International Journal of Educational Methodology
Volume 9, Issue 2, 297 - 307.
ISSN: 2469-9632
https://www.ijem.com/

Predictive Model for Clustering Learning Outcomes Affected by COVID-19 Using Ensemble Learning Techniques

Wongpanya Sararat Nuankaew
Rajabhat Maha Sarakham University, THAILAND

Pratya Nuankaew*
University of Phayao, THAILAND

Received: November 6, 2022 ▪ Revised: December 30, 2022 ▪ Accepted: February 22, 2023

Abstract: The influence of COVID-19 has caused a sudden change in learning patterns. Therefore, this research studied the learning achievement modified by online learning patterns affected by COVID-19 at Rajabhat Maha Sarakham University. This research has three objectives. The first objective is to study the cluster of learning outcomes affected by COVID-19 at Rajabhat Maha Sarakham University. The second objective is to develop a predictive model using machine learning and data mining technique for clustering learning outcomes affected by COVID-19. The third objective is to evaluate the predictive model for clustering learning outcomes affected by COVID-19 at Rajabhat Maha Sarakham University. Data collection comprised 139 students from two courses selected by purposive sampling from the Faculty of Information Technology at the Rajabhat Maha Sarakham University during the academic year 2020-2021. Research tools include student educational information, machine learning model development, and data mining-based model performance testing. The research findings revealed the strengths of using educational data mining techniques for developing student relationships, which can effectively manage quality teaching and learning in online patterns. The model developed in the research has a high level of accuracy. Accordingly, the application of machine learning technology obviously supports and promotes learner quality development.

Keywords: *Educational data mining, learning achievement, learning analytics, online learning model, student model.*

To cite this article: Nuankaew, W. S., & Nuankaew, P. (2023). Predictive model for clustering learning outcomes affected by COVID-19 using ensemble learning techniques. *International Journal of Educational Methodology*, *9*(2), 297-307. https://doi.org/10.12973/ijem.9.2.297

Introduction

COVID-19 has had a massive impact on the education system as schools and universities are where many students, teachers, and parents are gathered. Therefore, during the severe epidemic COVID-19, the learning management model is used instead of online and distance learning (Abuhammad, 2020; Al-Kumaim et al., 2021; Appiah-Kubi & Annan, 2020). The tools used in the teaching and learning process need to rely on communication technology and Internet networks. In addition, learners' learning styles must be adapted to self-directed control and self-regulated learning theories to achieve expected outcomes (Kizilcec et al., 2017; Lim et al., 2020; Nuankaew, 2020). Instructor monitoring is achieved through the design of learning activities focused on practice rather than traditional lectures. The problem with distance learning is that it is difficult to monitor and control, and many factors limit assistance for developing learners' knowledge (Giani & Martone, 1998; Pozdnyakova & Pozdnyakov, 2017). The downside is that the effectiveness of distance learning is not comparable to the quality of conventional classroom learning. Moreover, distance learning can create more significant disparities, leading to two major problems (a) learning loss; and (b) dropping out of the education system.

Learning loss is defined as the loss of knowledge discovered or the loss of specific skills that learners should have developed during the study (de Oliveira et al., 2021; Viloria & Pineda Lezama, 2019). Furthermore, many students face changes in their lifestyles and their impact on learning and skill development, resulting in further exclusion from the education system. Coupled with the epidemic situation of COVID-19, household incomes have declined. Children from low-income families bear four times the burden of higher education costs compared to high-income families. The problem of children falling out of the education system also affects economic and social development. At the higher education level, it was also affected by the transformation of learning styles from classroom learning to online learning. Learners and tutors who previously handled face-to-face learning must be transformed and mastered by the distance learning system. The learning activity pattern must, therefore, be adjusted accordingly. Such problems appeared in the Faculty of Information Technology at the Rajabhat Maha Sarakham University, Maha Sarakham, Thailand.

* **Corresponding author:**
Pratya Nuankaew, School of Information and Communication Technology, University of Phayao, Thailand. ✉ pratya.nu@up.ac.th

Machine learning and data mining techniques were applied to hybrid clustering and prediction outcomes for finding the problem in the education system. It helps researchers to improve the educational process and learning outcomes of students (Xu et al., 2021). Ensemble learning techniques have been used to enhance the predicting model of them (Badal & Sungkur, 2022; Karalar et al., 2021; Nachouki & Abou Naaj, 2022; Smirani et al., 2022).

Given the apparent impact of the COVID-19 epidemic on education systems (Abdel-Basset et al., 2021; Abuhammad, 2020; Al-Kumaim et al., 221; Dechsupa et al., 2020), researchers are encouraged to engage in research to find ways to design learning models that can promote and enhance learning which the online learning model has been modified. Therefore, this research has three primary objectives. The first objective was to study the context for clustering learning outcomes affected by COVID-19 at Rajabhat Maha Sarakham University. The second objective was to develop predictive models using machine learning and data mining techniques for clustering learning outcomes affected by COVID-19. Finally, the third objective was to evaluate a predictive model for clustering learning outcomes affected by COVID-19 at Rajabhat Maha Sarakham University. The population and the research sample were selected by selecting a specific example with purposive sampling from two subjects with 139 students.

