Journal of Applied Data Sciences
Vol. 3, No. 3, September 2022, pp. 128-136

Implementation of the K-Nearest Neighbor Algorithm for the Classification of Student Thesis Subjects

Adi Suryaputra Paramita [1,*], Indra Maryati [1], Laura Mahendratta Tjahjono [2]

[1] Information Systems Department, School of Information Technology, Universitas Ciputra Surabaya, Indonesia
[2] Informatics Department, School of Information Technology, Universitas Ciputra Surabaya, Indonesia
[1] adi.suryaputra@ciputra.ac.id *
* corresponding author

(Received: April 16, 2022; Revised: June 12, 2022; Accepted: August 18, 2022; Available online: September 30, 2022)

Abstract

Students who have studied for a considerable amount of time and will complete a lecture process must complete the necessary final steps. One of them is writing a thesis, a requirement for all students who wish to graduate from college. Each student's choice of topic or specialization will be enhanced if it not only corresponds to their interests but also to their skills. K-Nearest Neighbor is one of the classification techniques used. K-Nearest Neighbor (KNN) operates by determining the shortest distance between the data to be evaluated and the K-Nearest (neighbor) from the training data. K-Nearest Neighbor is utilized to classify new objects based on the learning data closest to the new object. Therefore, KNN is ideally suited for classifying data to predict student thesis topics. This research concludes that optimizing the k value using k-fold cross-validation yields an accuracy rate of 79.37% using k-fold cross-validation = 2 and the K-5 value. Based on the K-Nearest Neighbor Algorithm classification results, 45 students are interested in computational theory thesis (RPL) topics, 32 students are interested in artificial intelligence (AI) thesis topics, and 21 students are interested in software development topics.

Keywords: Artificial Intelligence, KNN, Data Mining, RPL

1. Introduction

The final project or thesis is one of the requirements that students at a university must carry out to be able to graduate to become a bachelor. In taking the thesis topic, the Faculty of Computer Science, provides several choices of topics or specializations that students can choose from. The choice of topic or specialization will be better if it is not only on the interests but also on the abilities of each student. Data mining is a collection of techniques to find previously unknown knowledge in an extensive database [1]. Data mining can be used to discover new knowledge or phenomena and increase our understanding of what we know. Data Mining (DM) has attracted much attention in data analysis, which can be used to extract valuable and meaningful knowledge from data [2]. K-Nearest Neighbor is one of the data mining methods used in classification [3]. Classification is vital to predicting a new variable's class [4]. The working principle of K-Nearest Neighbor (kNN) is to find the shortest distance between the data to be evaluated and the closest K-Nearest (neighbor) in the training data [5]. K-Nearest Neighbor aims to classify new objects based on learning data closest to the new object [6].

Several studies related to the use of kNN for predicting academic performance have been carried out by researchers. Tyas classifies students' academic achievement with kNN and kNN with a gain ratio. Research results with kNN produce an accuracy level of 74,068, while kNN with a Gain Ratio produces an accuracy level of 75,105 [7]. In the experiment using the KNN method, Nugroho produced a reasonably good accuracy [8]. From the results of experiments on 12 subjects in typing lessons conducted by Tanner, where this study used 15000 student data, the results showed that CNN could predict student performance accurately. Early tests on skills can be strong predictors of final scores in other skill-based courses. Furthermore, this method will be implemented as an early warning feature for typing course teachers to quickly focus their attention on the students who need the most help [9].

In this study, a classification process of student thesis topic specialization will be carried out based on student academic data, namely from the results of studies during the recovery process from semester one to semester seven. Almost all the courses held correlate with the topics or specializations that can be chosen. Therefore, an analysis can be carried out on student academic data, which can help determine the thesis topic according to their interests and abilities. With this research, which can determine the thesis topic according to the interests and abilities of students, the research results will be helpful for students and also for study programs at the computer science faculty, which is by the thesis topic taken by the student. And with the appropriate topics and interests.

2. Methods for the Development and Protection of ICH

2.1. Thesis in Definition

In conclusion, the thesis is a scientific paper resulting from research by undergraduate students who discuss the results of their research according to the research rules and the rules of the thesis report. The thesis is not just writing scientific papers. Of course, writing a scientific paper on this one is not just a graduation requirement. There is a purpose behind why students are required to be able to complete this final project.

