Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika is a national scientific journal that publishes scientific research papers/articles or reviews in the field of Computer Systems and Informatics. The journal is accredited “Sinta 2” according to the decree of DRPM Ministry of Research and Higher Education number 21/E/KPT/2018 dated July 9th, 2018 which is valid since vol 2 issue 1 until vol 6 issue 2. The scope of this journal includes software engineering, information systems development, computer systems and computer networking. The journal is published by Muhammadiyah University Press (MUP), Universitas Muhammadiyah Surakarta.

ISSN: 2621-038X e-ISSN: 2477-698X

khazanah informatika

Vol. 5, No. 2 | December 2019

Preface

Firstly, we'd like to express our great gratitude to Allah for His blessing. It is a great pleasure that volume 5 issue 2 of **Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika** has been published quite timely. The publication hopefully adds to the collection of articles and knowledge in the field of computer systems and informatics.

The current issue contains 8 articles, which are written by researches from various universities around the archipelago, either public universities or private institutions. Starting from this issue, Khazanah Informatika is published in English. We keep accepting submission in Bahasa, which, after the review process, will be translated into English by our team.

We'd like to thank all the authors that have submit their manuscripts to this journal. Through out the year of 2019, we have received no less than 134 manuscripts, of which 16 are published this year and 38 are still in the review process. The rest are rejected for various reasons. We are committed to do timely review process, which begins with an initial review in about a month after submission date. The initial review is then followed by peer review involving researchers from various universities. Time taken for review process depends on how long reviewers work, the authors respond, and the editors that handle a manuscript. Statistics for 2019 suggests that in average, it takes 40 days for a manuscript to get accepted, and 167 days to get published (see http://journals.ums.ac.id/index.php/khif/about/statistics).

Quality of this journal is recognized officially by Indonesian Ministry of Higher Education. It is accredited at level 2 (Sinta 2) according to the Decree of DRPM Ministry of Research and Higher Education number 21/E/KPT/2018 dated July 9th, 2018, which is valid for volume 2 issue 1 until volume 6 issue 2. The journal is indexed by DOAJ (Directory of Open Access Journal) and Google Scholar.

Kind regards,

Chief Editor

ISSN: 2621-038X e-ISSN: 2477-698X

khazanah informatika

Vol. 5, No. 2 | December 2019

Table of Contents

Effectiveness of SVM Method by Naïve Bayes Weighting in Movie Review Classification

Fadli Fauzi Zain, Yuliant Sibaroni 108-114

Architecture of Backpropagation Neural Network Model for Early Detection of Tendency to Type B Personality Disorders

Cynthia Hayat, Samuel Limong, Noviyanti Sagala 115-123

The Design of Exploratory and Preprocessing of Event Log Data in Online Learning Activities Based on Moodle LMS for Process Mining

Demaspira Aulia, Indra Waspada 124-133

Analysis of Slow Moving Goods Classification Technique: Random Forest and Naïve Bayes

Deny Jollyta, Gusrianty, Darmanta Sukrianto 134-139

A Virtual-Reality Edu-Game: Saving The Environment from the Dangers of Pollution

Dita Aluf Mawsally, Endah Sudarmilah 140-145

Knowledge Extraction on Reducing the Number of Students Using Explore, Elaborate and Execute Techniques

Juvinal Ximenes Guterres, Ade Iriani, Hindriyanto Dwi Purnomo 146-157

Silhouette Density Canopy K-Means for Mapping the Quality of Education Based on the Results of the 2019 National Exam in Banyumas Regency

Ridho Ananda 158-168

Case-Base Reasoning (CBR) and Density Based Spatial Clustering Application with Noise (DBSCAN)-based Indexing in Medical Expert Systems

Herdiesel Santoso, Aina Musdholifah 169-178

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Effectiveness of Naïve Bayes Weighted SVM Method in Movie Review Classification

Fadli Fauzi Zain
Informatics Study Program
Universitas Telkom
Bandung
fadlifzain@gmail.com
Yuliant Sibaroni
Informatics Study Program
Universitas Telkom
Bandung
yuliant@telkomuniversity.ac.id

Abstract-Classification of movie review belongs to the domain of text classification, particularly in the field of sentiment analysis. Popular text classification methods for the process include Support Vector Maching (SVM) and Naïve Bayes. Both methods are known to have good performance in handling text classification individually separately. Combination of the two may be expected to improve the classification performance compared to the performance of each individual method. This paper reports an effort to classify movie review using the combined method of SVM with Naïve Bayes as the weighting factor, which is commonly called NBSVM. Our work shows that higher accuracy is obtained when classification is done using NBSVM rather than using individual methods. Accuracy at the level of 88.8% is attained when using the combined feature of unigram and bigram with only data cleansing in the pre-processing stage.
Keywords: movie review, classification, NBSVM, Naïve Bayes, SVM

1. Introduction

a. Background

Text is a common media to deliver review not exceptionally in the case of movie review. Movie review is believed to give influence to consumers and film lovers in deciding whether or not to watch a screen [1]. Movie fans may first read reviews before deciding to watch or not to watch a movie in order to avoid disappointment of seeing under qualified play. The huge number of movies produced drives film lovers to get more selective in deciding which movie to see.

Movie review is beneficial not only to film consumers but also for film producers. People's sentiment towards films can be used by producers to infer which kind of movies people love and which they don't. Such a knowledge is useful for producers to make films that will find large audience and will fulfill the market demand.

People's sentiment towards a thing or an event may be inferred by sentiment analysis against comments or reviews on the topic. In computer science, sentiment analysis may be conducted using classification techniques. Classification in the context of text mining is a method to label texts into one of known categories or classes. In sentimen analysis, the categories may be positive, negative or neutral. Many works have used classification techniques to conduct sentiment analysis of movie reviews [1] - [3]. It is true that manual works are needed in the training stage to label texts for the classification algorithm to proceed. However, the later process of conducting sentiment analysis can be automatically run, eliminating the need to observe texts one at a time.

SVM (support vector machine) is one of text classification methods that is known to have high performance among many other classifiers [4]-[6]. On the other hand, Naïve Bayes is a classifier that is a simple and easy to implement [7]. The latter method in many cases of text classifications shows performance that is almost equal to that of the SVM method [6]. There is an expectation that combining the two methods will produce a better result than running each individual method in text classification.

The two aforementioned methods are combined by giving each a different role in the course of classification process. Classification of movie review proceeds through several preliminary stages including pre-processing, feature extraction, and feature weighting. Pre-processing stage includes raw text cleansing, stop-word removal, and lemmatization. Feature extraction stage processes text to produce n-gram features. Feature weighting stage is carried

out using Naïve Bayes probability model, while the main process of classification is conducted using the SVM method. The approach is known as NBSVM.

b. Topics and Limitations

This research is focused on the use of the Naïve Bayes method for calculating the weight and SVM method for classification, which is called in previous studies as the NBSVM method [8]. The use of the method may supported by initial data processing including stop-word removal and lemmatization [9]. The method is applied to movie review data that has 2 polarities, namely positive and negative. The data are obtained from English movie review of IMDB, which was collected in 2002 and has been used in many studies [10]. The data contain 2000 movie review records, consisting of 1000 positive and 1000 negative reviews.

This study aims at determining the performance of the NBSMV method in the movie review classification process. The baseline is the performance of the classification of movie review using separate individual SVM and Naïve Bayes methods. The NBSVM combination applies when Naïve Bayes is implemented for the weighting process of n-gram feature. Performance is measured based on the accuracy of the classification process. In addition to observing the performance of the classification method, this study also observes the performance of classification process for different pre-processing algorithms.

2. Related Studies

Research on the classification of movie reviews is closely related to the field of text mining that uses text as input data. Text mining or text data mining tries to find knowledge by analysing textual data. The process refers to the way of taking knowledge or information based on a pattern in the text [11].

Text mining first appeared in 1674 and was associated with the name Thomas Hyde for the library catalog process at Oxford University [12]. In 1958 a person named Luhn adapted an IBM 701 computer to produce document abstractions [12]. Research on text mining are still continue at this time to get deeper information discovery.

One branch of knowledge in text mining is text classification. Text classification can be applied in various fields such as topic detection, spam e-mail filtering, web classification and sentiment analysis [13]. Movie review classification in this study falls into the field of sentiment analysis because classification labels are only related to the emotions of the commentators, namely positive, negative or neutral.

Focus of research in text classification includes works such as developing methods in labeling texts, developing pre-processing methods (such as stemming, stop-word removal, and data cleansing), feature extraction, feature weighting methods, feature selection methods and also the invention of new classifier methods.

In the field of general text classification, Uysal and Gunal examined the effect of pre-processing on the performance of text classification. The study was conducted with e-mail data and online news and the languages used were Turkish and English. Their results showed that pre-processing affected performance of the classification of texts and the performance were influenced by domain and language used [13]. Other studies conducted by Dasgupta et al. [14] focused on the feature selection. Their research showed that strategies with provable performance guarantees give better results compared to other feature selection methods. Research to improve classification method has also been carried out, for example using particle swarm optimization which is claimed that it improves the process of identifying retinopathy [15].

In the field of sentiment analysis, particularly in the field of movie review classification, a number of studies have also been carried out. Research on classification of movie review is important because movie review turns out to influence consumers' decision to watch or not to watch a film [1]. However, according to Pentheny, this influence does not apply to all types of human personalities.

Multiple classifier strategy was used by Tsutsumi to classify movie reviews [2]. The results showed that the use of three classifiers with a voting mechanism gave better results than the use of a single classifier. In this observation, the classifiers tested were SVM, ME and score calculation.

Sahu and Ahuja focused more on multilabel classification, namely by classifying the polarity of movie reviews on 4 scales, from values 0 to 4 [3]. The structured N-Gram feature was also observed in this study and it proved to give the best accuracy.

Tripaty et al. in the movie review classification proposed a machine learning approach based on n-gram features [16]. The n-gram combination implemented in the study is for n = 1, 2 and 3. While the classifier tested in this study includes Naïve Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (GDE), and Support Vector Machine (SVM). The result showed that SVM classifier with complete n-gram features (n = 1, 2, 3) gave the best results, reaching a level of 88.94%. The study used the IMDB dataset.

3. Method

Our study begins with retrieving movie review data collected by Pang and Lee from the IMDB website [10]. Pre-processing is performed on the data which is then continued with the n-gram feature extraction process. The n-gram feature is then weighted using the Naïve Bayes probabilistic model. The results then become input for the SVM model during the learning process of the model.

a. Pre-processing

Pre-processing is implemented with the aim of reducing noise in the dataset, thereby it results in

increasing classification performance. In our research, pre-processing includes cleansing, stop-word removal and lemmatization.

Cleansing is conducted by removing symbols and numerical characters from texts in the dataset. The goal is to get rid of terms that have no meaning so that noise can be eliminated, which may reduces classification performance. As an example, the cleansing of phrase "when you see the scene on 13:56, urgh" will result in another phrase "when you see the scene on urgh".

Stop-word removal deletes words that are not considered important and do not add to the meaning of the sentence. Words that come out very often or conjunction are considered to have no meaning. For example, stop-word removal of the phrase "when you see the scene on urgh" will result in the phrase "you see the scene", because the words *when*, *the*, *on*, *urgh* belong to stop-words.

Lemmatization is conducted to return words to their basic form. It is assumed that a word has the same meaning even if it is in different forms. Lemmatization removes affixes to a word. A case of lemmatization is to revert past tense to simple tense, for example from "I wrote the letter" to "I write the letter".

In this research, cleansing is the only pre-processing method that is used in all experiments. On the other hand, stop-word removal and lemmatization become testing variables. We applied cleansing by removing symbols other than the alphabet, and we used NLTK library for stop-word removal and lemmatization.

N-gram Feature Extraction

One of the problems with text mining is the failure to take the combinatorial meaning of words that have different meanings when the words are separated. N-gram feature extraction is an effort to overcome this problem [17].

The n-gram feature extraction helps solve the problem by combining n words into one lexical or term, so that the meaning of a term or phrase like "good morning" can be interpreted better by machine as opposed to the separated words "good" and "morning". This greatly helps the task of Natural Language Processing in interpreting term or lexical units.

N-gram is conducted after pre-processing, so that the combination of terms occurs when the data is clean from noise. In this research, we use several n-gram, namely unigram, bigram, and trigram. Unigram breaks up sentences into one gram per term. For example, unigram feature extraction of the phrase "what do you want to say?" becomes "what", "do", "you", "want", "to", and "say?". The use of bigram feature extraction to the same sentence will produce features "what do", "do you", "you want", "want to", and "to say?". While the trigram feature extraction results "what do you", "do you want", "you want to", and "want to say?".

b. Naïve Bayes weighting

Naïve Bayes is one of the algorithms for classification process. It uses probability and statistical methods. Naïve Bayes algorithm predicts the likelihood of an event to occur by learning information that has been obtained previously. The probability theory involved is called the Bayes Theorem [18].

Naïve Bayes has various advantages. The algorithm is easy to implement because it has low complexity, it does not need too many training data, and it does not require model optimization. Attributes on training data that have independent assumptions are outside the scope of this study. If these conditions are not met, the performance of the Naïve Bayes method will diminish [18].

Naïve Bayes method is a classifier that uses a probabilistic model for the classification process. Probalilistic formula for an attribute X is

$$P(C \mid X) = \frac{P(X \mid C).P(C)}{P(X)} \quad (1)$$

where, P(C|X) is the probability of attribute X to be classified as class C, P(C) is the probability of class C to appear in all training data, while P(X) is the probability that attribute X appears in all training data and P (X|C) is the probability that attribute X to occur in class C.

Naïve Bayes weighting is conducted after the process of feature extraction with n-gram has finished. The process of weighting calculates the probability of occurrence of each term in each class, which produces a matrix containing the weighting value.

As an illustration, suppose we have the following two sentences in training data.

1) *"This movie is great, I like it"* – positive
2) *"I don't like the movie, the villain is too dumb"* – negative

Unigram feature extraction for sentence 1 and 2 (see Table 1) displays the number of unigram occurrences and their polarity. Naïve Bayes weighting calculates the probability of the occurrence of a unigram in a positive or negative class using equation (1) and the results is described in Table 2.

The result of weighting with Naïve Bayes probability makes the matrix data in Table 2 more detailed than the matrix data in Table 1 that does not use Naïve Bayes weighting. More detailed weight values are expected to support the SVM algorithm when building classifier models so that they can classify data better [8].

Table 1. Example of unigram matrix extracted from two sentences as discussed in the text

	dont	dumb	great	I	is	It	like	movie	the	This	too	villain
pos	0	0	1	1	1	1	1	1	0	1	0	0
neg	1	1	0	1	1	0	1	1	2	0	1	1

Table 2 Example of the Naïve Bayes weighting matrix

	dont	dumb	great	I	Is	It	like	movie	the	This	too	villain
pos	0	0	2.315	1.157	1.157	2.315	1.157	1.157	0	2.315	0	0
neg	0.578	0.578	0	1.157	1.157	0	1.157	1.157	0.385	0	0.578	0.578

c. SVM Classification

Support Vector Machine (SVM) can be said to be a semi-eager learner classification algorithm because it requires training. SVM also stores a small portion of the training data for reuse during the prediction process. Some of the data that is still stored is support vector so this method is called Support Vector Machine [19].

The basic idea of SVM is to separate support vectors between classes by creating restrictions on support vectors. The boundary is called a hyperplane. The delimiter is chosen based on the maximum margin (distance between delimiters). These hyperplane borders have different line shapes called kernels. The best known kernel is the linear kernel because it is easy to implement. Equation 2 below is an example of a linear kenel equation.

$$\overline{wx} + b = 0 \tag{2}$$

In the equation, $\overline{w}$ is a weight vector (*weight*), $\overline{x}$ is a vector of the attribute of the dataset, while b is a bias value.

Hyperplane with the selected kernel is then used as a model to predict the class of data. Prediction is obtained by mapping the vector data that is sought and reading the value of the support vector is located in which part of the whole class [19].

SVM has the advantage of being one of the most powerful and accurate methods among common methods and rarely overfitting when the model is right. However, computing from SVM is known to be heavy, because the more training data, the heavier the process of SVM is also [19].

In the sentiment analysis process, the SVM method is applied to the data in the weighted matrix with Naïve Bayes (Table 2). This study uses a linear SVM kernel that utilizes the LinearSVC module from Scikit-Learn with default parameters (C = 1, weight = 0). In this study, SVM parameter optimization has not been done.

d. Classification Performance

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Figure 2. Confusion matrix

The results of the classification by the SVM method are summarized in a format called a confusion matrix. Confusion matrices are commonly used to describe the performance of a classification method whose actual class is known [20]. The form of the confusion matrix for the classification of two classes can be seen in Figure 2.

The confusion matrix entry in Figure 2 is only a number illustration only as an example of calculating the accuracy value. The data in the confusion matrix shows the number of class predictions that correspond to the actual class. Accuracy which is a measure of classification performance is calculated using equation 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

TP or True Positive is the number of positive predictive results whose actual class is also positive, while TN or True Negative is the number of negative predictive results whose actual class is also negative. For the example data in Figure 2, the result of the calculation of the accuracy value is 150/165 = 90.9%.

e. Validation

Evaluation of the film review classification process is done using the k-fold cross validation method with k = 10. This evaluation method is common in several studies of text classification [21] - [23]. This method guarantees that the results obtained are more objective, and not obtained by chance because of the good data composition. This method is done by dividing the dataset into 10 parts, where 9 parts are used in the learning process and 1 part is used as testing. The choice of k = 10 is based on the results of previous studies to minimize overfit [8].

4. Results and Discussion

This study attempts to observe the performance of the SVM algorithm for the classification of movie reviews. Research parameters include the use of Naïve Bayes weighting, n-gram feature extraction, and pre-processing treatment. N-gram feature extraction testing is done with several n-gram ranges, namely {(1, 1), (2, 2), (3, 3), (1, 2), (1, 3), (2, 3) }. The range in question is a combination of n-gram features with a range from the initial value to the final value, for example range (1, 3) means the combination of unigram, bigram and trigram.

The results of the classification performance calculation for the Naïve Bayes method, Support Vector Machine and a combination of both (B

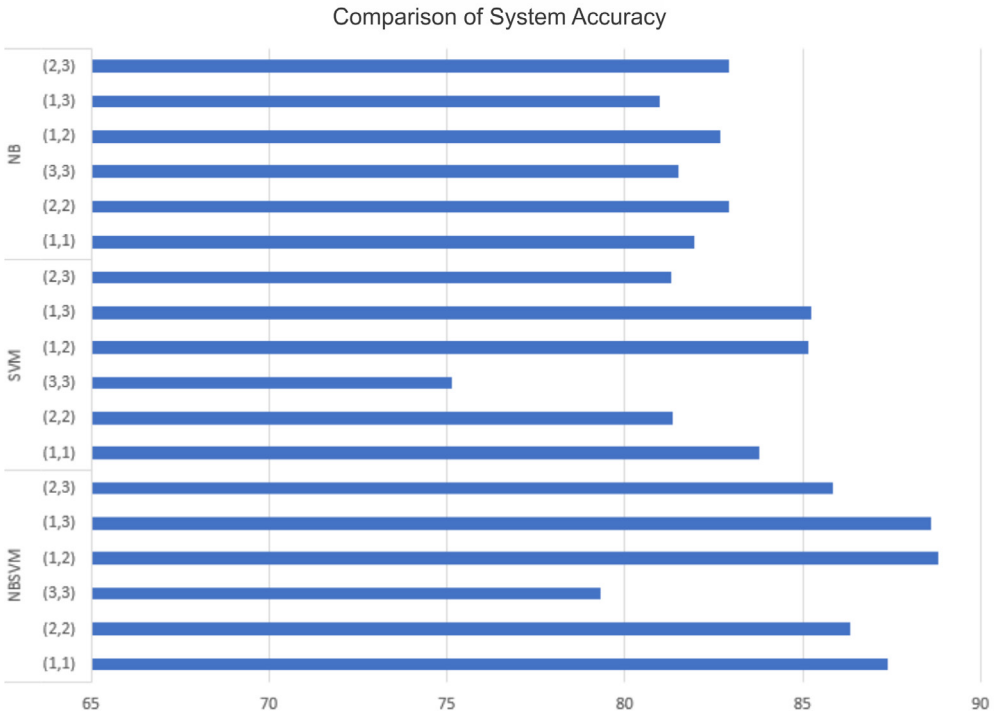


Figure 3. Accuracy of the classification process using the NB, SVM and NBSVM methods for various n-gram ranges

Table 3. Effect of Pre-processing treatment on NBSVM performance

Classifier	Range N-Gram	*Cleansing*	*Cleansing + Stop-word*	*Cleansing + Lemma*	*Cleansing + stop-word + Lemma*
NBSVM	(1,1)	87.4 %	86.6 %	87.1 %	86.4 %
	(1,2)	88.8 %	87.8 %	88.25 %	87.7 %
	(1,3)	88.6 %	87.4 %	88.05 %	86.5 %

NBSVM) is shown in Figure 3 which is presented in the form of a bar chart. There are 18 bars in the diagram, where the top 6 bars are the performance of the Naïve Bayes method, the middle 6 bars are the performance of the SVM method and the rest are the performance of the NBSVM method. Each bar represents a performance value for a different n-gram range as written on the label on the left. This diagram is obtained to treat pre-processing only in the form of data cleansing.

Figure 3 clearly shows that the NBSVM method shows better performance than the other two methods, for almost all n-gram ranges except range (3, 3). For range (3, 3), the Naïve Bayes method shows the best performance. The highest performance is obtained if the NBSVM method with n-gram range (1, 2) gives an accuracy value of 88.8%. This means that the use of the NBSVM method with unigram and bigram feature extraction together provides the highest classification performance.

Further observations were made on the NBSVM method to see the effect of pre-processing and n-gram range. The pre-processing process is varied to see the effect of each subprocess on classification performance. Table 3 shows the classification accuracy values for pre-processing which involve data cleansing only, cleansing with stop word removal, cleansing with lemmatization, and cleansing with both stop word and lemmatization.

Table 3 shows that the movie review classification provides the best performance when data cleansing is only done at the pre-processing stage. The stop word removal and lemmatization process does not improve the accuracy of the classification process. This phenomenon can occur if the deleted stop-word is actually an important word in the context of a movie review. The stop-word list used in this study is derived from the general NLTK module. It is suspected that some stops may need to be maintained and further observation is needed to verify this suspicion.

Table 4. Effect of range on NBSVM performance

Classifier	N-Gram	Accuracy
NBSVM	(1, 1)	87.4
	(1, 2)	88.8
	(1, 3)	88.6

The lemmatization process turns down the classification performance. The reason is probably the inadequate performance of the lemmatization method. The lemmatization process might produce words without affixes that have a different meaning than when the prefixes still exist. However, the change in accuracy due to

the stop word removal and lemmatization process is not too significant so it is recommended to use the simplest process, namely pre-processing with data cleansing only.

Subsequent observations were made on the effect of the n-gram range made for the NBSVM method. The results are shown in Table 4. The best accuracy is obtained when feature extraction is done by combining unigram and bigram, that is in the range (1, 2). The addition of the tri-gram extraction feature in our observations did not improve the performance of the movie review classification process. While the use of the unigram feature alone gives a relatively smaller performance. For this reason, it is recommended to use range (1, 2) in the film review classification process with the NBSVM method.

5. Conclusion

Based on the description in the Results and Discussion section it can be concluded that the best performance for movie review classification is obtained when the NBSVM method is used, which gives accuracy at a value of 88.8%. The method combines SVM method for text classification and the Naïve Bayes for weighting the n-gram extraction. The use of SVM and Naïve Bayes methods separately gives significantly lower accuracy.

The use of mere data cleansing at the pre-processing stage turns out to provide the best classification results. Classification performance does not improve when we included stop-word removal that cleaned data from unnecessary terms, nor when lemmatization which picked the basic form of words in the text. Classification performance is influenced by the addition of bigrams in the feature extraction process, but not affected by further addition of trigrams. Therefore, the authors recommend the use of unigram and bigram together during the feature extraction process.

References

[1] J. R. Pentheny, "The Influence of Movie Reviews on Consumers," University of New Hampshire, 2015.

[2] K. Tsutsumi, K. Shimada, and T. Endo, "Movie Review Classification Based on a Multiple Classifier *," *Proc. 21st pacific Asia Conf. Lang. Inf. Comput.*, no. 2007, pp. 481–488, 2007.

[3] T. P. Sahu and S. Ahuja, "Sentiment analysis of movie reviews: A study on feature selection & classification algorithms," in *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, pp. 1–6.

[4] S. K. Saritha, "Methods for Identifying Comparative Sentences," *Comput. Appl.*, vol. 108, no. 19, pp. 23–26, 2014.

[5] P. Das and S. Sharma, "An Entropy Based Effective Algorithm for Data Discretization," vol. 4, no. 3, 2017.

[6] H. Houngbo and R. E. Mercer, "An automated method to build a corpus of rhetorically-classified sentences in biomedical texts," in *Proceedings ofthe First Workshop on Argumentation Mining*, 2014, pp. 19–23.

[7] A. M. F. Al Sbou, A. Hussein, B. Talal, and R. A. Rashid, "A Survey of Arabic Text Classification Models," vol. 8, no. 6, pp. 4352–4355, 2018.

[8] S. Wang and C. Manning, "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification," *Proc. 50th Annu. Meet. Assoc. Comput. Linguist.*, vol. 94305, no. July, pp. 90–94, 2012.

[9] S. Vijayarani, J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining - An Overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[10] P. Bo and L. Lee, "Movie Review Data," 2004. .

[11] A.-H. Tan, "Text Mining: The state of the art and the challenges," *Proc. PAKDD 1999 Work. Knowl. Disocovery from Adv. Databases*, vol. 8, pp. 65–70, 1999.

[12] G. Miner, J. Elder, T. Hill, R. Nisbet, D. Delen, and A. Fast, *Practical Text Mining and Statistical Analysis for Non - Structured Text Data Applications*. Waltham: Elsevier, 2012.

[13] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014.

[14] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, and M. W. Mahoney, "Feature Selection Methods for Text Classification," *KDD*, pp. 230–239, 2007.

[15] T. Arifin and A. Herliana, "Optimasi Metode Klasifikasi Dengan Menggunakan Particle Swarm Optimization Untuk Identifikasi Penyakit Diabetes Retinopathy," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 2, pp. 77–81, 2018.

[16] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach _ Elsevier Enhanced Reader.pdf," *Expert Syst. with Appl.*, pp. 117–126, 2016.

[17] Y. Heights, "Class-Based n-gram Models of Natural Language Iwl)'" Pr (Wk Iw - -l). Wk," *Comput. Linguist.*, no. 1950, 1992.

[18] I. Rish, "An empirical study of the Naïve Bayes classifier," *Empir. methods Artif. Intell. Work. IJCAI*, vol. 22230, no. JANUARY 2001, pp. 41–46, 2001.

[19] H. Bhavsar and A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning," *Int. J. Soft Comput. Eng.*, vol. 2, no. 4, pp. 74–81, 2012.

[20] K. Markham, "Simple guide to confusion matrix terminology," *Data School*, 2014. .

[21] Kuspriyanto, O. S. Santoso, D. H. Widyantoro, H. S. Sastramihardja, K. Muludi, and S. Maimunah, "Performance Evaluation of SVM-Based Information Extraction using τ Margin Values," *Int. J. Electr. Eng. Informatics* -, vol. 2, no. 4, pp. 256–265, 2010.

[22] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," *Proc. EMNLP-06, Sydney, Aust.*, 2006.

[23] S. Teufel and A. Athar, "Detection of Implicit Citations for Sentiment Detection," *Proc. ACL-12 Work. Discov. Struct. Sch. Discourse, Jeju Island, South Korea, 2012*, no. July, pp. 18–26, 2012.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Architecture of Backpropagation Neural Network Model for Early Detection of Tendency to Type B Personality Disorders

Cynthia Hayat*, Samuel Limong, Noviyanti Sagala
Department of Engineering and Computer Science
Krida Wacana Christian University
Indonesia
*cynthia.hayat@ukrida.ac.id

Abstract-Personal disorder is a type of mental illness. People with personal disorder cannot respond changes and demands of life in normal ways. Women with type B personal disorder tend to have high risk of violence. It is important to make early detection of this personal disorder, so that it can be anticipated properly. This paper reports an architecture model of back propagation neural network (BPPN) for early detection of type B personal disorder. The back propagation process divided into two phases, training and testing. The training process used 43 data and the testing process used 34 data. The output classified into 4 diagnosis categories of type B personal disorder, namely: anti-social, borderline, histrionic, and narcissistic. The optimal parameters of BPPN model consist of maximum epoch of 1000, maximum mu of 10000000000, increase mu of 25, decrease mu of 0.1, and neuron hidden layer of 25. The MSE of training is 3.07E-14 and MSE of testing is 1.00E-03. The accuracy of training is 90.7%, while the accuracy of testing is 97.2%.

