World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:2, 2021

# Integration of Educational Data Mining Models to a Web-Based Support System for Predicting High School Student Performance

Sokkhey Phauk, Takeo Okazaki

***Abstract*—**The challenging task in educational institutions is to maximize the high performance of students and minimize the failure rate of poor-performing students. An effective method to leverage this task is to know student learning patterns with highly influencing factors and get an early prediction of student learning outcomes at the timely stage for setting up policies for improvement. Educational data mining (EDM) is an emerging disciplinary field of data mining, statistics, and machine learning concerned with extracting useful knowledge and information for the sake of improvement and development in the education environment. The study is of this work is to propose techniques in EDM and integrate it into a web-based system for predicting poor-performing students. A comparative study of prediction models is conducted. Subsequently, high performing models are developed to get higher performance. The hybrid random forest (Hybrid RF) produces the most successful classification. For the context of intervention and improving the learning outcomes, a feature selection method MICHI, which is the combination of mutual information (MI) and chi-square (CHI) algorithms based on the ranked feature scores, is introduced to select a dominant feature set that improves the performance of prediction and uses the obtained dominant set as information for intervention. By using the proposed techniques of EDM, an academic performance prediction system (APPS) is subsequently developed for educational stockholders to get an early prediction of student learning outcomes for timely intervention. Experimental outcomes and evaluation surveys report the effectiveness and usefulness of the developed system. The system is used to help educational stakeholders and related individuals for intervening and improving student performance.

***Keywords*—**Academic performance prediction system, prediction model, educational data mining, dominant factors, feature selection methods, student performance.

## I. Introduction

EDUCATION is considered as a key factor for the development and long-term economic growth of every country. The development of the developing countries relies mainly on the development of human resources in the education domain. The poor academic performance brings up with problem such as under education and lack of qualified human resources for the development of countries. This is why academic performance in educational institutions is important. Academic performance can be defined based on the score obtained at the end of their learning activities. One of the primary goals of any educational system is to enrich the quality of education to increase the best results and decrease the failure rate of at-risk students. At-risk students such as those who are highly possible to fail, drop out, or repeat classes due to their poor performance have become the worried-tasks in educational institutions [1]. Prediction of academic performance is one of the first and foremost challenging tasks for improving academic performance since at-risk students can only be accurately identified early enough through performance prediction [2]. Therefore, early prediction systems have been considered to be a powerful tool for early identification of students who are at risk of failure, drop out, repeated classes, and other target learning behaviors [3].

Several works have been conducting on predicting and evaluating academic performance. However, most of the study seems to be more available with higher education, while secondary education is considered as the background of higher education or carrier fields. To achieve effective methods of intervention and improving poor-performing students, an accurate prediction of their performance is required [4]. The main goal in educational institutions is to increase the passing rate and reduce the failure rate of students in their exams. Education institutions and the Ministry of Education oftentimes search for new strategies or policies to enhance educational performance. In the policy for the development of Cambodia, the Ministry of Education, Youth, and Sport (MoEYS, Cambodia) recently set out the strategies for the development of STEM (science, technology, engineering, and mathematics) fields and innovation to high school and university students [5]. Minister of MoEYS is trying to strengthen the education performance in high school by providing intervention before the final exam for improving student performance to increase the rate of passing the national exam (K-12), and enhancing student ability to respond to the STEM discipline and digital learning of innovation strategies. The new policies for developing the education system in Cambodia in terms of STI (science, technology, and innovation) are set out for 2020-2030 [6]. In general, early prediction systems are prediction models used to prevent expected failure at an early stage. Early prediction of student performance has a high impact on in-time intervention, managerial setting, scheduling, and planning in education institutions.

EDM is a disciplinary field of research utilizing data mining, machine learning, and statistics applied to uncover knowledge from data in educational environment [7]. Various managerial

Phauk Sokkhey is with Graduate School of Engineering and Science, University of the Ryukyus, 1 Senbaru, Nishihara, Okinawa 903-0213, Japan, on leave from Institute of Technology of Cambodia, Phnom Penh, Cambodia (e-mail: sokkheymath15@gmail.com).

Takeo Okazaki is with Department of Computer Science and Intelligent System, University of the Ryukyus, 1 Senbaru, Nishihara, Okinawa 903-0213, Japan (e-mail: okazaki@ie.u-ryukyu.ac.jp).

settings, planning, and scheduling in educational settings need more advanced techniques in EDM to investigate student learning patterns and give a prediction, and give new insights for improving academic performance [8], [9]. Romero and Ventura published the first EDM literature survey concerning data mining techniques in education in 2007 [10], which were improved in 2010 [11]. Many works continuously search for more accurate and advanced methods to give an analysis of various aspects of academic performance. Fig. 1 provides the cycle of applying EDM methods in the educational environment. Supervised and unsupervised methods of EDM such as clustering, classification, regression, relationship mining, pattern mining, and text mining are utilized by various educational platforms such as traditional classroom, e-learning systems, adaptive and intelligent web-based educational systems [7]-[13]. In recent decades, the more availability of educational data grows due to more availability of record systems, databases, and technology. However, the existing data in various educational database and data that are possible to obtain are oftentimes unstructured, bigger, more complicated. This is why EDM plays an important role importantly to deal with these types of datasets and resolve problems in educational issues [11].

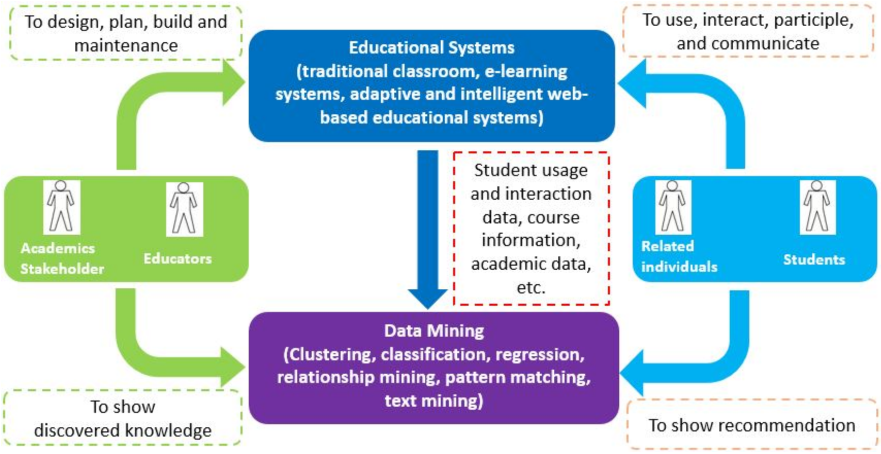

To design, plan, build and maintenance
Educational Systems (traditional classroom, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participle, and communicate
Student usage and interaction data, course information, academic data, etc.
Academics Stakeholder
Educators
Related individuals
Students
Data Mining (Clustering, classification, regression, relationship mining, pattern matching, text mining)
To show discovered knowledge
To show recommendation

Fig. 1 The cycle of applying data mining in EDM (adapted from [10])

Several studies [14]-[20] searched for effective methods to learn about students' information and give feedback for enhancing their performance. The purpose of this work is thus to develop an effective prediction system based on a web-based application using prediction models in EDM. The motivation is to design an effective intelligent system that can give an early prediction of student performance in high school education. The study aims to identify the poor performance group of students, understand their learning patterns so that educational stakeholders such as teachers, policymakers, and the related individuals can use the prediction results to give recommendations for intervention and improvement of academic performance.

*Research Questions*

In the context of academic poor-performance, EDM is used for timely prediction for intervention and improvement. From the rise of the advancement of this discipline, we proposed an APPS modeling to develop a web-based prediction system for early predicting student performance. The academic prediction modeling in the study of this literature is carried out to answer the following questions:

The modeling in this study is driven as the following research questions:

(i) Question 1: How can we accurately detect student learning patterns or factors that highly affect their academic performance (feature selection)?

(ii) Question 2: Is the proposed classifier or prediction models can generate the most successful results (the proposed EDM models)?

(iii) Question 3: What is the implication of the research? How can educational stakeholders utilize the proposed work (the designed APPS)?

In education settings, it is a challenging task to obtain information and knowledge from educational data for improvement. The remainder of the paper is organized into three main sections. The Literature Review of Related Works section gives reviews of literature that using feature selection methods and prediction models in EDM and developing early warning systems in education using EDM. APPS Modeling Using the CRISP-DM Methodology section presents the CRISP-DM process proposed for academic performance prediction modeling. Experimental Results: Materials and Methods in Developing APPS section describes the materials, methods, experimental results. Model Deployment: Design and Implementation of the APPS discuss the design of APPS and it implementation and usefulness for educational stakeholders. In the final section, we draw the conclusion of our proposed work.

