



# Evolutionary Feature Weighting Optimization and Majority Voting Ensemble Learning for Curriculum Recommendation in the Higher Education

Wongpanya S. Nuankaew<sup>1</sup> , Sittichai Bussaman<sup>1</sup> , and Pratya Nuankaew<sup>2</sup>  

<sup>1</sup> Rajabhat Maha Sarakham University, Maha Sarakham 44000, Thailand

<sup>2</sup> University of Phayao, Phayao 56000, Thailand

pratya.nu@up.ac.th

**Abstract.** The curriculum recommendation strategies are the engines that drive educational organizations. Therefore, this research has three main goals: 1) to explore the context of deploying text mining technology as a curriculum recommendation application, 2) to develop a prototype model for interaction between curriculum coordinators and interested parties, and 3) to evaluate the performance of the prototype model. Research tools are text mining techniques with the genetic algorithm for evolutionary feature weighting optimization and ensemble learning algorithms, including Naïve Bayes (NB), Neural Network (NN), and k-Nearest Neighbor (k-NN). Data collection is 1,592 transactions, with seven classes via the online chat platform of the Department of Information and Communication Technology at the Faculty of Information Technology, Rajabhat Maha Sarakham University. The results showed that the model developed with the majority voting technique had the highest accuracy of 91.65%, averaging 5% higher than that of the single split model. This research has discovered tools and methods to promote and support educational processes in higher education. Therefore, the adoption of text mining technology should be enabled in the education system to communicate with the learners to meet their needs and reduce the duplication of work.

**Keywords:** Curriculum recommendation · Ensemble learning · Feature weighting optimization · Majority voting · Metadata analytics

## 1 Introduction

Today, Information Technology and Internet networks have developed and grown widely. These modern technologies play an increasingly essential role in learners' daily life in all activities and show a marked change in consumer behavior at all educational levels. It reflects that communication channels for students to promote the curriculum proactively should use information technology and the Internet as tools.

Moreover, 21<sup>st</sup>-Century learners are familiar with using electronic tools and small devices. It is therefore imperative to choose applications and analytical techniques for

quick interactions that provide relevant information to the learner. Using agents to communicate through the application channel has attracted attention [1–7]. The process of text mining thus has a more significant impact on communication in the education system [6, 7], the classification of skills in line with Industry 4.0 following the text mining approach [5], and management of predictive maintenance with text mining techniques [3].

Successful research has incentivized, motivated, and intrigued researchers to apply text mining technology to create strategies for introducing the university curriculum. Thus, this research has three primary goals: the first primary goal is to explore the context of deploying text mining technology as a curriculum recommendation tool. The second primary goal is to develop a prototype model for interaction between curriculum coordinators and interested parties. The last primary goal is to evaluate the performance of the prototype model. Research data is 1,592 transactions via the online chat platform of the Department of Information and Communication Technology, at the Faculty of Information Technology, Rajabhat Maha Sarakham University. It is an inquiry and answer to the Bachelor of Science Program in Information Technology. Research tools and processes are divided into two parts: the first part uses text mining techniques to extract critical features. The second part uses ensemble learning algorithms and classification techniques to create forecasting models. The model performance assessment section uses split-validation techniques to divide the test data and uses a confusion matrix to measure the model prototype's performance.

The primary research hypotheses for this research consisted of two hypotheses: H1: Machine learning tools and text mining techniques can extract critical attributes of the interactions between curriculum coordinators and interested parties efficiently and appropriately. H2: The feature weighting optimization and ensemble learning algorithms can optimize model prototypes for efficient and appropriate predictions of interested parties. From the research objectives and hypotheses established, the researchers strongly believe this research will significantly benefit learners and educational institutions in Thailand, where the research process and its results are presented in the following sections.

## 2 Material and Methods

### 2.1 Research Definition

The purpose of explaining research definitions is to create mutual understanding between the researchers and the research reader. Research definitions include text mining, and ensemble learning.

