Educational Data Mining Applied For Predicting Students' ICT Literacy

Kanjana Haruehansapong, Suppat Rungraungsilp

Abstract: ICT literacy is essentially regarded as one of six strategies in the digital development plan for the Thai economy and society. The identification of students who are ICT literate and those that are not is therefore crucial. Educational institutions normally provide capability testing to classify ICT literacy of students; however, it is inconvenient to examine large groups by testing. This research proposed data mining techniques from historical student information for classification based on a decision tree, to build a model for the ICT literacy classification of the new students. In this way, the results of the ICT capabilities of students will be recognized with no need for knowledge examination by testing all students. If the result of prediction shows that students have low ICT literacy, they are required to attend an online course to improve their ICT literacy skills. As this research created a decision rule using the C4.5 algorithms and tested the predictive efficiency, the accuracy is 86.12%.

Index Terms: Classification, Data mining, Decision tree, Educational data mining, ICT literacy, Prediction

1 INTRODUCTION

ICT literacy is using digital technology, communications tools, and/or networks to access, manage, integrate, evaluate, and create information in order to function in a knowledge society [1]. ICT literacy development in youth is very important to the development of Thailand's economy and social fabric. The government focuses on the process of country renovation and development. The innovation of Thailand is called Thailand 4.0 or Digital Thailand. The Ministry of Digital Economy and Society is the main department to propel the digital development plan for the economy and society. This includes creating a practical plan to drive strategic development in cooperation with other departments, for the consistent development of the country. The use of technology is 1 of the 6 strategies of the digital development plan for the economy and society [2]. The development of human resources, to prepare for the digital economy and digital society is primal to this - build up people, build up works, build up inner strengths [3]. It is accepted that the more efficient people are in the digital world the more potential and more advantage there will be to society. Digital technology will enhance life values, enable innovations, and create super-efficient technology that will greatly move the economy forward and give the Thai people a better quality of life. In addition, the results of a research indicated that ICT skills had a more positive effect on student academic performance [4]. It is, therefore, necessary to reinforce the ICT literacy in youth and prepare to be a comprehensive, digital civil society in the future. Walailak University realizes the importance of the ICT literacy development of students by providing ICT courses and determining all students take the ICT test with the standard scores, in order to perceive the levels of ICT capabilities of students. We want to support and improve the low-potential students by having the assigned department to conduct and support the development of ICT literacy to enhance student's potential and advance their skills. In today's knowledge society, all students become ICT literate requires systematic integration of ICT literacy into the curriculum at the general education level and within each academic discipline [5]. This is the reason that most students today, are more skillful in digital knowledge, it is no longer necessary for all students to take the ICT course, except for the group of students who have a low-level of ICT capabilities. It, therefore necessary to have a classification procedure for student ICT literacy. Capability testing is one of the methods to classify ICT literacy of students; however, it is inconvenient to examine large groups. The advantage of data analysis technology is to be able to take the historical data of students, to make a 'mine of educational data', in order to analyze and research the aptitudes and abilities of students, to predict the result of student ICT capabilities without the need of classroom learning or testing. It is another procedure which provides efficiency and convenience for the educational authority.

This method of research looks at specific criteria of student data which can be used to create a research model for analyzing and classifying attributes that impact the ICT capabilities of students. This model can predict and identify the likelihood of individual student's needs for ICT literacy programs. The procedure includes some information from students, personal information, and historical ICT testing results, to further create a more powerful analysis model. The student ICT capabilities are identified using a decision tree algorithm. If the result of prediction shows that students have low ICT literacy, they are required to attend an online course and the university has developed the online curriculum for studying through MOOC (Massive Open Online Course). This helps to reveal students with a low-level of ICT literacy. The advantage of studying through the online curriculum in the form of MOOC is a widely-available educational method. It is growing rapidly and gaining popularity worldwide, as it is flexible in time and place to study [6, 7].

Based on the above information of this research, the following research questions are proposed:

RQ1. How to apply data mining techniques to extract knowledge as decision rules for predicting student ICT capability instead of the traditional examination?

RQ2. What the factors of student information including computer using behaviors that affect student ICT literacy?