## II. Literature Review of Related Works and the Current Study

### *A. Feature Selection Methods and Prediction Models Using EDM*

The first EDM literature survey was noticed in the work that was produced in 2007 by Romeo and Ventura [10], which was then improved in 2010 [11]. Then many systematic literature reviews are conducted to view the effectiveness of prediction models and what has been done so far. Various EDM methods range from statistical methods, machine learning (ML), and deep learning (DL) has gained increasing interest in applying its merits to the educational environment [2], [3].

This section summarizes related work on using EDM methods to predict academic performance. We investigate a wide range of literature focusing on three main points: prediction models or classifiers, feature selection methods, and early prediction systems or APPS. Table I summarizes the literature of related work.

TABLE I
THE SUMMARY OF THE LITERATURE REVIEW OF RELATED WORK

| References | Prediction Models or Classifiers | Feature Selection Methods | Data sets | Key Findings |
|---|---|---|---|---|
| Estrera et al. (2017) [14] | DT, NB, and KNN | CHI, IG, and GR | Student information records | DT algorithm combine with the FS produces the highest accuracy |
| Dimic et al. (2017) [15] | NB, AODE, DT, and SVM | IG, SU, REF, CB, Wrapper, and MI | Data from a blended learning environment | Wrapper and MI generate the most success classification results |
| Zaffar and Sativa (2018) [16] | BN, NB, NBU, MLP, SL, SMO DT, OneR, PART, JRip, DS, J48, RF, RT, and RepT | CB, CHI, Filtered, IG, PCA, and REF | Dataset1 obtained learning management system (LMS), and Dataset2 collected from three different colleagues in India | The overall results indicate that the FS methods improve the performance of prediction models |
| Saa. et. al. (2019) [17] | DT, ANN, RF, NB, LR, and GLM | Information gain (IG) | A dataset from high school records and university records | RF produces the most successful classification result |
| Hu et al. (2014) [18] | CART and AdaBoost | None | Learning management system (LMS), an early warning system exists | The early warning system successfully predict student performance |
| Akcapina et al. (2019) [19] | KNN | None | Computer Science course in a university and early warning system exists | The prediction accurately predict the final week of student performance with an accuracy of 89% |
| Lee and Chung (2020) [20] | RF and DT | None | National Education Information System (NEIS) of South Korea and dropout early prediction system | The dropout prediction system affecting prediction high school students utilizing data from the information system |
| Sokkhey and Okazaki (2019) [30] | SEM, LR, C5.0. RF, MLP, SVM, and DBN | None | Synthetics datasets are synthesized from benchmark datasets. | RF produce the most successful result follow by C5.0 and SVM |
| Sokkhey and Okazaki (2020) [31] | SEM, LR, Boosted C5.0. Bagged CART, RF, KNN, MLP, SVM, and DBN | IG, SU, CHI, and REF | Dataset collected from many high schools in Cambodia | RF, Boosted C5.0, Bagged CART, and KNN is the four best classifiers comparatively better than the rest models. |
| Sokkhey and Okazaki (2020) [32] | Hybrid RF, Hybrid C5.0, Hybrid NB, and Hybrid SVM | None | Dataset collected from many high schools in Cambodia and the other two synthetic datasets | Hybrid RF and Hybrid C5.0 are the two best models |
| Sokkhey and Okazaki (2020) [33] | LVQ, MLP, DBN, and IDBN | Information gain (IG) | One real dataset and four synthetics datasets | IDBN generates the most successful classification result. |

Classifier: DT: Decision tree, NB: Naïve Bayes, KNN: k-nearest neighbor, AODE: Aggregating one-dependence estimators, SVM: Support vector machine, BN; Bayesian network, NBU: Naïve Bayes updateable, MLP: Multilayer perceptron, SL: Simple logistic, SMO: Sequential minimal optimization, DS: Decision Stump, RT: Random tree, RepT; REP Tree, ANN: Artificial neural network, LR: Logistic regression, GLM: Generalized linear model, CART: Classification and regression tree, SEM: Structural equation modeling, LVQ: Learning vector quantization, DBN: Deep belief network, IDBN: Improved deep belief network

Feature Selection: IG: Information gain, GR: Gain ratio, REF: Relief, SU: Symmetric Uncertainty, CB: Correlation-based, PCA: Principal component analysis.

### B. Current Study

This study adopts the CRISP-DM methodology to develop prediction models and prediction systems to effectively predict student performance in high schools. Even though several works have been conducted for invesgating and predicting academic performance, a further study needs to be carried out to get more accurate prediction results with both better methodological contribution and applicability contribution. This work proposes a study of developing an APPS for timely-intervention to poor-performing students.

## III. APPS Modeling Using the CRISP-DM Methodology

The modeling in our study is conducted following the methodology stage in the Cross Standard Process for Data Mining (CRISP-DM) model [21]. As the name indicated, CRISP-DM is the standard process of data mining for extracting the information or knowledge in any business domain for new insight and fruitful results. The process consists of six main steps as shown in Fig. 2.

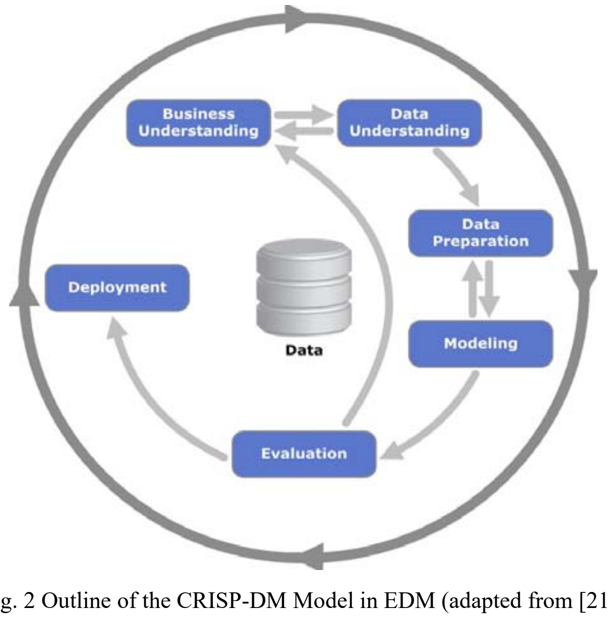


Fig. 2 Outline of the CRISP-DM Model in EDM (adapted from [21])

### *A. Business or Domain Understanding*

The prerequisite knowledge or understanding of the background of the domain is required in the first step. The problem can only be effectively solved with well understanding of the problem. Hence, by understanding the problem clearly, we then can convert it into a well-defined analytics problem and can carry out a brilliant strategy to solve it. In this study, understanding the domain, namely poor academic performance, was the initial stage of the EDM process. At this point the goal of the study was put into focus: to develop an intervention-level classifier model that predicts whether a student will require what levels of intervention to achieve passing marks in a high secondary school exit examination. Clearly, this is a classification problem; classification techniques of EDM were employed.

### *B. Data Understanding*

The data understanding phase of the CRISP-DM framework focuses on collecting the data, describing and exploring the data. This step is to know what can be expected and achieved from the data. It checks the quality of the data, in several terms, such as data completeness, values distributions, data governance compliance. At this point, data usefulness in terms of meeting the desired goal is also confirmed. In the context of this step in this research, we observed all variables which are related and influenced to academic performance. The details are discussed in the next section.

### *C. Data Preprocessing*

The best algorithm is useless unless your data are ready for the implementation. Coming up with features is difficult, time-consuming, and requires expert knowledge. Data preprocessing is an integral step data mining as the quality of data and the new insight can be obtained from it are highly influent on the performance of models. It is broadly known that there is no best model for poor quality data. Hence, data processing is an important step and a big part of data mining that contains several preprocessing tasks. Those main processing tasks include data cleaning, data transformation, feature encoding, data normalization, and feature engineering.

### *D. Data Mining or Modeling*

The selection of the modeling technique is the very first step to take. This step entails selecting the methods whether supervised or unsupervised learning (classification, regression, or clustering) on information or data in the business domain to obtain the new insight or wanted results. All the models are then assessed to make sure that they fall in line with the business initiatives. In this study, EDM models are used to predict student performance in high schools. Feature selection methods and prediction models are studied to get higher prediction results. The details are discussed in the next section.

### *E. Model Evaluation*

Evaluation of the data mining model is an integral task of modeling to confirm the effectiveness of the proposed methods. Hence, this stage of the CRISP-DM refers to the evaluation of the models via the experimental results before deploying the models for final usage. Through review and evaluation of various methods of EDM, we have investigated many effective baseline models and developed those models to get higher or more successful classification results. To answer the objectives of the problem, the evaluation of data mining results should be reached. In the concept of data mining techniques, evaluation metrics are used to measure the performance of the used models. The details are discussed in the next section.