#### **Text Mining.**

Text mining is searching for new knowledge and discovering the hidden facts in a series of texts. It can be said to be the process of analyzing the meanings nested in the message. Text mining focuses on two types of data: unstructured and semi-structured data. Text mining is a branch of data mining technique that discovers knowledge in databases: KDD. The text mining process focuses on building credibility, which consists of three steps: data selection, data preparation, and data indexing. These three steps acquire attributes

or factors for further forecasting prototyping. Text mining for knowledge acquisition is helpful in several dimensions, including topic detection and tracking, text summarization, text classification, text segmentation, and Q&A.

This research emphasizes on answering questions to provide information to those interested in the educational program of the Department of Information Technology, Faculty of Information Technology, at the Rajabhat Maha Sarakham University, Thailand.

### **Ensemble Learning**

Ensemble learning is a method of machine learning that uses to train multiple classifiers to solve the same problem and combines them to produce better predictions [9]. Ensemble learning reduces the likelihood of incorrect predictions and can expand the area for better approximation. Therefore, it is often used to increase the performance of the model. The majority voting ensemble is used in this research as a method by which the collected data trains are used in models created from different techniques. The data test was then put into the model to predict the outcome and voted the result with the most votes as the final answer.

### **Classification Technique**

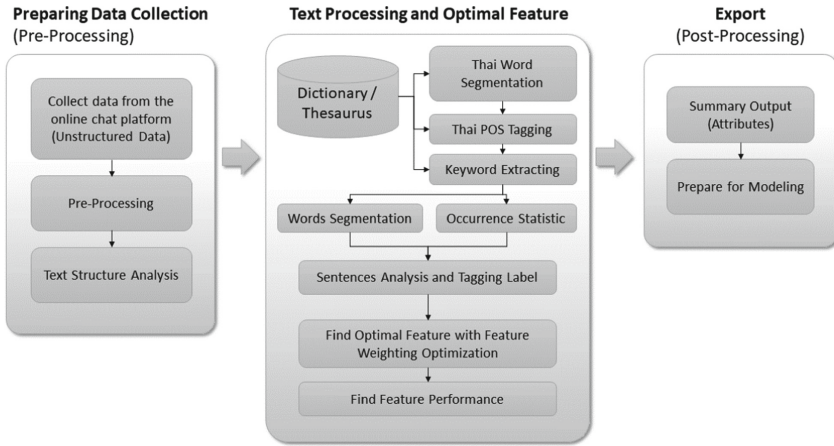
The classification techniques in this paper consisted of three techniques that were tested against a classification model based on the popular machine learning method in text classification, including Naïve Bayes (NB), Neural Network (NN), and k-Nearest Neighbor (k-NN) [9]. Naïve Bayes (NB) is the supervised learning technique used for classification that employs the principle of computational probability to infer the desired answer, also known as Bayes' Theorem. This technique is not complicated and often learns the problems that arise to create new classification conditions. The method assumes that the amount of interest depends on the probability distribution known as "Conditional Probability". The result of the selected NB calculation will be the class with the highest probability calculated and will be used as the prediction answer. Neural Network (NN) is a type of prediction network that is often used for forecasting tasks. It can help determine the critical node. It consists of a multilayer neural network, in which each layer receives and calculates the sum of the inputs and the weights of each associated point and forwards those values to the connected nodes in the next layer. k-Nearest Neighbor (k-NN) is calculated as the minimum distance between the data to be classified and the learning set, where K is the number to be chosen as the nearest neighbor. A property and advantage of using the k-NN method are that a group of similar datasets can be obtained.

## **2.2 Data Collection and Word Segmentation**

The data collected was 1,592 question-and-answer conversations about the Bachelor of Science in Information Technology program via an online chat platform of the Department of Information and Communication Technology, at the Faculty of Information Technology, Maha Sarakham Rajabhat University. It was used as data analysis to develop text mining. After collecting the data, the researcher considered and assigned each transaction with a category of classes. It consists of seven types. C1: Registration with 659 transactions, C2: Activities and Preparation in the university with 210 transactions,

C3: Occupation with 182 transactions, C4: Location and Roaming with 187 transactions, C5: Qualification with 112 transactions, C6: Expenses and Scholarships with 132 transactions, and C7: Dormitory with 110 transactions.