2 LITERATURE REVIEW

Educational data mining is a research field that is used to enhance education system [8]. Educational data mining is interesting because there is 'big data', in educational institutions. This can be used for deciding the benefits in the development of student potential. Research related to

- *Kanjana Haruehansapong, School of Informatics, Walailak University, Nakhon Si Thammarat, Thailand. E-mail: hkanjana@mail.wu.ac.th*
- *Suppat Rungraungsilp, School of Informatics, Walailak University, Nakhon Si Thammarat, Thailand. E-mail: suppat.ru@mail.wu.ac.th*

prediction and analysis of the educational information using data mining techniques has been carried out. For example, Sudirman and Utama applied to predict the GPA from existing datasets using data mining techniques with classification methods [9] and Wulansari et al. created the model that classifies student personality type, based on demographic factors and can identify a student's career interests [10]. Hatlevik et al. studied about personal characteristics and background contextual variables may affect students' ICT self-efficacy and Computer and information literacy and the analyses show that experience with technology, autonomous learning, and socioeconomic background explains the variations in ICT self-efficacy [11]. Hou et al. used data mining techniques to predict depression in university students on their reading habits and they found that logistic regression predicted the highest accuracy [12]. Mihai et al. proposed the prototype of a recommendation system based on association rules from LMS databases in order to extract the association rules and used as inference rules to provide personalized recommendations [13]. Siriporn and Puttiporn proposed a combined sampling technique to improve the performance of imbalanced classification of university student depression data [14]. A case study in KSU mathematics department is to create a model predicting students likely performance, in a programming course-based on their grades from other subjects and classification technique is used to evaluate the likely student performance and reduce dropout levels in a programming course by also helping to predict their likelihood of course success before students enroll in it [15]. A study of Apolinar-Gotardo made use of five classification algorithms to predict students' ICT competency level, wherein the Hoeffding Tree algorithm is recommended to be applied to the datasets used to predict the students' ICT competency level [16]. All researches will not only benefit students but also helps course instructors who will be able to enhance student performance, by better estimating their abilities to learn the subject matters and adjusting the teaching strategies and methods. The use of key algorithms in educational contexts, such as decision trees and clustering, can reveal relevant knowledge, including the attribute type that most significantly contributes to passing a course and the behavior patterns of groups of students who fail [17].

3 Background Knowledge

3.1 Data mining

Data mining is a popular research method as knowledge discovery from data (KDD). This is the analysis step of searching for knowledge or interesting patterns from large amounts of data which is used in many techniques [18]-[20]. For example, data classification, association rule mining, data clustering, building models to analyze or predict possibilities. The function of KDD is that it will take the results from Data mining to assess and identify useful knowledge for the next process. The KDD process contains 7 procedures as follows: [21]

1. Data cleaning means eliminating junk data or adjusting data to be complete or for the most accurate data in the data processing.
2. Data integration is the step of gathering information from many sources to keep together for further use.
3. Data selecting to select relevant information for searching for knowledge.
4. Data transforming to be in a form suitable for analysis.
5. Data analysis is the step of extracting patterns or knowledge from the relevant information.
6. Pattern evaluation means the process for assessing to select interesting and useful patterns or knowledge for further use.
7. Knowledge presentation is the process of presenting to users in a form that is easy to understand.

3.2 Data Classification

Data classification has been used for a long time. It is the process of finding a model that describes and distinguishes data classes. The model is derived based on the analysis of a set of training data. The model is used to predict the class label of objects for which the class label is unknown. There are several techniques in data classification such as Decision tree, Bayesian networks, K-nearest neighbor, Case-based reasoning, Genetic algorithm, and Fuzzy logic [22]. This research proposes rules from data classification in the tree-chart format which is called a decision tree. As it is a worldwide-interesting method, the effectiveness of this kind of data classification has been incessantly developing especially in data classification accuracy as well as speed and simplicity to interpret to understand the rules' form [23]. For the working procedures to build up a decision tree assisting in deciding from figure 1, training data will be formed in the first procedure. If all data are in the same class, the node becomes the leaf node and the classification results will be the class name. For the second step in the case that there is not any attribute remain, the node will be the leaf nodes and the classification results will be the class of the majority of data falling on that node. In the third step, the algorithms, such as the information gain value will be used as a heuristic method to select that which attribute will be the best data extractor to the class. The resulted attribute is the attribute used in testing and deciding. There is a creation of branches for each attribute that is used in the testing and dividing data in the fourth-step. The amount of branches is equal to the non-repetitive attribute. If there is a branch with no falling data, it is needed to form as the leaf node and the classification results will be the section's name of the majority of data that falls on the parent. If any branch has data, the algorithm will use the same process to duplicate in order to create a tree for data in each branch as step five to eight. The working process will be running in duplicate until all data in the branches get to the same section or there are no more attributes. For the attribute selection to use in data classification, the attribute containing the highest value of information gain will be selected. Information gain is calculated for a split by subtracting the weighted entropies of each branch from the original entropy. When training a decision tree using these metrics, the best split is chosen by maximizing Information gain. Information gain frequently employed to calculate or compute for a feature involves computing the entropy value of the class label to measures how important and relevant it is to the class label for the dataset [24, 25]. Information gain is defined as the following formula (1) [26].