Research instruments are divided into two components: tools for developing a predictive model and tools for testing model performance. The tools for building a predictive model of (a) the base classifiers include the Naïve Bayes (NB), Decision Tree (DT), and k-Nearest Neighbor (k-NN) techniques; and (b) the Majority Voting Ensemble (MVE). The techniques for testing a predictive model include cross-validation and confusion matrix techniques. The metrics for evaluating model performance include accuracy, precision, recall, and f1-score values. With the results of this research project, the researchers expect to support the education industry to create educational models that support and find solutions to improve learner quality. Moreover, the researchers expect this research to be pushed into the field of artificial intelligence technology to study further.

Definitions

In this research, we have developed models that build the hybrid clustering and classification techniques. The purpose of defining definitions is to restrict boundaries and devise coherent perceptions. It consists of three definitions: clustering learning outcomes, predictive model, and ensemble majority voting.

1) Clustering Learning Outcomes

Clustering learning outcomes (CLO) is the study of the relationship between activities that students participate in the learning process and affect their learning achievement. It is common for the sum of learning scores to be instrumental in determining learning outcomes. However, the researchers found that the learning behavior of learners in the learning information technology field received high practical scores but low grades. This research, therefore, develops an appropriate cluster predictive model to identify learners with a chance of not meeting the learning criteria or having low learning outcomes.

2) Predictive Model

A prediction model is a process of developing a scientific model using causal and outcome principles. In addition, a prediction model is an application of statistical techniques and artificial intelligence technologies using machine learning and data mining to predict and forecast probable future outcomes using historical and available data. It works by analyzing current and past data and anticipating what is learned from the models created to indicate potential impacts (Nuankaew & Nuankaew, 2021). Hence, this research aims to develop a model to predict learning outcome clusters to design learning suitable for learners according to individual learning behaviors.

3) Ensemble Majority Voting

Ensemble Majority Voting is a tool used in many powerful modeling techniques because its main goal is to optimize the model. Popular essential methods include Bagging and Boosting techniques (Ghoggali et al., 2022). Bagging is the creation of multiple models of the same data set to perform tests on a subset of data divided from the aggregate data set and then combine the prediction results of the different models. This learning algorithm includes Decision Tree, Random Forest, and Extra Trees. Boosting is the creation of multiple models while using the same data source to perform loop iteration tests, adjusting the weights to improve the model's prediction results. Examples of this learning algorithm are AdaBoost and Stochastic Gradient Boosting. This research uses a bagging ensemble learning model by selecting from various efficient models to create the model.

Literature Review

Various research has been applied the machine learning and data mining techniques in the educational field. They are mostly used for clustering, prediction, association, and relationship. This research looks on educational data mining and literature review during the COVID-19 pandemic. Researchers presented different models to find and solve the problem to give better results. Nuankaew et al. (2022) presented an improving predictive model for the prevention of Thailand students dropping out. They compared several classifiers for the Majority Voting Ensemble. Badal and Sungkur (2022) developed a model to predict and analyze students' grades and engagement. In the data preprocessing stage, the Feature Encoding and the Synthetic Minority Over-Sampling Technique is used to manage the imbalanced classes and normalize the range of values of the data. They compared the base classifiers, ensemble learning, and deep learning. An average accuracy for grade prediction was 85.13% and engagement prediction was 83.88% in the Random Forests. Smirani et al. (2022) presented a model to improve the classification of student failure. The first level combined three ensemble learning including Random Forests, Extreme Gradient Boosting, and Light Gradient Boosted Machine for base learners, and the second level used Multilayer Perceptron (MLP). The best result of the student success rates was 98.86%, and dropout rate declined in class Nachouki and Abou Naaj (2022) studied Random Forests to improve predict the student performance. The Pearson Correlation was used to evaluate the relationship between two variables and all variables. It received variables suitable for building the model. They showed different results between an actual and predicted using paired sample t-Test. Karalar et al. (2021) proposed an ensemble learning model to predict the student performance at risk for failure. It was combined from Extra Trees, Random Forest and Logistic Regression that has greater efficiency than another model. The class labels of their models were defined by scores, grades, and GPA (Grade Point Average). There may be mistakes in the analysis of some specific characteristics of some groups of learners.