### *F. Model Deployment*

This stage entails putting the discovered knowledge to use by incorporating it into a performance system. It could also involve just documenting the knowledge and passing it to the interested stakeholders. The aforementioned steps have been adopted in our study to generate a general research framework as discussed next. In this paper, we propose the developed EDM methods; it is subsequently integrated into the web-based application, named as the APPS. The development and implementation of the APPS are described in the last section. The design and implementation are mentioned in the next section.

## IV. Materials and Methods

### *A. Data Understanding*

The study focuses on evaluating the performance of students in secondary schools. To confirm the effectiveness of models and generalization of datasets, multiple datasets are utilized. The original/real dataset is obtained from many secondary schools in many provinces in Cambodia. Another dataset is a synthetic/artificial dataset that is synthesized from original datasets and many other educational benchmark datasets. The data consist of students' personal information, domestic or home factors, student or individual factors, and school factors. The proposed academic prediction can help in predicting students' final grades and performance levels so that the intervention can be implemented. The data were obtained using a questionnaire form. The questionnaire comprises 50 questions covering student's personal information, and the three main factors that affect students in the adolescent age as shown in Table II. Due to privacy, personal information is kept, and there are 43 input features to learning in this problem. The target of the modeling is the output or performance levels discretized based on the output score.

### *B. Data Preprocessing*

Data preprocessing is considered as the most boring task, yet important in data mining modeling. In data science community, it is an important task in data engineering which is carried out by a group of people since it concerns various data operations. Each operation is important since the quality of data relies on these tasks and it highly affects the ability of the proposed models. The tasks require the basic knowledge of exploratory data analysis (EDA) and statistical knowledge to clean and modify to derive high-quality data. Hence, we use R programming language as a tool for data processing and data modeling due to its richness in statistical modeling and ML.

TABLE II
THE PERFORMANCE METRICS ON ORIGINAL DATASET

| 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 |

### *C. Data Mining or Modeling: Feature Selection Methods*

Analysis of student's information, their learning behavior, and factors affecting students' academic performance is still a challenging task in educational institutes [22]. Many cognitive and non-cognitive factors affect the academic performance of children and adolescents. Several related domains weakly or highly influence on results and achievements of high school students. As dimensionality of domain expands, the number of features affects student performance increases. The analysis in this paper is to extract an optimal set of factors that are needed and sufficient to control the success of students and improve the prediction performance. The terminology that we defined for this optimal set is the dominant set. There are two main purposes for introducing an analysis of the dominant set in this study. The primary purpose is to enhance the quality of data with the analysis of the feature set. The high-quality data bring many advantages such as reducing computational cost, more understandable data, improve the performance of prediction models, and many more. The second purpose is to detect the student learning patterns and highly influencing factors that engage in students learning outcomes. By knowing their learning patterns and related factors, the right intervention and assistant can be implemented and adopted.

This section summarizes the feature selection method utilized for selecting the dominant set. Feature selection (FS) is one of the crucial methods in unsupervised learning. As the name indicated, FS is used to select the optimal subset of features in the preprocessing step. There are several FS approaches, however, it falls into three main categories: filter, wrapper, and hybrid or embed methods [23]. Wrapper and hybrid methods are more advanced methods for FS, however, it is computationally expensive to use when the dimension is too high [24]. The filter-based method is more preferably used due to its simplicity, lower cost, and highly effective in many cases of applications. The filter-based FS algorithm works independently of classifiers and more scalable than the two previous methods [25]. These characteristics increase its popularity among many dimensional reduction methods. This study introduces a comparative study of four existing FS methods and a developed FS method on each proposed prediction model.

#### 1. Information Gain (IG)

IG is a decision tree-based FS using entropy to split the level of importance of features. It is commonly used in many areas of application due to its simplicity and interpretability [15]-[17]. IG computes the relevance of an attribute or feature by splitting the training samples concerning the target or classes. The type of algorithm utilizes the entropy and information theory introduced by Shannon to rank the importance or level of relevance of feature set [24].

#### 2. Symmetrical Uncertainty (SU)

SU is a filter-based FS method that is widely used for selecting optimal subset in big data platforms [15]-[17]. Similar to the IG algorithm, SU works by determining the correlation of attributes to target variables or classes using entropy and information theory.

#### 3. Chi-Square (CHI)

CHI has been widely used in many research works for selecting the features of discrete type [24]. Datasets with data types of integer and categorical (ordinal or nominal) require this type of method to measure its relevance and level of importance. It is known as one of one the famous statistical tests used in statistical analysis and ML. As the name indicated, the method utilizes the CHI test to rank the importance of features [25].

#### 4. Mutual Information (MI)

Similar to IG and SU, MI uses the concept of information theory to calculate or measure the dependency between the input variable and the target variable [26]. While most of the FS methods can handle only a linear relationship between variables, MI is a symmetric measurement that extent its ability to handle non-linear relationships between two random variables. This increases the popularity of this algorithm.

#### 5. The Proposed FS Method (MICHI)

IG, SU, and MI are the commonly used filter-based FS methods used in many applications. Algorithms utilize the concepts of MI and information theory to compute the relevance and level of importance of feature sets [26], [27]. MI computes the statistical dependence between two variables and is the name given to IG when applied to variable selection. CHI is considered as one of the robust FS methods when applying to categorical data [25]. From a literature study, MI and CHI are the two effective methods and best fit with our dataset, however, trusting on the combined feature is better than on a single algorithm. To accomplish this, we proposed a method to combine these two FS method based on feature rank score.

The algorithm is called MICHI which is named based on the combination of MI and CHI algorithms using their vector scores. However, different algorithms utilized different concepts in different processes and result in different scales of feature score. Hence, first of all, we normalize the scores of the

Open Science Index, Computer and Information Engineering Vol:15, No:2, 2021 publications.waset.org/10011857.pdf

MI and CHI using (1):

$$\overline{MI} = \frac{MI_i - MI_{min}}{MI_{max} - MI_{min}}, \overline{CHI} = \frac{CHI_i - CHI_{min}}{CHI_{max} - CHI_{min}} \quad (1)$$

Secondly, we combine the normalized score as a vector that contains information on both MI and CHI algorithms. The vector of the combined score is indicated in (2):

$$MICHI = \begin{pmatrix} \overline{MI} \\ \overline{CHI} \end{pmatrix} \quad (2)$$

The vector is used to rearrange the new feature score from MI and CHI scores accordingly by computing the Euclidean norm of vector as in (3):

$$|MICHI_i| = \sqrt{\left(\overline{MI_i}\right)^2 + \left(\overline{CHI_i}\right)^2} \quad (3)$$

This means that the score of a feature in the CHIMI algorithm contains the vector of scores that are generated by the CHI and MI algorithms. The new feature rearranges the order of importance of feature, feature with bigger value of $|MICHI_i|$ will be ranked higher. Unlike other previous methods of combining scores from different techniques such as “AND” and “OR”, our approach gives a true metric on the space for score vector [28]. Some experimental results in earlier works reported a minor improvement or no improvement in classification performance when more than three FS methods were combined [29]. This method conducts a mathematical structure for detecting the vector space of combined scores.

The normalization of CHI and MI scores is to introduce a rank of input features based on the computed scores. This method may place the input features within their true rank and improves the higher possibility of certain significant features to being identified for selecting the dominant feature set.

### *D. Data Mining or Modeling: Classification Algorithms*

Various EDM methods obtained from the literature review in [14]-[20] were considered. The comparative study of prediction models on predicting student performance was conducted in [30]. The improvement version of the comparative study was conducted in [31]. The experimental results of both works indicated that k-nearest neighbor (KNN), C5.0 of decision tree, and RF are the optimal classifiers. The models were further developed in our earlier works [32], [33]. The analysis in this study is adopted using four classifiers:

#### 1. K-Nearest Neighbor (KNN)

KNN is one among the effective methods using in predicting academic performance. In [31], KNN is in the top four prediction models that generate high classification results and can be alternatively used instead of other optimal classifiers. However, the classification in KNN is sensitive with quality of data. Its classification ability is highly affected by the quality of the training data. Noisy and irrelevant features, as well as outliers and overlaps between data regions of different classes, lead to less accurate classification.

#### 2. Hybrid C5.0 and Hybrid RF

Hybrid C5.0 and Hybrid RF are developed models studied in our earlier work [32]. The study proposed the development of the optimal models obtained in the earlier studies [30], [31]. Four baseline models: support vector machine (SVM), the tree-based C5.0, RF, and the naïve Bayes (NB) were proposed. We combined these four models with the principal component analysis (PCA) and validated the models with k-fold cross validation of cross validation technique.