$$\text{Gain}(A) = I(s1, s2, \ldots, sm) - E(A) \quad (1)$$

A is the considered attribute and si is the number of data set S which is in class Ci by i is the number of classes with values from 1 to m classes. And information value (I) and entropy of

$$I(s1, s2, \ldots, sm) = -\sum_{i=1}^{m} \frac{Si}{S} \log 2 \frac{Si}{S} \quad (2)$$

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + \ldots + Smj}{S} I(S1j, \ldots, Smj) \quad (3)$$

attribute A or E(A) defined as formula (2)
In formula (3), v is the possible value of an attribute. The algorithm will calculate the values of information gain for each attribute in which the attribute with the highest-value of information gain will be selected as the tested attribute in set S before creating the node and the branch indicating the value in the attribute then continue to classify the data. The working procedure is being expressed in figure 1 [21].

Algorithm: Generate_decision_tree. (generate a decision tree from the given training data)

(1) create a node *N*;
(2) **if** *samples* are all of the same class, *C* **then** return *N* as a leaf node labeled with the class *C*;
(3) **if** *attribute-list* is empty **then** return *N* as a leaf node labeled with the most common class in *samples*; // majority voting
(4) select *test-attribute*, the attribute among *attribute-list* with the highest information gain; label node *N* with *test-attribute*;
(5) **for** each known value a_i of *test-attribute* // partition the samples
(6) grow a branch from node *N* for the condition *test-attribute*= a_i;
(7) let s_i be the set of *samples* in samples for which *test-attribute*= ai; // a partition
(8) **if** s_i is empty **then** attach a leaf labeled with the most common class in *samples*; **else** attach the node returned by Generate decision tree(s_i, *attribute-list, test-attribute*);

Fig. 1: The procedures of the algorithm to create decision tree

3.3 Algorithm C4.5
Algorithm C4.5 is the algorithm to generate a decision tree. It improves to extend from the ID3 that is used to build a decision tree as same as the CART algorithm [27]. The C4.5 algorithm uses the theory of information gain or entropy reduction to classify the nodes of the tree. The criteria that assist selecting attributes is the selection of each attribute to be root node and measuring gain value in which beset that which attribute can be the best to use in data classification. The best classification is to make leaf node all the same data and the attribute that has the highest-value of information gain is the best attribute to use to classify.

4 METHODOLOGY

4.1 Data preparation
Data preparation consists of 1) Data selection. This research uses 1,175 students in a bachelor's degree from Walailak University in the data analyzing process. The data used in the process consists of personal information, educational information at high school as well as bachelor's degree, ICT using behaviors which were received from the online questionnaire via google form and ICT testing result of students. 2) Cleaning the selected data by verifying for data redundancy, modifying some missing attribute, or the missing details of data which might come from the incorrect response of the questionnaire such as duplicate responding. The data therefore needs to be corrected or the missing part inserted. If there is incomplete or missing information, it will remove a transaction in the analyzing process. There will be only the use of the completed students' data for analysis. 3) Transform the data to display more general characteristics which are, the institution's type information, cumulative grade point average (gpax), skill level in using application software, computer usage level as well as an activity type in using the computer. The data used in analyzing consists of attributes displaying in table 1. Table 1 indicating details of all attributes including the possible values of each attribute in the data analyzing process. When it is passed, the abovementioned data preparation, there will be the data with the completed accuracy. After receiving the correct data, the data will be kept as an excel file which can be transformed into csv format that is the same category of excel file. The first line will be the name of the variable while the other lines will be data for the analyzing process according to the algorithm of data mining. The examples of csv file display in table 2.