Nuankaew and Nuankaew (2021) applied the Elbow method to K-Determination and K-mean clustering to create a model of academic success. They were not to test the exactness k value of the number cluster between their chosen. Francis and Babu (2019) proposed models to predict academic performance of students. The first layer selected the features and classified to identify which of the features that gave the best perform using the base classifiers of the machine learning. K-means clustering plus majority voting method was used to cluster of the features in the second layer. Their proposal obtained the best outperformance. Their research applied a machine learning algorithm using a hybrid algorithm of clustering and classification, to find the best model.

They did not identify the correlation analysis method of the variables that were implemented in the model (Nuankaew et al., 2022; Nuankaew & Nuankaew, 2021; Smirani et al., 2022). It was an important step in data analysis for building and evaluating the model. Including a hybrid clustering and classification of machine earning for the data set has been a challenge in these studies. This research study approaches the clustering and prediction model for clustering learning outcomes affected by COVID-19. It will help to develop the performance of the model that used one method.

Methodology

Populations and Samples

The research population was students from the Faculty of Information Technology at Rajabhat Maha Sarakham University during the first semester of the academic year 2020. During this period, Thailand was affected by the COVID-19 epidemic, making it unable to provide regular learning management, so it offered pedagogy in an online format only.

Research samples are classified into two groups. The first group consisted of 73 students enrolled in the course 7000103: Mathematics and Statistics for Information Technology. The second group comprised 66 students enrolled in the course 7011303: Data Warehouse and Data Mining. Students participating in the research consented to use students' academic results, which consisted of six activity scores, a midterm exam score, and a final exam score. The teacher and students' activities in class included (1) teacher activities: presenting principles and theories of the course, question-answer, teaching how to use tools, coding, case study, suggesting solving problems of the learner, assigning exams, and assessing students' activity and exams; and (2) students' activity: learning and studying the content of each of the topics, analyzing and designing problems, coding, summarizing problems with each of the topics, presenting the project of the group, discussing and reflecting on the projects, midterm exam testing, and final exam testing.

A practical learning content for course 7000103 included three contents for mathematical analysis, which are mathematical logic, graphs, and search tree. Three contents for statistical analysis and design included hypothesis testing, using statistical tools for data analysis, and interpretation. The practical learning content for course 7011303 included analyzing and designing models, pre-processing data, building a data warehouse, clustering models, association rule mining, and classification models. These contents were analysis, decision support, and implementation of business applications.

The contents and activities during class may cause learners to become confused. Therefore, there were exercises to review and increase understanding of the content of each topic, and the teacher summarized and reviewed, and answered the exercises of this. The preliminary analysis and summary of the collected data are shown in Table 1.

Table 1. Data Collection

7000103: Mathematics and Statistics for Information Technology					
Scores	**Max**	**Min**	**Mean**	**Median**	**Mode**
Six Activities (30 scores)	30.00	0.00	26.15	30.00	30.00
Midterm Exam (30 scores)	29.00	2.00	15.17	15.50	7.00
Final Exam (40 scores)	38.50	4.00	27.96	28.50	24.00
Grades: A = 24, B+ = 8, B = 4, C+ = 6, C = 26, D+ = 2, F = 3					
7011303: Data Warehouse and Data Mining					
Scores	**Max**	**Min**	**Mean**	**Median**	**Mode**
Six Activities (30 scores)	30.00	10.00	24.66	27.00	30.00
Midterm Exam (30 scores)	28.00	2.40	20.23	21.90	23.40
Final Exam (40 scores)	39.00	5.00	33.47	32.50	32.50
Grades: A = 31, B+ = 5, B = 12, C+ = 2, C = 15, F = 1					

Table 1 shows the collected data that were analyzed and statistically summarized. The research found that most students had a high level of academic achievement in both courses. In the course 7000103: Mathematics and Statistics for Information Technology, 24 students received A grades, representing 32.88%. For course 7011303: Data Warehouse and Data Mining, 31 students received A grades, representing 46.97%.

Online learning does not seem to hinder learning management, but a small ratio of students achieved low learning outcomes and did not meet the learning criteria. It is, therefore, essential to develop this research to control and manage students to meet the higher learning criteria.

Research Instruments

We study machine learning and data mining theory, and design instruments to enable this research to achieve its objectives. Research instruments are divided into two parts: model development tools and model testing tools. The description of this will be presented below.

1) Model Development Tools

Model development tools are classified into two parts with different objectives. The first part is intended to cluster learners, and clustering tools are an unsupervised learning technique (Wang & Biljecki, 2022). The learner clustering tool in this research chose a k-means approach to group learning behaviors of similar learners. K-means uses dimensional spacing calculations (attributes) to use the distance obtained to measure each record's similarity. To calculate the length of each member according to the k-means principle, use "Euclidean Distance (Euclidean Metric)" for calculation. We used variables to analyze modeling for improving the skills of learners (critical thinking, creativity, collaboration, communication, and programing analysis) such as six activities, midterm exams, and final exams.