The word segmentation process is to separate letters from text to determine the boundaries of each morpheme. This research uses a dictionary approach for collecting data and manipulating word segmentation by following Fig. 1.



**Fig. 1.** Data collection and word segmentation

Figure 1 demonstrates the data preparation process by applying management and analysis based on text mining principles. It has three essential steps. The first step is the text import, which is the pre-processing process. In addition, initial text structure analysis. The second step is text processing, whereas the purpose of this step is to create variables for defining the model development attributes. The final step is the text export, whereby the features are summarized through the analysis process and transformed into a modeling-ready state.

## 2.3 Research Tools

### Feature Weighting Optimization

At the end of the data collection and word segmentation process, Bag of Words (BoW) method is used to create the vectors of each word in the data set. Term frequency-Inverse document frequency (TF-IDF) is applied to the weight term of each word. The researchers found that 386 attributes with a massive number of words that needed to be considered were challenging to construct the predictive model. Therefore, finding a process for reducing the number of attributes is necessary. The technique used by the researchers is feature-weighted optimization.

A genetic algorithm (GA) is a metaheuristic optimization and search problem of the evolutionary algorithm (EA). The concept of GA starts with the beginning population as

a possible chromosome problem solution and evolution to a better solution. The mainly genetic processes of GA are reproduction, selection, crossover, and mutation. The fitness function is calculated, as in Eq. 1 [8].

$$\text{Fitness} = \text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{1}$$

GA is applied to optimize the feature weighting and reduce the number of attributes in this process. It is summarized as follows:

Step 1: Generate an initial population is a possible chromosome base on the operation parameters.

Step 2: Evaluate the fitness value of each chromosome in the population in Eq. 1. The fitness values help to choose the individuals that will be mated in the reproduction process.

Step 3: Select parent to mate for finding the best solution.

Step 4: The fitness chromosomes have a bigger chance to be chosen for reproduction in the crossover operators and mutation operators.

Step 5: Decode and evaluate the fitness.

Step 6: Generate a new population for the new parent generations.

Step 7: Select survivor and find the best solution.

The parameters of GA are setting mutation rate (value = 0.10), crossover (value = 0.80), population size (value = 50), maximum number of generation (value = 250), which are used to calculate in the GA process, and all final weight values are normalized between the range 0 to 1. After feature weighting optimization, it is necessary to test the feature optimally. Evaluation of the feature weighting optimization is used by Support Vector Machines (SVMs) with spilled validation (70:30). The researchers found 223 optimized features with a weight of each attribute greater than 0.97. It tested the efficiency of weighting optimization for the features, as the efficiency is shown in Table 1.

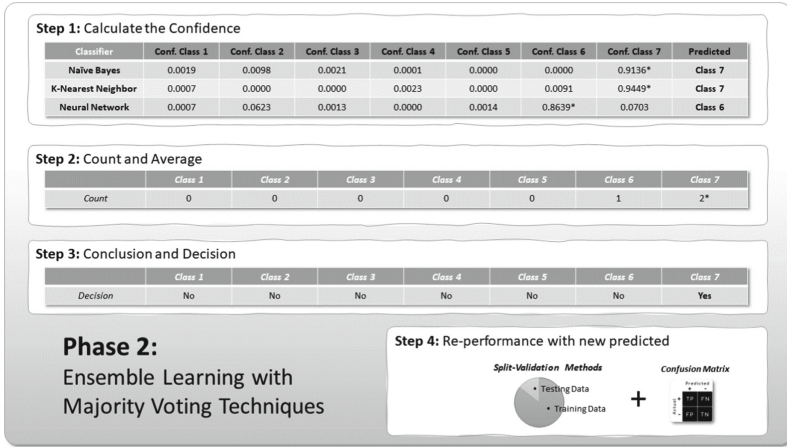
**Table 1.** Summary of the efficiency of weighting optimization features and use all features