Table 1. *Students' attributes used in building the model*

tribute's Name	Description
gender	gender: m male, f female
program_m	high school level's curriculum: 1 mathematics-science program, 2 arts program, 3 vocational program
school_t	school's type: 1 standard schools, 2 general schools
gpax_m	cumulative grade point average of high school levels: high, medium, low
gpax_t	cumulative grade point average of technology subjects: high, medium, low
comsubj_no	the amounts of ICT subjects studied in high school levels
program_u	program studied in university degree
faculty	Faculty in university
program_gr	group of faculties: health science, technology science, humanities and social sciences
gpax_u	cumulative grade point average of university degrees: good, fair, low
com_h	hour amounts of the computer using for each day: high, medium, low, very low
se_usage	frequency of using search engine: high, medium, low
doc_skill	word processing software potential: yes, no
graph_ skill	graphic potential: yes, no
com_freq	frequency of computer using: high, medium, low
fb_freq	frequency of using facebook: high, medium, low
email_freq	frequency of using email: high, medium, low
ict_score	ICT capabilities result is the data of class attribute or the categories which is needed to be classified: good, low or needed to develop

Table 2. *Examples of csv file for data classification*

ID	Gender	High school curriculum	School type	High school gpax	…	ICT score
1	female	liberal arts	standard	high	…	good
2	female	maths-science	general	high	…	good
3	female	maths-science	standard	high	…	good
4	female	liberal arts	general	high	…	low
5	female	maths-science	standard	high	…	good

4.2 Modeling using classification technique
In this research, the classification technique was chosen to

classify students' data that had been scrutinized to assist in the prediction of ICT capabilities of the new students. If the result is low, there will be an offering of training of ICT literacy, for these students. The decision tree was chosen in this research due to its simplicity to clarify to the users understanding the results of data analyzing as well as the software for data mining, RapidMiner Studio version 7 which is the tool for data analysis [28]. The conceptual process of this research displays in figure 2.

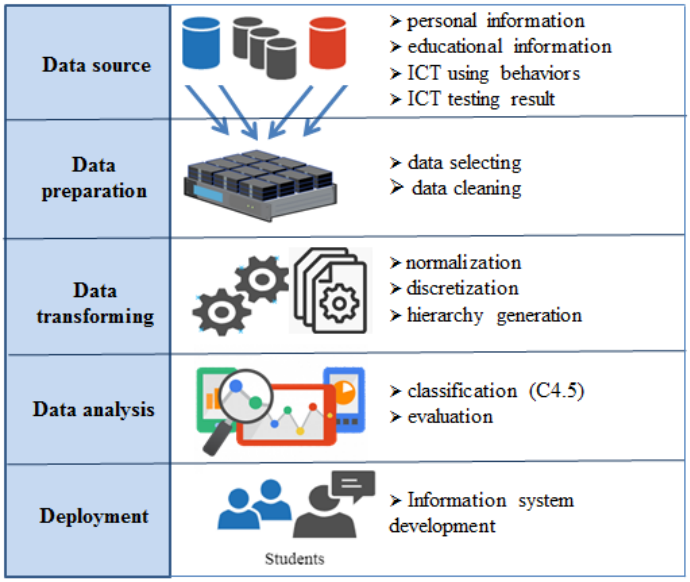

***Fig. 2**. The conceptual process of this research*

From all data which contain 18 attributes, attribute selection is a process of extracting the most relevant attributes from the student dataset. In this paper, we used information gain approach to calculate the weight of each attributes to select only important attributes. The attribute with the information gain value is upward 0.1 to be selected for building the model in this research. The calculation of each attribute's information gain will be used to reach the highest accuracy value in which the results display on Table 3 before forming the model by taking the selected attribute's data to the data classification in the format of a decision tree with C4.5 algorithm and RapidMiner Studio7 which is a software for the open source software available to be downloaded for free to create the model for the prediction of students' ICT literacy [18].