In part one, researchers have completed a published research process (Nuankaew & Nuankaew, 2021). The research concluded that the suitable cluster of the course 7000103: Mathematics and Statistics for Information Technology was k equal to 3. At the same time, the researchers found that the appropriate cluster for the course 7011303: Data Warehouse and Data Mining was also k equal to 3. The researchers applied this result to predict a suitable learning cluster, which presented the next part of the model development tool.

The second part is developing the predictive model for clustering learning outcomes. The tools used for model development in this section use supervised learning techniques. It consists of three techniques: Naïve Bayes (NB), Decision Tree (DT), and k-Nearest Neighbor (k-NN). All three techniques are popular predictive tools. The process for selecting and applying all three models is an ensemble learning by majority voting techniques, as presented in Fig. 1.

Step 1:
Build appropriate learner behavior clusters with k-Means for two courses.

Step 2:
Calculate the predictions' confidence in the individual learner for each classifier

	NB	DT	k-NN
Conf. Class 0	0.9904*	0.0002	0.0000
Conf. Class 1	0.0000	0.1213	0.7987*
Conf. Class 2	0.0096	0.8784*	0.2013
Prediction	**Class 0**	**Class 2**	**Class 1**

Step 3: Compute the classes and decide for each classifier with the highest confidence.

Step 4:
Count and calculate scores to conclude the majority.

	3 Classifiers (NB, DT, k-NN)		
	Class 0	*Class 1*	*Class 2*
Count	0+0+1	1+1+0	0+0+0
Compute	1/3	2/3	0/3
Avg.	0.33	**0.67***	0.00
Decision	No	**Yes**	No

Step 5:
Summary of prediction results with the majority voting

Step 6:
Test model's performance with cross-validation and confusion matrix techniques.

Cross-Validation

Confusion Matrix

Figure 1. The Predictive Model for Clustering Learning Outcomes

Figure 1 illustrates the six steps of the learning outcomes cluster prediction model development process. Step 1 is to organize the appropriate clusters of learners that have been published. Step 2 and Step 3 are to develop the model and select the most efficient model. After that, the confidence values of each model were calculated to find the appropriate answer class. Step 4 counts and calculates the votes to sum up, the majority vote. A count is when a response in a class gives a value of 1, and no response in a class provides a value of 0, then finds the percentage to be summarized in step 5. Step 5 summarizes the predictions of each new data record for each learner. In step 6, the new prediction data is tested for performance by this testing tool. This consists of two parts, as shown in the Model Testing Tools section.

2) Model Testing Tools

The model performance testing tool uses K-fold cross-validation and confusion matrix techniques to determine model performance with four metrics: Accuracy, Precision, Recall, and F1-Score values. In this research, the values of K are 5 and 10. The data for model testing are divided into two parts. The first data is used to develop the model (training set), and the remaining data is used to test the model (test set).

The four indicators from the confusion matrix are the Accuracy, Precision, Recall, and F1-Score values. The accuracy value is obtained from the total correct prediction result divided by the total amount of data. The precision value is received from all valid predictions in the class divided by the total amount of data in the class. The recall value is the total correct predicted actual value in the class divided by the total amount of data in the class. Finally, the F1-Score is the value used to determine the model's capabilities. They were testing the performance of all models in research in the testing process of these two techniques. All domain test results are presented in the following sections.

Findings / Results

Model Results to Course 7000103: Mathematics and Statistics for Information Technology

The model supports clustering learning outcomes for course 7000103: Mathematics and Statistics for Information Technology, classified into three parts. The first part presents the cluster results by offering the members in each cluster, as shown in Table 2. The second part summarizes the model performance for all classifiers, as indicated in Table 3. The third part presents the predictive cluster performance model for the course's learning outcomes with the ensemble learning technique, as illustrated in Table 4.

Table 2. Cluster Members for Two Courses

Cluster	7000103		7011303	
	Number	Percentage	Number	Percentage
Cluster_0	38	52.05%	12	18.18%
Cluster_1	3	4.11%	24	36.36%
Cluster_2	32	43.84%	30	45.45%
Total	73	100%	66	100%

Table 2 depicts the clustered members with a value of k equal to 3 for both courses. The researchers found that in course 7000103, the members were alienated in Cluster_1. It was ten times smaller than the other clusters. Students in this cluster received grades of F, which did not pass the course. While members 7011303 were distributed in a practical example. The researchers found that Cluster_0 has about half the number of other clusters. This is because students in this cluster have high scores, which account for 18.18% of the total number of students in the class.