Class	All features			Weighing optimization features		
	Precision	Recall	F1	Precision	Recall	F1
C1: Registration	96.52	97.98	97.24	98.00	98.99	98.49
C2: Activities and Preparation in the university	93.55	92.06	92.80	95.31	96.83	96.06
C3: Occupation	96.43	98.18	97.30	96.43	98.18	97.30
C4: Location and roaming	88.71	98.21	93.22	98.25	100	99.12
C5: Qualification	96.43	79.41	87.10	93.75	88.24	90.91
C6: Expenses and scholarship	94.74	90.00	92.31	100.00	95.00	97.44
C7: Dormitory	96.88	93.94	95.39	100.00	96.97	98.46
Accuracy	94.99			<b>97.49</b>		

From Table 1, the researchers found that the feature-weighted optimization method yielded higher overall accuracy and further contributed to the higher accuracy of each class. Therefore, preprocessing for feature manipulation is essential to the text mining process.

**Model Development**

There are two phases for model development. Phase one is the development of the voting component. The purpose of this phase is to create a voting tool. It consists of three classifiers: Naïve Bayes (NB), Neural Network (NN), and k-Nearest Neighbor (k-NN). The process is to develop a submodule with all three classifiers and apply it to the second phase of the majority voting.



**Fig. 2.** Prototype model development process and model performance testing

The second phase combines the models developed in the first phase using a majority voting technique to establish the most efficient model. There are four steps: The first step is calculating the confidence value and using the vote to count on the answer (class) with each classifier. The second step is to calculate the voting result based on the class with the highest mean that the class will use to answer that transaction. The third step is to draw conclusions and make decisions. The class with the highest mean is assigned to the prediction answer in that transaction. The final step is to test the voting results of each transaction to determine the effectiveness of the majority voting model. The process of this second phase is presented in Fig. 2. In addition, the results of the model performance analysis in this section are shown in Table 2.

**Model Performance Measurement**

Model performance measurement aims to find the best performance and deliver the prototype model. This research used the 70:30 data split-validation to develop and test the model, where the researchers conducted the model assessment with a confusion matrix. The basis of the measures in the confusion matrix table consists of four parts.

TP: True Positive, FP: False Positive, FN: False Negative, and TN: True Negative, as shown in Fig. 3. In addition, the indicator of the confusion matrix consists of four indicators: accuracy, precision, recall, and f1-score.

All four indicators are used as tools to determine the quality and properties of the model. The accuracy value results from all correct prediction calculations divided by the total number of transactions. The precision value is the result of calculating the accurate prediction in the predicted class divided by the total number of transactions in the predicted class. The recall value is the result of calculating the correct prediction in the class divided by the total number of transactions in the actual class. The F1-score is the harmonic mean between precision and recall as a single metric that measures model performance. The calculations for each indicator are presented in Fig. 3.

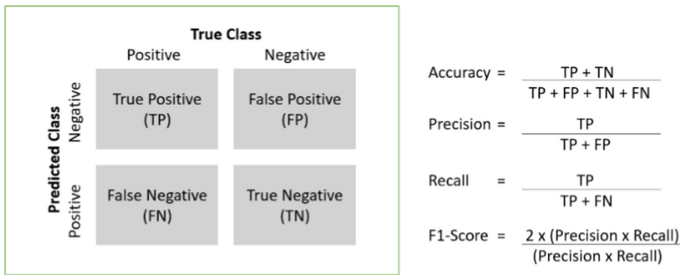


Fig. 3. The elements and calculations in the confusion matrix

This model’s methodology and performance indicators were used to assess the model’s performance, which is summarized in the research results and research discussion section.