Because it is not an easy thing to understand a problem, conduct research, analyze, get research results and compile it into a report. By compiling a thesis, students are expected to be able to think logically in describing and solving a problem and to be able to write down the results of their thoughts into a structured and systematic report. Later, the results of the thesis you make can present the results of scientific research findings that are useful for the development of science and the practical interests of state administration and communication.

2.2. Thesis Quality

According to Pierre and Simar, the definition of quality is the achievement of predetermined goals or conformity with predetermined standards. At the same time, Bourke states quality as a description of a product or work. Thus, quality covers various aspects regarding the criteria or characteristics of the work or activity. If it is associated with a student's thesis, then the quality of the thesis is related to the criteria that must be included.

Hont uses two approaches to understand the meaning of quality, namely (1) a descriptive approach and (2) a metaphysical approach. The descriptive approach assumes that quality is a description and characteristics of the results of a job. In this approach, something quality is seen as something of value. In the second approach, called the metaphysical approach, quality is seen as something that can not only be analyzed descriptively but also has criteria so that it can be measured.

A thesis is a form of scientific work of a student, which is usually prepared as one of the requirements for obtaining a degree in a Bachelor. The thesis is said to be a scientific work because the thesis is prepared based on the results of scientific research and is written systematically, consistent with the processes and steps of scientific thinking, and paying attention to the rules of scientific writing techniques. Thus, talking about the quality of the thesis means talking about the quality of the research. Research has a unique role when it is associated with certain activities. On the one hand, research is an educational tool associated with educational activities. On the other hand, research is very functional in finding knowledge or, more precisely, new information.

In thesis writing, the quality of the thesis is generally approached with a metaphysical approach. This is because the quality of the thesis is not only closely related to the value it contains, but it must also meet clear criteria. This is in line with Delors' opinion, which defines quality as related to a product, or the result of an activity that meets predetermined criteria. Furthermore, it is said that three things must be contained in the results of quality work, namely

1) reliability, namely the work results are by the established rules and as promised,
2) certainty, the work results are by expectations so that it creates trust, and
3) physical evidence, regarding the completeness of the evidence that must exist.

In terms of research, it is defined as a careful, systematic, and patient study in the field of knowledge to discover or prove facts and principles. Careful and systematic understanding is meant here because research must be carried out carefully and has a predetermined order and system. In other words, the working procedure can be described to others so that anyone can repeat the procedure to check the newly discovered information regarding validity and reliability as an indication of excellent or quality research.

Schulte in Billah states that research is said to be of quality if the research includes a series of processes from a sequence of successive stages which in outline consists of the stages of preparation, data collection, decomposition, and reporting of results. This means that the previous stages in the series of research work become the basis for the following stages. Meanwhile, Race stated that research is said to be of high quality if the research has reliable validity and reliability. About the validity of the research, Eisenhart, and Borko, suggest that there are standards that must be met, namely

1) the contribution of the research to knowledge in the field,
2) the suitability of research problems with data collection procedures and analytical techniques, and
3) the fulfillment of the requirements for the reliability of research instruments.

3. Methodology

In this research, the stages carried out starting from the analysis of the problem to the stage of concluding the results of the study. Details of the research stages can be seen in Figure 1 below.

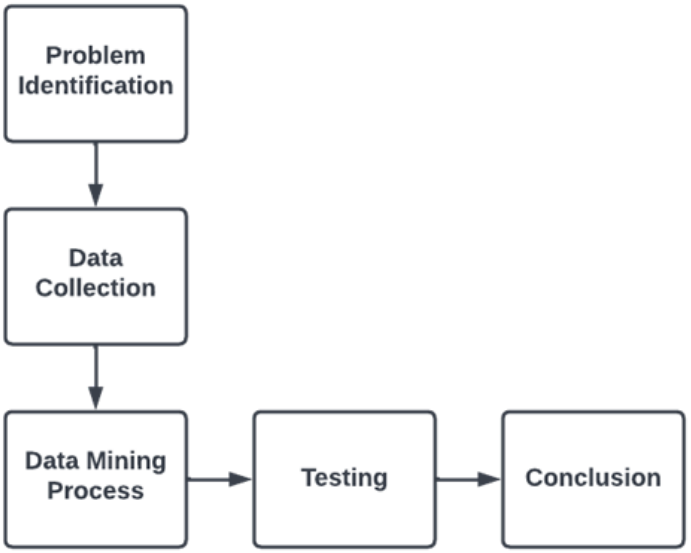

Figure 1. Research Stage

3.1. Problem Identification

Problem identification is carried out to understand the problem that has determined its boundaries and scope. By analyzing the problem, it is hoped that the problem can be understood correctly. The stages of analysis used are identifying the problems that occur, understanding the problems by collecting the required data, looking for the criteria used for the classification process, and collecting information about the needs needed in research.