Table 3. 7000103 Models Performance of All Classifiers

Cluster	Naïve Bayes (NB)					
	5-Fold			10-Fold		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Cluster_0	91.89%	94.44%	93.15%	97.06%	84.29%	95.65%
Cluster_1	60.00%	75.00%	66.67%	80.00%	66.67%	72.72%
Cluster_2	96.77%	90.90%	93.75%	91.78%	96.86%	93.94%
Accuracy		91.78%			93.15%	
Cluster	**Decision Tree (DT)**					
	5-Fold			10-Fold		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Cluster_0	94.59%	97.22%	95.89%	97.14%	94.44%	95.75%
Cluster_1	100.00%	60.00%	75.00%	75.00%	75.00%	75.00%
Cluster_2	93.75%	96.77%	95.24%	91.43%	94.12%	92.75%
Accuracy		94.44%			93.24%	

Table 3. Continued

Cluster	k-Nearest Neighbor (k-NN)					
	5-Fold			10-Fold		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Cluster_0	97.22%	97.22%	97.22%	92.31%	97.30%	94.74%
Cluster_1	75.00%	100.00%	85.71%	100.00%	75.00%	85.71%
Cluster_2	96.97%	94.12%	95.52%	96.77%	93.75%	95.24%
Accuracy		95.89%			94.52%	
Cluster	**Majority Voting (MVE)**					
	5-Fold			10-Fold		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Cluster_0	100.00%	97.39%	98.63%	97.30%	97.30%	97.30%
Cluster_1	75.00%	100.00%	85.71%	75.00%	100.00%	85.71%
Cluster_2	96.97%	96.97%	96.97%	96.88%	93.94%	95.38%
Accuracy		97.26%*			95.89%*	

Table 3 compares model performance with three classifiers and model combinations with a majority voting technique for course 7000103: Mathematics and Statistics for Information Technology. The researchers found that the model developed with the majority voting technique had the highest accuracy of 97.26% and 95.89% of the 5 and 10-Fold cross-validation, respectively. The details of the performance test with the confusion matrix technique and its indicators are shown in Table 4.

Table 4. 7000103 Model Performance of Majority Voting

Cluster	5-Fold cross-validation Accuracy = 97.26%			
	True Cluster_2	True Cluster_1	True Cluster_0	Precision Class
Predicted Cluster_2	32	0	1	96.97%
Predicted Cluster_1	1	3	0	97.00%
Predicted Cluster_0	0	0	36	100.00%
Recall Class	96.97%	100.00%	97.39%	
Cluster	**10-Fold cross-validation Accuracy = 95.89%**			
	True Cluster_2	True Cluster_1	True Cluster_0	Precision Class
Predicted Cluster_2	31	0	1	96.88%
Predicted Cluster_1	1	3	0	75.00%
Predicted Cluster_0	1	0	36	97.30%
Recall Class	93.94%	100.00%	97.30%	

Model Results to Course 7011303: Data Warehouse and Data Mining

The model supports clustering learning outcomes for course 7011303: Data Warehouse and Data Mining, are classified into three parts. The first part presents the cluster results by offering the members in each cluster, as shown in Table 2. The second part summarizes the model performance for all classifiers, as indicated in Table 5. The third part presents the predictive cluster performance model for the course's learning outcomes with the ensemble learning technique, as illustrated in Table 6.

Table 5. 7011303 Models Performance of All Classifiers

Cluster	Naïve Bayes (NB)					
	5-Fold			10-Fold		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Cluster_0	62.50%	45.45%	52.63%	85.71%	50.00%	63.16%
Cluster_1	73.08%	76.00%	74.51%	76.92%	86.96%	81.63%
Cluster_2	81.25%	86.67%	83.87%	83.87%	89.66%	86.67%
Accuracy		75.75%			81.25%	
Cluster	**Decision Tree (DT)**					
	5-Fold			10-Fold		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Cluster_0	91.89%	94.44%	93.15%	97.06%	94.29%	95.65%
Cluster_1	60.00%	75.00%	66.67%	80.00%	66.67%	72.73%
Cluster_2	96.77%	90.91%	93.75%	91.18%	96.88%	93.94%
Accuracy		91.78%			93.15%	

Table 5. Continued

Cluster	k-Nearest Neighbor (k-NN)					
	5-Fold			10-Fold		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Cluster_0	92.31%	85.71%	88.89%	92.86%	92.86%	92.86%
Cluster_1	88.46%	95.83%	92.00%	91.67%	95.65%	93.62%
Cluster_2	96.30%	92.86%	94.55%	96.30%	92.86%	94.55%
Accuracy		92.42%*			93.85%	
Cluster	**Majority Voting (MVE)**					
	5-Fold			10-Fold		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Cluster_0	87.50%	70.00%	77.78%	100.00%	84.62%	91.67%
Cluster_1	92.31%	96.00%	94.12%	91.67%	100.00%	95.65%
Cluster_2	90.63%	93.55%	92.06%	96.67%	96.67%	96.67%
Accuracy		90.91%			95.38%*	

Table 5 compares model performance with three classifiers and model combinations with a majority voting technique for course 7011303: Data Warehouse and Data Mining. The researchers found that the model developed with the majority voting technique had the highest accuracy of 95.38% of the 10-Fold cross-validation. The k-Nearest Neighbor technique had the highest accuracy of 92.42% of the 5-fold cross-validation. The details of the performance test with the confusion matrix technique and its indicators are shown in Table 6.