### 3 Research Results

#### 3.1 Model Performance Classified by Technique

The model performance evaluation results classified by each technique, showing the best model compared to each classifier, are shown in Table 2.

**Table 2.** Comparison of model performance classified by technique

Class	k-Nearest Neighbor			Naïve Bayes			Neural Network		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
C1	85.00	92.12	88.42	82.30	93.47	87.53	91.04	90.15	90.59
C2	86.89	81.54	84.13	84.62	73.33	78.57	64.13	80.82	71.52
C3	87.93	89.47	88.70	97.87	83.64	90.20	88.52	91.53	90.00
C4	90.00	90.00	90.00	84.21	90.57	87.27	87.10	90.00	88.52
C5	90.91	66.67	76.92	90.32	80.00	84.85	59.09	34.21	43.33
C6	84.62	78.57	81.48	88.57	75.61	81.58	70.59	76.60	73.47
C7	84.62	75.86	80.00	92.31	77.42	84.21	68.18	48.39	56.60
Accuracy	<b>86.34</b>			85.87			81.02		

From Table 2, the model with the highest accuracy was the k-Nearest Neighbor (k-NN) model with an accuracy of 86.34%. The model with the second highest accuracy was the Naïve Bayes (NB) model with an accuracy of 85.87%. The last model was Neural Network (NN) model with an accuracy of 81.02%.

### 3.2 Majority Voting Prototype Model

The performance of majority voting prototype model is shown in Table 3, where the model performance analysis with confusion matrix is also shown in Table 4.

**Table 3.** The majority voting prototype model performance

Class	Majority Voting Prototype Model		
	Precision	Recall	F1-Score
C1: Registration	86.73	98.99	92.46
C2: Activities and Preparation in the university	87.30	87.30	87.30
C3: Occupation	100.00	94.55	97.20
C4: Location and roaming	100.00	91.07	95.33
C5: Qualification	95.83	67.65	79.31
C6: Expenses and scholarship	97.30	90.00	93.51
C7: Dormitory	100.00	78.79	88.14
Accuracy	91.65		

Table 3 show that the model developed with the majority voting technique had greater accuracy than the single prediction model, with an accuracy of 91.65%. It has an average



**Table 4.** Majority voting prototype model performance

Majority Voting Prototype Model Performance: Accuracy = 91.65								
	True C1	True C2	True C3	True C4	True C5	True C6	True C7	Class Prec
Pred. C1	196	7	3	3	8	3	6	86.73
Pred. C2	2	55	0	2	3	0	1	87.30
Pred. C3	0	0	52	0	0	0	0	100.00
Pred. C4	0	0	0	51	0	0	0	100.00
Pred. C5	0	0	0	0	23	1	0	95.83
Pred. C6	0	1	0	0	0	36	0	97.30
Pred. C7	0	0	0	0	0	0	26	100.00
Class Rec	98.99	87.30	94.55	91.07	67.65	90.00	78.79	

increase in accuracy of 5%. Therefore, the majority voting model is reasonably chosen with the model performance test results shown in Table 4.

In this section, the researchers found that the development of the model with the majority voting technique had a positive impact on the model development, with the researchers found that the model obtained had an accuracy increase of approximately 5% over the previous model, with an accuracy of 91.65%, which was appropriate for further exploitation.

## 4 Research Discussion

The discussion in this section focuses on research objectives. This research was successful in all objectives. It consists of three main goals: 1) to explore the context of deploying text mining technology as a curriculum recommendation application, 2) to develop a prototype model for interaction between curriculum coordinators and interested parties, and 3) to evaluate the performance of the prototype model.

The context of educational data mining technology applications is becoming increasingly popular [10–14]. Using text mining to support education is part of improving the quality of education [3, 6, 15]. In the context of using text mining through online chat platforms, feature selection for outcome forecasting is a research competitive area [16–18]. This research has developed a feature selection process, as shown in Fig. 1, and applied the feature selection quality improvement process with the feature weighting optimization technique, as shown in the results in Table 1. The researchers found that the appropriate features for this research consisted of 223 features that were significant for the development of the curriculum recommendation in higher education.