3.2. Data Collection

At this stage, to obtain accurate data and information that can support the research process. In this study, data were collected on accessing library storage from various campuses in Indonesia. The data collected is 200 data in the form of abstracts from 5 different campuses with the same topic or significance. The data used to evaluate a thesis

generally requires all of the thesis documents. However, due to the limited access that the author has, the components that are analyzed are only abstracts from a thesis. This is not a limitation in this study, considering that the abstract is the conclusion of a thesis containing the main components of the thesis.

3.3. Data Mining Process

At this stage, the researcher selects the necessary data, such as training data and data to be classified, then applies the kNN algorithm to get predictions about the topic of the student thesis according to the interests and abilities of students based on student academic data. The accuracy of the prediction results is optimized with the k-fold cross-validation algorithm [4]. K-Fold cross-validation works by dividing the data set into several K parts, where each fold is used as a test set at some point. In the first iteration, the first fold is used to test the model, and the rest is used to train the model. In the second iteration and so on, the fold is used as a testing set, while the rest is used as a training set. This process is repeated until each fold has been used as a test set. The following illustrates the k-fold cross-validation process (Figure 2).

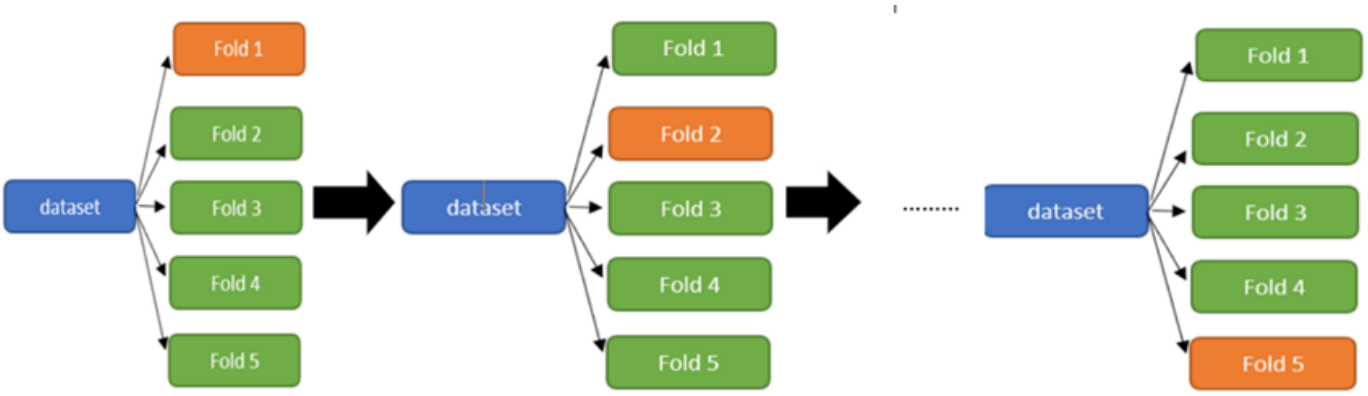

Figure 2. K-Fold Cross validation

3.4. Testing

At this stage, the testing process uses the KNIME application. This testing process is carried out to strengthen the results of the data mining process calculated manually in the previous stage.

3.5. Conclusion

At this stage, the authors can conclude from the results of research that has been done.

4. Result and Discussion

4.1. Data Training

For training data, use student data who have graduated and have a thesis topic. The training data used in this study amounted to 50 data. The training data used can be seen in the table below:

Table 1. Training data sample

No	Student ID	Computational Theory	Artificial Intelligence	Software Development	Thesis Topic
1	CS-452001	3,75	3,75	3,65	CT
2	CS-452002	3,80	4	3,50	CT
3	CS-452003	3,50	3,65	3,75	SD

4	CS-452004	3,35	3,65	3,25	AI
…	…	…	…	…	…
50	CS-452050	3,50	4	4	SD

Table 1 above is the training data that will be used in this study. The training data consists of 20 topics of computational theory thesis, 20 topics of artificial intelligence thesis and 10 topics of software development.

4.2. Data Testing

The testing data used in this study amounted to 150 data. The testing data used can be seen in the table below.