Table 6. 7011303 Model Performance of Majority Voting

Cluster	5-Fold cross-validation Accuracy = 90.91%			
	True Cluster_2	True Cluster_1	True Cluster_0	Precision Class
Predicted Cluster_2	29	1	2	90.63%
Predicted Cluster_1	1	24	1	92.31%
Predicted Cluster_0	1	0	7	87.50%
Recall Class	93.55%	96.00%	70.00%	
Cluster	**10-Fold cross-validation Accuracy = 95.38%**			
	True Cluster_2	True Cluster_1	True Cluster_0	Precision Class
Predicted Cluster_2	29	0	1	96.67%
Predicted Cluster_1	1	22	1	91.67%
Predicted Cluster_0	0	0	11	100.00%
Recall Class	96.67%	100.00%	84.62%	

Table 6 shows the details of the efficacy tests performed with the confusion matrix technique, which showed that all indicators showed a high level of efficacy. The researchers discussed these findings later to draw conclusions and make recommendations for further improvement of the quality of higher education in Thailand.

Discussion

Based on the research results, the researchers were able to summarize three critical points for discussion of the research results: participation in student learning activities, the development of predictive models, and the utilization of model development results. It was detailed as follows.

Participation in Student Learning Activities

In this research activity, the researchers designed a practical learning activity, and six scores were collected and used as part of the scores to determine grades. In addition, the researchers organized the exam to collect scores between the mid-term (midterm exam) and the end of the semester (final exam). The proportions of the various scores are presented in Table 1.

The researchers studied two courses over the same period to compare the learning behaviors of learners affected by the COVID-19 outbreak in Thailand. In Table 1, researchers found that online learning did not negatively affect learners from both courses, as most students achieved grade A. The percentage of students who received grade A in the 7000103 courses was 32.87 (24 of 73 students), and those who received grade A in the 7011303 courses was 46.97 (31 of 66 students).

However, an issue that requires special attention and concern is those who do not meet the learning criteria. The researchers found that four students were at the risk of dropping out of the education system, which is expected to be

affected by COVID-19. It is consistent with the many research studies (Casanova et al., 2021; de Oliveira et al., 2021; Karimi-Haghighi et al., 2022; Nuankaew, 2020; Nuankaew & Nuankaew, 2021; Viloria & Pineda Lezama, 2019), reflecting this research's importance that aims to develop a predictive model for clustering learning outcomes affected by COVID-19.

The Development of Predictive Models

Given the impact of the COVID-19 pandemic that may drive students out of the education system, it is logical for researchers to develop predictive models for clustering learning outcomes, as shown in Figure 1. The researchers developed a high-performance and highest-accuracy model for both courses.

The model developed by the researchers is an application of machine learning techniques with unsupervised learning and supervised learning to integrate the process. The researchers adopted a majority voting technique to enhance the model's predictive capabilities. The model development results revealed that both courses had a very high accuracy model. The model of course 7000103: Mathematics and Statistics for Information Technology has an accuracy of 97.26%, as summarized in the performance of each classifier in Table 3 and the best model performance in Table 4, respectively. Moreover, the model of course 7011303: Data Warehouse and Data Mining has high accuracy of 95.38%, as summarized in the performance of each classifier in Table 5 and details the best model performance in Table 6, respectively.

In the case of wrong prediction of the majority voting model, we found that the probability values were the same for all three classes (0.333), the activity 6 score was equal to 0, and the midterm score was less than or equal to 15 out of 30 points (7011303 course), which affects the processing of each class label to decide the final prediction. Therefore, the result of the majority voting model was predicted incorrectly because it can't vote for them. It was the weak point of this method (Nuankaew et al., 2023), while the k-Nearest Neighbor predicted more accurately than the majority voting of the 5-fold cross-validation. It works on complex conditions for computing the probability values in this dataset. We found that the variables used in the analysis affected the probabilities values of each class label. Therefore, data analysis and data preparation steps are important and have a direct effect on model performance.

We found that the result of the majority voting model predicted incorrectly for the 7000103 course. The students who had the activity 3 or 4 or 6 score equal to 0, the midterm scores and the final scores more than 80% of full marks. It affected predicting and reflected the students' participation in activities in the same cluster. With both developed models, researchers can use them to prepare for handling the students in high-risk clusters who are likely to drop out of the education system.

The Utilization of Model Development Results

As discussed, the impact of the COVID-19 epidemic in Thailand on the student in higher education, researchers have developed a predictive model for clustering learning outcomes affected by COVID-19. The developed model is highly effective, and it is reasonable to have the support of the head of the researchers' institution, which the researchers hope for when the research is published at the institution. The results of this research will be used to support the educational process in unusual situations.