Keywords: back propagation, early detection, neural network, personality disorders

1. Introduction

Mental Health of America defines personality disorder as a setback that occurs in the internal and external of the human self where the person tends to be inflexible, rigid and unable to respond to changes and the demands of life [1]. Meanwhile, according to Diagnostic and Statistical Manual (DSM) IV, personality disorders are patterns of experience and inner behavior that deviate significantly and causing disorder. [2]

The both definitions refer to the inflexibility of behavior patterns in thought so that it can produce an impact on personal social life. A study by Danie Martin de Barros and Antonio de Pádua Serafim, stated that personality disorders of antisocial and borderline are highly predicted to make violence [3]. Research conducted by Riikka Arola et al., found that borderline personality disorder is a factor of violence attacks by women [4].

Early detection of personality disorder is needed so that it can be anticipated properly. The lack number of experts in the field of clinical psychology causing difficulties for people to obtain the information, therapy, and treatment that they should have. Through this research, with the application of artificial neural networks, it is expected to be able to construct a modeling to initiate early diagnosis of the type B personality disorder tendency.

Artificial Neural Network (ANN) is a concept of artificial knowledge in the field of artificial intelligence that made by adopting the human nervous system in the brain. In the journal Disease-Free Survival Assessment by Artificial Neural Networks for Hepatocellular Carcinoma Patients after Radio Frequency Ablation [5] stated that the artificial neural network model can recognize patterns of data through the learning process and has been applied to attain medical decision support. The study conducted by Panpan Hu stated that artificial neural networks can be used to analyze patterns of psychological patients, where the ANN model developed has a high accuracy training level with an average of 98.2%. ANN has a good training ability with a sufficient prediction level and the method can support in determining the diagnosis based on the results of the analysis [6]. Elvia Budianita and Muhammad Firdaus reported the method of artificial neural networks use Learning Vector Quantization 2 (LVQ2) in diagnosing psychiatric illnesses in which has a good accuracy of 90% [7]. While Tri Nur Oktavia et.al, reported the expert system in diagnosing hysterical personality disorders. The results of the expert system are compared with the results of the experts with a percentage of accuracy of 83.01% [8]

The novelty of this study is the use of rule-based

interviews with the experts as input data during the training phase. And it is projected that this research can contribute to the discipline of clinical psychology by providing the results of analysis and diagnosis of type B personality disorder, so that preventive action can be taken as early as possible.

2. Research Methodology

a. BPNN Procedures

The procedure for developing the proposed BPNN model is as follows:

Step 1: Data preparation, conducted by collecting the necessary data. Data collection is performed in 2 ways, by collecting primary data and collecting secondary data. Primary data obtained from expert rule-based data. Whereas secondary data obtained from observation women at a Hospital in Jakarta and also through semi-structured interviews with experts, some psychiatric doctors and psychologists.

Step 2: Constructing a decision tree and rule based on the decision tree. Constructing rule-based is performed using if-then rules as in the expert system. Constructing a rule is based on data that has been collected.

Step 3: Training Stage: Symptom data will be analyzed to be the result of back propagation neural network training. Data from the analysis of symptoms are in the form of input data and output data which will become training data for artificial neural network architecture. The training data utilize the calculation of the average value to get the target value. The training process with a separate target value used for artificial neural networks can recognize each personality disorder separately [9]. The results of the training will be used to predict the tendency for personality disorders.

Step 4: The input value is changed to a random value with a range [0,1]. Inputs used in the system are data of symptoms that have been represented in the numerical form with variables 0 (No) and 1 (Yes).

Step 5: Determining the activation function and parameters used in the BPNN model. The activation function is used to control the output value to match a predetermined one. In this model, the activation function is binary *sigmoid*. Iteration will be stopped if the maximum value of the epoch is 1000, 3000, and 5000, the error goal is 0, the mean square error is close to 0 and the result of regressions is close to 1.

Step 6: Determining the number of hidden layer neurons from 1 hidden layer used. Hidden layer neurons are determined by trial and error, which means that the fastest and most precise learning outcomes will determine the number of hidden layer neurons. The variations in the number of hidden layer neurons used are 15, 25, and 35. The Mu Increase variations used are 5, 10, and 25. The Mu Decrease variations are 0.05, 0.1, and 1. The training function is *trainbr* (Bayesian Regulation Back Propagation)

Step 7: Testing parameters in the BPNN model. Where the weight of ANN from the training phase will be implemented into the testing phase by inputting the symptoms.

Step 8: The output result is a type of personality disorder tendencies.

b. Data Collection

This research is a qualitative research as the symptoms of type B personality disorder are obtained from interviews and processed into ruled based. The research instrument is a tool for collecting data needed for research. The type of research instrument used was interviews. While quantitative data are questionnaires in the form of tendencies of personality disorder type B in respondents. Data is collected in two ways including:

1. Primary data, obtained from rule-based data based on interviews with experts.
2. Secondary data, data of observations and semi-structured interviews conducted to two fields of expertise, psychiatric doctors and psychologists.

The data collection includes:

1) Type of Personality Disorders
Data of types of personality disorders are data according to the Diagnostic and Statistical Manual of Mental Disorders, presented in table 1. [2]

Table 1. Data on type B of personality disorders

Target Code	Type of Personality Disorders Type B
T1	Borderline
T2	Antisocial
T3	Narcissistic
T4	Histrionic

The type B of personality disorders are classified into 4 types, as follows:

a) Antisocial is a pattern of ignorance, and violation of the rights of others.
b) Borderline is a pattern of instability in interpersonal relationships, self-image and marked by impulsiveness.

c) Histrionic is a pattern of emotional and excessive attention.
d) Narcissistic is a pattern of grandeur, the need for admiration, and lack of empathy.

2) General Symptoms Data
General symptom data are symptom data that are visible and also felt by the patient [10]. General disorder data is presented in Table 2.

3) Type B Personality Disorder Symptoms Data
Personality disorder data are symptoms that are felt by the patient. These data are classified according to predetermined types of personality disorders. Data on personality disorders can be seen in Table 3.

Table 2. General Symptoms Data

1	The ability to understand yourself, others and events are weak
2	Emotional frequency range to the weak emotions (too far)
3	Weak interpersonal function
4	Weak emotional control
5	Inflexible in various fields of social life
6	Social life between people both from various environments is not good

Table 3. Symptoms Data

No.	Symptoms	Type of Disorder
G1	Excessive imagination or avoidance of reality	Borderline
G2	Pattern and intensity of unstable interpersonal relationships	
G3	Unstable self-identity	
G4	Impulsive, at least in two areas that have the potential to self-destruct (shopping, sex, drug abuse, careless driving, greedy)	
G5	Repeated suicidal behavior or behavior that hurts others/him/herself	
G6	Mood instability	
G7	Chronic empty heart feeling	
G8	Emotions that are not in the atmosphere/difficult to control anger	
G9	Severe dissociative symptoms	
G10	Often violates the norm	Antisocial
G11	Lies often, likes to use aliases, deceives others for personal gain	
G12	Impulsive when plans are not achieved	
G13	Easy to have bad-tempered and aggressive, especially in physical fights	
G14	Ignores the security of him/herself or others	
G15	Ignores the security of him/herself or others	
G16	Be indifferent	
G17	Exaggerates achievements and talents, hoping to be recognized as superior without commensurate achievement	Narcissistic
G18	Busy with fantasies of success, strength, brilliance, beauty, or unlimited love	
G19	Believes that he/she is “special” and unique, only wants to be recognized with other people or special status (or institutions)	
G20	Requires excessive admiration	
G21	Has the feeling of excessive rights	
G22	In exploitative inter-commercials, that is, utilizes other people to achieve their own goals	
G23	Lack of empathy: do not want to recognize the feelings and needs of others	
G24	Often envies others or believe that others envy him	
G25	Shows arrogant behavior	

No.	Symptoms	Type of Disorder
G26	Uncomfortable in a situation where she/he is not the center of attention	Histrionic
G27	Interaction with others is often characterized by seductive and provocative sexual behavior	
G28	Shows expressions of emotions that are fast shifting and superficial	
G29	Consistently uses physical appearance to attract attention	
G30	Has a style of speech that is too impressionistic and lacks detail	
G31	Shows self-dramatization, theatrics, and excessive emotional expression	
G32	Predictable, i.e. easily influenced by other people or circumstances	
G33	Thinksof a more intimate relationship/wants more intimacy	

c. Decision Tree & Rule-based

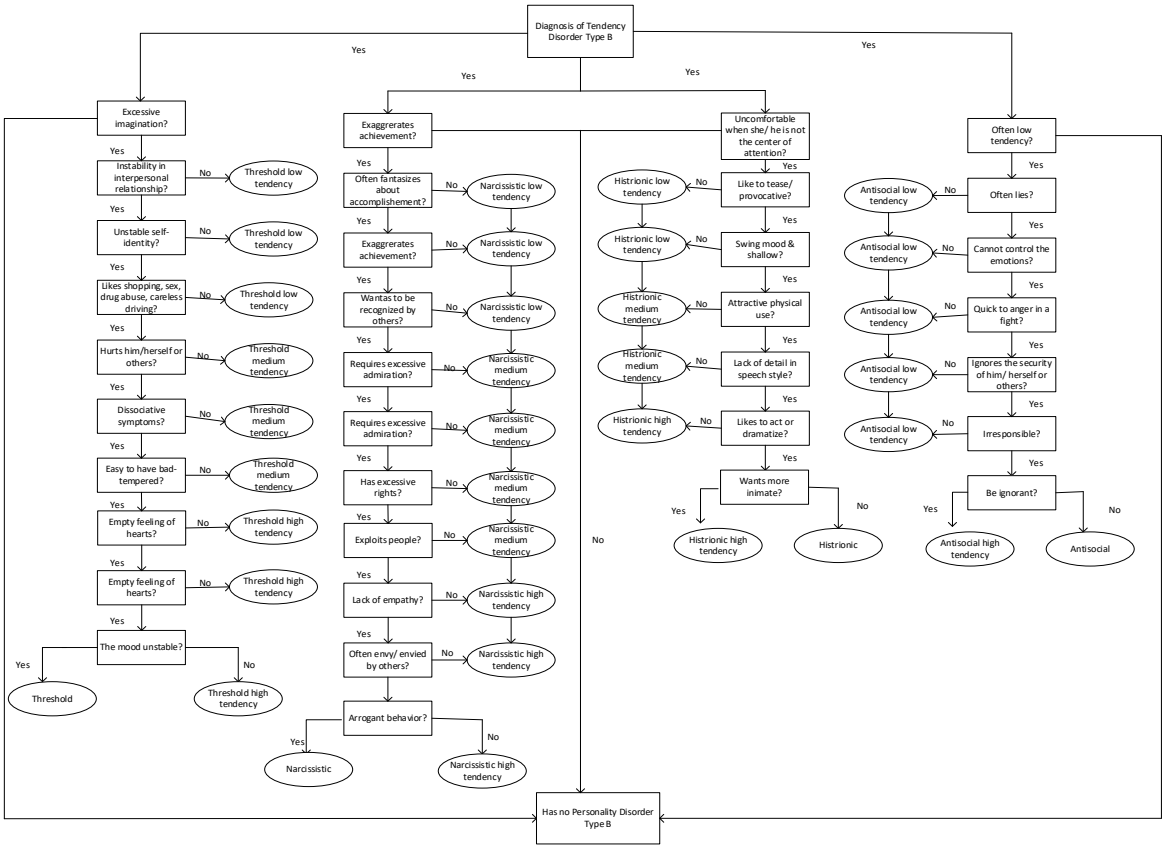


Figure 1. Decission Tree

The decision tree of symptoms and personality disorders is formulated as a tool in composing decisions (decision support tool) and identifying the relationship between the factors that influence a setback. Figure 1 illustrates the decision tree of symptoms and personality disorders.

The nodes in the decision tree are symptoms, while the bottom nodes are decisions. Furthermore, rule-based is made based on the decision tree. Rule-based construction will be used as an input layer in the architecture modeling of BPNN.

Rules-Based System

a) Rule 1
IF the ability to understand oneself, others and events is weak
OR the emotional frequency range in emotions is weak (too far)
OR the interpersonal function is weak
OR emotional control is weak
OR is not flexible in various fields of social life
OR social life between people from both different environments is not good
THEN checks for personality threshold/ antisocial/histrionic/narcissistic type
ELSE
THEN tends personality disorders

b) Rule 2: check borderline personality disorder
IF excessive imagination/avoid reality

OR unstable and intense patterns of marked interpersonal relationships that are unstable
OR self-identity is unstable
OR is impulsive, at least in two areas that have the potential to self-destruct (shopping, sex, drug abuse, careless driving, greedy)
OR repeats suicidal behavior/behavior that hurts others or him/herself
OR mood instability
OR chronic empty feeling of heart
OR emotions that are not in the atmosphere/ difficult to control anger
OR severe diss-associative symptoms
THEN likelihood of borderline personality disorder

b) Rule 3: check personality disorders
IF often violates the norm
OR often lies, likes to use aliases, deceives others for personal gain
OR is impulsive when plans are not achieved
OR is quick to anger and aggressive, especially in physical fights
OR ignores the security of him/herself or others
OR is not responsible
OR be indifferent
THEN antisocial personality disorder tendencies

c) Rule 4: check narcissistic personality disorder
IF exaggerates achievement and talent, hoping to be recognized as superior without commensurate achievement
OR is preoccupied with fantasies of success, strength, brilliance, beauty, or infinite love
OR believes that he/she is "special" and unique, only wants to be recognized with other people or special status (or institutions)
OR requires excessive admiration
OR has a feeling of excessive rights
OR is exploitative inter-commercial, which is utilizing other people to achieve their own goals
OR Lack of empathy: do not want to recognize the feelings and needs of others
OR often envies others or believes that others envy them
OR shows arrogant behavior
THEN tendency for narcissistic personality disorder

d) Rule 5: check histrionic personality disorder
IF is uncomfortable in situations where it is not the center of attention
OR interactions with others are often characterized by seductive and provocative sexual behavior
OR displays emotional expressions that are fast shifting and superficial
OR consistently uses physical appearance to attract attention
OR has a speaking style that is too impressionistic and lacks in detail
OR shows excessive self-dramatization, theatrics, and emotional expression
OR is predictable, that is easily influenced by other people or circumstances
OR consider relationships more intimate/want more intimate
THEN tendency for histrionic personality disorder

d. BPNN Architectural Model Design

Artificial Neural Network is a form of innovative learning, statistical models that mimic neuronal functions, able to identify patterns and separate them linearly by providing weight values numeric input for each and adjust it to the data sample [11]. Artificial Neural Networks can use parallel processing to predict solutions to complex variable data. The design of the BPNN architecture model for early detection of personality disorder tendencies is shown in Figure 2.

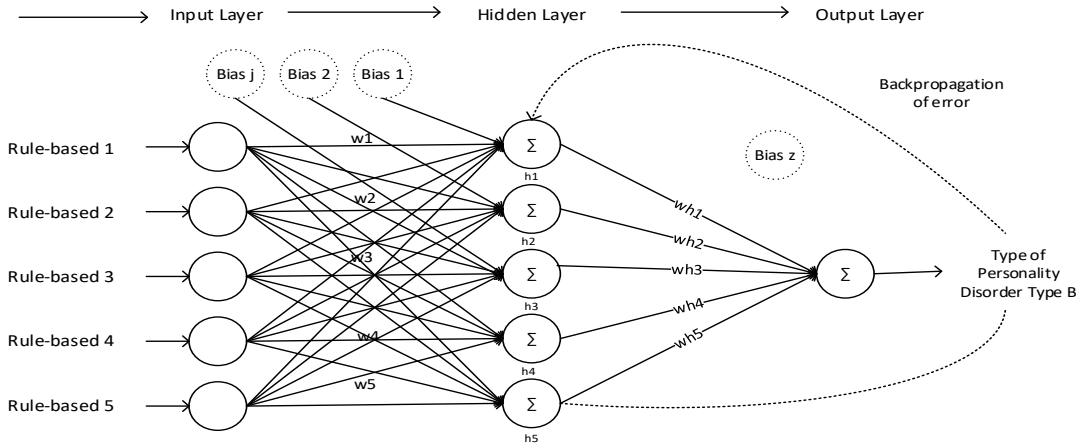


Figure 2 BPNN Architecture Model of Detection of Type B Personality Disorders

The architecture consists of three layers: input layer, hidden layer, and output layer. The back propagation learning algorithm consists of three stages: forward propagation of input signals, back propagation of errors, and weighting of updates. By spreading the input signal to the front (output), the first back propagation algorithm calculates the output for each vector included in the training data set. The next process is modifying the weights of each input vector. An error in the output neuron spreads in the reverse direction from the output layer to the input layer [12]

3. Discussion Results

a. Application of Artificial Neural Networks

The symptoms obtained from experts are processed into the training data. The processing of symptoms data becomes an input parameter for the training process. Meanwhile, the output data are in the form of personality disorder type B. Input data scale is in the form of 0 and 1, 0 for not having these symptoms and 1 for those who have these symptoms. For the output in the form of a scale from 0 to 1, the higher the value, the higher the tendency. In the target data, it is made into four targets, each of which represents symptoms of personality type B.

The data training process is the first step conducted to determine training data to get the results of the training data. The stage of the training divided into several processes, as follows:

1) Establishing training data, training data are obtained from the processing of symptoms.
2) Determining the activation function, the number of hidden layers and the training function. Determining the number of hidden layers according to Jeff Haeton. Hidden layers can be determined based on the complexity of the application, the rules are as follows:
   - Problems representing linear programs, so the number of hidden layers is only 0.
   - Problems representing functions that contain a mapping from space to another space, so the number of hidden layers is 1.
   - The problem represents that it can estimate all the mapping, then the number of hidden layers is 2. For more than 2 hidden layers, learning automation is needed.
3) Entering the maximum variable *epoch*, a number of *neurons*, *error goal*, *max fail*, *mu ins,* and *mu dec*.
4) It is a process of trial and error (repeating until getting optimal results), if we get poor training results, then the variables entered into process four and do the retraining process.
5) It is a training process.
6) Saving the results of the training in the form of weights for the training to be used in the application.

Table 4. Training and Testing Results

No	Epoch	Neuron Number	Mu Incr	Mu Decr	MSE Training	MSE Testing	Regr Train	Regr Test	Train Accuracy (%)	Test Accuracy (%)	Time
1	1000	15	5	0.05	1.04E-14	2.10E-03	1	0.9937	86	88.2	0:00:02
2	1000	15	5	0.1	4.704E-14	0.0022	1	0.9923	81.4	79.4	0:00:02
3	1000	15	5	1	8.62E-17	1.90E-03	1	0.9937	90.7	91.2	0:00:03
4	1000	15	10	0.05	1.51E-15	2.90E-03	1	0.9921	88.4	88.2	0:00:01
5	1000	15	10	0.1	1.37E-14	2.80E-03	1	0.9920	86	73.5	0:00:02
6	1000	15	10	1	1.83E-12	3.10E-03	1	0.9891	88.4	76.5	0:00:01
7	1000	15	25	0.05	2.28E-14	3.30E-03	1	0.9889	90.7	76.5	0:00:02
8	1000	15	25	0.1	1.72E-15	3.00E-03	1	0.9908	90.7	82.4	0:00:01
9	1000	15	25	1	1.24E-17	2.30E-03	1	0.9918	83.7	82.4	0:00:09
10	1000	25	5	0.05	6.41E-15	9.27E-04	1	0.9970	83.7	94.1	0:00:14
11	1000	25	5	0.1	4.62E-16	8.82E-04	1	0.9966	86	88.2	0:00:08
12	1000	25	5	1	2.28E-14	8.25E-04	1	0.9971	88.4	94.1	0:00:57
13	1000	25	10	0.05	1.33E-13	1.10E-03	1	0.9969	86	88.2	0:00:07
14	1000	25	10	0.1	4.95E-16	9.94E-04	1	0.9966	83.7	91.2	0:00:08
15	1000	25	10	1	1.90E-14	8.97E-04	1	0.9972	81.4	91.2	0:00:04
16	1000	25	25	0.05	6.84E-15	8.61E-04	1	0.9971	90.7	91.2	0:00:09
17	1000	25	25	0.1	3.07E-14	1.00E-03	1	0.9971	90.7	97.1	0:00:04
18	1000	25	25	1	1.23E-13	1.00E-03	1	0.9971	86	94.1	0:00:03
19	1000	35	5	0.05	1.15E-14	5.14E-04	1	0.9985	88.4	97.1	0:00:22
20	1000	35	5	0.1	1.59E-15	4.73E-04	1	0.9983	90.7	97.1	0:00:21

No	Epoch	Neuron Number	Mu Incr	Mu Decr	MSE Training	MSE Testing	Regr Train	Regr Test	Train Accuracy (%)	Test Accuracy (%)	Time
21	1000	35	5	1	1.14E-14	4.69E-04	1	0.9982	83.7	94.1	0:00:21
22	1000	35	10	0.05	2.69E-16	5.31E-04	1	0.9985	90.7	97.1	0:00:11
23	1000	35	10	0.1	4.35E-17	4.70E-04	1	0.9984	90.7	97.1	0:00:23
24	1000	35	10	1	5.39E-15	5.11E-04	1	0.9982	81.4	91.2	0:00:11
25	1000	35	25	0.05	1.51E-14	4.68E-04	1	0.9984	83.7	97.1	0:00:18
26	1000	35	25	0.1	3.37E-15	6.09E-04	1	0.9984	83.7	97.1	0:00:18
27	1000	35	25	1	4.97E-16	4.96E-04	1	0.9984	83.7	97.1	0:00:18
28	3000	15	5	0.05	3.44E-14	2.00E-03	1	0.9932	83.7	88.2	0:00:03
29	3000	15	5	0.1	1.81E-15	2.30E-03	1	0.9921	79.1	88.2	0:00:02
.											
.											
.											
75	5000	35	5	1	3.56E-15	4.94E-04	1	0.9982	88.4	94.1	0:00:35
76	5000	35	10	0.05	2.39E-14	5.94E-04	1	0.9984	86	97.1	0:00:11
77	5000	35	10	0.1	1.03E-16	5.74E-04	1	0.9978	79.1	91.2	0:00:23
78	5000	35	10	1	6.72E-15	4.78E-04	1	0.9984	86	94.1	0:00:17
79	5000	35	25	0.05	9.51E-15	4.93E-04	1	0.9985	86	97.1	0:00:19
80	5000	35	25	0.1	7.07E-15	5.14E-04	1	0.9986	86	97.1	0:00:10

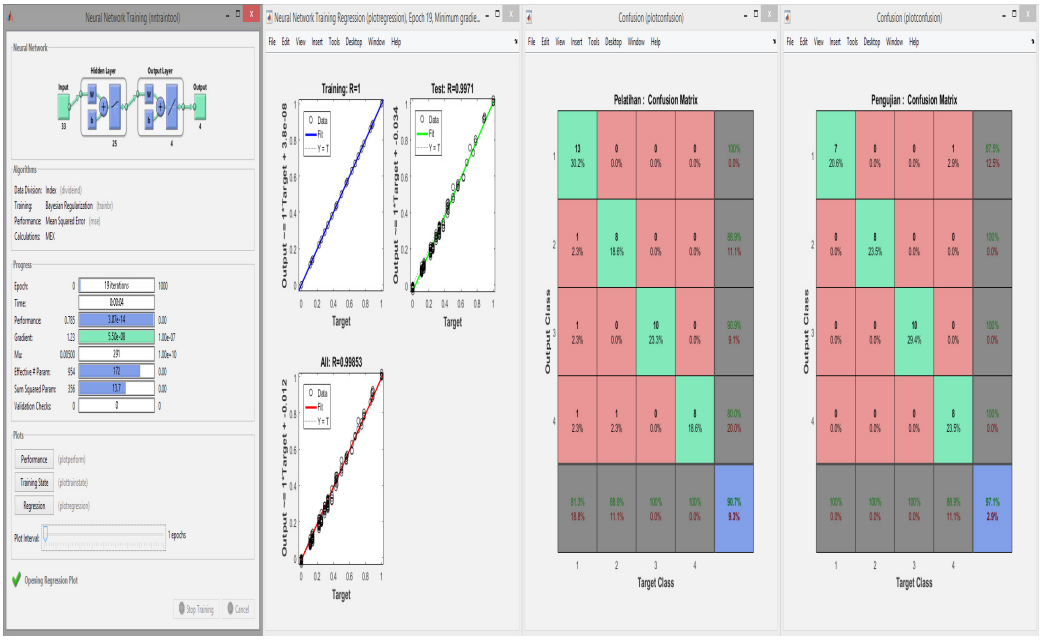


Figure 3. Training Results

Table 4 indicates the best value of the criteria with high regression values, low MSE values (close to 0), training accuracy values (close to 100%) and short training time. The training parameters used in the development of architectural models are optimal training parameters. Training and regression results are shown in Figure 3 and Figure 4.

Figure 4 presents the testing regression results. Regression of artificial neural network using *Matlab* is how much data that does not deviate, and data that is not outside the line (fit) is non-distorted. Data that deviates from the line (fit) is data that is not recognized by artificial neural networks. The graph can also be viewed as the results of these deviations, the data (round) which is in the blue line (fit) is data that does not deviate. These data are already parallel with the line (fit). It can be concluded that the training process is successful with a regression value of 1 of 1.

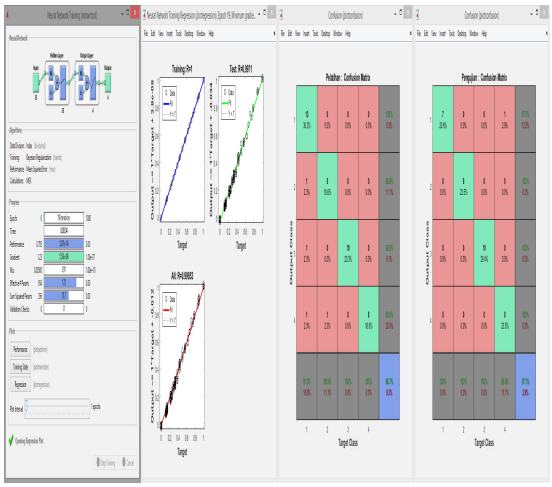


Figure 4. Testing Regression Result

Table 5. Optimal Training Parameters

Training Parameter	Value Parameter
Maximum *Epoch*	1000
Mu Max	10000000000
Mu increase	25
Mu decrease	0.1
Error goal	0
Max fail	0
Neuron Hidden Layer	25

The result of the training MSE is 3.07E-14, the result of the MSE Testing is 1.00E-03, the result of the training regression is 1, the result of the testing regression is 0.9971 and the taken time is 0:00:04 with the training accuracy level of 90.7% and the accuracy of the testing of 97.1%.

The optimal training and testing parameters will be the default parameters on the training page to make it easier for users. These parameters can be seen in Table 5.

4. Conclusion

The optimal value of BPNN architecture parameters for early detection of personality disorder in this research are: maximum epoch: 1000, mu max: 10000000000, mu increase: 25, mu decrease: 0.1, error goal: 0, max fail: 0, and neuron hidden layer: 25. MSE of training is 3.07E-14, MSE of testing is 1.00E-03. The training regression is 1, while the testing regression is 0.9971 with the taken time of 0:00:04. The accuracy of training is 90.7%, and testing accuracy of 97.2%, showing that the architecture of BPPN model provides very good accuracy.

Training on the multi-layer back propagation used activation function *sigmoid bayesian trainbr* that provides unlimited range of input value and limited range of 0 to 1 for output value. Thus, the data input has to be normalized to the range of 0 to 1, so the data output can be adjusted into the range of input value. Further research with larger data is needed to examine further the performance of the architecture of BPNN model.

Reference

[1] M. H. America, "Personality Disorder," http://www.mentalhealthamerica.net/conditions/personality-disorder, America, 12th of September 2017.

[2] A. P. Association, Diagnostic and Statistical manual of mental Disorders, fourth Edition: primary care version (DSM-IV-PC), Washington DC: American Psychiatric Association, 1995.

[3] D. Barros and A. Serafim, "Association between personality disorder and violent behavior pattern," *Forensic Science International,* vol. 179, no. 1, pp. 19-22, 2008.

[4] R.Arola, H.Antila, P.Riipinen, H.Hakko, K.Riala and L.Kantojarvi, "Borderline personality disorder and violent criminality in women; A population base follow-up study of adolescent psychiatric inpatients in Northen Finland," *Forensic Science International,* vol. 266, pp. 389-395, 2016.

[5] F. W. Chiueng, J. W. Yu, C. L. Po, H. W. Chih, F. P. Shinn and W. C. Hung, "Disease-Free Survival Assessment by Artificial Neural Networks for Hepatocellular Carcinoma patients after Radiofrequency Ablation," *Journal of The Formosan Medical Association,* vol. 116, no. 10, pp. 765-773, October 2017.