5 Results and Discussion

The result from calculating the information gain of each attribute to select the attribute with the highest value is from selecting 7 attributes with the higher value upwardly from 0.1 to build the model which are;

- school's type (school_t)
- word processing software potential (doc_skill)
- cumulative grade point average of university degrees (gpax_u)
- frequency of computer using (com_freq)
- frequency of using a search engine (se_usage)
- hour amounts of the computer using for each day (com_h)
- cumulative grade point average of technology subjects (gpax_t)

The results of the value analyzing display in Table 3 and the model formed by a decision tree with C4.5 algorithm displays in figure 3.

***Table 3.** The results of the information gain value of each attribute*

attributes	weight
school_t	0.309
doc_skill	0.217
gpax_u	0.212
com_h	0.109
se_usage	0.107
com_freq	0.104
gpax_t	0.102
gpax_m	0.090
gender	0.085
faculty	0.060
program_u	0.055
program_m	0.030
comsubj_no	0.012
program_gr	0.008
fb_freq	0.008
email_freq	0.005
graph_skill	0.003

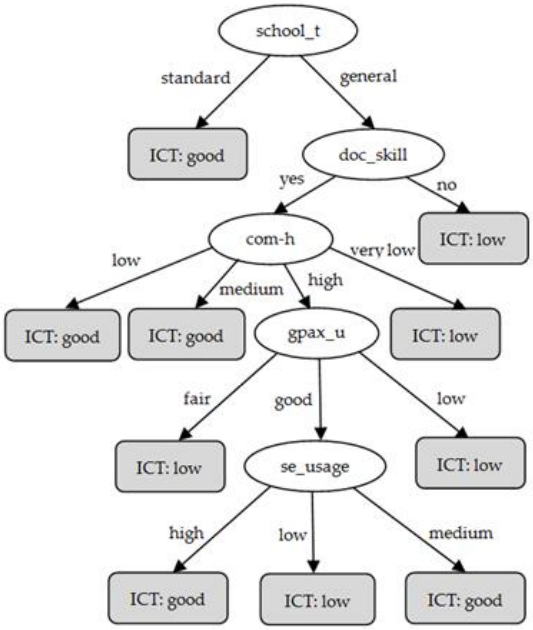

***Fig. 3**. Decision tree for ICT literacy classification*

From the model in figure 3, it can be described as the classification rules from trees as following:

IF school_t = "standard" THEN ICT literacy = "good"

IF school_t = "general" & doc_skill = "yes" & com_h = "low" THEN ICT literacy = "good"

IF school_t = "general" & doc_skill = "yes" & com_h = "medium" THEN ICT literacy = "good"

IF school_t = "general" & doc_skill = "yes" & com_h = "high" & gpax_u = "fair" THEN ICT literacy = "low"

IF school_t = "general" & doc_skill = "yes" & com_h = "high" & gpax_u = "good" & se_usage = "high" THEN ICT literacy = "good"

IF school_t = "general" & doc_skill = "yes" & com_h = "high" & gpax_u = "good" & se_usage = "low" THEN ICT literacy = "low"

IF school_t = "general" & doc_skill = "yes" & com_h = "high" & gpax_u = "good" & se_usage = "medium" THEN ICT literacy =

"good"
IF school_t = "general" & doc_skill = "yes" & com_h = "high" & gpax_u = "low" THEN ICT literacy = "low"
IF school_t = "general" & doc_skill = "yes" & com_h = "very low" THEN ICT literacy = "low"
IF school_t = "general" & doc_skill = "no" THEN ICT literacy = "low"
The evaluation of the test is conducted by measuring the accuracy of the model obtained from data classification to test the model's efficiency (Split-validation) with the 2 sets of data classified as the data to form the model (Training data) for 70% usage of all data as well as for 30% usage of the all data for the test (Testing data). The results from split-validation have tested the model formed by selecting 7 attributes with the information gain upward 0.1 values. The result of the accurate measuring of the model is 86.12% by the class predicting the good-level ICT capabilities of students will contain 86.3% of precision and 96.7% of recall while the prediction of the low-level ICT capabilities sections will contain 85.1% of precision and 55.2% of recall indicating on Table 4.