Table. 2. Testing data sample

No	Student ID	Computational Theory	Artificial Intelligence	Software Development
1	CS-452051	3,25	3,50	3,50
2	CS-452052	3,30	3,75	3,75
3	CS-452053	3,25	3,50	3,25
4	CS-452054	3,75	3,75	4
…	…	…	…	…
50	CS-452150	3,25	3,75	3,75

4.3. K-Value Optimization

The optimization of the value of k is done with the k-fold validation process algorithm. With the number of k-folds namely 3. The k-fold validation process algorithm can be seen in the flowchart of Figure 3 below.

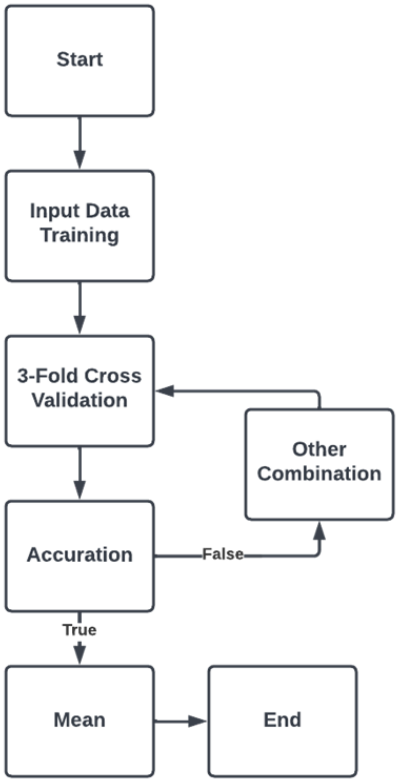

Figure 3. Flowchart Cross Validation

From the results of 2-fold cross-validation, the accuracy value is 56.67% with a k value of 5 (k=5). This value will then be used as the k value for the kNN process. Following are the results of the 2-fold cross-validation (Figure 4)

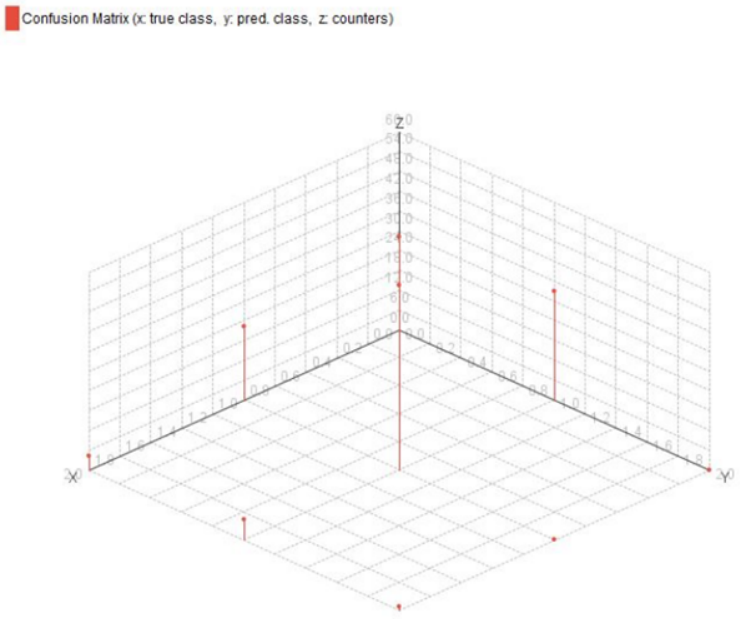

Figure 4. Graph of K Value Accuracy Level

4.4. KNN

The first step is to see what level of accuracy is generated from the 2-fold cross-validation with a value of k=5. The kNN process is carried out using KNIME. The stages of cross-validation with KNIME can be seen in Figure 5 below.