#### 3. Improved Deep Belief Network (IDBN)

The IDBN is the optimization version of deep belief network (DBN) model. In our previous work, we gave a study of an optimization approach of DBN concerning (i) FS method, (ii) optimization of hyper-parameter, and (ii) regularization method [33]. The proposed IDBN successfully achieves the high prediction performance when applying with larger datasets.

### *E. Model Evaluation Metrics*

As mentioned in the CRISP-DM process, the data mining model needs to be reviewed and evaluated before the final deployment of the models. In this study, we utilized two standard model evaluation metrics. These two metrics or measurements are Accuracy and Root Mean Square Error.

TABLE III
GRAPHICAL OF CONFUSION MATRIX

| | | 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 | $E_{41}$ | $E_{42}$ | $E_{43}$ | $TP_4$ |

Accuracy (ACC) is the leading metric in classification problems used for evaluating the rate or percentage of correct prediction. In Table III, we denote TP as the number of correct predictions and denote E as the error or incorrect predictions. Hence, the value of ACC can be calculated by using (4).

$$ACC = \frac{\text{Correctly predicted values}}{\text{Total values}} = \frac{\sum TP_i}{\sum TP_i + \sum E_{ij}} \quad (4)$$

The target in this classification is the performance levels of students categorized orderly based on the final score. Root mean squared error (RMSE) is utilized to predict prediction error in predicting student performance levels. The levels of student performance are categorized into four levels: high risk (HR) that need high intervention, medium risk (MR) need medium intervention, low risk (LR) that need less requirement of intervention, and no risk (NR) that no need intervention. It is represented by 1, 2, 3, and 4, respectively. Using a confusion matrix in Table II, RMSE can be computed using (5):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^a - y_i^p)^2} \qquad (5)$$

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.

## V. Experimental Results of Prediction Models

This section reported the performance evaluation of FS methods with optimal models. We executed the proposed optimal classifiers using subsets that were obtained from each FS method. The framework of the study is illustrated in Fig. 3.

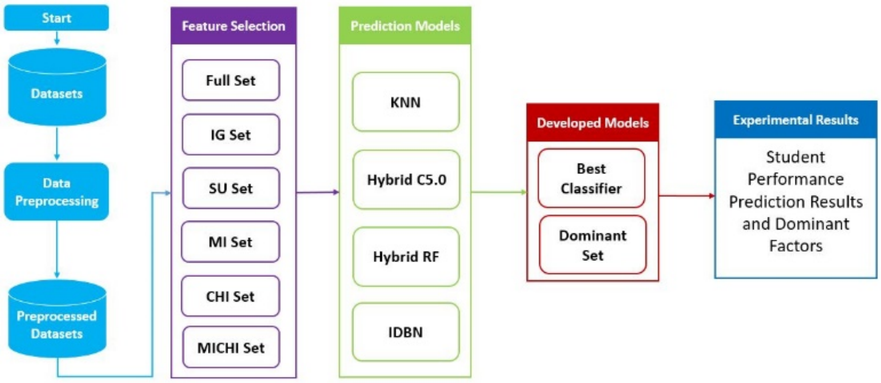


Fig. 3 Flowchart for the experiment in this study

The experiment was carried out with two phases. The first experiment was executed with the dataset ADS comprised of 1204 samples. The second experiment was with dataset GDS4 comprising 10000 samples. The second experiment was carried out with the context of a larger dataset to confirm the performance of IDBN and other proposed models. The experiment was made a minimal subset of 5 features to a fully selected set to extract the dominant set. To evaluate and compare the performance of prediction models, ACC and RMSE are measured. Recall that the higher value of ACC, the better model is. In contrast to ACC, the smaller value of RMSE, the better model is.

### A. Experimental Results with ADS

This section gives a comparative result of the four proposed classifiers on the feature set of each FS method with dataset ADS (1204 samples). Table IV illustrates the experimental results of the ACC and RMSE using the original/full dataset. Tables V-IX show the performance of the classifiers on subsets selected by five FS methods: IG, SU, CHI, MI, and the MICHI. The dominant set of each FS algorithm is observed and analysis is determined and compared.

TABLE IV
THE PERFORMANCE METRICS ON ORIGINAL DATASET

| Proposed Models | KNN | Hybrid C5.0 | Hybrid RF | IDBN |
|---|---|---|---|---|
| ACC (%) | 94.95 | 99.25 | 99.72 | 83.14 |
| RMSE | 0.261 | 0.073 | 0.041 | 0.759 |

TABLE V
PERFORMANCE METRICS USING SUBSETS FROM IG (29 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 95.35 | 0.257 | 5 | 97.35 | 0.163 |
| Hybrid C5.0 | 99.81 | 0.049 | 28 | 99.85 | 0.045 |
| Hybrid RF | 99.89 | 0.033 | 28 | 99.89 | 0.033 |
| IDBN | 86.55 | 0.571 | 28 | 86.55 | 0.571 |

The results presented in Table III demonstrate the performance of the four classifiers using datasets from IG FS. The performance of Hybrid C5.0 and Hybrid RF are comparatively better than other models.

TABLE VI
PERFORMANCE METRICS USING SUBSETS FROM SU (29 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 95.35 | 0.257 | 5 | 97.33 | 0.164 |
| Hybrid C5.0 | 99.81 | 0.049 | 28 | 99.85 | 0.045 |
| Hybrid RF | 99.89 | 0.033 | 28 | 99.87 | 0.033 |
| IDBN | 86.55 | 0.571 | 28 | 86.54 | 0.575 |

TABLE VII
PERFORMANCE METRICS USING SUBSETS FROM CHI (29 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 95.43 | 0.249 | 6 | 98.35 | 0.163 |
| Hybrid C5.0 | 99.85 | 0.041 | 29 | 99.85 | 0.041 |
| Hybrid RF | 99.95 | 0.015 | 29 | 99.95 | 0.015 |
| IDBN | 86.67 | 0.563 | 29 | 86.67 | 0.563 |

TABLE VIII
PERFORMANCE METRICS USING SUBSETS FROM MI (32 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 95.68 | 0.241 | 7 | 98.94 | 0.077 |
| Hybrid C5.0 | 99.89 | 0.035 | 32 | 99.89 | 0.035 |
| Hybrid RF | 99.97 | 0.012 | 32 | 99.97 | 0.012 |
| IDBN | 87.01 | 0.545 | 32 | 87.01 | 0.545 |

The results presented in Tables V and VI demonstrate the performance of the four classifiers using datasets from IG and SU algorithms. The performance of the two methods produces similar results. In both selected sets and dominant sets, Hybrid C5.0 and Hybrid RF are the two best models that outperformance the results for both RMSE and ACC.

TABLE IX
PERFORMANCE METRICS USING SUBSETS FROM MICHI (32 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 96.25 | 0.179 | 6 | 99.62 | 0.063 |
| Hybrid C5.0 | 99.90 | 0.033 | 31 | 99.90 | 0.033 |
| Hybrid RF | 99.97 | 0.012 | 31 | 99.98 | 0.011 |
| IDBN | 87.11 | 0.542 | 31 | 87.32 | 0.534 |

Table VII indicates the performance of the four classifiers on feature sets selected by the CHI algorithm. The performance of the KNN is significantly improved when using feature subset ranked by the CHI comparing to IG and SU. Hybrid RF stands out the other models and is followed by Hybrid C5.0. The dominant set of these two models are its full selected set.

Table VIII presents the experimental results of ACC and RMSE on feature subsets selected by the MI algorithm. The KNN generates the best result with the small feature set of the best 7 features. The dominant sets of Hybrid C5.0, Hybrid RF, and IDBN are their full selected set. Hybrid RF is still the best classifier in this classification problem when using MI's dataset.

The performance of four classifiers on the proposed FS method is illustrated in Table IX. KNN produces a higher result in the dominant set selected by MICHI comparing the other FS methods. The feature set rearranged by MICHI improves the performance of Hybrid C5.0, Hybrid RF, and IDBN. Hybrid RF stands out the rest models with respect to both ACC and RMSE.

### *B. Experimental Results with GDS4*

This section gives a comparative result of the four proposed classifiers on the feature set of each FS method with dataset GDS4 (10000 samples). Table X indicates the experimental results of the proposed classifiers with the original dataset. Tables XI-XV present the results of the four classifiers on feature subsets selected by IG, SU, CHI, MI, and MICHI, respectively. The dominant set is found by searching from the subsets of selected that produce equal or better classification results. The analysis of the combination of the FS method and classifiers are studied and compared for confirming the best classifier in our prediction problem.