Conclusion

The change in education patterns affected by the COVID-19 pandemic has forced learners and teachers to adapt their behaviors to the changes. Learners need to invest in technology to be able to attend the process required by educational institutions. At the same time, teachers must design learning activities that are consistent with online learning. These have severely impacted the education system. These are what inspired researchers to conduct this research with three key objectives. The first objective was to study the context for clustering learning outcomes affected by COVID-19 at Rajabhat Maha Sarakham University. The second objective was to develop predictive models using machine learning and data mining techniques for clustering learning outcomes affected by COVID-19. Finally, the third objective was to evaluate a predictive model for clustering learning outcomes affected by COVID-19 at Rajabhat Maha Sarakham University. The population and the research sample were selected by selecting a specific example with purposive sampling from two subjects with 139 students. This research applied the k-mean clustering modeling to cluster the learning outcomes of learners (Nuankaew & Nuankaew, 2021) and adopt a hybrid algorithm of clustering and classification (Francis & Babu, 2019). Nuankaew and Nuankaew (2021) were unable to identify and condition important features for the specific cluster and impossible to separate data for each activity of the cluster. In this research, the model can identify the cluster and condition of the features. It resulted in appropriate data clusters and the efficacy performance model.

Research instruments are divided into two components: tools for developing a predictive model and tools for testing model performance. The tools for building a predictive model include the Naïve Bayes (NB), Decision Tree (DT), and k-Nearest Neighbor (k-NN) techniques. The techniques for testing a predictive model include cross-validation and confusion matrix techniques. The metrics for evaluating model performance include accuracy, precision, recall, and F1-score values.

The results showed that the two courses' predictive models for clustering learning outcomes affected by COVID-19 were very effective. The first model for course 7000103: Mathematics and Statistics for Information Technology has an accuracy of 97.26%, as shown in the analysis results in Table 3 and Table 4. The second model for 7011303: Data Warehouse and Data Mining has high accuracy with 95.38% accuracy, as shown in the analysis results in Table 5 and Table 6. The process of developing both models was based on a scientific approach that concluded that they should be used to support further education in higher education institutions.

In further work, we will study an imbalanced data technique to manage this dataset, consider the data analysis with other techniques for improving result model, and test it with other courses. In addition, behaviors data in online and onsite learning system use to analyze it could be helped to improve a better performance model.

Limitations

The limitation of this research is the small number of students in each course. It may affect its application in the future if the number of learners is exponentially greater than the sample size. However, this research has positively impacted the integration and application of hybrid teaching and learning that allows learners and instructors to bridge the gap and use technology in a timely manner.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This research was supported by many advisors, academics, researchers, students, and academic staff from two organizations: the Faculty of Information Technology at the Rajabhat Maha Sarakham University, and the School of Information and Communication Technology at the University of Phayao. The authors would like to thank all of them for their support and collaboration in making this research possible.

Funding

This research project was supported by the Thailand Science Research and Innovation Fund and the University of Phayao (Grant No. FF66-UoE002).

Authorship Contribution Statement

Wongpanya S. Nuankaew: Conceptualization, design, data acquisition, analysis, and writing. Pratya Nuankaew: Editing/reviewing, supervision, and final approval.

References

Abdel-Basset, M., Chang, V., & Nabeeh, N. A. (2021). An intelligent framework using disruptive technologies for COVID-19 analysis. *Technological Forecasting and Social Change*, *163*, Article 120431. https://doi.org/10.1016/j.techfore.2020.120431

Abuhammad, S. (2020). Barriers to distance learning during the COVID-19 outbreak: A qualitative review from parents' perspective. *Heliyon*, *6*(11), Article e05482. https://doi.org/10.1016/j.heliyon.2020.e05482

Al-Kumaim, N. H., Mohammed, F., Gazem, N. A., Fazea, Y., Alhazmi, A. K., & Dakkak, O. (2021). Exploring the impact of transformation to fully online learning during COVID-19 on Malaysian University students' academic life and performance. *International Journal of Interactive Mobile Technologies*, *15*(5), 140-158. https://doi.org/10.3991/ijim.v15i05.20203

Appiah-Kubi, P., & Annan, E. (2020). A review of a collaborative online international learning. *International Journal of Engineering Pedagogy*, *10*(1), 109–124. https://doi.org/10.3991/ijep.v10i1.11678

Badal, Y. T., & Sungkur, R. K. (2022). Predictive modelling and analytics of students' grades using machine learning algorithms. *Education and Information Technologies*. *28*, 3027-3057. https://doi.org/10.1007/s10639-022-11299-8