[6] P. Hu, "Identification of Psychological Paterns using Neural Networks Approach," Digital

Commons University of Nebraska Lincoln, Nebraska, 2010.

[7] E. Budianita, S. Sanjaya, F. Syafria and Redho, "PenerapanMetode Learning Vector Quantization (LVQ) untuk mementukan gangguan kehamilan trisemester I," *JurnalSains, teknologi, danindustri* , vol. 15, no. 2, pp. 144-151, 2nd of June2018.

[8] T. Oktavia, D.H.Satyareni and E. Jannah, "Rancang Bangun Sistem Pakar untuk Mendiagnosis Gangguan Kepribadian Histerik Menggunakan Metode Certainty Factor," *Jurnal Ilmiah Teknologi Sistem Informasi,* vol. 1, no. 1, pp. 15-23, January 2015.

[9] S.A.Oyewole and O.O.Olugbara, "Product Image Classification using Eigen Colour Feature with Ensemble Machine Learning," *Egyptian informatics Journal,* vol. 19, no. 2, pp. 83-100, July 2018.

[10] L. M. James and J. Taylor, "Impulsivity and Negative Emotionality Associated with Subtance use Problems and Cluster B Personality in College Students," *Addictive Behaviors,* vol. 32, no. 4, pp. 714-727, 2007.

[11] A. A. Pradika, J. Jusak and J. Lemantara, "Sistem pakar untuk Mendiagnosis gangguan jiwa skizofrenia menggunakan metode fuzzy expert system (studi kasus RS.Jiwa Menur Surabaya)," *Jurnal JSIKA,* vol. 1, no. 1, 2012.

[12] H.takizawa, T.Chida and H.Kobayashi, "Evaluating Computational Performance of Backpropagation Learning on Graphics Hardware," *Electronic Notes in Theoretical Computer Science,* vol. 225, pp. 379-389, 2008.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

The Design of Exploratory and Preprocessing of Event Log Data in Online Learning Activities Based on Moodle LMS for Process Mining

Demaspira Aulia*, Indra Waspada
Departemen Ilmu Komputer/Informatika, Fakultas Sains dan Matematika
Universitas Diponegoro
Semarang
*demaspira@student.undip.ac.id

Abstract-Process Mining is one of the sub-studies of Data Mining that focuses on the events of a system. An area that benefits from process mining is education, especially online learning. This study used Moodle as a platform to provide online event activity log data in online learning. Moodle-based process mining requires several stages that are not easily understood directly by teachers. As a solution, some efforts are needed to integrate Moodle with process mining. This study built an application that could contribute to the Preprocessing and Exploratory Data Analysis (EDA) stages of Moodle event log data – as an important part of the process mining stage. Preprocessing was implemented by using the simple heuristic filtering method, while EDA was employed through visualization using flow control and dotted charts. Eventually, the application built in this study successfully performed preprocessing in Moodle event log data and could display the results visually, as a tool of control flow analysis and dotted chart analysis.

Keywords: exploratory data analysis, Moodle, process mining

1. Introduction

The use of Data Mining (DM) has been often used in several fields. One of the fields is the educational field. The goal of implementing the DM process is to find interesting patterns from large data [1]. The use of DM in the educational field is also known as Educational Data Mining (EDM). EDM has two types of objectives which are to improve the learning process and to gain an understanding towards the learning phenomenon [2].

In addition to using DM, process mining has been recognized to be used in various fields, especially in business field. Process mining is one of the new research disciplines between machine learning and data mining using process modelling and analysis. The objectives of process mining included finding, monitoring, and improving the processes that occur by taking existing knowledge from the event log obtained from the system [3]. From these objectives, the process mining can also be applied to other fields that are in education – known as Educational process mining.

EDM employs an event log that is obtained to find, monitor, and improve the educational process that is applied [4]. In the implementation of EDM, the perspective that is the most often used is control flow perspective – focuses on the sequence of existing activities [5]. By using control flow perspective, we could see the application of the patterns in teaching and learning process. However, the control flow used in this study is still limited to EDA and does not cover the process mining part. Another perspective that is frequently applied in educational process mining is a performance perspective that is depicted in the form of a dotted chart. The dotted chart is used to observe the flow of events occurs based on the event log [6].

One of the implementations of online learning can be applied to the Learning Management System (LMS). One of LMSs that is often used is Moodle. Moodle is designed to help teachers who try to make quality online learning. Moodle is used in various universities, schools, and companies throughout the world. Moodle helps teachers to arrange lessons in various ways and integrate lessons with collaborative activities [7]. Moodle saves data in a MySQL database – storing an event log that occurs on the system is done into several tables in the database [8]. By accessing the database, the event log can be obtained from the learning process.

A previous study conducted by Slaninova *et al.* [8] uses data from event log Moodle to analyze students' behavior in Moodle; group them based on behavioral similarities between students; and also visualize the relationship between students and the groups. Besides, Bogarin *et al.* [9] used clustering algorithms on students' interaction data

in Moodle to improve the existing process mining models. Another study conducted by Bogarin *et al.* [10] also use clustering and process mining to find navigation paths of students in Moodle. A research conducted by Juhanak *et al.* [11] analysed the patterns of students' behaviour and interaction – which were different in each quiz attempt in LMS. Therefore, by using process mining, the sequence of activities carried out by students at the time of quizzes can be obtained.

Exploratory Data Analysis (EDA), according to Willems [12] is used to answer questions and business assumptions as well as to make hypotheses for further analysis. On top of that, the function of EDA is to prepare data for Modelling. It is known that by having good knowledge regarding the data used could provide the answers needed or could build intuition to interpret the results of modelling that were conducted. Stages in EDA include describing the data to be used, taking samples from the data, overcoming problems in the data, understanding the features in the data, and understanding the pattern of the data.

The role of EDA and preprocessing as the initial stages of using process mining is considered important [13][14], especially because there are several problems in the implementation of process mining including the poor quality data and other issues related to data quality which had not been frequently raised by researchers [15]. The learning process that became a focus in this research was the process of quizzes for students or participants in Moodle. Thus, we could find out the flow and order of the quiz occurred in Moodle.

Although there have been several studies that use process mining in education, especially those using Moodle, there was no integration between process mining and Moodle which results in the difficulty for teachers and researchers to obtain data and information from process mining results on the Moodle data.

The contribution given to this research is in the form of application design as an initial stage of the integration of process mining on Moodle. The integration conducted in this study was still simulated and not directly integrated into Moodle. Additionally, the data employed still referred the exported data from Moodle. In this application, EDA was performed on the event log data of Moodle, which was done as the initial stage before the conducting process mining. The stages in building the application commenced from data identification, requirement analysis and definition, case studies identification, application development result, application testing and experiments with specified scenarios to prove that the application can be applied to original data from Moodle.

2. Method

a. Application Architecture

The application developed in this study is a part of the system that was built to integrate Moodle with process mining. The aim of this application was to carry out the EDA and preprocessing in the event log prior to the implementation of process mining. The data used was obtained by downloading the event log through Moodle and uploading it to the application.

This application is divided into two parts – including web server and client. The client is used as an intermediary between users and web servers for preprocessing and EDA. Moreover, the client also has a function to display graphs based on data sent by the webserver. The web server acts as a data processor and handles all other processes that are being executed.

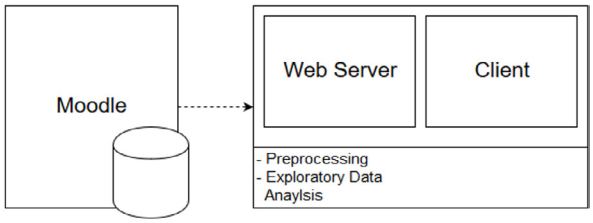


Figure 1. Application Architecture

b. Data Identification

The data used in this application were online learning data obtained in the form of an event log from Moodle. The data used were in the form of CSV files. The number of attributes needed to use this application did not have a maximum limit but had at least three main attributes as mentioned by Aalst [3]: case id, event, and timestamp. The case id and event can change according to the context [15]. By this, this study has two types of scenarios using different case id and event. An explanation of the attributes in the data can be seen in Table 1.

Table 1. Description of the event log data attributes from Moodle

Attributes	Descriptions
Time	Time marker when the event occurs
User full name	Complete name of the event user
Affected user	Name of the event target
Component	Marker of event types in general
Event context	Marker of the event conducted
Event name	Name of the event conducted
Description	Description of the event
Origin	Origin of the event
IP address	IP address of the event users

From the data, there was a problem with the Time attribute in which the value on that attribute could not be directly used as timestamp since the format was not related to ISO standard. Consequently, it made the application difficult to use the value of the Time. Therefore, in the preprocessing stage of the application, there was an option offered to change the format of

Time attribute to conform to ISO standard.

c. Requirements Analysis and Definition

To achieve the goals and solve the problems that is found in data identification, it is necessary to identify the required features that were used in this application.

Table 2. Results of features identification of ProM

No	Name of Feature	Description
1	Import data in csv format	Receiving the Moodle event log data in csv format
2	Data column setting	Choosing three required main columns including case id, event, and timestamp
3	Summary of data statistic	Displaying the information in the form of the statistic of start event, end event, and whole event of preprocessed data
4	Simple heuristic filtering	Conduct filtering using simple heuristic filtering
5	Control flow analysis	Visualization using *control flow perspective*
6	Dotted chart analysis	Visualization using the performance's perspective of the *dotted chart*. Visualization given consisted of two types of *dotted chart* including *dotted chart* based on absolute time and relative time.

Table 3. Additional features based on problems in data

No	Name of feature	Description
1	Preprocessing data	Preprocessing on data – consisting of time format conversion, alias and initial filter naming, column combination, quiz attempt calculation, column deletion, and data column setting.

The first step was analyzing the features of the existing process mining tool called ProM. ProM is an open source framework that supports various forms of process mining techniques. The version of ProM analyzed in this study was ProM Lite version 1.2. Based on the analysis results, the features of ProM that can be implemented in the application is informed in Table 2.

Table 4. SRS lists of the application

No	ID SRS	Description
1	SRS-F-01	Receive event log data in csv format
3	SRS-F-02	Preprocess data
4	SRS-F-03	Display summary of data statistic
5	SRS-F-04	Analyse control flow
6	SRS-F-05	Analyse dotted chart

Based on the results of the feature identification in Table 2 and Table 3, the features to be implemented will be used as a reference to determine the functional requirements of the application. The functional requirements are represented in the form of Software Requirement Specification (SRS) in which the SRS list of applications can be seen in Table 4.

The requirements of preprocessing data (SRS-F-02) includes the following parts:

1. Time format conversion
2. Data alias presentation
3. Column combination
4. Quiz attempt calculation
5. Column deletion
6. Data column setting
7. Filtering using simple heuristic filtering.

d. Case Study

The data used was lecture data in Basic System course in semester 1 of 2018/2019 Academic Year at the Department of Informatics, Diponegoro University. The lecture utilized the Moodle platform used by Undip Informatics, which could be accessed at https://ioclass.if.undip.ac.id/.

The data included event log from all activities conducted by lecturers and students. However, this study was focused on quiz work at IOClass only. The data used contained data starting from the beginning of lectures to the end of semester exams. The data has 9 columns and 178920 rows. The examples of the scores from the data used is informed in Table 5.

Table 5. The example of the content of event log data

Time	User full name	Affected user	Event Context	Component	Event Name	Description	Origin	IP address
24/09/2018, 15:45	SDAW	SDAW	Quiz: Quiz 5: Boole Algebra and Combinational Logic	Quiz	Quiz Attempt Viewed	The user with id '845' has viewed the attempt with id '9206' belonging to the user with id '845' for the quiz	web	182.1.68.53
24/09/2018, 15:44	SDAW	SDAW	Quiz: Quiz 5: Boole Algebra and Combinational Logic	Quiz	Quiz Attempt Viewed	The user with id '845' has viewed the attempt with id '9206' belonging to the user with id '845' for the quiz	web	182.1.68.53

Time	User full name	Affected user	Event Context	Component	Event Name	Description	Origin	IP address
24/09/2018, 15:44	SDAW	SDAW	Quiz: Quiz 5: Boole Algebra and Combinational Logic	Quiz	Quiz Attempt Viewed	The user with id '845' has viewed the attempt with id '9206' belonging to the user with id '845' for the quiz	web	182.1.68.53
24/09/2018, 15:43	SDAW	SDAW	Quiz: Quiz 5: Boole Algebra and Combinational Logic	Quiz	Quiz Attempt Viewed	The user with id '845' has viewed the attempt with id '9206' belonging to the user with id '845' for the quiz	web	182.1.68.53

The quiz setting applied to "Basic Systems" course was that each student was allowed to attempt many times on each quiz – yet there was a specified time limit. In case a quiz attempt reaches the time limit, it was automatically collected into the system. By applying this setting, it was possible for students to conduct the same quiz repeatedly. Also, the final quiz scores taken were the highest scores from the quiz's attempts conducted. This setting applied to all quizzes, in addition to the midterm and final semester exams.

e. Experiment Scenarios

The experiment used Moodle event log data from "Basic Systems" course. The experiment that was carried out consisted of two scenarios. The first scenario produced data that contained case id where participants with the whole quiz data as event and producing data containing case id where each quiz attempted by each participant and quiz data separated considered as event.

Table 6. Attribute mapping on scenario 1

Attribute	Form	Example
case id	participantName	Demaspira Aulia
event	eventName	Quiz attempt started

Table 7. Attribute mapping on scenario 2

Attribute	Form	Example
case id	participantName_ noAttempt	Demaspira Aulia_12
event	quizName_ eventName	Quiz 1_Quiz attempt started

The objectives of these two scenarios were to find possible quiz patterns on the Moodle platform and to analyse whether the patterns are in accordance with the actual events. Other objectives of both scenarios were to observe the duration the quizzes attempted by students and to find patterns that are considered abnormal.

The mapping of the two scenarios is informed in Table 6 and Table 7. The visible difference between the two scenarios is that in scenario 1, general data without any specific quiz details was the only visible result in scenario 1. Whereas, in scenario 2, specific data for each experiment attempted on each quiz was clearly observed.

3. Results

a. The Result of Application Development

The results of the application design explained in the previous chapter were then implemented into a web-based application using the Python and Typescript programming languages. Both programming languages are chosen based on their performance and ease of application on the web platform. Python is used to implement the data processing functions that contained in the application. Whereas, Typescript is utilized to display the data that was processed using Python into an easier and more understandable form.

The needs for receiving data in CSV format (SRS-F-01) was implemented into the web page which was used to upload the event log data employed. The data was stored on a storage server later on and was used as a reference for the further process of initial data. The interface – as the result of SRS-F-01 – can be seen in Figure 2.

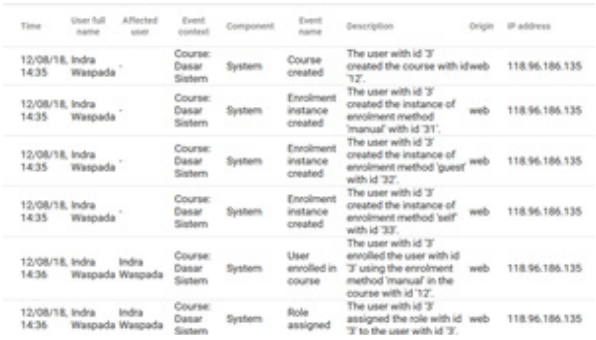
The needs of data preprocessing (SRS-F-02) are implemented into a web page consisting of six parts including time conversion, data alias assignment, attempt quiz calculation, merging two columns, columns deletion, and three main columns selection that was used as case id, event, and timestamp. The time conversion section has a function to convert the time format from a column into ISO format, data alias assignment has a function to pre-filter the values from a column that are not required. In addition, another name or abbreviated name was given to the values of a column to facilitate the user in reading the data. Subsequently, the function of the quiz attempt calculation was to calculate the number of quiz attempts conducted by participants.

This is intended to distinguish the same quiz attempt conducted by certain participants. Furthermore, merging two columns benefited to merge the scores from two different columns. Also, column deletion was used to eliminate unneeded columns and the selection of three main columns was intended to select the case id, event, and timestamp of the data used. Hence, other features of this application could be applied as well. In this case, filtering was also conducted by employing simple heuristic filtering – to select the values used at the start event, end event, and all events. This is intended to eliminate data

that may include outliers from existing data. The interface – result of SRS-F-02 – is illustrated in Figure 3 and the interface of simple heuristic filtering is shown in Figure 4.

The need of displaying the statistics of data (SRS-F-03) was used to show statistical information of data that had already been subjected to preprocessing and/or filtering. The information displayed included the number of existing cases, the total number of existing events, the number of classes of events, the frequency of start events, end events, and the whole events (all events) of existing cases. The interface of SRS-F-03 is depicted in Figure 5.

The need of control flow analysis (SRS-F-04) and dotted chart analysis (SRS-F-05) was used to provide an overview of the existing event flow using a control flow chart and dotted chart. The dotted chart consisted of two types including dotted chart using absolute time that was used to show how and when each case begins and how the flow occurs. Moreover, dotted chart using relative time benefited to observe the performance of the time of each existing case. The interface implementation of SRS-F-04 and SRS-F-05 are shown in Figure 6 and Figure 7, respectively.



Time	User full name	Affected user	Event context	Component	Event name	Description	Origin	IP address
12/08/18, 14:35	Indra Waspada		Course: Dasar Sistem	System	Course created	The user with id '3' created the course with id '12'.	web	118.96.186.135
12/08/18, 14:35	Indra Waspada		Course: Dasar Sistem	System	Enrolment instance created	The user with id '3' created the instance of enrolment method 'manual' with id '31'.	web	118.96.186.135
12/08/18, 14:35	Indra Waspada		Course: Dasar Sistem	System	Enrolment instance created	The user with id '3' created the instance of enrolment method 'guest' with id '32'.	web	118.96.186.135
12/08/18, 14:35	Indra Waspada		Course: Dasar Sistem	System	Enrolment instance created	The user with id '3' created the instance of enrolment method 'self' with id '33'.	web	118.96.186.135
12/08/18, 14:36	Indra Waspada	Indra Waspada	Course: Dasar Sistem	System	User enrolled in course	The user with id '3' enrolled the user with id '3' using the enrolment method 'manual' in the course with id '12'.	web	118.96.186.135
12/08/18, 14:36	Indra Waspada	Indra Waspada	Course: Dasar Sistem	System	Role assigned	The user with id '3' assigned the role with id '3' to the user with id '3'.	web	118.96.186.135

Figure 2. Result interface of SRS-F-01

case_id	task	timestamp	User full name	Event context	Component	Event name	n_attempt
147_1	Quiz: quiz 1 - pengenalan dunia digital_qas	2018-08-21 07:00:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qas	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:00:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:01:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:01:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:01:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1
147_1	Quiz: quiz 1 - pengenalan dunia digital_qav	2018-08-21 07:02:00	147	Quiz: quiz 1 - pengenalan dunia digital	Quiz	qav	1

Figure 3. Result interface of SRS-F-02

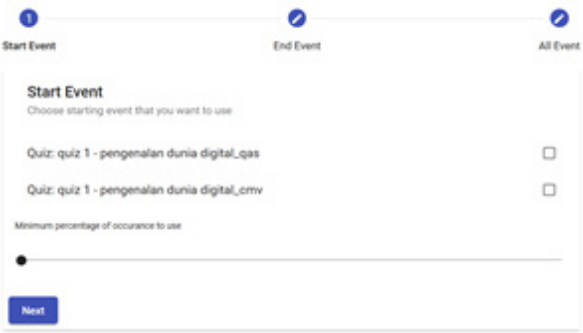


Figure 4. Filtering start event interface in SRS-F-02

Figure 5. Statistic interface of start event of SRS-F-03

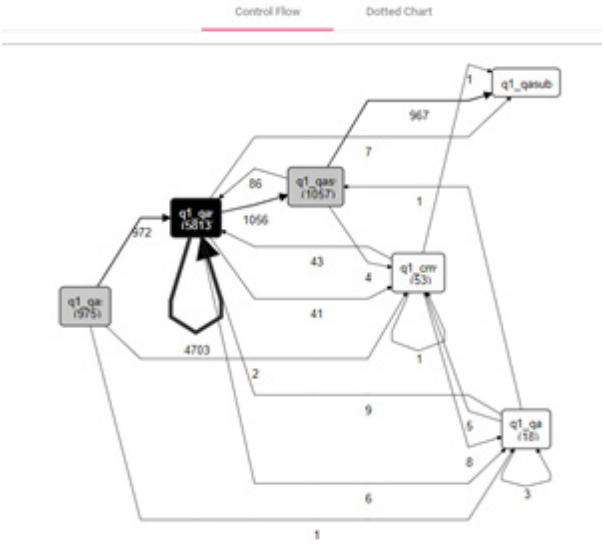


Figure 6. Interface implementation of SRS-F-04 (control flow analysis)

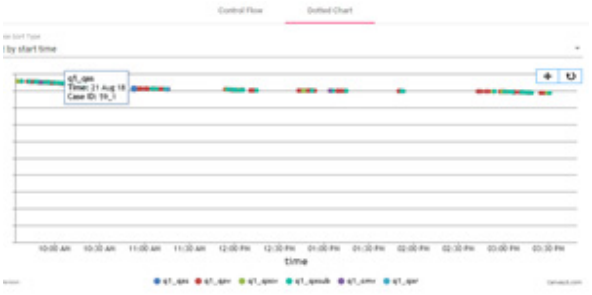


Figure 7. Interface implementation of SRS-F-05 (dotted chart analysis)

b. Experiment

The experiment was conducted using original Moodle data taken from online lecture data from IOClass Undip Informatics "Basic System" course. The experiment was conducted to prove that this application was able to run in accordance with predetermined needs using real data. Table 8 informs the steps conducted, the process worked on, and the features used in each trial scenario.

Table 8. Table of experiment steps of scenario 1 and 2

Scenario	*Preprocessing*	Used features	*Simple Heuristic Filtering*
Scenario 1	1. Converting time in the time column 2. Filtering the component column by simply filling in the Quiz value 3. Filtering user full name column by removing grades from a lecturer, assistant, or administrator 4. Giving alias to the value in the event name column 5. Giving alias in the form of a number to value in the user full name column 6. Deleting affected user column and the IP address 7. Selecting user full name column as the case id, event name column as event and time as the timestamp	- Time format conversion - Giving alias names - Column deletion - The setting of the data column	No filtering conducted in simple heuristic filtering
Scenario 2	1. Converting time in the time column 2. Filtering the component column by simply filling in the quiz value 3. Filtering user full name column by removing grades from the lecturer, assistant, and administrator 4. Giving alias to the value in the event name column 5. Giving alias to value in the event context column and conducting filtering process by only taking values of certain quiz 6. Giving aliases in the form of sequential numbers in the value in the User full name column 7. Calculating the number of attempts using user full name column as base column, event name column as count column, selecting quiz attempt started as start event, quiz attempt submitted as end event, and generating the n_attempt column 8. Merging the event name column with Event context using underscore delimiter (_) and then saving it with the Event name 9. Merging the user full name column with n_attempt column using underscore delimiter (_) and then saving it with case id name 10. Deleting affected user and the IP address columns 11. Selecting case id column as case id, event column as event, and time column as the timestamp	- Time format conversion - Giving alias names - Calculation of quiz attempts - Columns merging - Column deletion - Column setting - Simple heuristic filtering	The minimum occurrence value of the event that was used in start event and end event, was 90%. Also, for all event, the entire events were used

Table 9. Table of Aliases event name column

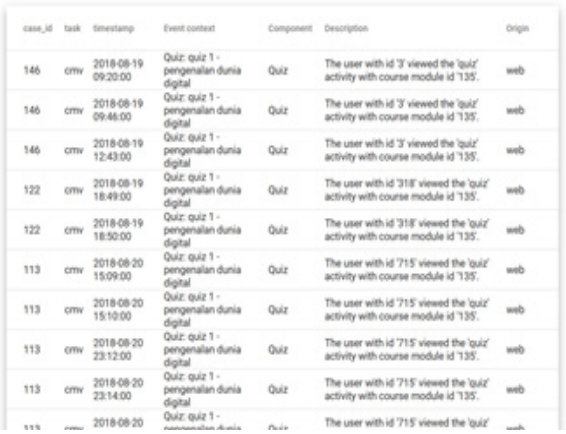
No	Initial score	Alias	Description
1	*Quiz attempt started*	Qas	Quiz is started
2	*Course module viewed*	Cmv	Participants view the module
3	*Quiz attempt viewed*	Qav	Working on the quiz
4	*Quiz attempt submitted*	Qasub	Quiz results are submitted
5	*Quiz attempt summary viewed*	Qasv	The summary of the quiz before it is submitted
6	*Quiz attempt abandoned*	Qaban	Quiz attempt aborted
7	*Quiz attempt reviewed*	Qar	Review the results of the quiz

Giving alias names in certain columns is intended to ease data reading. For the User full name column, the alias name is intended to maintain the confidentiality of quiz participants' identity. Details of the aliases for the event name column and event context column are shown in table 9 and table 10, respectively.

Table 10. Table of aliases in event context column

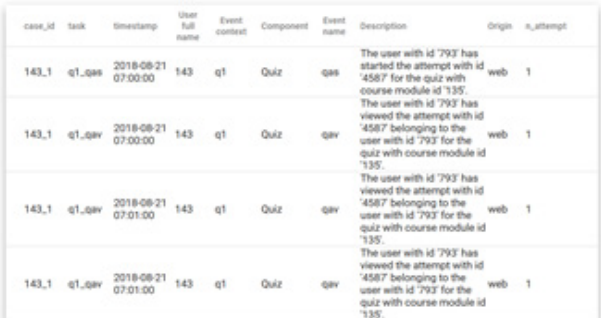
No	Initial Score	Alias
1	Quiz: quiz 1 – introduction to digital world	q1
2	Quiz: Quiz 2 : Digital system	q2
3	Quiz: Quiz 3: number conversion	q3
4	Quiz: Quiz 4: Basic logic	q4
5	Quiz: Quiz 5: Boole Algebra and combinational logic	q5
6	Quiz: Quiz 6: Comparator	q6
7	Quiz: Quiz 7: Combinational logic adder substractor	q7
8	Quiz: quiz 8: mux & ndcoder	q8
9	Quiz: Quiz 9: sequencial logic - FF	q9
10	Quiz: Kuis 10: Register, Counter, ROM	q10
11	Quiz: Responsi 1	qr1
12	Quiz: Responsi 2	qr2
13	Quiz: UTS 1	quts
14	Quiz: UTS 2	quts

It is observed in Table 16 that each Quiz responsi and UTS Quiz are given the same alias name – Quiz responsi is given the alias name qr1 and qr2 while UTS Quiz is given the alias name quts.



case_id	task	timestamp	Event context	Component	Description	Origin
146	cmv	2018-08-19 09:20:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '3' viewed the 'quiz' activity with course module id '135'.	web
146	cmv	2018-08-19 09:46:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '3' viewed the 'quiz' activity with course module id '135'.	web
146	cmv	2018-08-19 12:43:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '3' viewed the 'quiz' activity with course module id '135'.	web
122	cmv	2018-08-19 18:49:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '318' viewed the 'quiz' activity with course module id '135'.	web
122	cmv	2018-08-19 18:50:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '318' viewed the 'quiz' activity with course module id '135'.	web
113	cmv	2018-08-20 15:09:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '715' viewed the 'quiz' activity with course module id '135'.	web
113	cmv	2018-08-20 15:10:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '715' viewed the 'quiz' activity with course module id '135'.	web
113	cmv	2018-08-20 23:12:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '715' viewed the 'quiz' activity with course module id '135'.	web
113	cmv	2018-08-20 23:14:00	Quiz: quiz 1 - pengenalan dunia digital	Quiz	The user with id '715' viewed the 'quiz' activity with course module id '135'.	web
		2018-08-20	Quiz: quiz 1 -		The user with id '715' viewed the 'quiz'	

Figure 8. Preprocessing results of scenario 1



case_id	task	timestamp	User full name	Event context	Component	Event name	Description	Origin	n_attempt
143_1	q1_qas	2018-08-21 07:00:00	143	q1	Quiz	qas	The user with id '793' has started the attempt with id '4587' for the quiz with course module id '135'.	web	1
143_1	q1_qav	2018-08-21 07:00:00	143	q1	Quiz	qav	The user with id '793' has viewed the attempt with id '4587' belonging to the user with id '793' for the quiz with course module id '135'.	web	1
143_1	q1_qav	2018-08-21 07:01:00	143	q1	Quiz	qav	The user with id '793' has viewed the attempt with id '4587' belonging to the user with id '793' for the quiz with course module id '135'.	web	1
143_1	q1_qav	2018-08-21 07:01:00	143	q1	Quiz	qav	The user with id '793' has viewed the attempt with id '4587' belonging to the user with id '793' for the quiz with course module id '135'.	web	1

Figure 9. preprocessing results of scenario 2

The resulted data of preprocessing stage is shown in figure 8 and figure 9. It is observed that the difference between the two scenarios lies on the case id and the event used.