TABLE X
THE PERFORMANCE METRICS ON ORIGINAL DATASET

| Proposed Models | KNN | Hybrid C5.0 | Hybrid RF | IDBN |
|---|---|---|---|---|
| ACC (%) | 95.12 | 98.55 | 98.88 | 97.01 |
| RMSE | 0.193 | 0.163 | 0.161 | 0.195 |

TABLE XI
PERFORMANCE METRICS USING SUBSETS FROM IG (29 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 95.50 | 0.257 | 5 | 98.30 | 0.122 |
| Hybrid C5.0 | 99.61 | 0.059 | 28 | 99.65 | 0.055 |
| Hybrid RF | 99.73 | 0.052 | 28 | 99.79 | 0.049 |
| IDBN | 99.65 | 0.052 | 28 | 99.67 | 0.050 |

TABLE XII
PERFORMANCE METRICS USING SUBSETS FROM SU (29 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 95.50 | 0.257 | 5 | 98.25 | 0.129 |
| Hybrid C5.0 | 99.61 | 0.059 | 28 | 99.63 | 0.057 |
| Hybrid RF | 99.73 | 0.052 | 28 | 99.75 | 0.050 |
| IDBN | 99.65 | 0.052 | 28 | 99.67 | 0.050 |

TABLE XIII
PERFORMANCE METRICS USING SUBSETS FROM CHI (29 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 95.51 | 0.262 | 6 | 98.95 | 0.089 |
| Hybrid C5.0 | 99.71 | 0.047 | 29 | 99.71 | 0.047 |
| Hybrid RF | 99.82 | 0.047 | 29 | 99.82 | 0.047 |
| IDBN | 99.77 | 0.047 | 29 | 99.77 | 0.047 |

TABLE XIV
PERFORMANCE METRICS USING SUBSETS FROM MI (32 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 95.65 | 0.243 | 7 | 99.52 | 0.062 |
| Hybrid C5.0 | 99.75 | 0.045 | 32 | 99.75 | 0.045 |
| Hybrid RF | 99.84 | 0.045 | 32 | 99.84 | 0.045 |
| IDBN | 99.81 | 0.045 | 32 | 99.81 | 0.045 |

TABLE XV
PERFORMANCE METRICS USING SUBSETS FROM MICHI (32 FEATURES)

| Models | Selected set | | Dominant set | | |
|---|---|---|---|---|---|
| | ACC | RMSE | N | ACC | RMSE |
| KNN | 96.35 | 0.175 | 6 | 99.67 | 0.059 |
| Hybrid C5.0 | 99.75 | 0.045 | 31 | 99.75 | 0.045 |
| Hybrid RF | 99.85 | 0.043 | 31 | 99.85 | 0.043 |
| IDBN | 99.83 | 0.044 | 31 | 99.83 | 0.044 |

Table X indicates that the experiment of Hybrid RF and Hybrid C5.0 with the original dataset gives the best results. The two developed tree-based models generate the highest ACC and lowest RMSE.

The results presented in Tables XI and XII demonstrate the performance of the four classifiers using datasets from IG and SU algorithms. The two algorithms in dataset GDS4 produce similar results. The performance of the two methods produces similar results. In both selected sets and dominant sets, Hybrid C5.0 and Hybrid RF are the two best models that generated the optimal results concerning both RMSE and ACC.

Table XIII presents the performance of the four models on the set selected by the CHI methods. The performance of KNN is confirmed to be improved when applying the dominant sets containing the best 6 features. Hybrid RF and IDBN produces are found to produce the most classification result. The dominant set of CHI algorithm significantly improves the performance of proposed models.

Table XIV presents the experimental results of ACC and RMSE on feature subsets selected by the MI algorithm. Hybrid RF and IDBN outperform the other models when using the selected set. However, the performance of the IDBN and Hybrid C5.0 are comparatively improved when considering the dominant sets.

Table XV demonstrates the performance of the proposed classifiers with the input feature subsets from the proposed MICHI method. The performance is significantly improved when using the dominant sets. Hybrid RF and IDBN are comparatively better than other models.

### *C. Summary and Discussion*

This study aims to boost up the performance of the proposed classifiers to reach the most classification results. Hence, the optimal models are then combined with dominant sets, which is believed to significantly improve the performance of prediction models and selected the highly influencing factors in academic performance.

From Tables IV-XV, we can compare the performance of KNN, Hybrid C50, Hybrid RF, and IDBN on dominant sets of IG, SU, CHI, MI, and CHIMI methods. Figs. 4 and 5 summarize the value of ACC and RMSE of each classifier on each FS method on data ADS and GDS4, respectively.

Fig. 4 represents the values of ACC and RMSE of the four classifiers with the five FS methods using the ADS dataset. Concerning prediction models, Hybrid C5.0 and Hybrid RF are comparatively better than the other two models. Regarding the FS methods, the performance of IG and SU methods are not statistically different. The performance of CHI and MI methods stands out the performance of IG and SU. The figure reports that the proposed CHIMI successfully improves the performance of the four prediction models and standout the performance of the four FS methods.

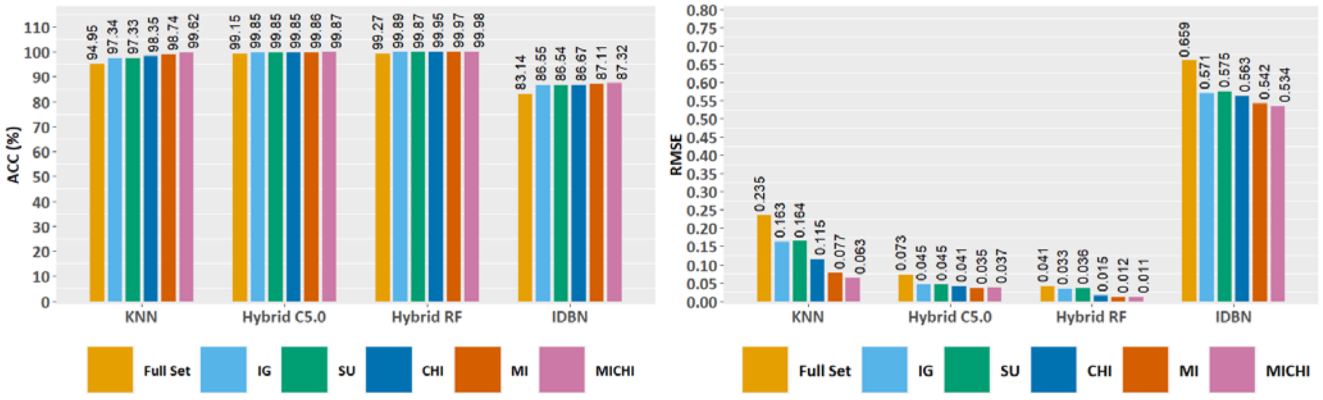


Fig. 4 ACC and RMSE comparison using ADS dataset

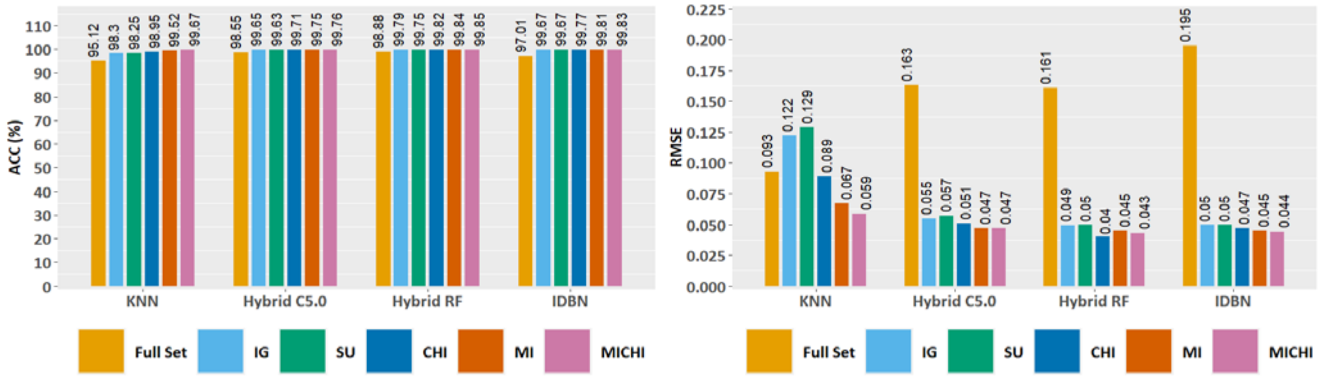


Fig. 5 ACC and RMSE comparison using GDS4 datasets

Fig. 5 graphically demonstrates the performance of the four classifiers with five FS methods using the GDS4 dataset. In the context of a larger dataset, the performance of IDBN is significantly improved. The performance of IDBN and Hybrid are not statistically different and the two models stands out the performance of Hybrid C5.0 and KNN. The proposed CHIMI plays its role as the best FS method in selecting the dominant factors for improving the performance of the prediction models. Hence, the experimental results indicate that Hybrid RF is the best prediction model in this prediction. The model achieves the most successful classification results. However, in the context of a larger dataset, the proposed IDBN can be alternatively used as an optimal model in predicting academic performance.