Table 4. *Accuracy values in data classification obtained from the model*

Accuracy: 86.12%			
	True good	True poor	Class precision
Pred. Good	846	134	86.3
Pred. Poor	29	165	85.1
Class recall	96.7	55.2	

The research has found the factors used in the classification of students' ICT capabilities include the following;

- what type of school did the student attends
- ability to use computer programs such as word processing program
- ability to use a search engine application
- time spent on a computer each day
- the types of high school will obviously affect the classification of students' ICT abilities.
- the students who graduated from standard schools will have good-level ICT capabilities, this may because those majority standard schools are supporting in ICT education.
- in addition, high-skill software usage (such as word processing software and frequency of use of search engines) are also essential factors.
- it was found that students with high-skill software experience were classified as having better ICT literacy results than those students who did not have such experience.
- it is not found the relation of the students' gpax in high school levels which can be used in the classification of digital capabilities.
- it is found that the students' gpax in university degree is related to the students' ICT capabilities, this may be because the gpax in high school levels of each school may contain a different standard, it is hard to classify.
- apart from this, it is found that time spent using computers is related to the classification of the students' ICT capabilities in which most of the students with too much or too little computer experience will have low-level ICT literacy.

6 Conclusion

This research proposed methods for searching knowledge from data mining of student's personal information, the result of ICT tests, and the information from online questionnaires regarding student's behaviors of 1,175 students from Walailak University by using a decision tree to assist the data classification. The procedures begin with the data assembling for 18 attributes with the information gain value is upward 0.1 values are selected to form the model for the classification of students' ICT capabilities. The obtained knowledge is used to develop the information system in order to assist the prediction of ICT literacy results and advise students with the low-potential to study further in the online course. The obtained result - the model accuracy evaluation, is a satisfactory level of 86.12%. The knowledge rules from the abovementioned model are therefore used in the students' classification. The information system is developed for new students, to classify their ICT potential from their own information, GPAX, and their ICT familiarity and exposure. However, the university has developed the online course, WU-MOOC with the subject entitled Information Technology in the digital era in order to support the students with low-level ICT literacy for further study before the real test procedure. In this way, this process can reduce the amounts of students who failed the test as it is found in 2020. The amounts of students who failed the test have decreased to 12.29% compared to 2019 in which the percentage of students failing the test was 26.23%.The limitation of this research is related to the overall accuracy in the prediction from this paper is not very high because the amount of data used for analysis is a small dataset. In the future, further studies will be carried out collecting more datasets and using other decision tree algorithm to compare with algorithm C4.5 in order to improve the performance of ICT capabilities prediction.

Acknowledgment

This research has received funding from Research Institutes, Division of Planning and Strategy, Walailak University, contract number WUDPL 61005**.** All the gainful advice in conducting this research is received thankfully from the institutional research committees, Walailak University**.**

References

[1] Educational Testing Service Policy Information Center & International ICT Literacy Panel. Digital Transformation: A Framework for ICT Literacy. Available online: http://www.ets.org/Media/Research/pdf/ICTREPORT.pdf (accessed on 16 June 2020).
[2] Ministry of Information and Communication Technology. Digital Thailand. Available online: https://www.ega.or.th/upload/download/file_9fa5ae40143e13a659403388d226efd.8pdf (accessed on 16 June 2020).
[3] The Office of the National Digital Economy and Society Commission, Ministry of Digital Economy and Society. Thailand Digital Economy Society Development Plan. Available online: https://file.onde.go.th/assets/portals/1/ebookcategory/23_Digital_Thailand_pocket_book_EN/ (accessed on 16 July 2020).
[4] X. Hu, Y. Gong, C. Lai, FKS. Leung, "The relationship between ICT and student literacy in mathematics, reading,

