indicate that Hybrid RF outperformed the other three classifiers with the superior classification results. The developed prediction model and dominant factors are integrated into the web-based prediction system.

The finding of this work confirms the effectiveness of the prediction model and the usability of the APPS. The system illustrates operation flows consist of the method of faster and more comfortable way of data collection, dominant factors that have a significant impact on the student performance, and results of prediction. According to the results from APPS, it is informative for educational institutions to carry out the right intervention for improving student performance. The developed prediction system will help educational stakeholders such as teachers, educational administrators, and policymakers, and related individuals to improve academic performance in educational institutions. The teachers can quickly adjust their teaching methods and adopt adaptive teaching approaches to meet the needs of students. Educational stakeholders and related individuals can figure out the weak points and the solution to make improvements. Therefore, overall learning quality and learning performance can be improved greatly and reduce the failure rate of poor-performing students.

Acknowledgment

The authors are grateful to the staff in the Ministry of Education, Youth, and Sport (MoEYS, Cambodia), senior researchers in the educational field, and directors, administrators, and teachers in high schools in Cambodia for guidance, assistance, help for data collection and support to make this research work realizable. This work was supported by JSPS Grant-in-Aid for Scientific Research(S) (Grant Number JP17H06115).

References

[1] G. Ackcapinar, G, M. N. Hasine, R. Majumdar, B. Flanagan, and H. Ogata, “Developing early system for spotting at-risk students by using eBook interaction logs,” Smart Learning Engineering, vol. 6, Issue 4, pp. 1-15, March 2019.

[2] S. Slater, S. Joksimovic, V. Kovanovic, R. S. Baker, and D. Gasevic, “Tools for educational data mining,” Journal of Educational and Behavioral Statistics, vol. 10, Issue 3, pp. 85-106, 2017.

[3] C. Romero and S. Ventura, “Data mining in education,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, issue 1, pp. 12-27, 2013.

[4] C. Romero and S. Ventura, “Educational data mining: A Survey review of the state of the art,” IEEE Transaction on System, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, issue 6, pp. 601-618, 2010.

[5] P. Thakar, A. Mehta, and Manisha, “Performance analysis and prediction in educational data mining,” International Journal of Computer Application, vol. 110, no. 15, pp. 60-68, 2015.

[6] A. Pena-Ayala, “Educational data mining: Survey and a data mining-based analysis of recent works,” Expert Systems with Application, vol. 41, pp. 1432-1462, 2014.

[7] A. A. Mazidi and E. Abusham, “Study of general education diploma students’ performance and prediction in Sultanate of Oman, based on data mining approaches,” International Journal of Engineering Business Management, vol. 10, pp. 1-11, 2018.

[8] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 40, issue 6, ID: 21355, 2020.

[9] OECD, Low-performing students: Why they fall behind and how help them succeed, PISA, OECD Publishing, Paris, 2016.

[10] P. J. M. Estrera, P. E. Natan, B. G. T. Rivera, and F. B. Colarte, “Student performance analysis for academic rankings using decision tree approach in the University of Science and Technology of Southern Philippine senior high school,” International Journal of Engineering and Technology, vol. 3, Issue 5, pp. 147-153, 2017.

[11] G. Dimic, D. Rancic, I Milentijevic, P. Spalevic, and K. Plecic, “Comparative study: Feature selection methods in blended learning environments,” Facta Universitatis, Series: Automatic Control and Robotics, vol. 16, no. 2, pp. 95-116, 2017.

[12] M. Zaffar and K.S. Savita, “A study of feature selection algorithms for predicting students’ academic performance,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 5, 2018.

[13] A. A. Saa, M. Ai-Emran, and K. Shaalan, “Mining student information system records to predict students’ academic performance,” Springer Nature Switzerland AG 2020, AMLTA 2019, AISC 921, pp. 229-239, 2019.

[14] Y. H. Hu, C. L. Lo, and S. P. Shih, “Developing early warning systems to predict students’ online course learning performance,” Computers in Human Behavoirs, vol. 36, pp. 469-478, 2014.

[15] G. Ackapinar, A. Altun, and P. Askar, “Using learning analytics to develop early warning systems for at-risk students,” International Journal of Educational Technology in Higher Education, vol. 16, issue 40, pp. 1-20, 2019.

[16] S. Lee and J. Y. Chung, “The machine learning-based dropout early warning systems for improving performance of dropout prediction,” Journal of Applied Science, vol. 9, issue 15, pp. 3093-4016, 2019.

[17] A. Hellas et al., “Predicting academic performance: A systematic literature review,” Proceeding of Companion in Computer Science Education, Larnaca, Cyprus, pp. 175-199, July 2-4, 2018.

[18] P. Jinal and D. Kumar, “A review on dimensional reduction techniques,” International Journal of Computer Applications, vol. 173, no. 2, pp. 42-46, 2017.

[19] L. Ma et al., “Evaluation of feature selection methods for object-based land cover machine classifiers,” International Journal of Geo-Information, vol. 6, no. 51, 2017.

[20] P. Yildirim, “Filter base feature selection methods for prediction of risk in hepatitis disease,” International Journal of Machine Learning and Computing, vol. 5, no. 4, pp. 258-263, 2015.

[21] S. Bassine, “Feature selection using an improved Chi-square for Arabic text classification,” Journal of King Saud University-Computer and Information Science, vol. 32, no. 2, pp. 225-231, 2020.

[22] D. H. Mazumder and R. Vilumuthu, “An enhanced feature selection filter for classification of microarray cancer data,” WILEY ETR Journal, vol. 41, no. 3, pp. 358-370, 2019.

[23] A. Bummert, X. Sun, B. Bischa, J. Rahnenfuhrer, and M. Lang, “Benchmark for filter methods for feature selection in high-dimensional classification data,” Computational Statistics & Data Analysis, vol. 143, 2020.

[24] P. Sokkhey and T. Okazaki, “Comparative study of prediction models for high school student performance in mathematics,” Journal of IEIE Transactions on Smart Processing and Computing, vol. 8, no. 5, pp. 394-404, 2019.

[25] P. Sokkhey and T. Okazaki, “Multi-models of educational data mining for predicting student performance: A case study of high schools in Cambodia,” vol. 9, no. 3, pp. 217-229, 2020.

[26] P. Sokkhey and T. Okazaki, “Hybrid machine learning algorithms for prediction academic performance,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, pp. 32–41, 2020.

[27] P. Sokkhey and T. Okazaki, “Development and optimization of deep belief networks for academic prediction with larger datasets,” Journal of IEIE Transactions on Smart Processing and Computing, (Accepted 20-April-2020).