In the part of the second and third objectives, the researchers acted in parallel. The researchers selected three machine learning techniques to develop a prototype model. It consists of Naïve Bayes (NB), Neural Network (NN), and k-Nearest Neighbor (k-NN). Regarding model performance evaluation, the researchers used data split-validation techniques to test the prototype and the confusion matrix assessment process to measure the

quality of the prototype model. The model development results with all three techniques showed that the model developed by the k-Nearest Neighbor (k-NN) method had the highest accuracy with 86.34% accuracy. The Naïve Bayes (NB) method is as accurate as 85.87%, which has the second highest accuracy. The last model is the Neural Network (NN), with 81.02% accuracy, as shown in Table 2. Moreover, researchers believe the model could be developed with more precision. Therefore, it was decided and applied the majority voting technique to improve the prototype. The results of the model development by using a majority voting technique to improve the prototype showed a very high increase in the model's accuracy, with an accuracy of 91.65%, as shown in Table 3 and Table 4. Significantly for the improvement of the model, it was found that the model had a very high increase in accuracy of about 5%.

Based on the research findings and discussion of the results, the researchers concluded that this research is consistent with the first hypothesis that machine learning tools can extract critical attributes for creating strategies for curriculum recommendation in higher education. The evidence supporting the conclusion is the outcome of the first objective. In addition, the conclusions from the results of the second and third objectives support the second research hypothesis. The researchers found that using the feature weighting optimization and ensemble learning algorithms could improve the model's efficiency better than the previous model. Therefore, researchers mentioned that the research objectives have been achieved and can be used to support the research results to promote the quality of education in higher education, with all research findings answering all research hypotheses.

## 5 Conclusion

This research has aims to apply the body of knowledge and artificial intelligence technology as a tool to support the development of education quality in higher education. There are three main objectives. The researchers extracted 223 critical attributes analyzed by the feature weighting optimization technique for the first objective. In addition, the researchers found that the feature selection test for determining the number of features had very high accuracy, with an accuracy of 97.49%, as shown in Table 1. The second objective was to develop a prototype model and the final objective was to determine the effectiveness of the model where the last two objectives are performed simultaneously.

The researchers developed the prototype model in two phases. The first phase was the development of the voting component. The researchers used three classifiers as an indicator of the vote: k-Nearest Neighbor (k-NN), Naïve Bayes (NB), and Neural Network (NN). The practical model of each classifier is summarized in Table 2. However, the researchers believed it could improve the higher-performing model, leading to a second phase in which the model was developed with a majority voting technique. The researchers found that it had increased efficiency with an accuracy of 91.65% and an average accuracy increase of approximately 5% over previous models.

All research results reflect the success of the research in which the researchers demonstrate a step-by-step process. Therefore, the researchers believe that the results of this research help apply text mining to improve the quality of education. However, some limitations and recommendations will be an issue in future research studies. As this

research process took place during the COVID-19 pandemic, the impact of government regulators, including changing lifestyles, could affect research findings. Therefore, in the future, the data collection period should be extended under normal circumstances to compare with the results of this research.

**Acknowledgements.** This research project was supported by the Thailand Science Research and Innovation Fund and the University of Phayao (Grant No. FF65-UoE006). The authors would like to thank all of them for their support and collaboration in making this research possible.