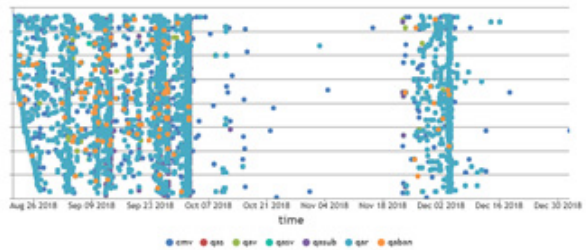


Figure 10. Dotted chart using absolute time in scenario 1

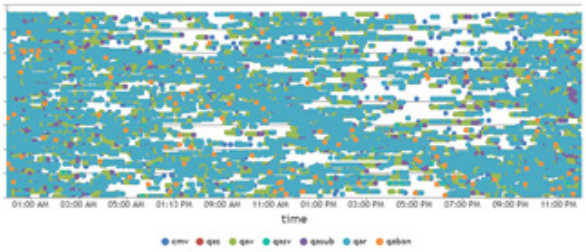


Figure 11. Dotted chart using relative time in scenario 1

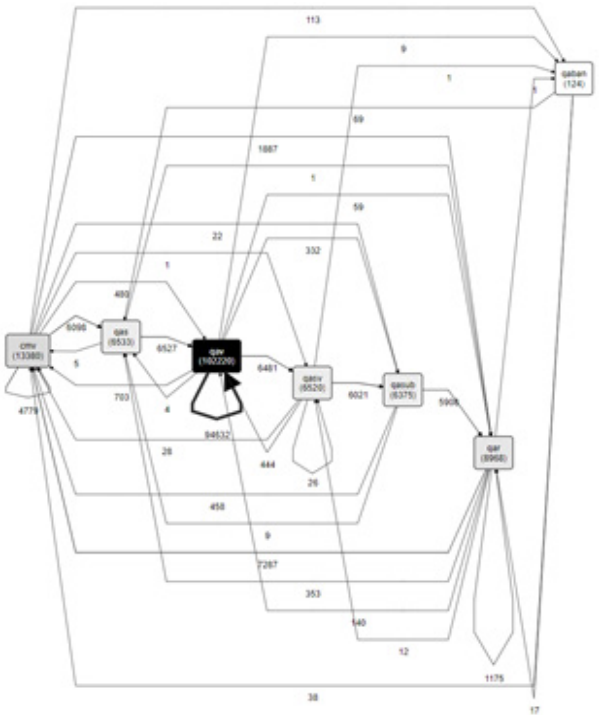


Figure 12. Control flow of scenario 1

By using scenario 1, we can see the visualization of the process from the flowchart. However, we hardly find information from the figure 10 and figure 11. Additionally, from figure 12, the flow of the quiz was observed. It was obvious that the most common pattern is cmv - qas - qav - qasv - qasub - qar. The produced flow was the expected result since it represented the normal attempt of the quiz. For the second scenario, only control flow and dotted charts were shown using event context with a value of Quiz: Quiz 1 – Introduction to Digital World. The calculation of the number of quiz attempts which was conducted during

preprocessing focuses on the starting time of the quiz. The starting time of the quiz was marked with an Event name – labelled *Quiz attempt started* and ended with *Quiz attempt submitted*.

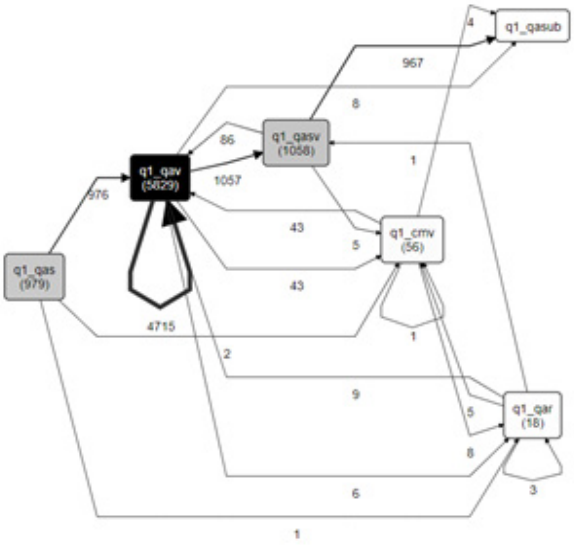


Figure 13. Control flow of scenario 2

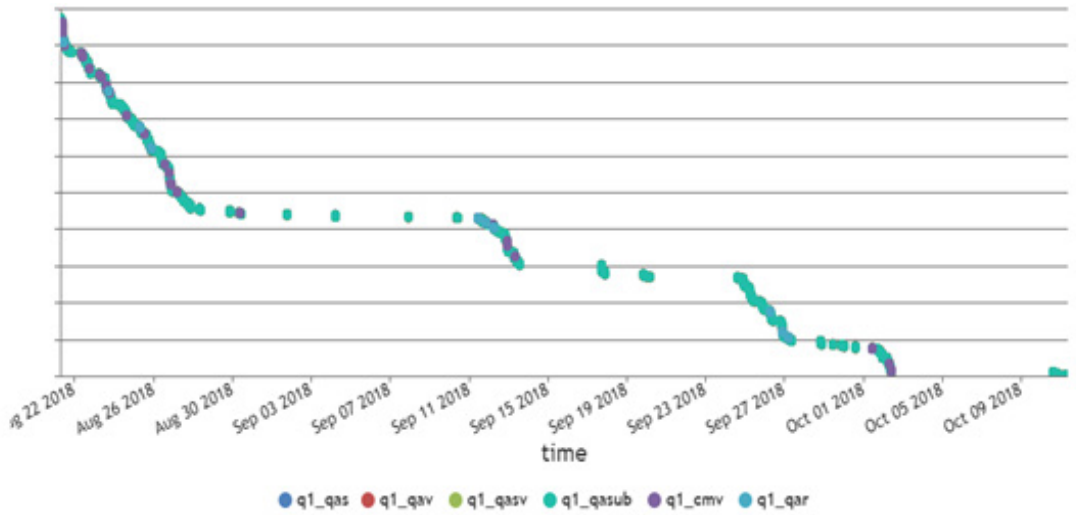


Figure 14. Dotted chart using absolute time in scenario 2

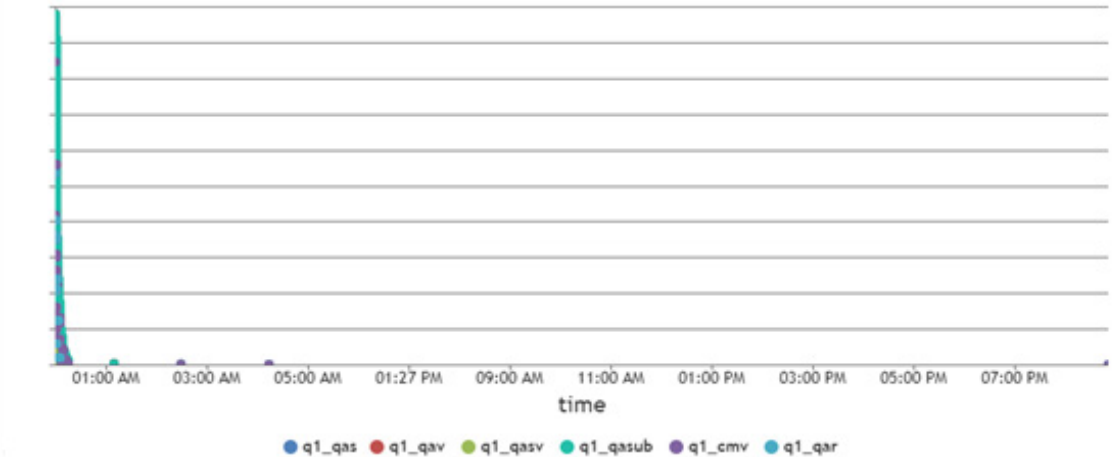


Figure 15. Dotted chart using relative time in scenario 2

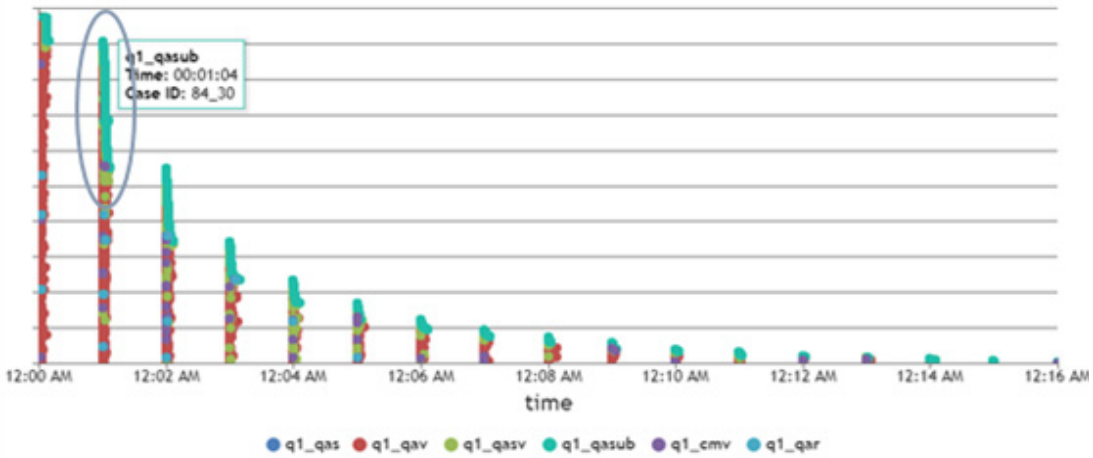


Figure 16. Dotted chart of the results of the anomaly deletion in scenario 2

The control flow using scenario 2 in figure 13, differs from the control flow in scenario 1. The reason was that the scenario 2 focused on the starting point of the quiz – when the Quiz attempt started. It was shown that the control flow obtained from scenario 2 is easier to read since the case id taken focuses on only one quiz. The most common pattern appeared was similar to scenario 1 because it was the expected flow.

By looking at the dotted chart with absolute time in scenario 2 shown in figure 14, it was seen that the highest frequency of quiz 1 is found at the beginning of the lecture in one-week span. Then, the quiz attempt became active again around 11 to 13 September and before the midterm examination (UTS), which was on September 24 to October 2. It was concluded that the participants used the quiz in Moodle to learn and recall the lessons that had been learned for exam preparation.

In figure 15, it is shown that the resulted dotted chart is still quite difficult to read. The reason was that the data anomalies were still occurred in the data – with case id 78_1, 43_1, 60_1 and 26_1. In those case ids, the quiz took more than 1 hour. It should not be possible since the time of the quiz attempt had been set at the beginning. Consequently, data deletion was carried out on the four case ids to earn the more appropriate results. The deletion was done by using the alias naming feature available in preprocessing. The results of the deletion is depicted in Figure 16.

Based on the dotted chart with the relative time in scenario 2 – shown in figure 16, the maximum range of the graph is 100% and each line on the Y-axis had the distance of about 10%. Therefore, it was obvious in the section inside the blue circle, there are around 35% of quiz attempts completed within 1 minute. This was likely due to the quiz resubmission conducted by the participants who had already known the questions after the first attempt and continuously tried to earn better scores in each trial.

4. Conclusion

The conclusion obtained from this study is that the integration of the preprocessing and EDA stages of the process mining was successfully carried out and the system built was able to be used properly. By using this system, teachers and researchers could carry out preprocessing and EDA on Moodle data. For example, based on experiments conducted, it was found that the highest attempt of quiz 1 was found at the beginning and mid-term of the semester. Besides, it was also found that the most time spent in quiz 1 was 1 minute

5. Further Research

For the further research, process mining algorithms will be applied to this system using data that have been processed in this study, so that teachers and researchers can use this system to conduct process mining of data from Moodle. Also, the further research might integrate data retrieval from Moodle into the application by using a web service so that it can reduce the processing time required to export data from Moodle and to upload it back into the system.

References

[1] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. 2012.

[2] C. Romero and S. Ventura, "Educational Data Mining : A Review of the State of the Art," *IEEE Trans. Syst. Man. Cybern.*, vol. 40, no. 6, pp. 601–618, 2010.

[3] W. M. P. Van Der Aalst, *Process Mining Data Science in Action*, 2nd ed. Heidelberg: Springer, 2016.

[4] K. Grigorova, E. Malysheva, and S. Bobrovskiy, "Information Technology and Nanotechnology," 2017.

[5] A. Bogarín, R. Cerezo, and C. Romero, "A survey on educational process mining," *WIREs Data Min. Knowl. Discov.*, vol. 8, no. February, pp. 1–17, 2018.

[6] D. R. Ferreira, *A Primer on Process Mining*. 2017.

[7] A. M. Momani, "Comparison between two Learning Management Systems : Moodle and Blackboard," *Inf. Syst. Behav. Soc. Methods eJournal*, vol. 2, no. 54, pp. 1–10, 2010.

[8] K. Slaninova, J. Martinovic, P. Drazdilova, and V. Snasel, "From Moodle Log File to the Students Network," 2014, pp. 641–650.

[9] A. Bogarín, C. Romero, R. Cerezo, and M. Sánchez-santillán, "Clustering for improving Educational Process Mining," 2014, pp. 11–15.

[10] A. Bogarín, C. Romero, and R. Cerezo, "Discovering students' navigation paths in Moodle," in *8th International Conference on Educational Data Mining*, 2015, pp. 556–557.

[11] L. Juhanak, J. Zounek, and L. Rohlíkov, "Using process mining to analyze students' quiz-taking behavior patterns in a learning management system," *Comput. Human Behav.*, vol. 92, pp. 496–506, 2019.

[12] K. Willems, "Python Exploratory Data Analysis Tutorial (article) - DataCamp," 2017. [Online]. Available: https://www.datacamp.com/community/tutorials/exploratory-data-analysis-python. [Accessed: 10-Dec-2018].

[13] M. Fani Sani, S. J. van Zelts, and W. M. P. Van Der Aalst, "Repairing Outlier Behaviour in Event Logs," in *International Conference on Business Information Systems*, 2018, vol. 320, pp. 115–131.

[14] N. Tax, N. Sidorova, and W. M. P. Van Der Aalst, "Discovering more precise process models from event logs by filtering out chaotic activities," *J. Intell. Inf. Syst.*, vol. 52, no. 1, pp. 107–139, 2019.

[15] R. P. J. C. Bose, R. S. Mans, and W. M. P. Van Der Aalst, "Wanna Improve Process Mining Results ? It ' s High Time We Consider Data Quality Issues Seriously," in *IEEE Symposium on Computational Intelligence and Data Mining*, 2013, pp. 127–134.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Analysis of Slow Moving Goods Classification Technique: Random Forest and Naïve Bayes

Deny Jollyta*[1], Gusrianty[1], Darmanta Sukrianto[2]
[1]Program of Informatics
Sekolah Tinggi Ilmu Komputer Pelita Indonesia
Pekanbaru, Indonesia
[2]Program of Informatics
AMIK Mahaputra Riau
Pekanbaru, Indonesia
*deny.jollyta@lecturer.pelitaindonesia.ac.id

Abstract-Classifications techniques in data mining are useful for grouping data based on the related criteria and history. Categorization of goods into slow moving group or the other is important because it affects the policy of the selling. Various classification algorithms are available to predict labels or class labels of data. Two of them are Random Forest and Naïve Bayes. Both algorithms have the ability to describe predictions in detail through indicators of accuracy, precision and recall. This study aims to compare the performance of the two algorithms, which uses testing data of snacks with labels for packaging, size, taste and category. The study attempts to analyze data patterns and decides whether or not the goods fall into the slow moving category. Our research shows that Random Forest algorithm predicts well with accuracy of 87.33%, precision of 85.82% and recall of 100%. The aforementioned algorithm performs better than Naïve Bayes algorithm which attains accuracy of 84.67%, precision of 88.33% and recall of 92.17%. Furthermore, Random Forest algorithm attains AUC value of 0.975 which is slightly higher than that attained by Naïve Bayes at 0.936. Random Forest algorithm is considered better based on the value of the metrics, which is reasonable because the algorithm does not produce bias and is very stable.

Keywords: slow moving; random forest; naïve bayes;

1. Introduction

Goods can be classified based on its circulations over a certain period of time and goods with very slow circulation are called slow moving goods [1]. Slow moving goods have be stored in warehouses in large quantity. Slow moving goods are materials that circulate with the speed of one item within a year [2]. Classification problems associated with slow moving goods occur due to lack of analysis of previous data [3]. Analysis can be conducted using classification algorithms of data mining. Classifications create patterns through analysis of the closeness of labels or attributes that construct item data. The resulting patterns are the predictions of slow moving goods.

In this study, Random Forest and Naïve Bayes are the classification algorithms that are used, which work on data of packaged snacks. Both algorithms were chosen because they can produce accurate predictions with descriptions that highly agree with actual situations. Many studies have been carried out that relate to the two algorithms for classification. In [4], Random Forest was used to analyzing multispectral images by classifying points in images. Taxonomy of Random Forest algorithm has been described in [5] through several parameters such as the base of classifications, size division, number of tracks, combination of strategies, number of attributes, criteria, cut-off ability, additional classifications, and number of datasets used in training phase. In addition, Random Forest algorithm has highlighted the advantages and benefits in prediction on large datasets [6].

The ability of Naïve Bayes algorithm has been tested in various data predictions including to predict the behavior of the purchase on transaction time [7]. The pattern shows that more buyers make transactions in the afternoon, particularly on Sundays. Naïve Bayes algorithm has been used to group blogger data [8] and banking product marketing data [9] - [10] to assist banks to find potential customers. The performance of Naïve Bayes algorithm has been compared with other classification algorithms such as K-Nearest Neighbor (KNN) algorithm and Decision

Tree [11] to group data of school students who consume alcohol. Despite the differences, Naïve Bayes' performance has shown better accuracy than the other two algorithms.

The results of previous studies in using Random Forest and Naïve Bayes algorithms motivate an attempt to observe both algorithms in classification of slow moving goods. The result is valuable for decision makers to implement policies related to such goods. A comparison is required for a clear picture of the performance of both algorithms.

2. Theory

a. Random Forest Algorithm

Random Forest algorithm is an ensemble model that was created and developed by Tin Kam Ho [12]. It belongs to supervised learning and works based on calculations of various models to obtain results [6]. As an ensemble model, Random Forest is able to build decision trees and uses its rules for the calculation of the final result, following formula (1) [13].

$$hj(X,\Theta j)$$
(1)

Having processed training data, predictions are obtained from the average results of all trees, using formula (2).

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$
(2)

In its various applications, Random Forest algorithm is widely used for its advantages, such as better accuracy, resistance to various disturbances, speed, and convenience in implementation [14].

b. Naïve Bayes Algorithm

Naïve Bayes algorithm is a simple probabilistic classification technique based on the application of the Bayes theorem with strong assumptions [15]. Naïve Bayes is applied to a limited number of data to get the appropriate parameters of classification. Naïve Bayes formula is expressed using formula (3) [11].

$$P(H \mid E) = \frac{P(H)\prod_{i=1}^{a} P(E_i \mid H)}{P(E)} \quad (3)$$

P (H|E) = data probability with vector E in class H.
P (H) = initial probability of class H
$\prod_{i=1}^{a} P(E_i \mid H)$ = independent probability of class H from all features in vector E

The advantages of the Naïve Bayes algorithm include the ability to handle quantitative and discrete data, resistance to isolated noise points, sufficiency of small number of training data, ability to handle missing values by neglecting instances during the calculation of estimated probability, speed, efficiency in space, and robust against irrelevant attributes.

3. Method

a. Training Data

The use of the two algorithms in this research was administered by employing two different tests using RapidMiner application. RapidMiner is an application in the field of data mining such as machine learning, information mining, and content mining [16]. In this study, RapidMiner is used to display the performance of the two algorithms using the data of packaged snacks. Data items were taken randomly for as many as 150 data. Data have to pass a selection process in accordance with the stages of Knowledge Discovery in Database (KDD). The data were arranged based on several attributes considered to affect most on the speed of items transactions, such as packaging, size, taste, and category. Attributes description is shown in Table 1.

Table 1. Attributes of Training Data

Attribute	Description
Item Code	Code of each packaged snack
Item Type	Type of snacks such as candies, jelly, gum drop, etc.
Item Name	Name of packaged snack.
Taste	Taste of the packaged snack such as sweet, spicy, sweet and sour, etc.
Packaging	Shape and material of packaging, namely plastic, bottle, and can.
Size	Size of package such as small, medium, and large.
Category	Label regarding snack resistance, such as premature spoilage, fragile, and resistant.
Slow Moving	Label regarding transaction flow, such as Yes for slow moving and No for fast moving

c. Research Framework

To produce a result of prediction of slow moving goods, the research goes through several steps as shown in Figure 1. The first stage is data preparation, which follows the stages of the Knowledge Discovery in Database (KDD). Data selection consumes a huge amount of time in order to adjust with the classification algorithms, i.e. Random Forest and Naïve Bayes. Data have to pass the KDD stages to obtain proper quality of training data. The KDD selection produces training data as described in Table 1.

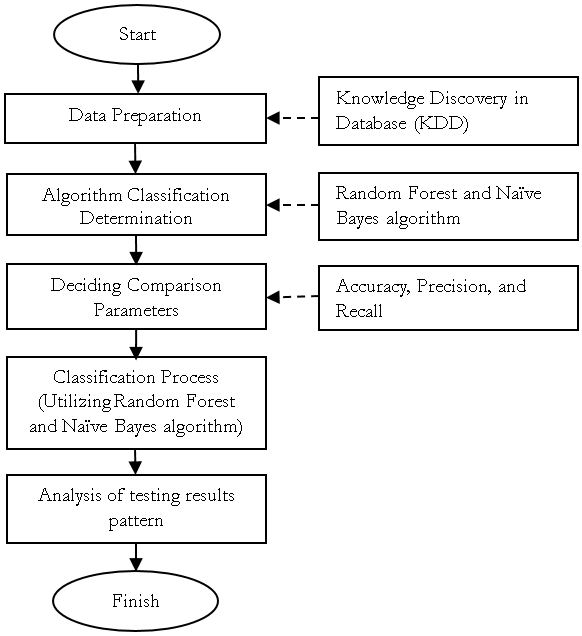


Figure 1. Research Framework

The next step is the selection of parameters as a measure to compare the performance of test results. We choose accuracy, precision, and recall, which are taken from the Gain Ratio criteria in RapidMiner. The choice of gain ratio as the comparison results parameter concerns more about its ability to calculate every data in the available sample space. The parameter selection is intended to see the comparison of the results of testing the two algorithms. In addition to the three parameters, the test results are displayed in accordance with the algorithm features. Training produces patterns that may be analyzed to obtain predictions that reflect the actual situation. This step provides the best prediction results for each of the two algorithms.

4. Results and Discussion

Research results on the two algorithms are described in the following two sections: the prediction results and the parameter results.

a. Prediction Results

Item data were tested on both algorithms by using 10 iterations. Each iteration produced a different tree structure. Confidence was displayed to indicate the level of confidence of each attribute in producing the decision whether the items belong to either slow moving or non–slow moving category. The gain ratio criterion was used as a measure to read the test results of Random Forest. There are some discrepancies in the test results, especially on the target attribute "No". This is because confidence value of the "No" is higher than the target attribute "Yes". Of 150 data, there arises 19 discrepancy data, which implies a value of 12.67% error rate of the calculation results. The rules resulting from the calculation of the Random Forest algorithm are described in Table 2.

Table 2. The Slow Moving Goods Prediction Rules of Random Forest

No	Rules
1	If item type=candy ^ taste=sweet and sour ^ packaging=plastic ^ size=small ^ category=premature spoilage, then slow moving=Yes
2	If item type=candy ^ taste=spicy ^ packaging=can ^ size=big ^ category=resistant, then slow moving=Yes
3	If item type=candy ^ taste=spicy ^ packaging=plastic ^ size=large ^ category=premature spoilage, then slow moving=Yes
4	If item type=dried sunflower seeds ^ taste=sweet ^ packaging=plastic ^ size=small ^ category=premature spoilage, then slow moving=Yes
5	If item type=sponge cake ^ taste=salty ^ packaging=can ^ size=large ^ category=resistant, then slow moving=Yes

Table 3. The Slow Moving Goods Prediction Rules of Naïve Bayes

No	Rules
1	If item type=candy ^ taste=sweet and sour ^ packaging=plastic ^ size=small ^ category=premature spoilage, then slow moving=Yes
2	If item type=candy ^ taste=sweet ^ packaging=bottle ^ size=large ^ category=fragile, the slow moving=Yes
3	If item type=candy ^ taste=sweet ^ packaging=bottle ^ size=large ^ category=resistant, then slow moving=Yes
4	If item type=candy ^ taste=spicy ^ packaging=can ^ size=large ^ category=resistant, then slow moving=Yes
5	If item type=candy ^ taste=spicy ^ packaging=can ^ size=small ^ category=resistant, then slow moving=Yes
6	If item type=candy ^ taste=spicy ^ packaging=plastic ^ size=large ^ category=resistant, then slow moving=Yes
7	If item type=candy ^ taste=sweet ^ packaging=bottle ^ size=small ^ category=fragile, then slow moving=Yes
8	If item type=candy ^ taste=sweet ^ packaging=bottle ^ size=large ^ category=fragile, then slow moving=Yes
9	If item type=snack ^ taste=spicy ^ packaging=plastic ^ size=small ^ category=premature spoilage, then slow moving=Yes
10	If item type=snack ^ taste=spicy ^ packaging=can ^ size=large ^ category=resistant, then slow moving=Yes
11	If item type=sponge cake ^ taste=sweet ^ packaging=plastic ^ size=large ^ category=resistant, then slow moving=Yes
12	If item type=sponge cake ^ taste=salty ^ packaging=can ^ size=large ^ category=resistant, then slow moving=Yes
13	If item type=biscuit ^ taste=milky ^ packaging=plastic ^ size=large ^ category=resistant, then slow moving=Yes

Based on Table 2, the calculation shows an accuracy of 87.33%. This value is the conclusion of the

accumulation of each decision tree produced through Random Forest. Calculation of gain ratio for positive class = Yes is 45.71%, while for positive class = No is 100%.

The prediction for the Naïve Bayes algorithm using the gain ratio shows a lower error rate at 8.67%. There are 13 different data. The differences between the prediction and initial data are mostly on data with attributes "No" which become "Yes" according to Naïve Bayes calculation. This means that items, which are not originally included in the slow moving category, fall into the category. The confidence value is lower here so it changes data with class attributes "No" to become "Yes". Naïve Bayes produces a model of slow moving attributes into 2 classes previously mentioned with respective value of 0.767 for the "No" class and 0.233 for the "Yes" class. The rules resulting from the calculation of the Naïve Bayes algorithm are in Table 3.

Based on Table 3, Naïve Bayes produces an accuracy of 84.67%. Calculation of gain ratio for positive class = Yes is 60% while for positive class = No is 92.17%.

b. Accuracy, Precision, and Recall Parameters

The implementation of gain ratio criteria in this training stage produces detailed calculations in the form of confusion matrix. Both algorithms reveal patterns that are hidden in the training data. Running RapidMiner with operator Performance produces results in values of metrics in 3 parameters: Accuracy, Precision and Recall, as shown in Table 4.

Table 4. Parameter Calculation Results

Parameter	Random Forest Algorithm	Naïve Bayes Algorithm
Accuracy	87.33%	84,67%
Precision	85.82%	88.33%
Recall	100%	92.17%

Entries of Table 4 show that accuracy and recall parameters of the Random Forest algorithm are higher than the Naïve Bayes algorithm. However, the precision of the Naïve Bayes algorithm is higher. Hence the Random Forest algorithm is superior in two of three metrics against Naïve Bayes algorithm. To further decide which classification algorithm is better, we need to observe the Receiver Operating Characteristic (ROC) curve and calculate the Area under the ROC Curve (AUC) [7]. An ROC curve expresses confusion matrix data, in which the horizontal line represents false positive (FP) values and the vertical line represents true positive (TP) values.

Figure 2 is an ROC curve obtained from the calculation of the Random Forest algorithm with the acquisition of AUC values of 0.975. In [8], AUC was used to measure discriminative performance by predicting the possibility of the emergence of output from random samples for positive and negative populations. The greater the AUC, the firmer the classification be recommended. AUC is part of the square unit area, AUC value will always be between 0.0 and 1.0.