## VI. Model Deployment: Design and Implementation of the APPS

The final step of the CRISP-DM model is the deployment of the model. The proposed models in the previous section are integrated into a system for predicting student performance. To answer our last research question regarding the applicability contribution or implication of our study, we design a web-based application for educational stakeholders for predicting student performance. Since our experiment was carried out in the R language, we design the web application using a Shiny App. Shiny App is a web application developed by Shiny package, Shiny dashboard package, and some other related packages. The packages contained the concepts and the built-in framework of HTML, CSS, and JavaScript.

The basic architecture of our APPS is indicated in Fig. 6. Educational stakeholders can access our given web-based application from their individual workplace via the internet. The first section in the user interface (UI) is the introductory part about the background of the study, Mind map, student performance levels, and a sample of descriptive statistics of educational data using in our prediction. The prototype of the UI of the APPS is shown in Fig. 7.

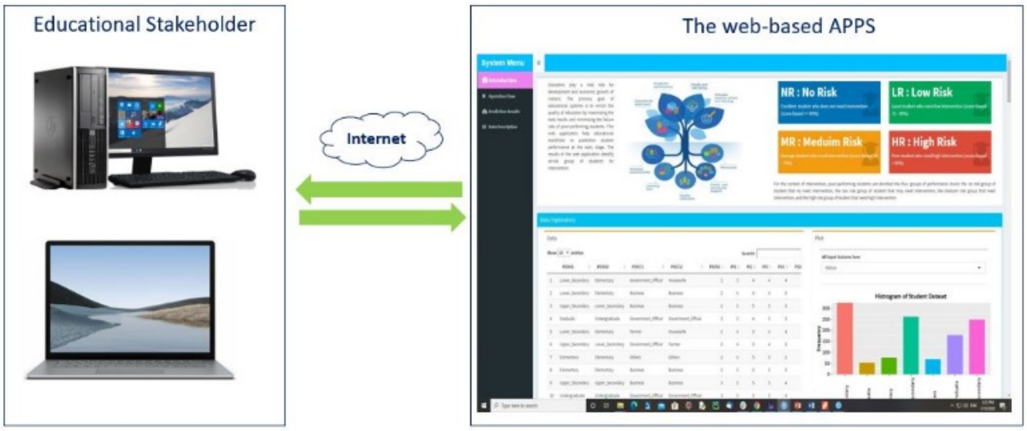
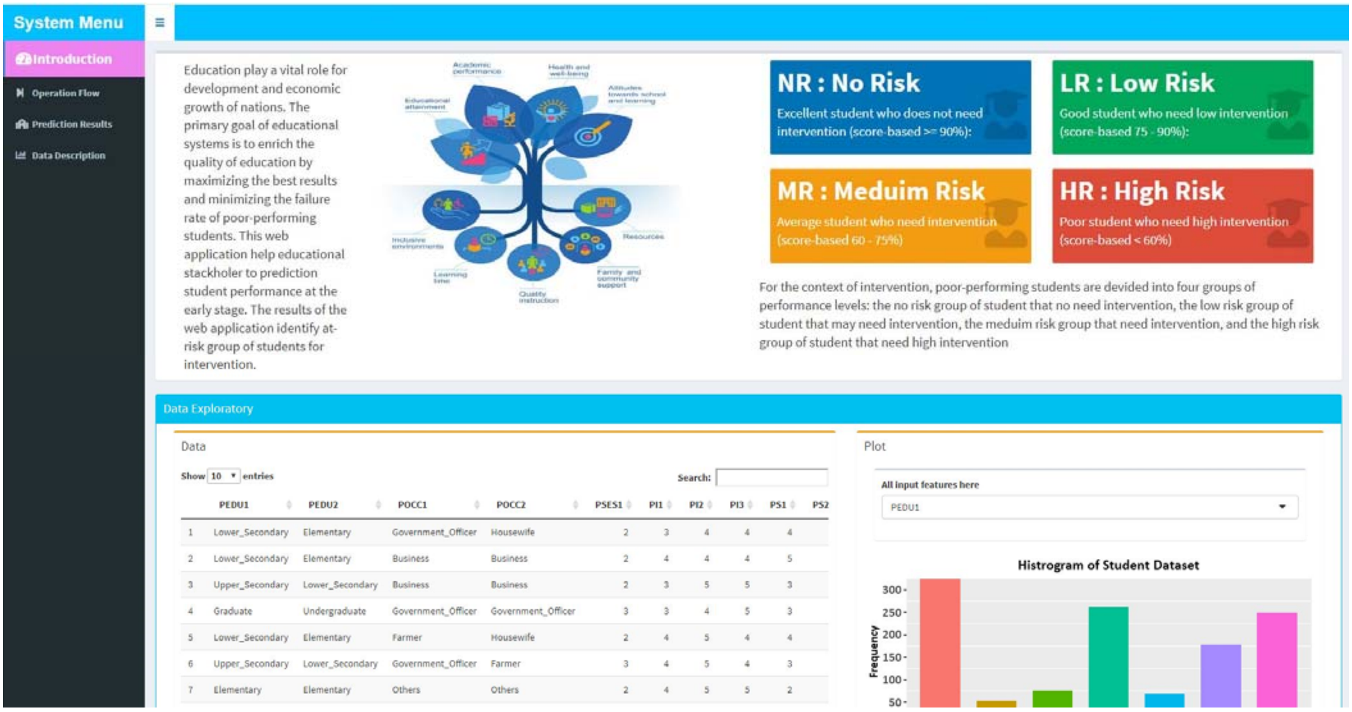


Fig. 6 APPS architecture



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 context of intervention, poor-performing students are devided 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 meduim risk group that need intervention, and the high risk group of student that need high intervention

Data Exploratory

Data

Show 10 entries

Search:

| | PEDU1 | PEDU2 | POCC1 | POCC2 | PSES1 | PI1 | PI2 | PI3 | PS1 | PS2 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Lower_Secondary | Elementary | Government_Officer | Housewife | 2 | 3 | 4 | 4 | 4 | |
| 2 | Lower_Secondary | Elementary | Business | Business | 2 | 4 | 4 | 4 | 5 | |
| 3 | Upper_Secondary | Lower_Secondary | Business | Business | 2 | 3 | 5 | 5 | 3 | |
| 4 | Graduate | Undergraduate | Government_Officer | Government_Officer | 3 | 3 | 4 | 5 | 3 | |
| 5 | Lower_Secondary | Elementary | Farmer | Housewife | 2 | 4 | 5 | 4 | 4 | |
| 6 | Upper_Secondary | Lower_Secondary | Government_Officer | Farmer | 3 | 4 | 5 | 4 | 3 | |
| 7 | Elementary | Elementary | Others | Others | 2 | 4 | 5 | 5 | 2 | |