and science across 44 countries: A multilevel analysis," Computers & Education, vol. 125, pp. 1-13, 2018.
[5] L.S.J. Farmer, "ICT literacy integration: Issues and sample efforts," 10.4018/978-1-5225-2000-9.ch004, 2016.
[6] M.H. Baturay, "An overview of the world of MOOCs," Procedia - Social and Behavioral Sciences, vol. 174, pp. 427-433, 2015.
[7] B. Liu and H. Chen, "A study on the role of MOOCs in computer basic teaching in universities," Proceeding of the 15th International Conference on Computer Science and Education (ICCSE 2020), Delft, Netherlands, 18-22 August 2020, pp 235-238.
[8] J.M. Amala, S.I. Elizabeth, "Role of educational data mining in student learning processes with sentiment analysis: A survey," International Journal of Knowledge and Systems Science, vol. 11, no. 4, pp. 31-44, 2020.
[9] I.D. Sudirman and I.D. Utama, "Predicting GPA in Entrepreneurship Study Program by Using Data Mining Technique," Universal Journal of Educational Research, vol. 8, no. 7, pp. 3259-3273, 2020.
[10] A.D. Wulansari, H.S. Kumaidi, M. Saleh, Friyatmi, "Detection of Students' Interest with the Logistics Model," TEM Journal., vol. 8, no. 2, pp. 564-571, 2019.
[11] O.E. Hatlevik, I. Throndsenb, M. Loi, G.B. Gudmundsdottir, "Students' ICT self-efficacy and computer and information literacy: Determinants and relationships," Computers & Education, vol. 118, pp. 107-119, 2018.
[12] Y. Hou, J. Xu,Y. Huang, X. Ma, "A big data application to predict depression in the university based on the reading habits," Proceeding of the 3rd International Conference on Systems and Informatics (ICSAI), Shanghai, China, 19-21 November 2016; pp. 1085-1089.
[13] G. Mihai, "Recommendation System Based On Association Rules For Distributed E-Learning Management Systems," ACTA Universitatis Cibiniensis, vol. 67, no. 1, pp. 99-104, 2015.
[14] S. Sawangarreerak, P Thanathamathee, "Random Forest with Sampling Techniques for Handling Imbalanced Prediction of University Student Depression," Information, vol. 11, no. 11, pp. 519, 2020.
[15] G. Badr, A. Afnan, A. Hanadi, A. Manal, "Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department," Procedia Computer Science, pp. 80-89, 2016.
[16] M. Apolinar-Gotardo, "classification algorithm analysis of students' ict competency level using data mining technique," International Journal of Scientific and Technology Research, vol. 9, no. 3, pp. 3256-3258, 2020.
[17] L.N.M. Bezerra and M.T. Silva, "Educational data mining applied to a massive course," International Journal of Distance Education Technologies, vol. 18, No. 4, pp. 17-30, 2020. doi:10.4018/IJDET.2020100102
[18] P. Eakasit, An Introduction to Data Mining Techniques. Asia Publisher: Bangkok, Thailand, pp. 50-75, 2015.
[19] H.W. Ian, F. Eibe, A.H. Mark, Data Mining: Practical machine learning tools and techniques, 3rd Ed., Morgan Kaufmann Publishers: San Francisco, USA, pp. 4-76, 2011.
[20] T.L. Daniel, Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons: New Jersey, USA, pp.1-39, 2014.
[21] J. Han, M. Kamber, J. Pei, Data Mining Concepts and Techniques, 3rd Ed.; Morgan Kaufmann Publishers: San Francisco, USA, pp. 327-391, 2012.
[22] S. Vikram and N. Midha, "A Survey on Classification Techniques in Data Mining," International Journal of Knowledge Management Studies, vol. 16, no. 1, ISSN (Online) 2231-5268, 2015.
[23] J.R. Quinlan, "Induction of decision tree," Journal of Machine Learning Research, vol. 1, no. 1, pp. 81-106, 1986.
[24] C.S. Dhir, N. Iqbal, S. Lee, "Efficient feature selection based on information gain criterion for face recognition," Proceedings of the International Conference on Information Acquisition, Jeju, Korea, 8–11 July 2007, pp. 523–527.
[25] T.A. Alhaj, M.M. Siraj, A. Zainal, H.T. Elshoush, F. Elhaj, "Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation," PLoS ONE, vol. 11, no. 11, 2016. https://doi.org/10.1371/journal.pone.0166017.
[26] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining. Pearson: Boston, USA, pp. 25-43, 2005.
[27] J.R. Quinlan, C4.5: Programs for Machine Learning. Kaufmann Publishers: San Francisco, USA, pp. 5-13, 1993.
[28] P. Eakasit, Advanced Predictive Modeling with R & RapidMiner Studio 7. Asia Publisher: Bangkok, Thailand, 2018.