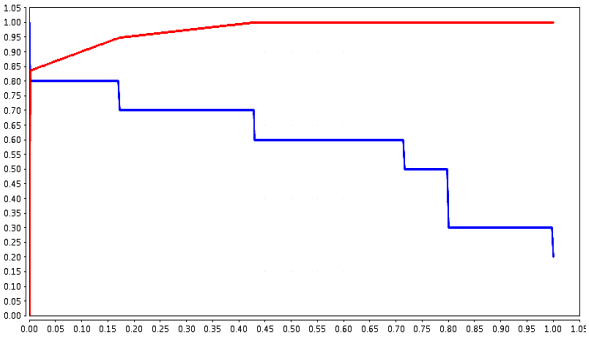


Figure 2. Random Forest Area under ROC Curve (AUC)

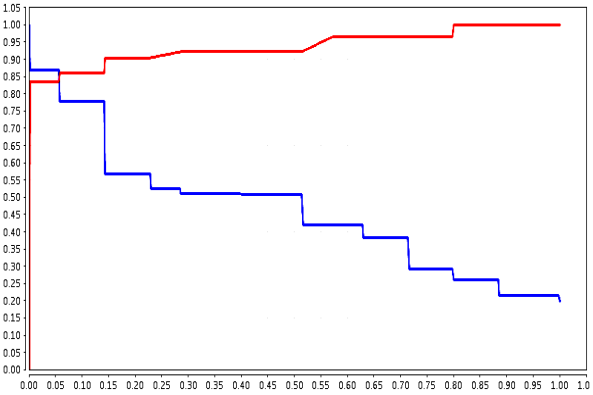


Figure 3. Naïve Bayes Area under ROC Curve (AUC)

Figure 3 is the ROC curve from the calculation of the Naïve Bayes algorithm with AUC of 0.936. The AUC for Random Forest algorithm at 0.975 is slightly higher. However, both algorithms behave as a nearly perfect classification model with AUC values close to 1.00.

c. Discussion

Both algorithms show good performance but with different results. The overall results of testing the item data with the Random Forest and Naïve Bayes are in the following Table 5.

Metrics in Table 5 show that the performance of the Random Forest algorithm is generally better. Random Forest algorithms produces a tree structure in each iteration that is easy to compare with structures in other iterations. The most results from each structure become the final result. The ability of Random Forest algorithm to analyze the results of each decision tree in 10 iteration has apparently produce higher accuracy than the Naïve Bayes algorithm. The dominantly similar rule in every iteration is one of the advantages of Random Forest, which may support its performance to achieve a high accuracy [17]. The recall value reaching 100% and the AUC value of 0.975 have brought the Random Forest as the best choice for classification of slow moving goods. Therefore, attributes that are considered responsible to cause a goods become slow moving are taken from those identified by Random Forest algorithm.

Table 5. Comparison of Random Forest and Naïve Bayes Performance

No	Indicator		Random Forest	Naïve Bayes
1	Number of Rules		5	13
2	Prediction Error	Total of Data	19	13
		Percentage	12.67%	8.67%
		Accuracy	87.33%	84.67%
3	Parameters	Precision	85.82%	88.33%
		Recall	100%	92.17%
4	Gain Ratio	Positive class Yes	45.71%	23.3%
		Positive Class No	100%	76.7%
5	AUC Value		0.975	0.936

4. Conclusion

We have observed two algorithms: the Random Forest and Naïve Bayes algorithms to classify data on packaged snacks and to identify which attributes supports the class label of slow moving. Calculation using RapidMiner on both algorithms give predictions with almost similar accuracy. The difference in the precision value of the two algorithms of 2.51% suggests that Naïve Bayes algorithm has better accuracy in slow moving goods in the training data. This is shown by the smaller prediction errors than that of Random Forest algorithm, and because the confidence values tend to be identical. However, Random Forest algorithm is more reliable to get a precise prediction because it may be obtained from several decision trees. This research shows that Random Forest algorithm provides better predictions to reflect actual conditions with a limited number of data. A total of 5 rules were produced, showing perfect compatibility with the actual situation of packaged snacks, which is 100%.

Acknowledgment

Acknowledgment is addressed to Sekolah Tinggi Ilmu Komputer Pelita Indonesia Pekanbaru which supports the completion and publication of this research.

References

[1] Rajahstan, *Reading Material Drug Store Management Rational Drug Use For Medical Officers , Nurses & Pharmacists*, no. December. 2010.

[2] D. Janari, M. M. Rahman, and A. R. Anugerah, "Analisis Pengendalian Persediaan Menggunakan Pendekatan Music 3D (Muti Unit Spares Inventory Control- Three Dimensional Approach) Pada Warehouse Di PT Semen Indonesia (PERSERO) TBK Pabrik Tuban," *Teknoin*, vol. 22, no. 4, pp. 261–268, 2016.

[3] G. Chodak, "The Nuisance of Slow Moving Products in Electronic Commerce," *MPRA Munich Pers. RePEc Arch.*, vol. 70141, no. 3, pp. 1–7, 2016.

[4] B. Lowe and A. Kulkarni, "Multispectral Image Analysis Using Random Forest," *Int. J. Soft Comput.*, vol. 6, no. 1, pp. 1–14, 2015.

[5] V. Y. Kullarni and P. K. Sinha, "Random Forest Classifier: A Survey and Future Research Directions," *Int. J. Adv. Comput.*, vol. 36, no. 1, pp. 1144–1156, 2013.

[6] N. Horning, "Random Forests: An algorithm for image classification and generation of continuous fields data sets," in *International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences 2010*, 2010, pp. 1–6.

[7] Susanto, E. D. S. Mulyani, and I. R. Nurhasanah, "Penerapan Data Mining Classification Untuk Prediksi Perilaku Pola Pembelian Terhadap Waktu Transaksi Menggunakan Metode Naïve Bayes," in *Konferensi Nasional Sistem dan Informatika (KNS&I)*, 2015, pp. 313–318.

[8] Ardiyansyah, P. A. Rahayuningsih, and R. Maulana, "Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner," *J. Khatulistiwa Inform.*, vol. VI, no. 1, pp. 20–28, 2018.

[9] I. Oktanisa and A. A. Supianto, "Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank Direct Marketing," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, pp. 567–576, 2018.

[10] N. H. Niloy and M. A. I. Navid, "Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients," *Am. J. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 1–12, 2018.

[11] N. Sagala and H. Tampubolon, "Komparasi Kinerja Algoritma Data Mining pada Dataset Konsumsi Alkohol Siswa," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 2, pp. 98–103, 2018.

[12] A. K. Mishra and B. K. Ratha, "Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis," *Int. J. Adv. Electr. Comput. Eng.*, vol. 3, no. 4, pp. 5–7, 2016.

[13] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," in *Random Forest*, no. January, 2011, p. 21.

[14] E. Goel and E. Abhilasha, "Random Forest: A Review," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 7, no. 1, pp. 251–257, 2017.

[15] S. Taheri and M. Mammadov, "Learning the naive bayes classifier with optimization models," *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 4, pp. 787–

795, 2013.

[16] S. Dixit and S. Kr, "Collaborative Analysis of Customer Feedbacks using Rapid Miner," *Int. J. Comput. Appl.*, vol. 142, no. 2, pp. 29–36, 2016.

[17] K. . Ghose, R. Pradhan, and S. S. Ghose, "Decision Tree Classification of Remotely Sensed Satellite Data using Spectral Separability Matrix," *Int. J. Adv. Comput. Sci. Appl.*, vol. 1, no. 5, pp. 93–101, 2010.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

A Virtual-Reality Edu-Game: Saving The Environment from the Dangers of Pollution

Dita Aluf Mawsally*, Endah Sudarmilah
Program Studi Informatika
Universitas Muhammadiyah Surakarta
Surakarta, Indonesia
*Ditaaluf11@gmail.com

Abstract-Our virtual reality educational game –themed "saving the environment from the dangers of pollutions"– aims to help children aged 10-13 years in learning about pollution and its reduction. This game uses smartphone and VR Box. This edu game is designed using the SDLC (Software Development Life Cycle) model of the waterfall model. Assets contained in this game are obtained from the Unity Asset Store which will simulate learning techniques with several educational games. This can be seen as a development from the traditional learning method. Based on a black box test, this game runs well on the target device. On the usability test using the System Usability Scale (SUS), this game gets a score of 71.58 which is in the "good" criteria.

Keywords: Educational Games, Pollution, Virtual Reality

1. Introduction

The rapid development of Information Technology drives many people to use applications as tools to facilitate their daily work, as entertainment gadgets, and even educational devices. Utilization of smartphones for use in various fields is done by developing applications that support activities in that field. The same thing happens in the field of education, which uses smartphones as learning media [1]. Learning using digital media must still pay attention to the curriculum, materials, methods, and students' learning abilities in achieving learning goals in school [2].

Learning media always follow the development of existing technology, such as 3D animation or science fiction films, which make technology more interesting and interactive [3]. One particularly popular technology right now is virtual reality educational games that are loved by children, adolescents, and even adults [4]. Virtual reality devices depicts a three dimensional environment constructed by computers in such a way that it the devices can interact with the users [5]. Our VR-based educational game, named "Earth Hero", makes users feel the VR experience by combining gyroscope-equipped mobile devices with Google Cardboard and VR Box. VR technology is useful in the fields of education, property, medicine, transportation, architecture, entertainment, etc. [6].

The theme "saving the environment from the dangers of pollution" was chosen because many human behavior intentionally or unintentionally cause pollution on earth. Pollution is a threat that can damage the ecosystem of terrestrial and marine life. Air pollution, or changes in the composition of the air element from normal conditions, can result in changes in temperature and damages to the environment [7].

The adverse effects of air pollution and water pollution for human health cannot be refuted anymore. Marine pollution has long been recognized as a threat to the growth of marine biota [8]. This educational game, called "Earth Hero", was created to simulate learning methods in the classroom related to the environment. It is hoped that students will be more aware of the impacts and dangers of pollution so they can better maintain the cleanliness of the environment and the earth.

2. Method

This study aims to design VR-based educational games as learning media for elementary school students aged 10 to 13 years, namely grades 4 to 6 elementary school. This game was built using the SDLC (Software Development Life Cycle) method [9], [10]. The system development is done in order, starting from the requirement analysis, design, coding, testing, and implementation [11]; as shown in Figure 1.

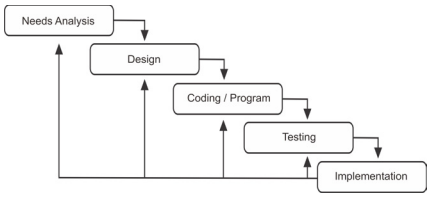


Figure 1. Flowchart of research system flow

a. Requirement Analysis

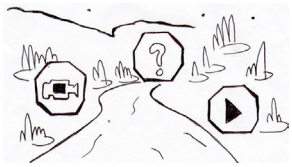
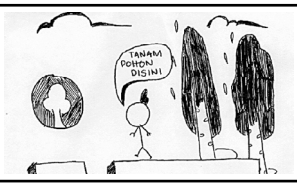
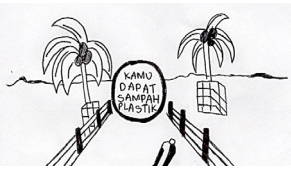
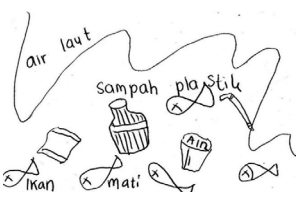
Hardware and software requirements needed in making the game "Earth Hero" are presented in Table 1.

Table 1. Hardware and Software Needs

Software	*Hardware*
a. Unity b. Adobe Premiere Pro CC 2017 c. Corel Draw d. Notepad ++	a. Laptop ASUS A456U, CPU Intel ® Core ™ i5 7200U up to 3.1GHz, RAM 4GB, HDD 1TB b. Android OS smartphone with gyroscope sensor c. Virtual Reality Box + Remote

b. Design

Table 2. Game storyboards

Picture	Information
	The main menu when starting the Edu Game "Earth Hero". (Play, Video, Help)
	Video scene containing educational film about pollution
	Game 1: Plant trees in the surrounding environment so that the environment is cooler and beautiful
	Game 2: Clean up trash on the streets to reduce air pollution
	Game 3: Clean up trash in the sea, to reduce sea pollution

The design stage is an advanced stage after the analysis stage where the story board is drawn and various things such as assets, music, sound, and video are assembled. Table 2 shows the storyboard.

c. *Coding*

Some code is used to control assets and scenes. Among them are codes for movement, movement, up and down, and disappearing. The programming language used is C# and written using Notepad++ code editor [12].

d. Testing

Testing is the final stage after the application is complete. The application is tested to find out if this program can be accepted by the user and to ensure that there are no errors in the application and there are no difficulties felt by the user when playing the game. The test method used is the black box test and the reusability test [13].

e. Implementation

The last stage of SDLC is the implementation of the game in the hope that the results will be good and in accordance with the initial objectives [14]. At this stage there was also an observation of the use of the application when played by respondents who participated in the test.

3. Results

This research resulted in an VR-based educational game "Earth Hero - Save the Earth from Pollution" that can help students learn in an effective and fun way.

a. Game Start Page

The initial appearance of the game is designed with 3 main buttons, namely the video, play, and help buttons which can be seen in Figure 2.

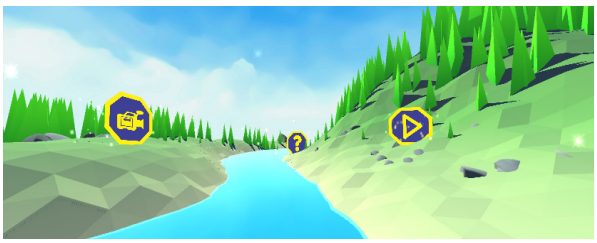


Figure 2. Initial Display of the Game

The game's display of the menu section - when viewed using an android smartphone with a gyroscope sensor accompanied by special tools - is shown in Figure 3.

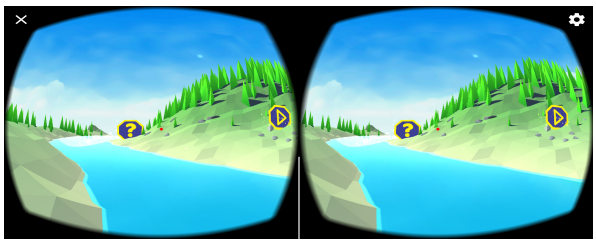


Figure 3. Menu display when viewed on a smartphone with certain capabilities.

b. Educational Video

Videos are added to present new experiences in learning. There are several obstacles in video development. Among them are the limited time the teachers have to develop the media and the lack of mastery of video editing software [15]. To help teachers, relevant educational videos are added in this application. When pressing the video button on the start menu, one will see a 2 minute video as depicted in Figure 4.

Figure 4. The educational video when playing on a VR-abled smartphone

In the video menu display, there is a play button which if the reticle pointer is directed for 2 seconds (clicked), will bring the player into Game 1. The button can be clicked at any time: when finished watching the video, in the middle of the video running, or in the beginning of the video.

c. Game 1

The player is tasked with planting trees around the park environment to increase oxygen so that air pollution caused by vehicle fumes can be reduced. This game is done by players who will walk around the park and find the button to plant trees. If the tree has been planted, a pop up will appear containing an educational message in the Game 1 environment as shown in Figures 5, 6, and 7.

Figure 5. Environment of the Game 1

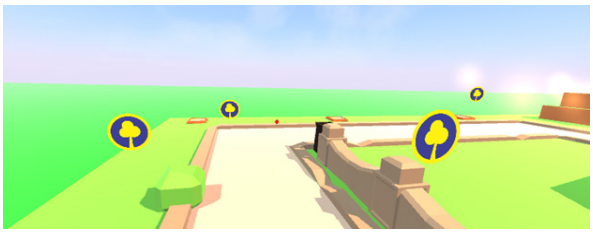


Figure 6. Blue button for planting trees

Figure 7. Trees that have been planted and the corresponding pop up.

d. Game 2

The player is tasked with picking up trash on the road so that the road is clean from rubbish and so that soil pollution is reduced. When the player passes the trash, the player will pick it up and an educational popup appears. The game's Display 2 can be seen in Figures 8 and 9.

Figure 8. Initial entry into Game 2

Figure 9. Garbage to be taken in Game 2

e. Game 3

The task of the players in Game 3 is to pick up trash from the sea that aims to reduce marine pollution and prevent the extinction of the marine ecosystem. When the player passes the trash, the player will take it until the sea environment is clean.

Figure 10. Game Display 3

4. Discussion

a. Application Testing

Testing is the final stage after the application is complete. The user tests whether the application can be accepted and to ensure that there are no errors and no difficulties felt by the user when running the application. Black box testing is carried out on several user interfaces, sound, and controls in the game, as well as testing on smartphones [16]. The focus of the black box test on the main features of the system is done to find functions that are not running properly, interface errors, errors in data structures and database access, performance errors, and initialization / termination errors [17], [18]. Tests was carried out using an Android smartphone branded Nougat Oppo type F5. The test results show that the "Earth Hero" application can run well on the target device.

b. Usability Test

Usability Test is done by observing the object when running the application and recording events when the object encounters an error in the application that [19]. The test was conducted on 30 elementary school students aged 10-13 years by demonstrating the game which is played using VR Box and an Android smartphone. The assessment system uses the System Usability Scale (SUS), which is a questionnaire to measure the usability of a computer system according to the user's subjective point of view. SUS was developed by John Brooke since 1986 and is still often used today. Respondents were asked to give an assessment of the system starting from "Strongly disagree", "Disagree", "Neutral", "Agree", and "Strongly agree" on the 10 items SUS statement. Each statement item has variables R1 through R10 [20] [21]. The overall SUS score is obtained from the average individual SUS score using the following Equation 1 formula:

$$SUS = ((R1 - 1 + (5 - R2) + (R3 - 1) + (5 - R4) + (R5 - 1) + (5 - R6) + (R7 - 1) + (5 - R8) + (R9 - 1) + (5 - R10)) * 2,5)) \quad (1)$$

Table 3 is the results of questionnaire data usability test using SUS formula calculation with statement criteria:

1 : I love Edu Game Earth Hero so I will play it many times.
2 : In my opinion Edu Game Earth Hero is too complicated to play.
3 : I think Edu Game Earth Hero is easy to use
4 : I need help from others to play Edu Game Earth Hero
5 : I consider Edu Game Earth Hero parts to be played well
6 : I think the way to play Edu Game Earth Hero is confusing
7 : I think other people will learn to play Edugame Earth Hero very quickly
8 : I consider Edu Game Earth Hero impractical (difficult) to play
9 : I feel that I can play Edu Game Earth Hero
10 : I need to learn a lot to be able to play Edugame Earth Hero

Table 3. Calculation results with the SUS formula

Respondent Number	Item Question										Total	SUS Score (Total * 2.5)
	1	2	3	4	5	6	7	8	9	10		
1	3	4	3	3	3	4	3	4	4	3	34	85
2	3	3	4	1	3	3	4	4	4	1	30	75
3	2	0	4	0	3	2	4	1	4	4	24	60
4	4	3	3	4	4	3	3	4	3	2	33	82.5
5	2	2	2	1	2	2	0	2	2	2	17	42.5
6	4	3	2	2	3	2	4	4	3	2	29	72.5
7	1	3	3	1	3	1	3	3	3	1	22	55
8	3	2	3	0	4	4	4	4	4	1	29	72.5
9	4	4	4	4	4	4	4	4	4	4	40	100
10	4	4	4	4	4	4	4	4	4	1	37	92.5
11	4	3	4	3	4	3	4	3	4	0	32	80
12	4	4	2	0	3	3	2	2	3	2	25	62.5
13	2	1	2	1	2	3	1	3	2	0	17	42.5
14	4	4	4	0	4	0	0	4	4	4	28	70
15	3	4	3	2	4	4	2	4	3	2	31	77.5
16	4	4	4	4	4	4	4	4	4	4	40	100
17	2	4	3	2	3	4	2	4	4	4	32	80
18	4	2	3	1	4	1	0	0	0	2	17	42.5
19	4	4	4	3	3	4	4	4	4	0	34	85
20	3	3	2	3	2	3	2	1	3	2	24	60
21	4	4	4	4	4	4	4	4	4	4	40	100
22	2	3	3	3	3	2	3	4	3	2	28	70
23	3	1	2	1	4	3	4	4	3	1	26	65
24	4	3	4	3	4	5	3	4	3	2	35	87.5
25	3	2	3	2	3	2	3	2	2	1	23	57.5
26	2	3	3	1	4	2	4	4	4	2	29	72.5
27	3	2	3	3	4	2	3	3	4	0	27	67.5
28	4	3	3	2	3	3	4	3	3	3	31	77.5
29	3	0	2	1	4	1	2	3	0	0	16	40
30	4	2	4	2	4	2	2	2	4	3	29	72.5
TOTAL												**2147.5**

c. Calculation Chart Results Using the SUS Formula

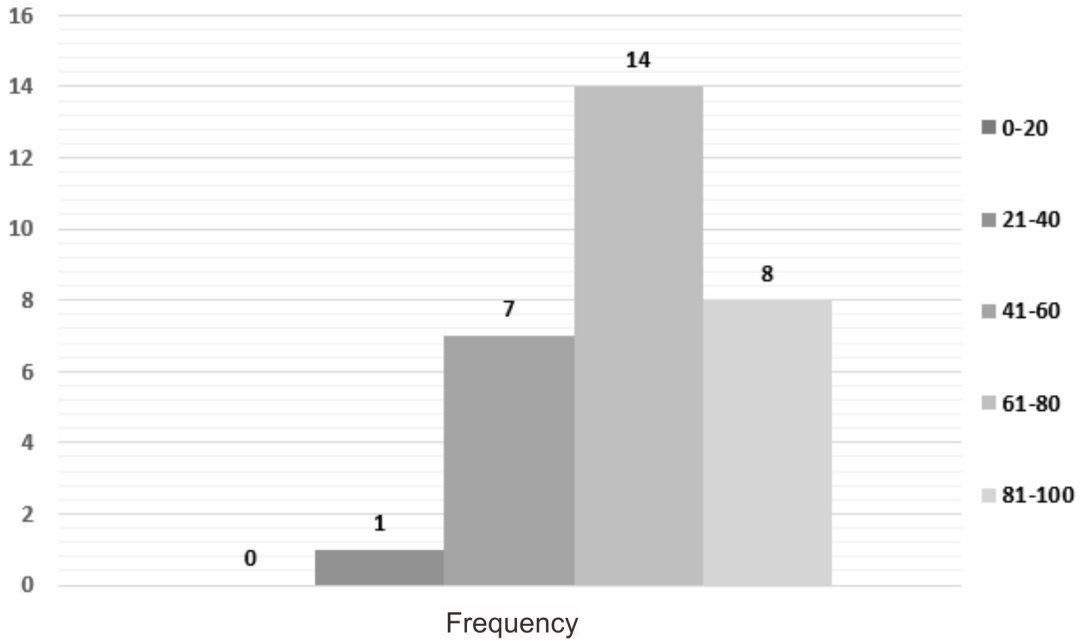


Figure 11. Calculation graph with the SUS formula

The formula for calculating the average value using Equation (2):

$$\text{Average Value} = \sum_{i=1}^{n} \frac{x_i}{N} \quad (2)$$

Where xi is: Respondent Score Value, and N: Number of Respondents

The average value = 2147.5 / 30 = 71.58

Calculation of the average value using Equation 2 yields 71.58. This is within the range of 61 - 80, which is "good".

5. Conclusions

Edu Game based on Virtual Reality "Earth Hero" can increase knowledge about environmental pollution in students aged 10-13 years. Based on the black box test, this educational game can run according to its function. And based on the usability test using the SUS calculation formula, the game "Earth Hero" gets "good" criteria.

References

[1] H. Supriyono, A. N. Saputra, and E. Sudarmilah, "Rancang bangun aplikasi pembelajaran hadis untuk perangkat mobile berbasis android," *J. Inform.*, vol. 8, no. 2, pp. 907–920, 2014.

[2] E. Sudarmilah and M. G. Negara, "Augmented Reality Edu Game Senjata Tradisional Indonesia," *Khazanah Inform.*, vol. 1, no. 1, pp. 12–15, 2015.

[3] E. Sudarmilah, H. Supriyono, F. Yasin, A. Irsyadi, and A. Fatmawati, "Prototyping AR Edu Game untuk anak-anak : belajar budaya Indonesia," *MATEC Web Konf. 197, 03012 AASEC 2018*, vol. 03012, pp. 2–5, 2018.

[4] L. Sigit, M. Dicky, and S. Hendri, "Penerapan Algoritme Fisher Yates pada Game Edukasi Eco Mania Berbasis Unity 3D," no. x, pp. 1–12, 2015.

[5] Z. Tuma, J. Tuma, R. Knoflíček, P. Blecha, and F. Bradác, "The Process Simulation Using by Virtual Reality," *Procedia Eng.*, vol. 69, pp. 1015–1020, 2014.

[6] F. Setiawan Riyadi, A. Sumarudin, and M. Sari Bunga, "Aplikasi 3D Virtual Reality Sebagai Media Pengenalan Kampus Politeknik Negeri Indramayu Berbasis Mobile," *J. Inform. dan Komput.*, vol. 2, no. 2, pp. 75–82, 2017.

[7] Ismiyati, D. Marlita, and D. Saidah, "Pencemaran Udara Akibat Emisi Gas Buang Kendaraan Bermotor," *J. Manaj. Transp. Logistik*, vol. 01, no. 03, pp. 241–248, 2014.

[8] M. Haward, "Plastic pollution of the world's seas and oceans as a contemporary challenge in ocean governance," *Nat. Commun.*, vol. 9, no. 1, pp. 9–11, 2018.

[9] Y. dan U. Firmansyah, "Penerapan Metode SDLC Waterfall Dalam Pembuatan Sistem Informasi Akademik Berbasis Web Studi Kasus Pondok Pesantren Al-Habi Sholeh Kabupaten Kubu Raya , Kalimantan Barat," *J. Teknol. Manaj. Inform.*, vol. 4, no. 1, pp. 185–191, 2018.

[10] R. Scroggins, "SDLC and Development Methodologies," *Glob. J. Comput. Sci. Technol. C*

Softw. Data Eng., vol. 14, no. 7, pp. 0–2, 2014.

[11] M. A. Zaus, R. E. Wulansari, S. Islami, and D. Pernanda, "Perancangan Media Pembelajaran Listrik Statis dan Dinamis Berbasis Android," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 1, no. 1, pp. 1–7, 2018.

[12] U. Al Faruq, "Rancang Bangun Aplikasi Rekam Medis Poliklinik Universitas Trilogi," *J. Inform.*, vol. 9, no. 1, pp. 1017–1027, 2017.

[13] W. A. Kusuma, V. Noviasari, and G. I. Marthasari, "Analisis Usability dalam User Experience pada Sistem KRS Online UMM menggunakan USE Questionnaire," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 5, no. 4, pp. 294–301, 2017.

[14] A. Mufa and E. Sudarmilah, "Game Anti Narkoba Berbasis Multi-Platform," *Khazanah Inform.*, vol. 2, no. 2, pp. 95–98, 2016.

[15] M. Fadhli, "Pengembangan Media Pembelajaran Berbasis Video Kelas IV Sekolah Dasar," *J. Dimens. Pendidik. dan Pembelajaran*, vol. 3, no. 1, pp. 24–29, 2015.

[16] H. Supriyono, R. F. Rahmadzani, M. S. Adhantoro, and A. K. Susilo, "Rancang Bangun Media Pembelajaran Dan Game Edukatif Pengenalan Aksara Jawa ' Pandawa ,'" *Pros. 4thUniversity Res. Colloq. 2016*, pp. 1–12, 2016.

[17] R. A. Zulfikar and A. A. Supianto, "Rancang Bangun Aplikasi Antrian Poliklinik Berbasis Mobile," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 3, p. 361, 2018.

[18] M. S. Mustaqbal, R. F. Firdaus, and H. Rahmadi, "Pengujian Aplikasi Menggunakan Black Box Testing Boundary Value Analysis," *J. Ilm. Teknol. Inf. Terap.*, vol. I, no. 3, pp. 31–36, 2015.

[19] M. Kumar, S. K. Singh, and R. . Dwivedi, "A Comparative Study of Black Box Testing and White Box Testing Techniques," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 3, no. 10, pp. 32–44, 2015.

[20] E. Febriyanto, U. Rahardja, A. Faturahman, and N. Lutfiani, "Sistem Verifikasi Sertifikat Menggunakan Qrcode Pada Central Event Information," *Techno.COM*, vol. 18, no. 1, pp. 50–63, 2019.