Plot

All input features here

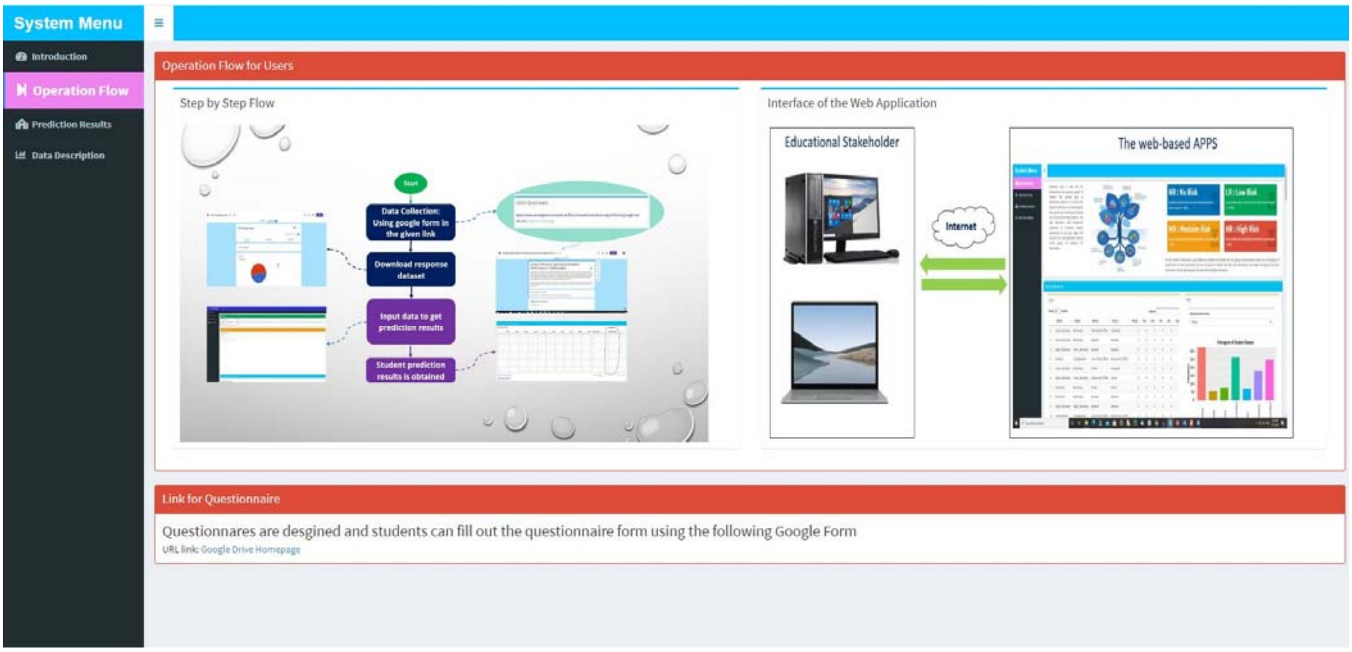
PEDU1

Fig. 7 A prototype of the Introduce interface of the APPS

The prototype illustrated in Fig. 8 presents the operation flow of the process in achieving the final result of academic prediction. The flow chart gives introductory details of each step in using this APPS. Once, educational stakeholders get access to the system, they can see the introduction part and link that can assess an existing questionnaire. The questionnaire was designed properly using Google form. They subsequently can share it with their students to fill in and get back in the predefined deadline since it takes only a few minutes to complete. Educational stakeholders finally can get the data in the right format and input into the system (*File Upload* button).

The statistics of students learning patterns and informative features are interactively displayed in Fig. 9. Teachers and policymakers can view the data to understand their students' learning patterns, related factors that highly affect their students' learning outcomes. Accurate prediction results are stored in Fig.

10. At the bottom of the interface shown in Fig. 10, there is a button where users can download these results and save these data in the CSV format for further usage.



Operation Flow for Users

Step by Step Flow

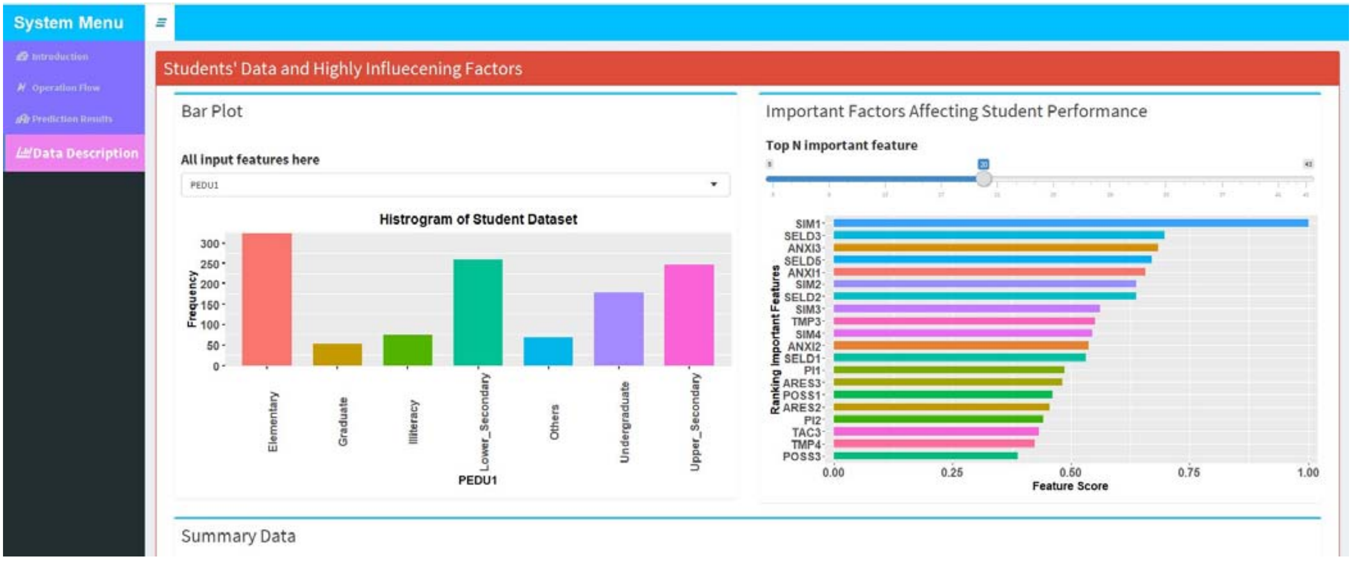
Interface of the Web Application

Link for Questionnaire

Questionnares are desgined and students can fill out the questionnaire form using the following Google Form

URL link: Google Drive Homepage

Fig. 8 Operation flowchart to instruct users of the overall process to get prediction results



Students' Data and Highly Influecening Factors

Bar Plot

All input features here

PEDU1

Important Factors Affecting Student Performance

Top N important feature

20

Summary Data

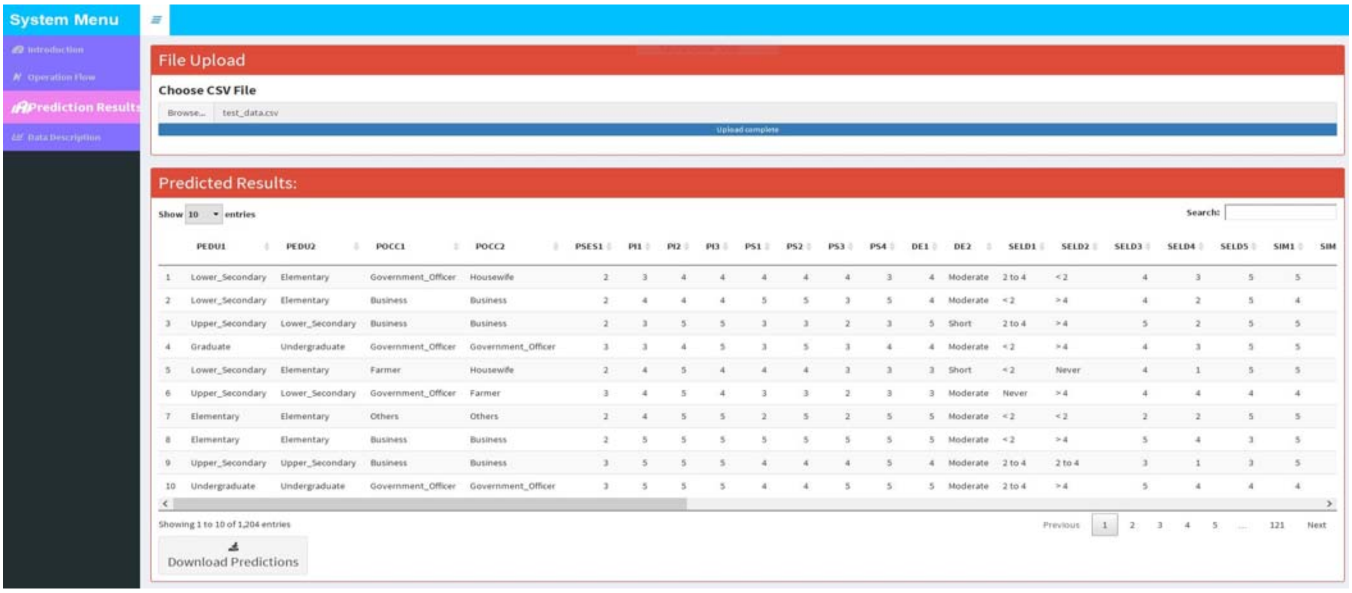
Fig. 9 Summary of the dominant factors and ranking the highly influencing factors

*A. The Deployment of the APPS*

The APPS is a web-based application that can be easily shared or distributed to users. The users have two options to get this application to use. The system is distributed in a web page or browser stored in the Shinyapp.io. By clicking on the given link, educational stakeholders can get the application in a web browser. This is the most commonly used method since users can use their knowledge in a programming language. They can navigate to the APPS from their individual workplace via the internet in any web browser. Additionally, if the users have knowledge in using programming languages and web development, they access to the GitHub via the given link to get R scripts. From this option, they can modify the interface of the APPS or model coding as they wish.

*B. The System Evaluation*

To evaluate the performance of the designed APPS, we designed a subjective questionnaire for educational stakeholders to evaluate the properties and characteristics of the data. The questionnaire consists of ten questions describe ten characteristics of the APPS: useful, motivating, user-friendly, relevant, reliable, efficient, organized, time cost, adaptable, and sophisticated [34]. For giving opinion or evaluation, the question of a 5-point Likert scale type is designed. The possible answer is the ordinal numbers 1, 2, 3, 4, and 5, which represent strongly disagree, disagree, neutral, agree, and strongly agree, respectively. The survey consists of 67 participants including 57 high school students and 10 high school teachers. The outcome of the survey is summarized in Table XVI. When writing 11 (1) mean 11 students and 1 teacher agree with the given statement.