Casanova, J. R., Gomes, C. M. A., Bernardo, A. B., Núñez, J. C., & Almeida, L. S. (2021). Dimensionality and reliability of a screening instrument for students at-risk of dropping out from higher education. *Studies in Educational Evaluation*, *68*, Article 100957. https://doi.org/10.1016/j.stueduc.2020.100957

Dechsupa, S., Assawakosri, S., Phakham, S., & Honsawek, S. (2020). Positive impact of lockdown on COVID-19 outbreak in Thailand. *Travel Medicine and Infectious Disease*, *36*, Article 101802. https://doi.org/10.1016/j.tmaid.2020.101802

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2021). How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. *Big Data and Cognitive Computing*, *5*(4), Article 64. https://doi.org/10.3390/bdcc5040064

Francis, B. K., & Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of Medical Systems*, *43*, Article 162. https://doi.org/10.1007/s10916-019-1295-4

Ghoggali, N., Douak, F., & Ghoggali, W. (2022). Towards a NIR spectroscopy ensemble learning technique competing with the standard ASTM-CFR: An optimal boosting and bagging extreme learning machine algorithms for gasoline octane number prediction. *Optik*, *257*, Article 168813. https://doi.org/10.1016/j.ijleo.2022.168813

Giani, U., & Martone, P. (1998). Distance learning, problem based learning and dynamic knowledge networks. *International Journal of Medical Informatics*, *50*(1), 273–278. https://doi.org/10.1016/S1386-5056(98)00080-X

Karalar, H., Kapucu, C., & Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education*, *18*, Article 63. https://doi.org/10.1186/s41239-021-00300-y

Karimi-Haghighi, M., Castillo, C., & Hernández-Leo, D. (2022). A causal inference study on the effects of first year workload on the dropout rate of undergraduates. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 15–27). Springer. https://doi.org/10.1007/978-3-031-11644-5_2

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers & Education*, *104*, 18–33. https://doi.org/10.1016/j.compedu.2016.10.001

Lim, C. L., Ab Jalil, H., Ma'rof, A. M., & Saad, W. Z. (2020). Peer learning, self-regulated learning and academic achievement in blended learning courses: A structural equation modeling approach. *International Journal of Emerging Technologies in Learning*, *15*(3), 110-125. https://doi.org/10.3991/ijet.v15i03.12031

Nachouki, M., & Abou Naaj, M. (2022). Predicting student performance to improve academic advising using the random forest algorithm. *International Journal of Distance Education Technologies*, *20*(1), Article 2. https://doi.org/10.4018/IJDET.296702

Nuankaew, P. (2020). Clustering of mindset towards self-regulated learning of undergraduate students at the University of Phayao. *Advances in Science, Technology and Engineering Systems*, *5*(4), 676–685. https://doi.org/10.25046/aj050481

Nuankaew, P., Nasa-Ngium, P., & Nuankaew, W. S. (2022). Improving predictive model to prevent students' dropout in higher education using majority voting and data mining techniques. In O. Surinta & K. Kam Fung Yuen (Eds.), *Multi-disciplinary trends in artificial intelligence* (pp. 61–72). Springer. https://doi.org/10.1007/978-3-031-20992-5_6

Nuankaew, W., & Nuankaew, P. (2021). Educational engineering for models of academic success in Thai universities during the COVID-19 pandemic: Learning strategies for lifelong learning. *International Journal of Engineering Pedagogy*, *11*(4), 96–114. https://doi.org/10.3991/ijep.v11i4.20691

Nuankaew, W., Nuankaew, P., Doenribram, D., & Jareanpon, C. (2023). Weighted voting ensemble for depressive disorder analysis with multi-objective optimization. *Current Applied Science and Technology*, *23*(1), 1–20. https://doi.org/10.55003/cast.2022.01.23.015

Pozdnyakova, O., & Pozdnyakov, A. (2017). Adult students' problems in the distance learning. *Procedia Engineering*, *178*, 243–248. https://doi.org/10.1016/j.proeng.2017.01.105

Smirani, L. K., Yamani, H. A., Menzli, L. J., & Boulahia, J. A. (2022). Using ensemble learning algorithms to predict student failure and enabling customized educational paths. *Scientific Programming*, *2022*, Article e3805235. https://doi.org/10.1155/2022/3805235

Viloria, A., & Pineda Lezama, O. B. (2019). Mixture structural equation models for classifying university student dropout in Latin America. *Procedia Computer Science*, *160*, 629–634. https://doi.org/10.1016/j.procs.2019.11.036

Wang, J., & Biljecki, F. (2022). Unsupervised machine learning in urban studies: A systematic review of applications. *Cities*, *129*, Article 103925. https://doi.org/10.1016/j.cities.2022.103925

Xu, F., Li, Z., Yue, J., & Qu, S. (2021). A systematic review of educational data mining. In K. Arai (Ed.), *Intelligent computing* (pp. 764–780). Springer. https://doi.org/10.1007/978-3-030-80126-7_54