**Conflict of Interest.** The authors declare no conflict of interest.

## References

1. Masood Khan, A., Rahat Afreen, K.: An approach to text analytics and text mining in multilingual natural language processing. *Materials Today: Proceedings*. (2021). <https://doi.org/10.1016/j.matpr.2020.10.861>
2. Jung, H., Lee, B.G.: Research trends in text mining: Semantic network and main path analysis of selected journals. *Expert Systems with Applications*. **162**, 113851 (2020). <https://doi.org/10.1016/j.eswa.2020.113851>
3. Nota, G., Postiglione, A., Carvello, R.: Text mining techniques for the management of predictive maintenance. *Procedia Computer Science*. **200**, 778–792 (2022). <https://doi.org/10.1016/j.procs.2022.01.276>
4. Zarindast, A., Sharma, A., Wood, J.: Application of text mining in smart lighting literature - an analysis of existing literature and a research agenda. *Int. J. Info. Manage. Data Insights* **1**, 100032 (2021). <https://doi.org/10.1016/j.jjime.2021.100032>
5. Chiarello, F., Fantoni, G., Hogarth, T., Giordano, V., Baltina, L., Spada, I.: Towards ESCO 4.0 – Is the European classification of skills in line with industry 4.0? A text mining approach. *Technological Forecasting and Social Change*. **173**, 121177 (2021). <https://doi.org/10.1016/j.techfore.2021.121177>
6. Urushima, A.Y.F., Tokuchi, N., Hara, S.: Text mining assessment of sustainability learning topics at higher education in Japan. In: 2021 9th International Conference on Information and Education Technology (ICIET), pp. 91–97 (2021). <https://doi.org/10.1109/ICIET51873.2021.9419584>
7. Liu, W.: Collaborative innovation of online ideological education platform with data mining and text recognition algorithms. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1372–1375 (2021). <https://doi.org/10.1109/ICCMC51019.2021.9418306>
8. Tao, P., Sun, Z., Sun, Z.: An improved intrusion detection algorithm based on GA and SVM. *Ieee Access* **6**, 13624–13631 (2018). <https://doi.org/10.1109/ICDAR.2001.953980>
9. Onan, A., Korukoğlu, S., Bulut, H.: A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications*. **62**, 1–16 (2016). <https://doi.org/10.1016/j.eswa.2016.06.005>
10. Antonenko, P.D., Toy, S., Niederhauser, D.S.: Using cluster analysis for data mining in educational technology research. *Education Tech Research Dev*. **60**, 383–398 (2012). <https://doi.org/10.1007/s11423-012-9235-8>

11. Baker, R.S.J.D., Inventado, P.S.: Educational data mining and learning analytics. In: Larusson, J.A., White, B. (eds.) *Learning Analytics: from Research to Practice*. Springer, New York, NY (2014)
12. Jeong, H., Biswas, G.: Mining student behavior models in learning-byTeaching environments. In: *Educational Data Mining*, pp. 127–136 (2008)
13. Nuankaew, P., Teeraputon, D., Nuankaew, W., Phanniphong, K., Imwut, S., Bussaman, S.: Perception and attitude toward self-regulated learning in educational data mining. In: *2019 6th International Conference on Technical Education (ICTechEd6)*, pp. 1–5 (2019). <https://doi.org/10.1109/ICTechEd6.2019.8790875>
14. Nuankaew, P., Nuankaew, W.S.: Student performance prediction model for predicting academic achievement of high school students. *Student Performance Prediction Model for Predicting Academic Achievement of High School Students* **11**, 949–963 (2022). <https://doi.org/10.12973/eu-jer.11.2.949>
15. Yuensuk, T., Limpinan, P., Nuankaew, W., Nuankaew, P.: Information systems for cultural tourism management using text analytics and data mining techniques. *Int. J. Interact. Mob. Technol.* **16**, 146–163 (2022). <https://doi.org/10.3991/ijim.v16i09.30439>
16. Chen, J., Huang, H., Tian, S., Qu, Y.: Feature selection for text classification with naïve bayes. *Expert Systems with Applications*. **36**, 5432–5435 (2009). <https://doi.org/10.1016/j.eswa.2008.06.054>
17. Jovic, A., Brkic, K., Bogunovic, N.: A review of feature selection methods with applications. In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205 (2015). <https://doi.org/10.1109/MIPRO.2015.7160458>
18. Ramaswami, M., Bhaskaran, R.: A study on feature selection techniques in educational data mining (2009). <https://doi.org/10.48550/ARXIV.0912.3924>