[21] S. M. Muyasaroh and E. Sudarmilah, "Game Edukasi Mitigasi Bencana Kebakaran Berbasis Android," *J. PROtek*, vol. 06, no. 1, pp. 31–35, 2019.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Knowledge Extraction on Reducing the Number of Students Using Explore, Elaborate and Execute Techniques

Juvinal Ximenes Guterres*, Ade Iriani, Hindriyanto Dwi Purnomo
Information Systems Masters Program, Faculty of Information Technology
Universitas Kristen Satya Wacana Salatiga
Central Java
*guterresmenex@gmail.com

Abstract-The decline in the number of students at East Timor's private universities can create new problems. This will be a burden for universities and other private institutes to develop. The financial and financing components are the factors that determine the implementation of teaching and learning activities that are used, among others, for the cost of facilities and teaching equipment. This study aims to extract knowledge of the decline in the number of students that occur every year at UNITAL. The method used is knowledge capture with Explore, Elaborate, and Execute techniques. Data collection techniques carried out by observation, interviews, documentation and questionnaires. Explore techniques to investigate results data, observations, interviews, documentation and questionnaires related to the reduction in the number of students arrested in the form of causes and effects of problems, then elaborate techniques describe interview, observation, brainstorming, and documentation data. The execute process is the stage of executing data from knowledge capture techniques based on explore and elaborate techniques and then produces a tacit and explicit knowledge from the actors or actors. The results obtained from this research are able to identify that the decline in students is caused by the main factors, namely, the quality of human resources consisting of lecturers, staff and technicians, service quality, facilities and infrastructure, buildings, classrooms, laboratory facilities, libraries and UNITAL academic information systems. Externally, there is a lack of cooperation with universities, both domestic and foreign, promotion to the public both directly and through social media. The results of knowledge capture is wealth that can be stored in a repository to be shared to help the processing of knowledge with the help of Information Technology.

Keyword: Knowledge Capture; Explore; Elaborate and Execute; knowledge management; Student Decline.

1. Preliminary

The education sector has a very important role to support the development of a country because it is a place for the creation of creative and innovative human resources, which will contribute to the progress of a country both now and in the future, essentially every college is an organization where knowledge created and can be used continuously. With routine activities to serve the interests of students, parents, lecturers, employees and community users of graduates [1].

Democratic Republic of Timor Leste (RDTL), the Government was formed in 2002. The Government of Timor-Leste through the ministry of education continues to strive hard to improve the quality of human resources. In addition to the construction of physical facilities and infrastructure, human resource development continues through the process of providing education. This is a clear evidence that after declaring independence unilaterally for 17 years, there are 15 tertiary institutions, both public and private. With the increase in the number of tertiary institutions in Timor-Leste it will provide more opportunities for the community to choose tertiary institutions according to their choices and economic conditions.

East Timor Oriental University (UNITAL) is one of the private universities that always pays attention to all learning processes and quality in service to consumers, namely students. Quality of service becomes a very important role for the continuation of an educational institution, for the achievement of student satisfaction and loyalty. Steps that need to be taken by educational institutions to improve these services is to optimize the ability of human resources and improve facilities and infrastructure that support the smoothness of educational services [2]. To realize professional education and training, it takes a solid commitment among all academic community members by mutually contributing by all management,

lecturers, staff and students to follow the rules that have been set together. The high number of students must be balanced with good service, synchronization of academic regulations and facilities and infrastructure is needed to strengthen the commitment of all stakeholders to carry out and evaluate their respective activities.

Services provided to students are the top priority in providing quality academic services capable of providing student satisfaction. It also produces quality services so that students can evaluate the services they receive. Both parties have a reciprocal relationship so that each party gets the same satisfaction.

Services to students as the most important element in educational institutions that need to be listened to, whether the service is in line with expectations or not because the quality of services provided to students needs to be evaluated regularly. Also provided facilities in accordance with the needs of students in order to know what is actually expected, what satisfaction is felt by students to be able to increase their loyalty to the institution. High loyalty will reflect the behavior in order to maintain the good name of the institution during the lecture process.

To produce high-quality student graduates, it is inseparable from high-quality academic services, good services can be supported by infrastructure that provides security and comfort for students during the learning process, where human resources are lecturers, guardians, administrative staff and technicians . [3]

Based on interviews and observations conducted on the leadership and related parties, it is found that there has been a decline in the number of students in the last four years, a decrease in the number of students occurring for prospective new students who register and active students, the temporary hypothesis that the cause of the decline in students is a big factor (brand image) UNITAL is damaged in the community, lack of facilities and infrastructure, curriculum offered, non-strategic location factors, lack of good service to students. UNITAL student data are shown in tables 1 and 2.

The data above shows the number of new students found in each faculty during 2014-2018. From this data the faculty of education is the faculty that has the most number of students compared to 7 other faculties. However, every year there is always a decrease in prospective new students who register, from the last four years continuously and the decline is also experienced by all faculties. Table 2 shows data on the number of new students at UNITAL universities.

Table. 1 UNITAL New Student Data 2014/2015 -2017/2018

School year		Faculty							
		S1	S1	S1	S1	S1	S1	S1	S1
		Economics	Agriculture	Law	Political	Education	Technique	Health	Mining
2014/2015	Interest	1100	1000	500	480	1400	1200	1000	900
	Be accepted	1050	900	500	480	1350	1140	550	880
	Rejected	50	40	0	0	50	60	450	200
	Re-registration	1030	855	500	450	1350	1135	550	880
2015/2016	Interest	1000	450	210	170	790	400	800	400
	Be accepted	800	450	210	170	790	400	500	400
	Rejected	10	0	0	0	0	11	300	0
	Re-registration	700	400	200	150	600	355	450	370
2016/2017	Interest	500	330	210	110	466	300	400	370
	Be accepted	500	330	210	110	466	300	350	370
	Rejected	0	0	0	0	0	0	50	0
	Re-registration	400	260	200	99	400	288	350	350
2017/2018	Interest	400	199	100	76	250	299	350	277
	Be accepted	400	199	100	76	250	299	320	277
	Rejected	0	0	0	0	0	0	30	0
	Re-registration	270	155	88	66	250	255	320	268

Source: BAA UNITAL University Year 2015-2018

Table 2. New Student Data on UNITAL Programs from 2014/2015 - 2017/2018

School year	Faculty							
	Economics		Agriculture		Law		Political	
	Be accepted	Re-registration	Be accepted	Re-registration	Be accepted	Re-registration	Be accepted	Re-registration
2014/2015	1050	1030	900	855	500	500	480	450
2015/2016	800	700	450	400	210	200	170	150
2016/2017	500	400	330	260	210	200	110	99
2017/2018	400	270	199	155	100	88	76	66

Tahun Ajarang	Fakultas							
	Education		Technique		Health		Mining	
	Be accepted	Re-registration	Be accepted	Re-registration	Be accepted	Re-registration	Be accepted	Re-registration
2014/2015	1350	1350	1140	1135	550	550	880	880
2015/2016	790	600	400	355	500	450	400	370
2016/2017	466	400	300	288	350	350	370	350
2017/2018	250	250	299	255	320	320	277	268

The decline in the number of students at private universities in East Timor in recent years has not only occurred at UNITAL but other private universities have also experienced the same decline. This decline in the number of students can create new problems for UNITAL and also other private universities because it will become a burden for universities to develop because the financial and financial components are the factors that determine the exact teaching and learning activities that are used, among others, for the costs of facilities and teaching equipment..

Reduction in funding caused poor service conditions, neglected building maintenance, and decreased academic quality. Student satisfaction becomes a reference for organizations to meet student needs so that they excel in sustainable competitiveness. Students choose the services of an organization to survive based on information from friends, family, service user institutions after comparing the services they experience with what is expected [4].

The decrease in the number of students can indicate that there are issues that must be taken seriously, the lack of facilities, human resources and quality of service and good relations with external in meeting the needs of students will become obstacles in the learning process, all the knowledge possessed by the leadership and staff in handling student data is not well documented or still tacit.

Handling the problem of student decline is done by having stakeholders who have gained a lot of knowledge (knowledge creating) and shared knowledge (knowledge sharing) within the university organization through evaluation meetings. Knowledge capture modeling will describe the problem situation, capture knowledge in the process, tacit knowledge pooling and the process of its application. expressing knowledge that was formerly a personal knowledge for each person will be a wealth of university organizations to strategically build in handling the decline of both prospective students and drop out students, and the capture of knowledge can utilize information technology [5].

From the description of the problems that have been raised, it is necessary to create knowledge in the process of handling the problem of reducing the number of students that can be documented so that it becomes a wealth of knowledge and can be utilized, then it can be applied in the concept of Knowledge Capture. Explore, elaborate and execute techniques will be applied as a method to capture the knowledge possessed by stakeholders, especially leaders and related parties who fight in handling student services. This three-way amalgamation technique becomes one to deal with the problem. Therefore, researchers will discuss this problem with the problem formulation "How to analyze the decrease in the number of students with Explore, Elaborate and Execute techniques"

2. Method

a. The method used

1) Knowledge Management

Knowledge Management is a collection of principles, processes, organizational structures and technology applications that help people share and improve knowledge to meet business goals in the organization through the application of information and communication technology so that it can be shared with all employees. In essence there are two main types of knowledge, namely tacit knowledge is the experience of every individual that involves several factors such as perspective, beliefs, experience and personal values, and tacit knowledge is difficult to transfer. While explicit knowledge is knowledge that can be transferred and created in books, data, newspapers, magazines. [6].

2) Knowledge Management Component

The most important components contained in KM are four, namely:

a) **Human**. Human knowledge management is the main actor in the process, and is also a source of knowledge management knowledge.

b) **Technology**. As the main media for distributing knowledge and making it easier to use information and knowledge through technology.

c) **Process**. Is an activity that consists of capturing, filtering, validating, transforming, knowledge throughout the organization equipped with carrying out certain procedures and processes.

d) **Content**. It is information and knowledge along with documents needed by people in implementing their obligations in knowledge management.

3) Model Socialization, externalization, Combination, Internalization (SECI)

Converting two types of knowledge, namely tacit and explicit through four kinds of conversion processes namely SECI shown in Figure 1.

	Tacit Knowledge	*Explicit Knowledge*
Tacit Knowledge	*Socialization*	*Exterrnalization*
Explicit Knowledge	Internalization	Combination

Figure 1: SECI Model

a) Socialization is the process of sharing and creating tacit knowledge through direct interaction and experience from individual to individual.

b) Externalization is the articulation of tacit knowledge into explicit knowledge through a process of dialogue and reflection, tacit knowledge in a more general form so that it can be understood by other people's explicit knowledge.

c) Combination is the explicit conversion of knowledge into more subsets.

d) Complex through systematics and the application of explicit information and knowledge from groups to organizations.

e) Internalization is the conversion of explicit knowledge into tacit knowledge of organizational members, which is spread throughout the organization through its own experience, so that it becomes new tacit knowledge from the organization to individuals. [7].

Research conducted by Ammar A. Ali Zwain et al., That in this information age mentioned that almost all organizations are driven by knowledge to achieve or maintain competitive advantage. KM in education can be defined as a tool that gives instructions to managers and staff in organizations. KM helps educational organizations to realize the advantages and beauty of creating knowledge sharing as a means to improve teaching and learning. Because higher education is currently subject to market pressures that are required to make big changes to compete. Universities think like business to the point that students are now treated as customers, so Universities have the responsibility to produce graduates who are able to accommodate challenges that arise in society, such as producing high-quality graduate profiles and competencies in their respective professions [8].

4) Knowledge Capture

Capturing knowledge is the process of thinking of an expert that can be captured properly. To produce knowledge management, a knowledge management application maker can work together with the relevant experts to translate it into a pre-programmed application.

Research conducted by Silwattananusarn and Tuamsu, namely that data mining in developments in business organizations there are four findings in this study. Knowledge sources, knowledge types, knowledge datasets, data development tasks, and data development applications used in knowledge management. Can conclude in an organization, knowledge is an important resource. Knowledge resource management has become a strong demand for development and finding useful knowledge and also management in decision making [9].

5) Customer Satisfaction

Consumer satisfaction is the main goal of every company. Universities as institutions that provide educational services to students, therefore must uphold student satisfaction as a way to retain students as consumers, subjective evaluation of students from what is felt and this experience will be formed in a sustainable manner in everyday life in the relevant university environment with education. For companies engaged in customers, it can cause consumer dissatisfaction with disappointment, anger, and protest so that consumers can leave the company, while customer satisfaction can increase company profits and create consumer loyalty to the company [10]. According to Kotler 2000, service quality is the basis for service marketing, because the core product being marketed is performance (quality), and performance is purchased by customers, therefore service quality performance is the basis for service marketing. The concept of good service will provide opportunities for companies to compete in winning consumers. While the good performance (quality) of a service concept creates a competitive situation where it can be implemented through strategies to convince customers, strengthen the image of sales and pricing [11].

6) Explore, Elaborate, *and* Execute

Exploration is the first step in increasing understanding and building knowledge through a phenomenon, so in this exploration cycle to build their own knowledge of the stimuli obtained, exploration activities do not only focus on what is found but arrive at the process of knowledge

excavation that is somehow related to the previous material or information is completely new. Elaborate is the actualization of various meaningful activities and works from a series of activities. Elaboration means completion and diligence and execution is the application or implementation. execution are some of the methods commonly used in conducting research [12].

b. Research Phase

This research is a qualitative research with the model used is knowledge capture with techniques. Explore (exploration), Elaborate (elaboration), and Execute (execution). Data collection is done by means of, observation, interviews, documentation and questionnaires. The problem solving process uses steps, i.e.

1. Case Study Analysis Phase,
2. Knowledge Capture Modeling Phase

The following is an illustration of the problem solving process as shown in Figure 2.

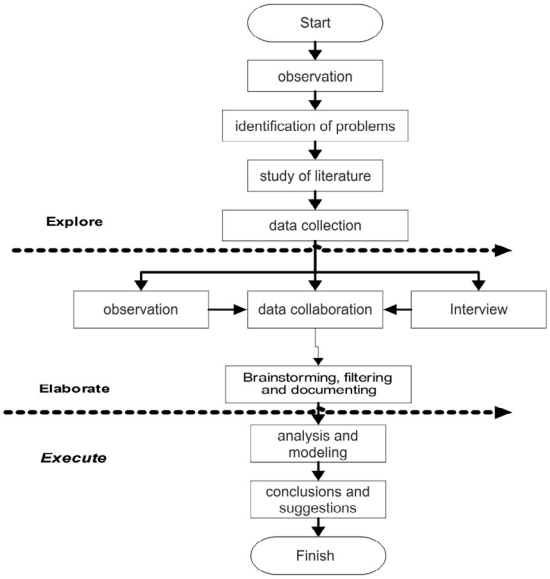


Figure 2. Framework for Resolving Problems [5]

In this study there are several processes to solve the problem, namely the exploration process, the analysis process and the knowledge capture modeling process. The researcher began by observing at the UNITAL research location to identify problems related to the decrease in the number of students and to see firsthand all the activities carried out by the actors. After making observations, it can be identified on several things that cause a decrease in the number of students and can also identify that the number of students decreased significantly from the last four years.

Interviews were conducted to find out the tacit knowledge that was known by the actors in handling the problem of decreasing the number of students. Interviews were conducted by the chancellor, vice chancellor (WR), deans, study program heads, lecturers, and students as service users who needed to be captured. From the capture of this knowledge the researchers initially conducted a study of related literature so that the results of the research could become credible scientific work. Data collection through observation and interviews and then data is collaborated to find out what causes the decrease in the number of students that occur each year. In the exploration stage is the initial stage of processing data that has been collected through collecting data, elaborate is resistant to data collection by actualizing the data collected, the execute stage is comparing brainstorming information, documenting that is done with actors to produce models in knowledge capture.

Knowledge Capture Model

In dealing with the problem of decreasing the number of students from Explore, Elaborate and Execute techniques, a new model of knowledge creating is formed, the process of final execution, arrest has resulted in explicit knowledge of the actors. Illustrated in Figure 3.

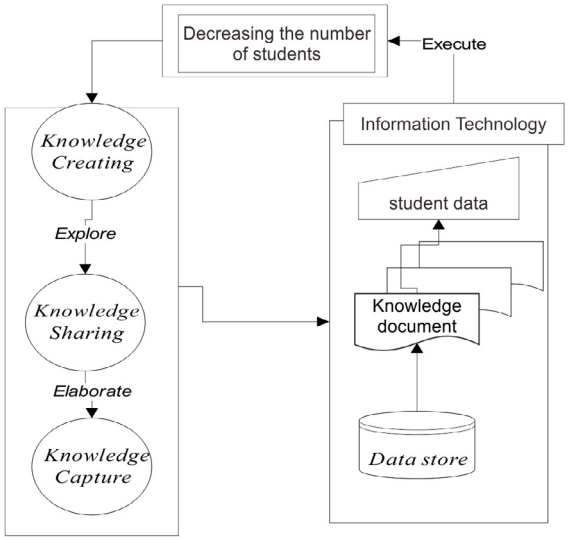


Figure 3. Knowledge Capture Model

The method applied in problem solving is to capture the knowledge associated with a decrease in the number of students using explore, elaborate, and execute techniques. During dealing with the problem, a knowledge creating by the actor was formed, where the capture of knowledge by exploring technique was carried out to find the cause of the decrease in the number of students and the knowledge of the actors regarding the decrease in the number of students.

During the process that was passed, in the course of this handling a tacit knowledge was formed that had not yet been structured and stored. That knowledge should be stored as the wealth of the organization to increase knowledge in decision making for the future. To be more structured, efficient and meaningful, knowledge must be shared by conducting elaboration techniques through knowledge sharing to accommodate all tacit knowledge from superiors, lecturers, employees and technicians, then filtering and documenting more structured.

The results of elaborate as a knowledge capture to be stored in a database are used to help the processing of knowledge with the help of information technology. The knowledge captured in the form of stored knowledge documents is a wealth of organization that always pays attention to the quality of service to students. This is also to anticipate a decrease in the number of students and to strive for prospective new students and active students who have gone on to survive to complete their education.

3. Results

In the field research, several causes can be identified to decrease the number of students, both new students who register and also active students who drop out every year. Based on observations and interviews conducted with the chancellor, WR, deans, program leaders, lecturers and students as the main actors in the academic process. The following are the results of the interview:

a. Interview Results Data

Table 3: Results of the Rector and WR Interviews, Deans, Study Program Heads and Lecturers.

	Factors That Affect Student Declines
Rector	Professional attitude, friendliness, empathy, and respect for students by lecturers, employees, and technicians. The financial system is in the form of burdensome increases and dispensations. Academic hospitality for filling in and compiling KHS and KRS, distributing and recapitulating active and inactive student lecturer data, facilitating examinations, documenting alumni, library management, and laboratories. the recruitment of lecturers, staff, and technicians does not meet the standard operating procedures (SOP)
Vice Rector	Domestic government political policy, institutional competition with facilities provided by other private state universities, the supply of physical facilities such as buildings, classrooms, tables, chairs, and furniture to provide security and comfort, public facilities and infrastructure. The academic system operates easily understood and easy to implement, clarity of information, and guarantee the confidentiality of information.
Dean	Promotional services to the public with an updated, safe and credible web information system, providing services to complaints from students and the public. Location and public transportation facilities that access UNITAL.
Head of the study program	The value of study program accreditation and curriculum renewal in accordance with the changing times. The environment around the building is clean, ensuring student comfort, and managing parking in an orderly and safe manner.
Lecturer	Availability of public facilities, computer standards in the laboratory, collection of books provided in the library, availability of free internet for students, e-library system.

b. Student Interview Data

Table 4: Student Interview Results

	Student Dissatisfied
Student	Timeliness of service to students for validation of SPP receipts, Community Guidance by custodian lecturers, faculty physical buildings and study programs with other supporting facilities. A careful and cooperative administrative system. The presence of lecturers provides courses, thesis guidance, modules used when giving lectures. Free Wifi for students, lecturers, employees, laboratory availability according to lecture needs.
Student	Clarity of information Employees to students, lecturers are ready to help students if needed, other programs that encourage students to be creative in improving abilities, such as: seminars, workshops, sports, dance, music. Friendliness of employees in serving students, understanding material by students when attending lectures. The method used by the lecturer can make it easier for students and lecturers to master the material taught, academic administration hours, finance, and faculties that are in accordance with student needs.
Student	Using information technology in the learning process and other information services, strategic location of the campus, library clerk working hours in accordance with the specified working hours, friendliness, willing to help when needing, explaining the book search system, information clarity, using information systems to manage the library. improve computers and other practical tools in the laboratory.

c. Observation Data

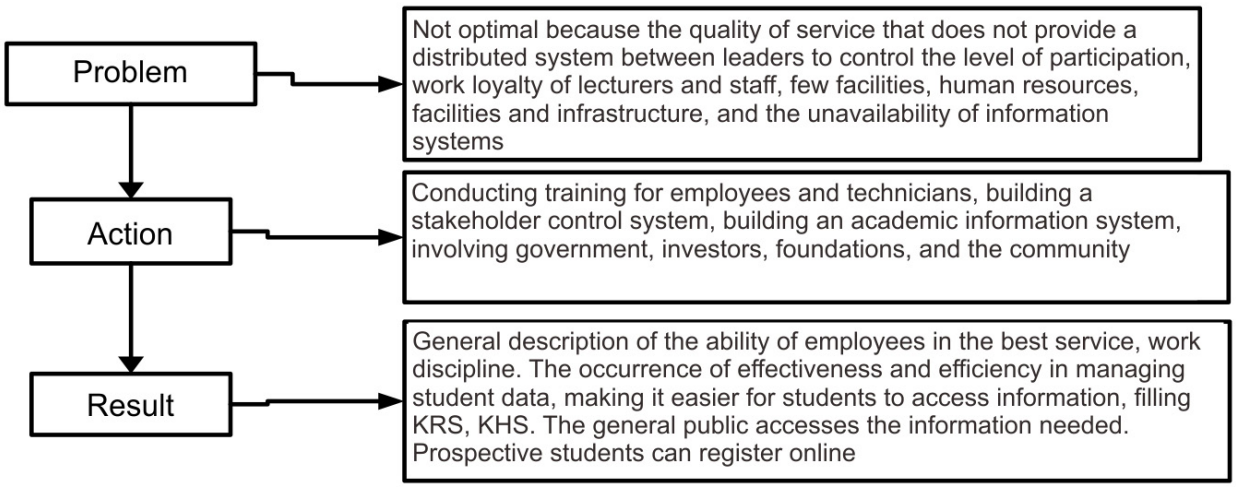


Figure 4. Thinking Framework for Observation Data

From observations made that internal and also external factors that influence so that problems arise in the decline of students. The learning process especially UNITAL is currently not optimal, namely the quality of services provided to students, due to the absence of a distributed system between superiors to control their subordinates so that the current human resources do not maximize their skills to build each faculty and study program, lack of participation rates, work loyalty of lecturers and employees and technicians. Besides that, from facilities, human resources (HR) facilities and infrastructure and information systems are a series of interrelated activities and are not separated from each other, if one is abandoned then the learning process cannot run properly so that learning objectives cannot be achieved. The lack of institutional collaboration with other universities both domestically and abroad, companies, non-government organizations and the community. After making observations and taking and researchers want to examine more deeply. Here is a framework for thinking from research conducted by researchers.

d. Brainstorming

Tacit Knowledge is knowledge that is rooted in one's actions and experiences so that the knowledge possessed by the individual is still categorized as intuition and conjecture [13]. To document the knowledge of tacit into explicit is by brainstorming. The brainstorming stage is as shown below.

The process of data retrieval using brainstorming techniques is an effort to extract the contents of the knowledge possessed by actors which encourages the emergence of many good ideas, bad ideas, criticisms, suggestions and creative ideas and then documents all the knowledge in the form of tacit related to the problem of decreasing the number of students. Braingstorming produces a document which is then filtered and selected as tacit knowledge related to the material, after filtering that knowledge to be applied in the knowledge of each tacit.

Table 5. Brainstorming Results

Actor	Knowledge
Actor 1	Responsive staff lecturers, and technicians are willing to help and have the capacity to solve urgent problems, employees have the capacity to solve problems, and respond to student complaints responsibly.
Actor 2	Guaranteed universality in service to students who are friendly, polite. The university's responsibility to provide information technology systems to facilitate both managing research, managing grades, student data, lecturers, staff and other resources, promotion information, which is effective and safe.
Actor 3	Academic elements such as: qualifications and professional lecturers, staff, and technicians, development planning and creativity in study programs, curriculum, instructions from superiors regarding methods, synchronizing assessments and evaluations for students, university financial management with transparency systems.
Actor 4	The academic department plans, accreditation programs, graduations, access for alumni, outside parties concerned. Striving for a campus atmosphere that provides comfort, campus clinics, student counseling services.

c. Problem Analysis Schema

Based on the results of interviews, observations, documentation and brainstorming can be found above then do an analysis of the consequences, reasons, and field conditions that have been carried out as follows:

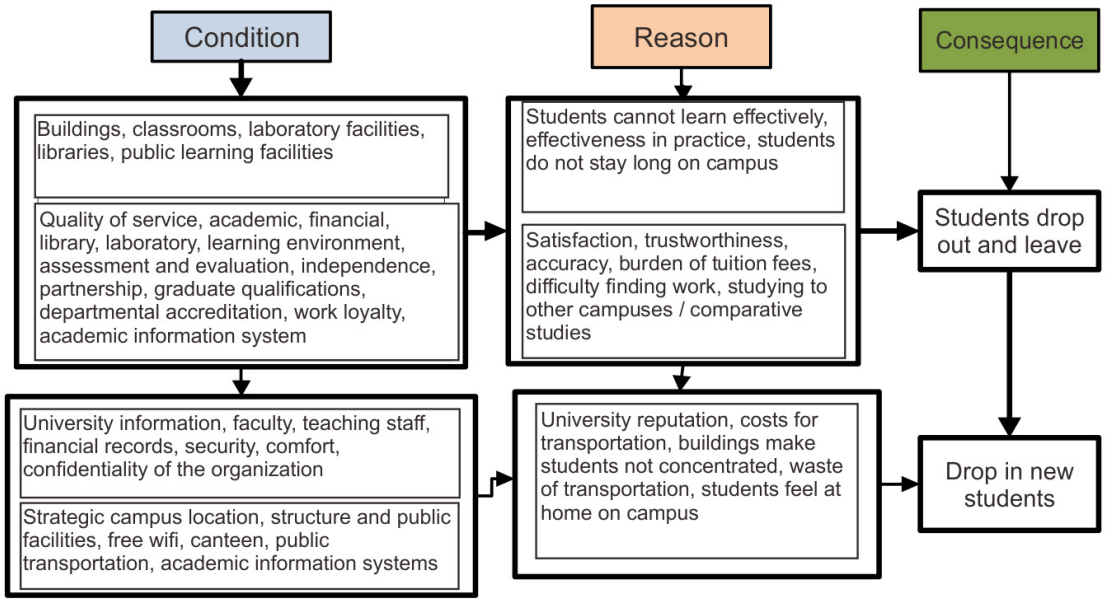


Figure 5. Schematic Analysis of Problems

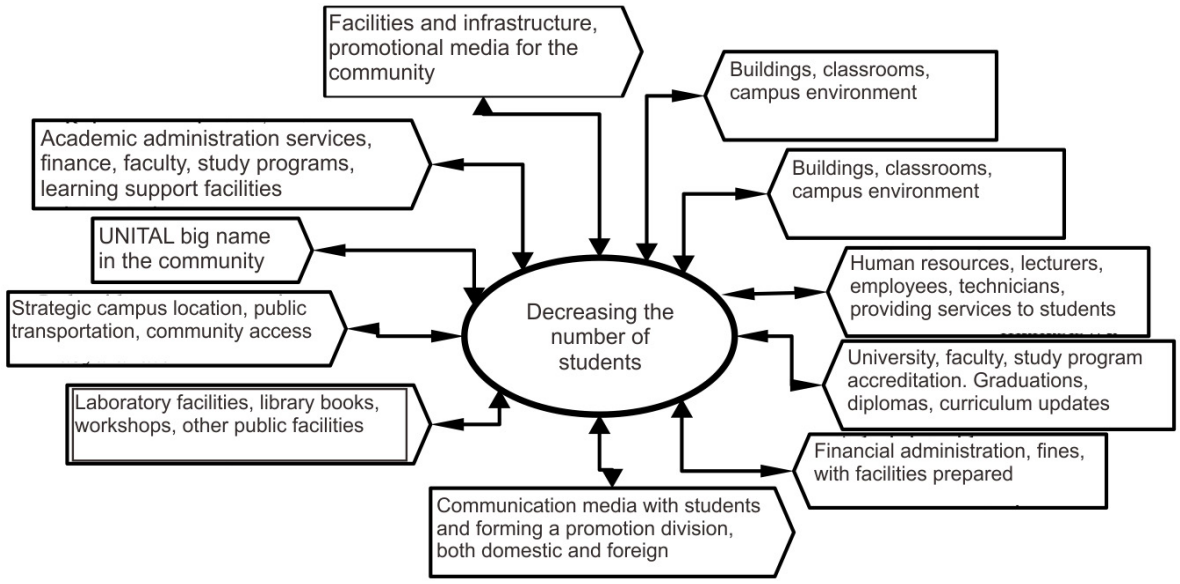


Figure 6. Causes of Student Decline

In the scheme above it can be found the cause of the decline in students due to several reasons such as the big name of UNITAL in the community, the safety and comfort of students during college, the strategic location of the campus. Other causes of the decline in active students who drop out are the lack of facilities to support the teaching and learning process, such as classrooms and furnishings, availability of facilities and infrastructure for students, laboratories and student practice hours, availability of books in the library, professionalism in academic services, HR (lecturers, employees and technicians), a campus atmosphere that provides student comfort, accreditation of study programs.

d. Explore

With the Explore data interview technique, observation and documentation conducted can show that there is indeed a very significant decrease in the number of students, from the knowledge captured, it can be identified that there are various reasons as a cause of the decrease in the number of students shown in Figure 6.