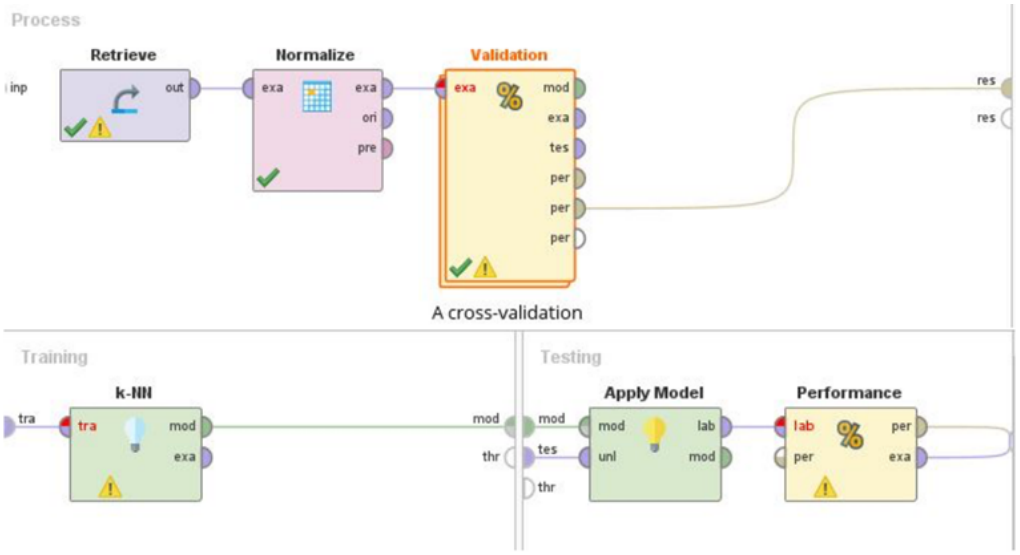

Figure 5. Accuracy level stage

From the results of the 2-fold cross-validation above with a value of k = 5, the accuracy level is 56.67% with the Confusion Matrix, as shown in Table 3 below.

Table 3. Confusion matrix

Accuracy: 79.37%	True. Computational Theory	True. Artificial Intelligence	True. Software Development	Class precision
Pred. Computational Theory	**45**	5	12	72.5%
Pred. Artificial Intelligence	12	**32**	5	65.3%
Pred. Software Development	5	13	**21**	53.8%
Class recall	72.5%	64%	55.2%	

4.5. Prediction

The prediction test is done using KNIME. At this prediction stage, the testing data will be tested with training data to see the prediction results on academic achievement with a value of k=5. The flow diagram of this process can be seen in Figure 6 below.

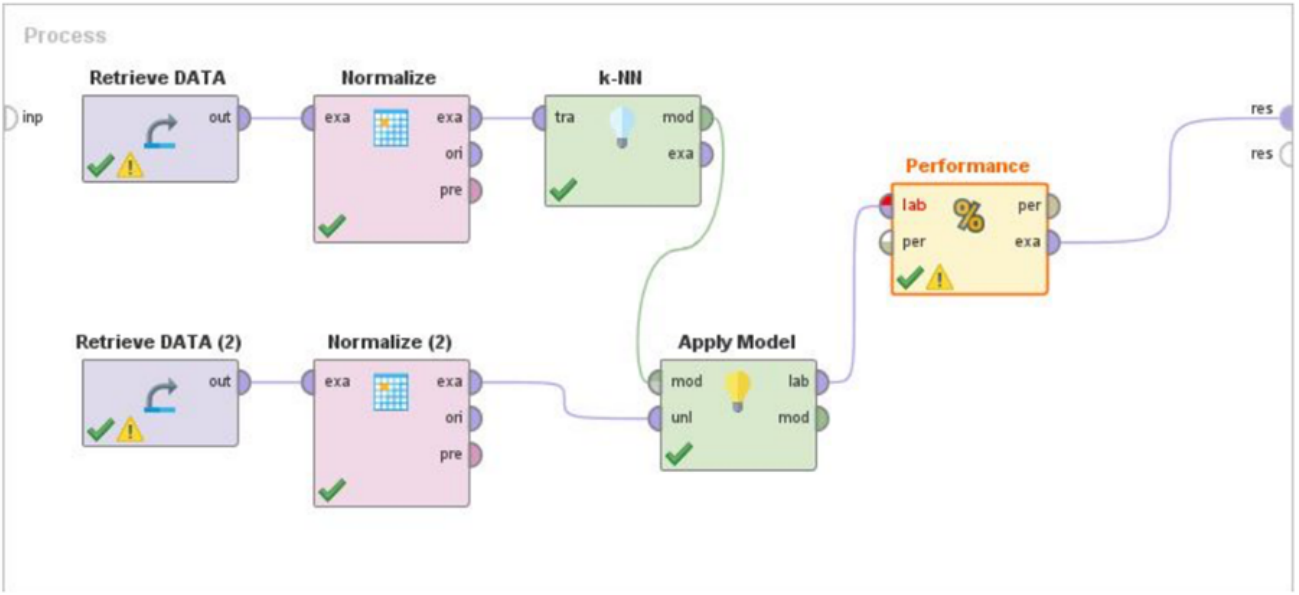

Figure 6. Data prediction step

The process flow from Figure 6 starts from entering the training data and the resulting testing data, followed by the normalization process of the two data. Next, the training data will be processed with the kNN algorithm; these two data will be applied in a model. From the results of this application, the resulting performance value will be calculated in the form of accuracy values, confusion matrix tables, and prediction tables for thesis topics. The results of the prediction test can be seen in table 4 below:

Table 4. Prediction result

No	Student ID	(Confidence)			Thesis Topic
		Computational Theory	Artificial Intelligence	Software Development	
1	CS-452051	0.000	0.300	0.460	SD
2	CS-452052	0.000	0.872	0.240	AI
3	CS-452053	0.000	0.430	0.140	SD
4	CS-452054	0.170	0.240	0.940	SD
…	…	…	…	…	…
50	CS-452150	0.600	0.323	0.473	CT

From table 3, the prediction test results can be seen in the confidence value of each class label which consists of three class labels, namely network, artificial intelligence, and software engineering. The predictive value of each variable is the highest confidence value obtained from the highest of each class variable.

5. Conclusion

From the results of the description in making this report, the authors can conclude as follows, Optimizing the value of k using k-fold cross validation produces a classification accuracy level for prediction that is 79.37% with a value of k-fold cross validation = 2 and a K-5 value. From the results of the classification using the K-Nearest Neighbor Algorithm, the result is that 45 students are interested in taking computational theory (CT) thesis topics, 32 students are interested in taking artificial intelligence (AI) thesis topics and 21 students are interested in software development (SD) thesis topics that this study predicted correctly.

References

[1] P. Prabhu, P. Valarmathie, and K. Dinakaran, "To Forecast Learner Using Mining Classification To Evaluate Tertiary Education," J. Crit. Rev., vol. 7, no. 4, pp. 329–330, 2020.

[2] K. Sunday, P. Ocheja, S. Hussain, S. Oyelere, B. Samson, and F. Agbo, "Analyzing student performance in programming education using classification techniques," Int. J. Emerg. Technol. Learn., vol. 15, no. 2, pp. 127–144, 2020.

[3] L. Lu and J. Zhou, "Research on mining of applied mathematics educational resources based on edge computing and data stream classification," Mob. Inf. Syst., vol. 2021, 2021.

[4] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," J. Med. Syst., vol. 43, no. 6, pp. 1–15, 2019.

[5] S. Hussain, R. Atallah, A. Kamsin, and J. Hazarika, "Classification, clustering and association rule mining in educational datasets using data mining tools: A case study," in Computer Science On-line Conference, 2018, pp. 196–211.

[6] M. A. Prada et al., "Educational data mining for tutoring support in higher education: a web-based tool case study in engineering degrees," IEEE Access, vol. 8, pp. 212818–212836, 2020.

[7] Y. K. Salal, S. M. Abdullaev, and M. Kumar, "Educational data mining: Student performance prediction in academic," IJ Eng. Adv. Tech, vol. 8, no. 4C, pp. 54–59, 2019.

[8] C.-C. Kiu, "Data mining analysis on student's academic performance through exploration of student's background and social activities," in 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), 2018, pp. 1–5.

[9] R. Ferreira‐Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 9, no. 6, p. e1332, 2019.

[10] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 10, no. 3, p. e1355, 2020.

[11] A. V. Manjarres, L. G. M. Sandoval, and M. S. Suárez, "Data mining techniques applied in educational environments: Literature review," Digit. Educ. Rev., no. 33, pp. 235–266, 2018.

[12] A. S. M. Al-Rawahnaa and A. Y. B. Al Hadid, "Data mining for Education Sector, a proposed concept," J. Appl. Data Sci., vol. 1, no. 1, pp. 1–10, 2020.

[13] C. Jalota and R. Agrawal, "Analysis of educational data mining using classification," in 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 243–247.

[14] L. Fan, "The method of interest text recommendation in English education based on data mining," Int. J. Contin. Eng. Educ. Life Long Learn., vol. 32, no. 3, pp. 374–388, 2022.

[15] P. Sokkhey, S. Navy, L. Tong, and T. Okazaki, "Multi-models of educational data mining for predicting student performance in mathematics: A case study on high schools in Cambodia," IEIE Trans. Smart Process. Comput., vol. 9, no. 3, pp. 217–229, 2020.