File Upload

Choose CSV File

Browse... test_data.csv

Upload complete

Predicted Results:

Show 10 entries

Search:

| | PEDU1 | PEDU2 | POCC1 | POCC2 | PSES1 | PI1 | PI2 | PI3 | PS1 | PS2 | PS3 | PS4 | DE1 | DE2 | SELD1 | SELD2 | SELD3 | SELD4 | SELD5 | SIM1 | SIM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Lower_Secondary | Elementary | Government_Officer | Housewife | 2 | 3 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | Moderate | 2 to 4 | <2 | 4 | 3 | 5 | 5 | |
| 2 | Lower_Secondary | Elementary | Business | Business | 2 | 4 | 4 | 4 | 5 | 5 | 3 | 5 | 4 | Moderate | <2 | >4 | 4 | 2 | 5 | 4 | |
| 3 | Upper_Secondary | Lower_Secondary | Business | Business | 2 | 3 | 5 | 5 | 3 | 3 | 2 | 3 | 5 | Short | 2 to 4 | >4 | 5 | 2 | 5 | 5 | |
| 4 | Graduate | Undergraduate | Government_Officer | Government_Officer | 3 | 3 | 4 | 5 | 3 | 5 | 3 | 4 | 4 | Moderate | <2 | >4 | 4 | 3 | 5 | 5 | |
| 5 | Lower_Secondary | Elementary | Farmer | Housewife | 2 | 4 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | Short | <2 | Never | 4 | 1 | 5 | 5 | |
| 6 | Upper_Secondary | Lower_Secondary | Government_Officer | Farmer | 3 | 4 | 5 | 4 | 3 | 3 | 2 | 3 | 3 | Moderate | Never | >4 | 4 | 4 | 4 | 4 | |
| 7 | Elementary | Elementary | Others | Others | 2 | 4 | 5 | 5 | 2 | 5 | 2 | 5 | 5 | Moderate | <2 | <2 | 2 | 2 | 5 | 5 | |
| 8 | Elementary | Elementary | Business | Business | 2 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | Moderate | <2 | >4 | 5 | 4 | 3 | 5 | |
| 9 | Upper_Secondary | Upper_Secondary | Business | Business | 3 | 5 | 5 | 5 | 4 | 4 | 4 | 5 | 4 | Moderate | 2 to 4 | 2 to 4 | 3 | 1 | 3 | 5 | |
| 10 | Undergraduate | Undergraduate | Government_Officer | Government_Officer | 3 | 5 | 5 | 5 | 4 | 4 | 5 | 5 | 5 | Moderate | 2 to 4 | >4 | 5 | 4 | 4 | 4 | |

Showing 1 to 10 of 1,204 entries

Previous 1 2 3 4 5 ... 121 Next

Download Predictions

Fig. 10 Prediction results identifying the at-risk levels of poor-performing students

TABLE XVI
SURVEY RESULTS FOR REVALUATING ACADEMIC PREDICTION SYSTEMS (ADAPTED FROM [34])

| 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) |

63% of participants are male (83% are students and 17% are teachers) and 37% are female (88% are students and 12% are teachers). The analysis gives the positive opinion on the system. Most of the participants reply positively (agree and strongly agree) on the characteristics of the APPS. The survey result reported that 82.08% of the participants agree that the system is useful (55.22% agree, 26.86% strongly agree), 83.58% supported that the APPS is motivating (62.68% agree, 20.89% strongly agree), 91.04% stated that the interface of the system is friendly (82.08%agree, 8.95% strongly agree), 85.07% agreed that the information show in the system is relevant (58.20% agree, 14.92% strongly agree), and 73.13% of participants believed that the system is reliable (62.68% agree, 20.89% strongly agree). In addition, 82.08% appreciated with efficient of the APPS (55.12% agree, 26.86% strongly agree), 74.62% stated that the system was well-organized (58.78% agree, 20.89% strongly agree), 91.04% reported about that speed (time cost) of the system was fast (64.17% agree, 26.86% strongly agree). Among that, 88.05% of support about the adaptability of the system (62.68% agree, 20.89% strongly agree), and 92.58% admired about the sophistication of the system (77.61% agree, 14.92% strongly agree). The analysis of evaluation is shown in Fig. 11. The overall results conclude the effectiveness and usefulness of the system.

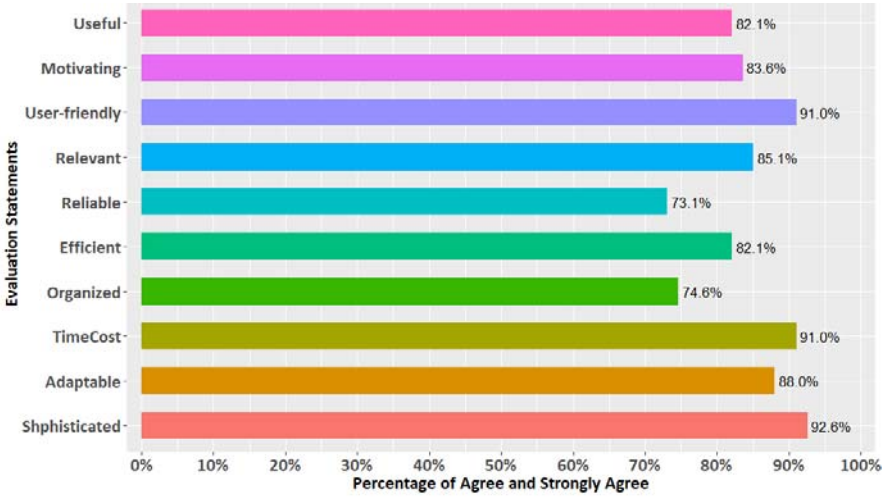


Fig. 11 Users feedback rating the characteristics of the APPS

## VII. Conclusion

This work aimed to develop more accurate prediction models of EDM and their implication by designing a web-based academic prediction system for educational stakeholders. The study proposed the hybrid ML model and optimization of deep belief networks. Hybrid RF and the improved deep belief network (IDBN) generate the most accurate prediction. Simultaneously, we developed an FS method, MICHI, to improve the performance of the classifiers and select the dominant factors for academic performance prediction.

The developed prediction models are integrated in a designed web-based application, called APPS. The system is designed for educational stakeholders and related individuals to give prediction of their students at the early stage for intervention.

Our findings confirm the effectiveness of the prediction model and the usability of the APPS. The developed APPS is reported to be useful for educational stakeholders and related individual to accurately predict their students performance. The system also gives the details of students learning patterns and highly influencing factors. Hence, teachers and related individuals can adapt their learning methodlogy, set up the right intervention and policy to improve academic performance.

## Acknowledgment

The authors would like to give big thanks to their teachers and students in high schools in Cambodia, MoEYS (Minisitry of education, youth, and sport in Cambodia) for mentally and physically support in this work.

## References

[1] Jonhson, E. S., Smith, L. and Harris, M. L. How RTI works in secondary school. Thousand Oaks, CA, 2009.

[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] 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.

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

[5] Ministry of Education, Youth, and Sport (MoEYS). Education in Cambodia: Finding from Cambodia's Experience in PISA for Development. Phnom Penh: Author, 2018.

[6] MoEYS, Policies on Science, Technology, and Innovation, 2020-2030. Phnom Penh, Cambodia, 2019.

[7] Barnes, T., Dessmaris, M., Romero, C., & Ventura, S. (2009, July 1-3). Educational data mining 2009. *Proceedings of the 2nd International Conference on Educational Data Mining*, Cordoba, Spain.

[8] 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.

[9] 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.

[10] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expertise Systems with Application*, 33(1), 135–146, 2006.

[11] 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.

[12] 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.

[13] C. Romero, C. and S. Ventura. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 40(6), e1355, 2020.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," J Data Warehousing (2000), vol. 5, pp. 13—22, 2020.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] C. F. Tsai and Y. C. Hsiao, "Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches," Decis Support System, vol. 50, no. 1, 258-269, 2010.

[29] A. Thubaity, N. Abanumay, S. Al-Jerayyed, A. Alrukban, Z. Mannaa, "The effect of combining different feature selection methods on Arabic text classification," IEEE: The 14th ACIS International Conference Software Engineering, Artificial Intelligent, Networking and Parallel/distributed Computing (SNPD), 211-216.

[30] 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.

[31] 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.

[32] 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.

[33] 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.*

[34] P. Sokkhey and T. Okazaki, "Developing web-based support system for predicting poor-performing students using educational data mining techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, pp. 23–32, 2020.