The decrease in the number of students each year, both prospective new students who want to register and active students who drop out are obstacles to the learning process. The causes of the decrease in students are internal management factors, facilities, facilities and infrastructure for students, lecturers and staff, books and library furniture, workshop laboratories for practice, studios, clinics, faculty building construction, strategic locations for public transportation access, political stability in country, HR factor, communication media to bridge students with academics as a credible promotion media, study program accreditation, UNITAL good name. The external factor is cooperation with other universities both

domestically and abroad, government institutions and the community.

e. Elaborate

To solve the problem in accordance with the elaboration of the framework in solving the problem at this elaboration stage can collaborate between four stages of research carried out namely Observation data collaboration, interview, brainstorming and documenting. The final result of the collaboration of observation and interview data obtained is the result of a discussion of a decrease in the number of students, namely exploration. From the data collected from the interview, the services provided to students are from the attitude of friendliness, clarity of information, certainty of information from lecturers, staff and technicians, the supply of special and public facilities such as laboratories, books and furniture in the library, buildings that have not been evenly distributed as a whole. all faculties.

The absence of an academic information system to provide accurate and up to date information to students and the public, the lack of human resources, and government politics. Observations, which are carried out to be able to observe directly that the decline in students can be caused by the convenience of students while on campus, the attitudes and loyalty of lecturers to work, time discipline, lack of a system to control all lecturers, employees and technicians, the lack of cooperation between universities, government and private institutions both inside and outside the country to conduct training for lecturers, staff and technicians. Brainstoring, accuracy of information held by lecturers, staff and technicians, because information is not centralized, qualifications and professional lecturers, staff, and technicians, development planning and creativity in study programs, curriculum, synchronization of assessments and evaluations for students, financial management of universities with systems transparency, information technology to facilitate both managing research, managing grades, data of students, lecturers, staff and other resources, promotion information, which is effective, and information that guarantees its safety.

Documentary, from the documentation can show that the decline in students occurs every year, and field conditions occur lack of facilities both public facilities, both physical and non-physical facilities.

The elaboration process is the result obtained based on the results of exploration, the material contained in this elaboration process is about the problem of decreasing the number of students. The question is why there is a decrease in the number of students not only for prospective new students but also for active students who drop out and leave and are no longer active. From the data in the field can show several problems and as a solution to the problem. This problem concerns the decrease in the number of new students enrolled and active students dropping out. This problem and the internal factors caused by external factors can be explained in Figure 7.

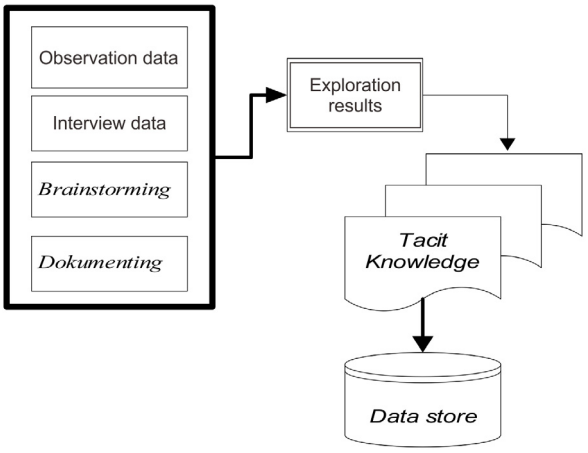


Figure 7. Elaboration Process

After finding material in the elaboration process, brainstorming with actors is carried out to gain the knowledge and experience they have regarding the problem of decreasing the number of students, the amount of experience they have is in accordance with the material prepared at the beginning.

f. Execute

This process is the third stage of the explore and elaborate process or the final stage can be illustrated in Figure 8.

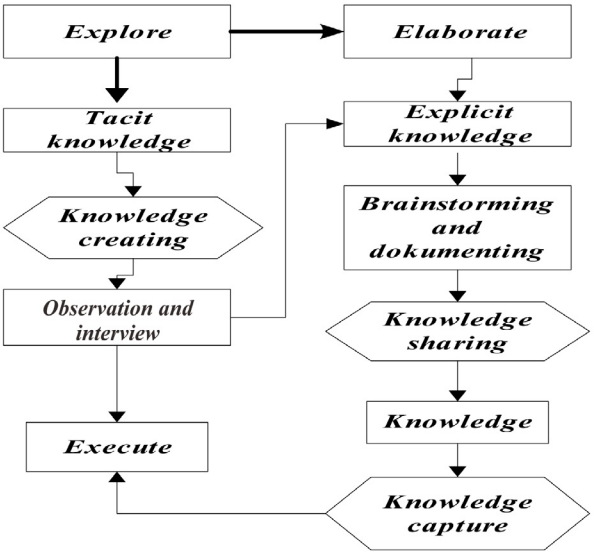


Figure 8. The execute process

The process of exploring and elaborating is done with the results obtained is that the material contained contains a decrease in the number of students. The process of exploration results obtained from data collaboration, observation and interviews which are the result of discussion of the problem of decreasing the number of students and elaborating data on Brainstorming and Documenting. The execute process is the last stage of knowledge capture techniques based on explore and elaborate techniques that produce explicit knowledge from actors.

g. Proposed Solution

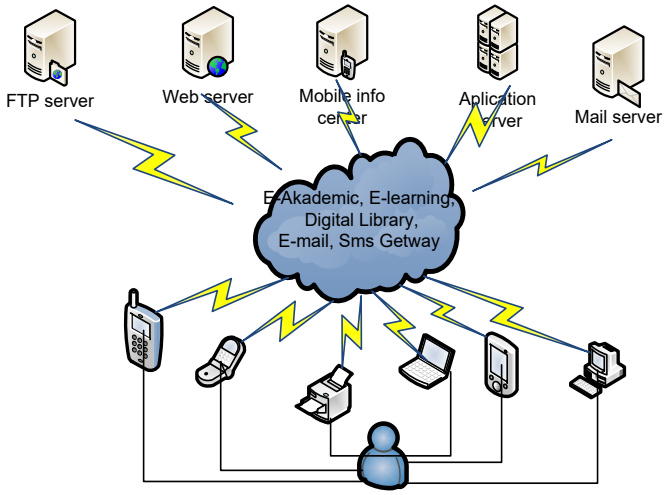


Figure 9. Concept of UNITAL Digital Campus

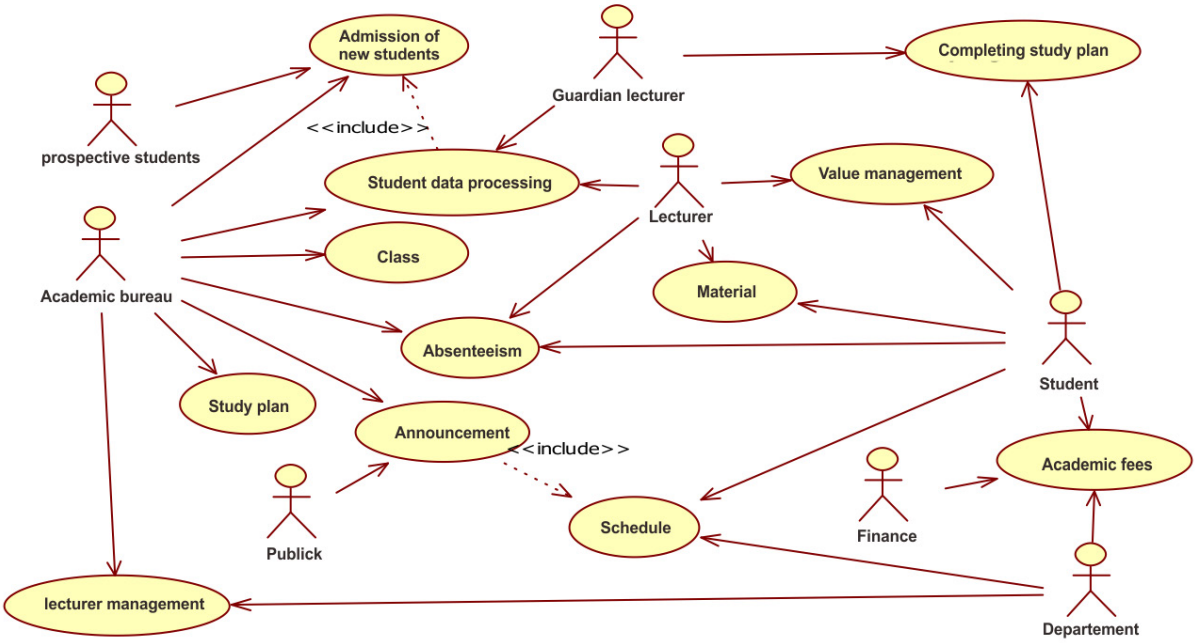


Figure 10. (Use Case) of the Academic System Proposed

The decrease in the number of students is the responsibility of all stakeholders. With competition in the world of education especially private universities becoming increasingly fierce, stakeholders are required to have strong intellectual capital and adhere to a system that is open to all and responds to all the desires and needs of effective and efficient stakeholders, especially for students and the community. Organizational structure must encourage the creation of a creative, innovative and efficient intellectual learning culture in the process of sharing knowledge between individuals so that they realize and understand their overall role. UNITAL is able to compete and even win the competition if it is able to offer quality products or value-added services desired by students and the community. The solution is to use Knowledge Management System (KMS) that refers to innovative technologies such as the internet, extranets, and data warehouses to facilitate and facilitate KM communication between inside and out.

The concept of digital-based campus will not be separated from supporting components such as the internet, computers, and websites. Digital campus will emphasize more services, namely the management of education administration that is effective and efficient. Online-based Student Information Systems will be a new way of recording transaction management and processing that is efficient in processing student information and will

help for administrative staff, academic staff, grant providers or stakeholders, and can also produce student data [14].

Supporting factors for the achievement of Information Systems Technology infrastructure such as academic web, sms center, mobile access, e-mail, e-learning and e-library. Internet network inventory will be able to facilitate users in accessing information through various devices such as laptops, computers, phones and tabs.

The schema below can illustrate that to meet the needs of students in computerized services and can improve performance in the service of quality human resources and improve competitiveness, academic information is needed in managing student grades, management data of faculties / majors, teaching staff data (lecturers) and also other relevant information. With the information system, BAAK can facilitate the management of student data, dosages, employees, KRS, announcements, attendance management, PMB registration, class management. Lecturers can access teaching schedules, KSR guidance, filling material, filling student grades. The finance department can carry out financial management with academic well. Students can access lecture schedule information on grades, lecture material and other academic information. Prospective students can access PMB information, information about public campus facilities and register online. The general public can access information about UNITAL in general. Implementation of information systems with KM, then an organizational and individual process becomes easily coordinated and systematic. Technology can provide added value through innovation thinking. And through Information Technology all knowledge can be inclusive and elaborated to develop short-term, medium-term and long-term strategies [15]. The proposed network scheme is shown in Figure 11.

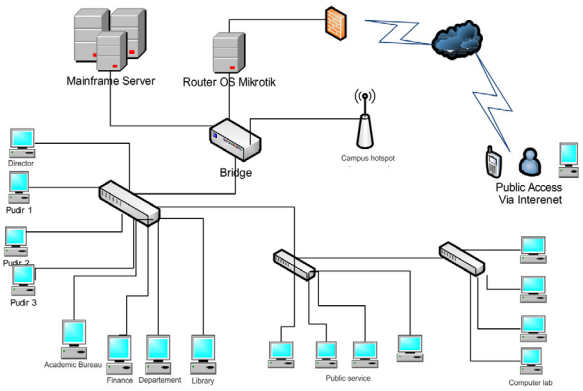


Figure 11. Proposed Overall Network Schema

4. Conclusion

To handle the case of the decline in the number of students, the chancellor, vice chancellor (WK) chancellor and lecturer, and employees already have Knowledge creating, as shared knowledge while in UNITAL.

In knowledge sharing through evaluation meetings at the rector and faculty level, knowledge is not captured to make important documents for the organization to be used in decision making. By using knowledge capture methods and explore, elaborate and execute techniques, it has succeeded in describing the conditions of the problem, the process of capturing knowledge and the process of pooling knowledge..

It is recommended for universities to improve the responsiveness of management to provide services to student complaints, seek regular training to all staff, technicians and educators. Building information systems that provide information, leadership, management systems, quality assurance, students and graduates, human resources, curriculum, academic atmosphere, financing, facilities and infrastructure, research information systems, community service, cooperation and Information Technology services.

References

[1] Retnoningsih Endang, "Knowledge management system (kms) dalam meningkatkan inovasi lppm perguruan tinggi",Evolusi Vol. I No.1. 2013

[2] Rinala Nyoman, Yudana Made., et al. **"**Pengaruh kualitas pelayanan akademik terhadap kepuasan dan loyalitas mahasiswa pada sekolah tinggi pariwisata nusa dua bali", program pascasarjana universitas pendidikan ganesha. Program studi administrasi pendidikan. Vol 4, No 1 (2013). 2013

[3] Shifia Amna Aulia, wiwik. "Analisis kepuasan mahasiswa terhadap layanan akademik di institusi xyz" Jurnal Teknik dan Ilmu Komputer, Vol. 05 No. 18, Apr – Jun 2016.

[4] Sumarno, Efendi Ahmad. "Dampak Biaya Kuliah Tunggal Terhadap Kualitas Layanan Pendidikan. Jurnal manajemen pendidikan", Magister Manajemen Pendidikan, FKIP Universitas Kristen Satya Wacana, e-ISSN 2549-9661, Volume: 4, No. 2, Juli-Desember 2017.

[5] Victor, Manongga, et al. "*Knowledge capture* menggunakan teknik *explore, elaborate* dan *execute* untuk bagian kesiswaan sekolah", jurnal informatika jurnal pengembangan IT(JPIT) vol 03, no. 03 september 2018, DOI: 10.30591/ Jpit. V313.1002. 2018

[6] Iskandar, tony. "Perancangan knowledge management system pada it bina nusantara menggunakan blog, wiki, forum dan document", Computer Science Department, School of Computer Science, Binus University, ComTech Vol. 5 No. 1 Juni 2014: 110-122. 2014

[7] Smith and lyles. *Handbook organization learning annd knowledge management*. The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom. 2014

[8] Ammar A. Ali Zwain. "Knowledge Management

Processes and Academic" Performance in Iraqi HEIs: An Empirical Investigation, International Journal of Academic Research in Business and Social Sciences June 2012, Vol. 2, No. 6 ISSN: 2222-6990.

[9] Silwattananusarn Tipawan, Tuamsuk Kulthida. "Data Mining and Its Applications for knowledge Management" Khon Kaen University, Thailand Vol.2, No.5, September 2012 DOI : 10.5121/ijdkp.2012.2502 13.2012

[10] Ina ratnasari. Pengaruh kualitas pelayanan dan citra institusi terhadap kepuasan mahasiswa yang berdampak pada *word of mouth* (studi kasus pada mahasiswa universitas singaperbangsa karawang.. *Value Journal of Management and Business, ISSN 2541-397X, Vol. 1 No. 1 Oktober 2016.*

[11] Kotler philip. Maketing management the millenium edition. Ten edition. USA: prentice-hall,inc.)2000

[12] Pramono, Nia Ariani. "Kemampuan Guru Melaksanakan Kegiatan Eksplorasi, Elaborasi dan Konfirmasi dalam Pembelajaran Bahasa Indonesia SD Negeri 182/I Hutan Lindung". 2018.

[13] Karto iskandar; tony; claudia henlly phankova; wongso agustino "Perancangan knowledge management system pada it bina nusantara menggunakan blog, wiki, forum dan document" Computer Science Department, School of Computer Science, Binus University ComTech Vol. 5 No. 1 Juni 2014: 110-122 no 5).

[14] Eileen Bayangan-Cosidon. "Student Information System for Kalinga State University-Rizal Campus". International Journal of Management and Commerce Innovations ISSN 2348-7585 (Online) Vol. 4, Issue 1, pp: (330-335), Month: April 2016 - September 2016.

[15] Muhammad Jawad Iqbal, Amran Rasli, at., el.. "*Academic staff knowledge sharing intentions and university innovation capability* Vol.5 (27), pp. 11051-11059.2011

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Silhouette Density Canopy K-Means for Mapping the Quality of Education Based on the Results of the 2019 National Exam in Banyumas Regency

Ridho Ananda
Faculty of Industrial and Informatics Engineering
Institut Teknologi Telkom Purwokerto
Indonesia
ridho@ittelkom-pwt.ac.id

Abstract-Mapping the quality of education units is needed by stakeholders in education. To do this, clustering is considered as one of the methods that can be applied. K-means is a popular algorithm in the clustering method. In its process, K-means requires initial centroids randomly. Some scientists have proposed algorithms to determine the number of initial centroids and their location, one of which is density canopy (DC) algorithm. In the process, DC forms centroids based on the number of neighbors. This study proposes additional Silhouette criteria for DC algorithm. The development of DC is called Silhouette Density Canopy (SDC). SDC K-means (SDCKM) is applied to map the quality of education units and is compared with DC K-means (DCKM) and K-means (KM). The data used in this study originated from the 2019 senior high school national examination dataset of natural science, social science, and language programs in the Banyumas Regency. The results of the study revealed that clustering through SDKCM was better than DCKM and KM, but it took more time in the process. Mapping the quality of education with SDKCM formed three clusters for social science and natural science datasets and two clusters for language program dataset. Schools included in cluster 2 had a better quality of education compared to other schools.

Keywords: Density canopy, K-means, Quality mapping, Silhouette.

1. Introduction

National Examination (UN) is a national-scale examination activity with the reference of graduate competence standard [1]. UN is held at the elementary level (SD), junior high school (SMP), and senior high school (SMA) or equivalent level in certain subjects. The government institution obliged to carry out the UN is Badan Standar Nasional Pendidikan (BNSP) aimed at measuring the fulfillment of graduate competence. The results of the UN then can be used as a mapping of the quality of education units [2].

The mapping of the quality of the education unit program is important to help education stakeholders in making education-related policies. BNSP has mapped the quality of the education unit program based on the UN results [1]. The quality is then classified into four criteria: (a) "excellent" if , (b) "good" if (c) "satisfactory" if , and (d) "poor" if . These criteria certainly have weaknesses when applied to the average UN results since the average calculation is sensitive to extreme values or outliers [3]. Therefore, we need a specific method that can provide the mapping of the quality of the education unit program concerning UN results where the calculation does not directly use the average score. Clustering is considered to solve this problem.

Clustering is a statistical classification technique to determine the classification of an individual of a population to be grouped into the same group or different group based on the quantitative comparison of the measured variables [4]. A clustering algorithm is needed to apply the clustering process. One of the classic and well-known algorithms is K-means, proposed by MacQueen[5]. K-means is included in the the unsupervised learning group category. The main principle of the K-means is classifying objects based on Euclidean distance measurement. The use of K-means is quite popular in the field of technology [6] [7], health [8], education [9][10][11] and other fields. In the process, K-means requires initial centroids in the first step. A centroid is the central data in certain cluster. This determination is often done randomly which affects the accuracy of the clustering results. To overcome this problem, some scientists have carried out some research and development of the K-means and proposed new algorithms such as canopy [12], K-means++ [13], K-means-u [14], and DCKM [15]. All these algorithms have been tested by [15] with the conclusion that DKCM is the most effective algorithm and can overcome the

existence of extreme data or outliers. In DC, the first centroid is chosen from the object which has the most neighbors based on its Euclidean distance. This leads to the formation of less effective initial centroids due to the possibility that the first centroid is most likely not far from the data centroid and has the most members. Whereas the other centroids are obtained from the periphery of data that are not neighbors to the original centroid.

Based on the description above, research related to mapping the quality of the education unit program was conducted with the basis of the results of the UN. The data used in this study were data of 2019 senior high school UN results in Banyumas Regency. This current study used the DC algorithm as preprocessing of K-means for its ability to choose the initial centroid and its optimal location. On the other hand, the algorithm was appropriate since data were numerical. This study proposes a modification to the DC algorithm, which in the algorithm process the centroid determination should also consider Silhouette criteria. This study is expected to provide information for education stakeholders related to the mapping of the quality of the education unit program based on the results of the UN at the SMA level in Banyumas Regency in 2019. It is also hoped that the proposed modification can contribute to meaningful knowledge especially in clustering.

2. Research Method

a. Research Procedure

The research procedure for each dataset is shown in figure 1.

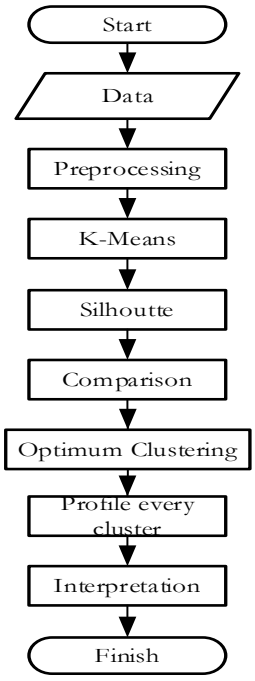


Figure 1. The Diagram of the Research Procedure

The computational process used Matlab 2014a software. In the preprocessing stage, this study used a modified density canopy and density canopy. Preprocessing is an initial step performed before the main step (clustering process). Then the clustering process was done with K-means and the Silhouette value was calculated. Furthermore, the clustering results of modified density canopy k-means were compared with the results of density canopy k-means and regular k-means. The optimum results will be used to map the quality of the education unit in Banyumas Regency.

b. Algoritme Density Canopy

Density canopy (DC) algorithm, proposed by [15], is the development of the Canopy algorithm [12]. The algorithm is a preprocessing of K-means to determine the initial centroids. DC selects the first centroid based on maximum density (number of neighbors) as shown in Figure 2. The figure provides an illustration of centroid selection on the DC algorithm. The steps of the DC algorithm are as follows:

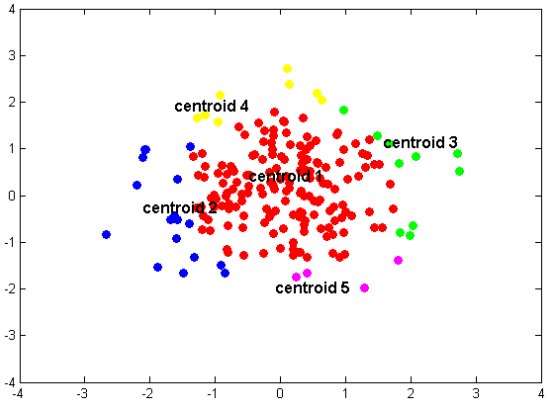


Figure 2. Distribution of initial centroids of the DC algorithm on random data with normal distribution

Step 1. Suppose matrix dataset sized The first step, calculate the average distance between objects with the formula (1) and the density of each object with the formula (2).

$$\bar{d}_E = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_E(\mathbf{x}_i, \mathbf{x}_j). \quad (1)$$

is the row vector of the object-i and is Euclidean distance of the object-i to object-j or vice versa using formula (9).

$$\rho(i) = \sum_{j=1}^{n} f\left(d_E(\mathbf{x}_i, \mathbf{x}_j) - \bar{d}_E\right) \quad (2)$$

where valued 1 for and for the rest. To each object-j that fulfills is considered as the closest neighbor of the object-i and the matrix of neighboring objects is formed as shown in equation (3).

$$\mathbf{N}_i = \left[\mathbf{x}_i, \mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \cdots, \mathbf{x}_{j_{\rho(i)}}\right] \quad (3)$$

Step 2. Select object-i with maximum density (maximum) as the initial centroid, then objects which enter are deleted from the dataset.

Step 3. The rest of the objects on the dataset are calculated and the neighbors are determined or to each object-i. The next the calculation of the average distance between objects in the neighbor matrix of the object-i so that is obtained And then calculating local density with the formula (4).

$$q(i) = \begin{cases} \min d_E(\mathbf{x}_i, \mathbf{x}_j) & if\ \exists j\ \ni\ \rho(i) < \rho(j) \\ \\ \text{maks}\ d_E(\mathbf{x}_i, \mathbf{x}_j) & if\ \forall j\ \ni\ \rho(i) > \rho(j) \end{cases} \quad (4)$$

Step 4. Calculate the weight of the object with the formula (5) for .

$$w(i) = \rho(i) \times \frac{1}{a(i)} \times q(i). \quad (5)$$

Step 5. Object with the highest score is chosen as the next centroid and objects in are deleted from the dataset.

Step 6. Repeat step 3 until no objects left in the dataset.

The obtained centroids are used as initial centroids in the grouping process with K-means.

From figure 2, it can be seen that the distribution of the centroids indicates unequal neighbor distribution. Centroids formed earlier have more neighbors than centroids come afterward. This is because DC is greedy in the process in which the centroid is chosen from the most neighbors consequently the next centroid is obtained from the rest of the dataset. There is a possibility that the last centroid has no neighbor because its neighbors have been claimed by previous centroids. Considering this condition, the current study proposes additional criteria to determine centroid with DC. The proposed additional criteria are inter-cluster distance and nearest cluster to the centroid candidate using the Silhouette formula.

c. Algoritme Silhouette Density Canopy

The underlying condition for the proposed additional criteria in the density canopy algorithm is the centroids selection based on the number of neighbors they have. The additional criteria proposed in this study are the Silhouette criteria. Figure 3 is the concept of the added criteria. The distance of a centroid to its neighbors is called intracluster distance, while the distance of centroid candidate to the objects that are not its neighbors is called the nearest cluster distance. The choice for Silhouette criteria is because it can determine many optimum clusters [16]. The best results of clustering are obtained when the Silhouette value is maximum. Based on this information, the multiplication operation will be used to combine the Silhouette value with maximum neighbor selection. Further, the proposed algorithm is named Silhouette density canopy (SDC). The step of the SDC are as follows:

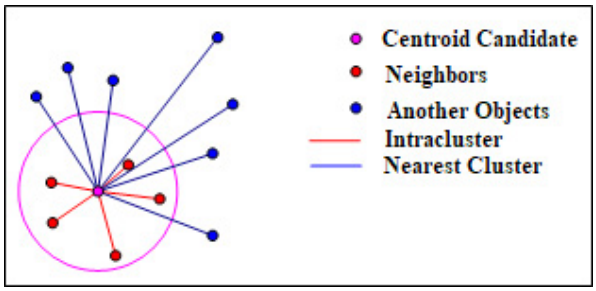


Figure 3. The Centroid cadidate based on Silhouette criteria on SDC

Step 1. Suppose matrix dataset sized The first step, calculate the distance between objects with the formula (1) and the density of each object with the formula (2). The next, create the neighbor matrix of the object-i with equation (3).

Step 2. Calculate the Silhouette criteria of the object-i with the formula (6).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (6)$$

Where is the average distance of the object-i to its neighbors. is the average distance of from all other objects out of its neighbors.

Step 3. Select the object-i with maximum criteria as initial centroid then objects in are deleted from. Maximum criteria are obtained from formula (7).

$$\text{argmax}_i(p(i) \times s(i)) \quad (7)$$

Step 4. The ρ*(i)* of the rest of the objects on the dataset are calculated and their neighbors are determined or for each object-i. Then, calculate the average distance between objects on the neighbor matrix of the object-i and their Silhouette criteria . The next, calculate the local density with the formula (4).

Step 5. Calculate *w(i)* of each object with formula (8) for.

$$w(i) = \rho(i) \times s(i) \times \frac{1}{a(i)} \times q(i) \quad (8)$$

Step 6. The object with the highest score assigned as the next centroid and the objects in are deleted from the dataset.

Step 7. Repeat step 3 until no objects left in the dataset.

d. K-means

K-means (KM) algorithm is a grouping algorithm that minimizes the distance between the objects of the same group and maximizes dissimilarity between objects of different groups. The dissimilarity size used is the Eucladian distance [17] which is calculated with the formula (9).

The steps of the KM algorithm are as follows:

step 1. By using centroid obtained from Density canopy algorithm, for example, , every object is assigned to the group closest to its centroid.

step 2. Determine the new centroid of the average member. Every object is allocated to the group closest to the centroid formed. Each object can move to other groups if the new centroid makes it closer to the previous centroid.

step 3. Repeat step 2 iteratively until there are no changes in the grouping.

e. Model Validation

The validation is needed to measure the quality of clustering. There are two types of validation, those are external validation and internal validation [18]. In this study, the researchers only used internal validation. The external validation was not performed due to the absence of initial group information. It is said that the accuracy of the internal validation is higher than the external validation [18][19]. The internal validation used was the Silhouette index using formula (6). In the Silhouette validation, is the Silhouette size of the object-i. is the average distance between and other objects within a cluster (intracluster). is the average distance between and the objects in the nearest cluster. The higher the value the more appropriate the placement of the objects in the group. Silhouette index is obtained from the average size of the silhouette. Silhouette index ranges from -1 to 1, where the value closer to 1 indicates the object is well matched to its cluster [20]. The ability of the Silhouette index to validate the results of grouping is considered to be better than some validations in other fields [21]. The use of Silhouette validation has also been used, among others to validate clustering data of the automatic dependent surveillance-broadcast [22], clustering the province in Indonesia based on rice production [23], and the clustering of dengue-prone areas [24].

3. Results and Discussion

a. Data

The data used in this study are the report of UN results of the SMA level for natural science, social science, and language programs in Banyumas Regency in 2019. The data were obtained from the official report of the Ministry of Education and Culture. The data are in the form of matrix sized for natural science, for social science, and for a language program. Table 1 and table2 provide information on the list of high schools in the dataset.

Table 1. The list of high schools with a language program in dataset

Language Program	
Variable	**Senior High School**
B1	SMA N Ajibarang
B2	SMA N 1 Purwokerto
B3	SMA N 2 Purwokerto
B4	SMA N 5 Purwokerto

Table 2. The list of high schools with natural science and social science programs in dataset

Natural Science Program		Social Science Program	
Variable	**SMA**	**Variable**	**SMA**
A1	SMA N Ajibarang	S1	SMA N Ajibarang
A2	SMA N Banyumas	S2	SMA N Banyumas
A3	SMA N Baturraden	S3	SMA N Baturraden
A4	SMA N Jatilawang	S4	SMA N Jatilawang
A5	SMA N Patikraja	S5	SMA N Patikraja
A6	SMA N 1 Purwokerto	S6	SMA N 1 Purwokerto
A7	SMA N 2 Purwokerto	S7	SMA N 2 Purwokerto
A8	SMA N 3 Purwokerto	S8	SMA N 3 Purwokerto
A9	SMA N 4 Purwokerto	S9	SMA N 4 Purwokerto
A10	SMA N 5 Purwokerto	S10	SMA N 5 Purwokerto
A11	SMA N 1 Rawalo	S11	SMA N 1 Rawalo

Natural Science Program		Social Science Program	
Variable	**SMA**	**Variable**	**SMA**
A12	SMA N 1 Sokaraja	S12	SMA N 1 Sokaraja
A13	SMA N Sumpiuh	S13	SMA N Sumpiuh
A14	SMA N Wangon	S14	SMA N Wangon
A15	SMA Brunderan	S15	SMA Brunderan
A16	SMA Diponegoro Simpiuh	S16	SMA Budi Utomo Sokaraja
A17	SMA Ma'arif NU 1 Ajibarang	S17	SMA Diponegoro 1 Purwokerto
A18	SMA Ma'arif NU 1 Kemranjen	S18	SMA Jendral Sudirman
A19	SMA Ma'arif NU 1 Sokaraja	S19	SMA Karya Bakti Jatilawang
A20	SMA Muh. 1 Purwokerto	S20	SMA Ma'arif NU 1 Ajibarang
A21	SMA PGRI Gumelar	S21	SMA Ma'arif NU 1 Kemranjen
A22	SMA Yos Sudarso	S22	SMA Muh. 1 Purwokerto
A23	SMA Al Irsyad	S23	SMA Muh. Sokaraja
A24	SMA Muhammadiyah BSZ	S24	SMA Muh. Tambak
A25	SMA Islam Andalusia Kebasen	S25	SMA PGRI Gumelar
A26	SMA Nasional 3 BPH	S26	SMA Al Irsyad
A27	MAN 1 Banyumas	S27	SMA Muhammadiyah BSZ
A28	MAN 2 Banyumas	S28	SMA El-Madani Rawalo
A29	MAN 3 Banyumas	S29	SMA Islam Andalusia Kebasen
A30	MA Al-Ikhsan Beji	S30	SMA Nasional 3 BPH
A31	MA Ma'arif NU 1 Kemranjen	S31	MAN 1 Banyumas
A32	MA Miftahul Huda Rawalo	S32	MAN 2 Banyumas
A33	MA PPPI Miftahussalam	S33	MAN 3 Banyumas
A34	MA Wathoniyah Islamiyah	S34	MA Al-Ikhsan Beji
A35	MA Al-Falah Jatilawang	S35	MA Ma'arif NU 1 Kemranjen
A36	MA Ar-Ridlo Pekucen	S36	MA Muhammadiyah Purwokerto
A37	MA Ma'arif NU 1 Cilongok	S37	MA PPPI Miftahussalam
		S38	MA Wathoniyah Islamiyah
		S39	MA Ma'arif NU 1 Kebasen
		S40	MA Ar-Ridlo Pekucen
		S41	MA Al-Hidayah Purwojati

Table 3. The display of UN 2019 dataset

School	Indonesian	Englis	Math	Physics	Chemistry	Biology
SMA N Ajibarang	86.47	69.19	56.15	66.11	62.5	68.56
SMA N Banyumas	86.71	75.33	58.59	60.42	70.09	70.7
SMA N Baturraden	76.17	53.06	36.61	53.61	46.35	55.71
SMA N Jatilawang	85.46	66	54.01	58.3	68.75	66.67
…	…	…	…	…	…	…

The variables contained in the dataset of the results of the national examination of the natural science program are the results of examinations in Indonesian, English, Mathematics, Physics, Chemistry, and Biology. For the social science program, the variables are Indonesian, English, Mathematics, Economics, Sociology, and Geography. In the language program, those are Indonesian, English, Mathematics, Indonesian literature, Anthropology, and Indonesian language and literature. Table 3 illustrates the data used in the study

b. Preprocessing Results

In the preprocessing stage, the process is run from the SDC and DC algorithm. In the SDC algorithm, the Silhouette criterion value of each object will be observed based on the number of neighbors as shown in figure 4. The figure is the visualization of the Silhouette criterion value of ordered objects from the fewest to the most. Figure 4 provides information that the Silhouette criteria are not proportional to the number of neighbors they have. It can be seen that the object with the fewest neighbor does not necessarily has the lowest Silhouette criterion and vice versa. This indicates that Silhouette criteria will affect clustering results.

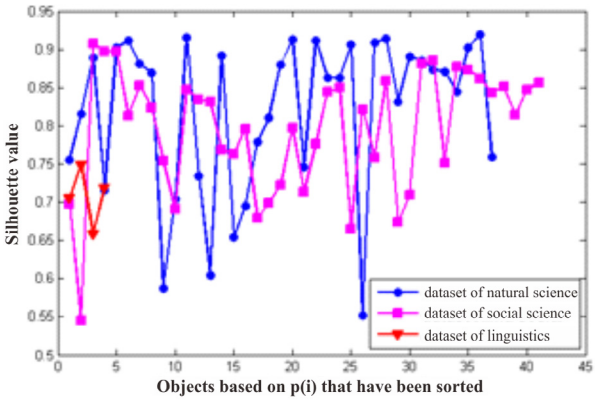


Figure 4. Silhouette criteria value based on ordered

Figure 5 is the visualization of the number of centroids and their location from the SDC and DC algorithm generated from principal component analysis(PCA, Principal Component Analysis). All visualizations have a component value of more than 88% based on the number of components 1 and 2. This indicates that visualization can maintain more than 88% of the information contained in the dataset so that it has relatively high quality. The location and the number of centroids in the natural science dataset obtained from the SDC algorithm (Figure 5a) and the DC algorithm (Figure 5d) have different values. It can be seen in the picture that SDC provides fewer initial centroids compared to DC. While in other datasets, the results of distribution and the number of initial centroids generated from SDC and DC are the same. The picture also provides information that SDC tends to choose most neighbors as initial centroids even though it has been offset by other criteria. Furthermore, the location and the number of initial centroids of the SDC and DC algorithms will be used as a prerequisite of the K-means algorithm.

c. Clustering Results

In the next stage, the clustering process is done by using K-means with SDC (SDCKM), K-means with DC (DCKM), and K-means without preprocessing algorithm (KM). Table 4 provides information on the number of clusters, Silhouette value, and the average time needed by each algorithm. The average time and Silhouette validation values are obtained from the iteration of each algorithm for 100 times in each dataset. The visualization of time and Silhouette validation of each iteration can be seen in figure 6.

Table 4. The number of clusters, Silhouette value, and the average time of each algorithm

Dataset	Algorithm	Number of clusters	Silhouette	Time
Natural science	SDCKM	3	0.6895	6.938102
	DCKM	4	0.4475	1.189253
	KM	3	0.5013	0.485832
	KM	4	0.4148	0.622489
Social science	SDCKM	3	0.6340	8.629198
	DCKM	3	0.6340	0.962059
	KM	3	0.4472	0.600157
Language program	SDCKM	2	0.7079	0.083598
	DCKM	2	0.7079	0.075903
	KM	2	0.4303	0.048455

Table 4 shows that the SDCKM algorithm has a higher Silhouette value compared to other algorithms in the natural science dataset with three clusters formed.

Whereas in the social science and natural science dataset, SDCKM has the same value as the DCKM algorithm. Figures 6a, 6b, and 6c show that SDCKM and DCKM have a consistent Silhouette validation value for each iteration, it is known that this algorithm is deterministic. It is different when it comes to K-means where its validation values are not consistent since the determination of the initial centroids is done randomly. This creates a possibility that there is an opportunity element on the clustering results with K-means. The results of the Silhouette obtained by SDCKM in each dataset are optimum results with a Silhouette value higher than 0.5. Table 4 also shows the weakness of the SDCKM which is it needs the longest average time compared to other algorithms for each dataset, as shown in figure 6d, 6e, dan 6f.

d. Interpretation of Clustering Results

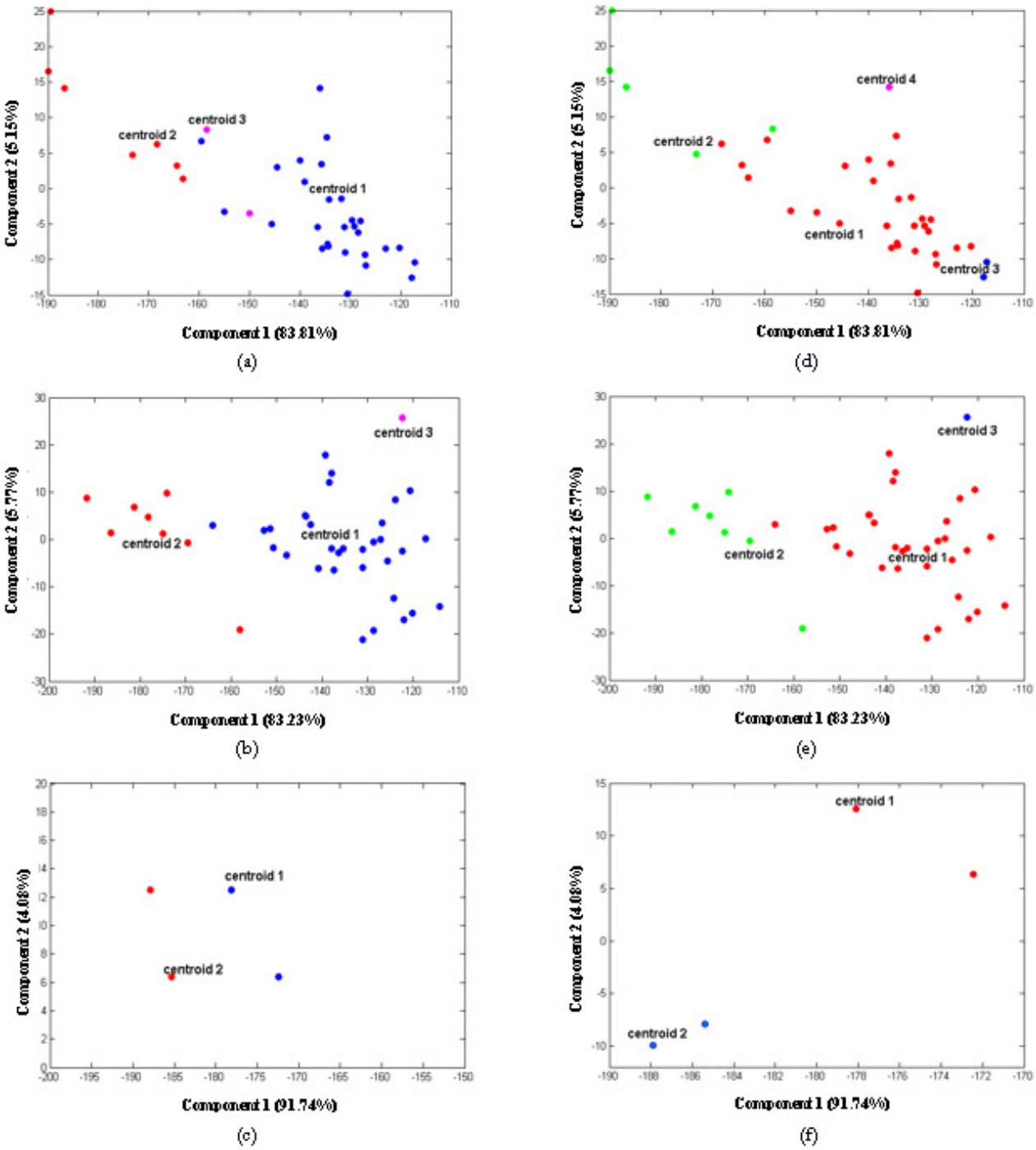


Figure 5. Visualization of centroids distribution of the SDC algorithm in the dataset (a) natural science, (b) social science, and (c) language program and DC algorithm in dataset (d) natural science, (e) social science, and (f) language program.

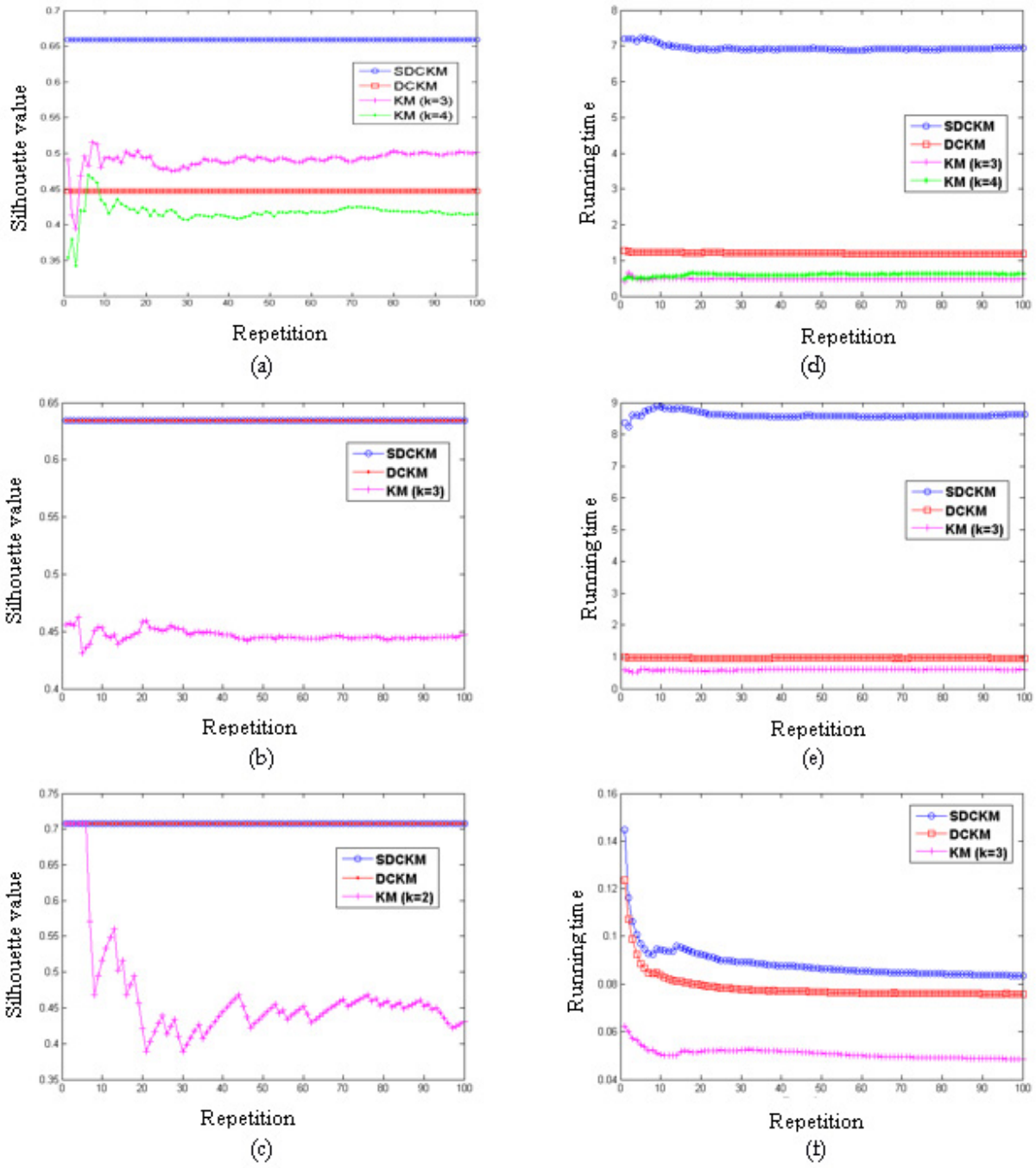


Figure 6. Visualization of the iterative Silhouette validation in the dataset (a) natural science, (b) social science, and (c) language program and running time in the dataset (d) natural science, (e) social science and (f) language program for each algorithm

Table 5. List of schools in certain cluster

Program	*Cluster* 1	*Cluster* 2	*Cluster* 3
Natural Science	A3, A5, A8, A11, A12, A14, A16, A17, A18, A19, A20, A21, A22, A24, A25, A27, A28, A29, A30, A31, A32, A33, A34, A35, A36, A37	A1, A2, AA6, A7, A23	A4, A9, A10, A13, A15, A26
Social Science	S3, S5, S8, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S21, S22, S23, S24, S25, S27, S28, S29, S31, S32, S33, S34, S35, S36, S37, S38, S39, S41	S1, S2, S4, S6, S7, S9, S10, S26, S30	S40
Language	B2, B4	B1, B3	

Table 6. The comparison of the algorithms

Algorithm	Strength	Weakness
SDCKM	• It can determine the location and number of centroids optimally. • The determination of centroids considers 2 criteria which clustering results, those are the number of neighbors and Silhouette criteria.	• It requires a quite long time for the clustering process.
DCKM	• It can determine the location of centroids and the number of initial centroids • It requires a relatively short time.	• The determination of centroid is done based on the number of neighbors so that the formation of the initial centroid is not optimal.
KM	• It requires a relatively short time. • Simple algorithm	• The clustering results are not optimal since the determination of initial centroid and its number is done randomly.

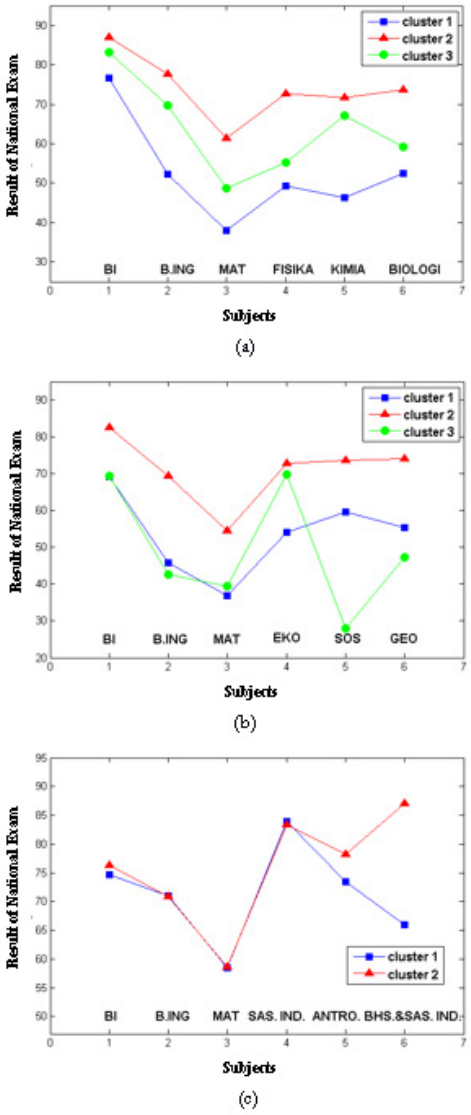


Figure 7. The visualization of cluster profile for the dataset (a) natural science, (b) social science, and (c) language program

The interpretation of dataset mapping is done in the clustering process with the optimum result, which is the SDCKM result. The cluster's profile shown in figure 7. Figure 7a provides information on the cluster profile from the centroid of the clustering results on the natural science dataset. The results indicate that SMA in cluster 2 is better in quality compared to schools in cluster 1 and 3. This can be seen from the score obtained of the subjects tested, those are Indonesian (BI), English (B,ING), Mathematics (MAT), Physics, Chemistry, and Biology. The scores obtained by cluster 2 are higher than other clusters. while SMA in cluster 3 has better quality than schools in cluster 1. Figure 7b provides information that in social science dataset, the quality of SMA in cluster 2 is better than other schools for each subject such as Indonesian (BI), English (B. ING), Mathematics (MAT),

Economics (EKO), Sociology (SOS), and Geography (GEO). Cluster 1 is considered to have relatively better quality than cluster 3since there are two subjects namely Sociology (SOS) and Geography (GEO) that show cluster 1 is better than cluster 3 with significant difference, on the contrary cluster 3 is only a way better than cluster 1 in one subject, that is Economics(EKO). For other subjects, cluster 1 and cluster 3 obtain relatively the same results. Figure 7c provides information that cluster 2 has a better quality of education unit than cluster 1 based on two subjects, namely Anthropology (ANTRO) and Indonesian language and literature (SAS. & BHS. IND.), these show that cluster 2 is better than cluster 1. Whereas in other subjects, such as Mathematics(MAT), Indonesian (BI), English (B.ING), and Indonesian literature (SAS. IND.), cluster 1 and cluster 2 relatively have the same results.

To find out the grouping of certain schools into a certain cluster, it can be seen in table 5 with information referring to table 1 and 2. From the table, it is obvious that for natural science schools included in cluster 2 are SMA N Ajibarang, SMA N Banyumas, SMA N 1 Purwokerto, SMA N 2 Purwokerto, and SMA Islam Teladan Al Irsyad Al Islamiyyah. Schools in cluster 3 are SMA N Jatilawang, SMA N 4 Purwokerto, SMA N 5 Purwokerto, SMA N Sumpiuh, SMA Bruderan Purwokerto, and SMA Nasional 3 Bahasa Putera Harapan. The rest of the schools which are not mentioned included in cluster 1. Meanwhile, in social science dataset, the schools included in in cluster 2 are SMA N Ajibarang, SMA N Banyumas, SMA N Jatilawang, SMA N 1 Purwokerto, SMA N 2 Purwokerto, SMA N 4 Purowkerto, and SMA N. The next, school included in cluster 3 is MA Ar- Ridlo Pekuncen only. Other schools that are not mentioned are in cluster 1. In the language program dataset, schools included in cluster 1 are SMA N 1 Purwokerto and SMA N 5 Purwokerto. While schools in cluster 2 are SMA Negeri Ajibarang and SMA N 2 Purwokerto. At the end of the discussion, table 6 displayed to provide information on the comparison of algorithms, such as SDCKM, DCKM, and KM, which are used in this study.

4. Conclusion

Based on the results and discussion presented in the previous parts, several conclusions are addressed. First, SDCKM has a better ability than CDKM in the clustering process of the 2019 UN dataset in Banyumas Regency. This can be seen from the Silhouette validation of the clustering results. Second, the time needed by the SDCKM algorithm is relatively long, so that it is not good enough to cluster a big dataset. Based on these conclusions, the proposed modification which includes Silhouette in the Density Canopy K-Means algorithm yield a better clustering result based on Silhouette validation. However, the addition of Silhouette criteria makes the clustering process to be longer than without the criteria. The results of the mapping of the quality of education concerning 2019 UN results in Banyumas Regency show that in the dataset of natural science, social science, and language program, the schools in cluster 2 have the best education quality compared to schools in cluster 1 and 3.

References

[1] Badan Standar Nasional Indonesia, "Prosedur operasional standar (POS) penyelenggaraan ujian nasional tahun pelajaran 2018/2019," BNSP, indonesia, 2018.

[2] Badan Standar Nasional Indonesia, "Prosedur operasi standar ujian nasional sekolah menengah pertama, madrasah tsanawiyah, sekolah menengah pertama luar biasa, sekolah menengah atas, madrasah aliyah, sekolah menengah atas luar biasa, dan sekolah menengah kejuruan tahun pelajaran 2010/2011," BNSP, Indonesia, 2011.

[3] A. Asra, Rudiansyah, *Statistika terapan untuk pembuat kebijakan dan pengambil keputusan*, 2nd Ed, In Media, 2014.

[4] A. J. Jain, "Data clustering: 50 years beyond k-means", in *the 19th International Conference on Pattern Recognition, 1967.*

[5] J. B. MacQueen," Some Methods for Classification of High Dimensional Data Sets with Application to Reference Matching", in *Proceeding of 5th Berkeley Symposium on Mathematical Statistics and* Probability, 1967.

[6] D. S. Wardiani, N. Merlina, "Implementasi data mining untuk mengetahui manfaat RPTRA menggunakan metode k-means clustering," *Jurnal PILAR Nusa Mandiri*, vol. 15, no. 1, pp. 125-132, 2019.

[7] A. A. Hassan, W. M. Shah, A. M. Husein, M. S. Talib, A. A. J. Mohammed, M. F. Iskandar, "Clustering approach in wireless sensor networks based on k-means: limitations and recommendations," *IJRTE*,vol. 7, no. 6S5, pp. 119-126, 2019.

[8] U. Yelipe, S. Porika, M. Golla, "An Efficient Approach for Imputation and Classification of Medical Data Values Using Class-Based Clustering of Medical Records," *Computers and Electrical Engineering*, vol. 66, pp. 487-504, 2018.

[9] Mardalius, "Implementasi algoritma k-means clustering untuk menentukan kelas kelompok bimbingan belajar tambahan (Studi kasus: siswa sma negeri 1 ranah pesisir)", in *Proceding SEMILOKA ROYAL 2017 "Teknologi Mobile"*, 2017.

[10] H. Yuwafi,F. Marisa,I. D. Wijaya, "Implementasi Data Mining untuk Menentukan Santri Berprestasi di PP. Manarulhuda dengan Metode Clustering Algoritma K-means," *Jurnal SPIRIT*,

vol. 11, no. 1, pp. 22-29, 2019.

[11] Y. S. Nugroho, S. N. Haryati, "Klasifikasi dan klastering penjurusan siswa sma negeri 3 boyolali," *Khazanah informatika*, vol. 1, no. 1, pp. 1-6, 2015.

[12] A. Kumar, Y. S. Ingle, A. Pande, P. Dhule, "Canopy clustering: a review on pre-clustering approach to k-means clustering," *IJIACS*, vol. 3, no. 5, pp. 22-29, 2014.

[13] J. Yoder, C. E. Priebe, "Semi-supervised K-means++," *The Journal of Statistical Computation and Simulation*, 2016.

[14] B. Fritzke, "The k-means-u* algorithm: non-local jumps and greedy retries improve k-means++ clustering[J]. 2017.

[15] G. Zhang, C. Zhang, H. Zhang, "Improved K-means Algorithm Based on Density Canopy," *Journal Knowledge-based Systems*, vol. 145, pp.289-297, 2018.

[16] A. R. Mamat, F. S. Mohamed, M. A. Mohamed, N. M. Rawi, M. I. Awang, "Silhouette index for determining optimal k-means clustering on images in different color models," *International Journal of Engineering & Technology*, vol. 7, no. 2, pp. 105-109, 2018.

[17] R. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th Ed, New York Pearson Education, 2007.

[18] E. Rendon, I. M. Abundez, C. Gutierrez, S. D. Zagal, A. Arizmendi, E. M. Quiroz, H. E. Arzate, "A comparison of internal and external cluster validation indexes," in *Proceeding of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications*, 2011.

[19] E. Rendon, I. Abudez, A. Arizmendi, E. M. Quiroz, "Internal Versus External Cluster Validation Indexes," *International Journal of Computers and Communications*, vol. 5, no. 1, pp. 27-34, 2011.

[20] L. Vendramin, R. J. G. B. Campello, E. R. Hruschka, "On the comparison of relative clustering validity criteria," in *Proceedings of the 2009 SIAM International Conference on Data Mining*, 2009.

[21] J. Baarsch, M. E. Celebi, "Investigation of Internal Validity Measures for K-means Clustering," in *Proceedings of the IMECS,* 2012.

[22] A. Saiful, J. L. Buliali,"Implementasi particle swarm optimization pada k-means untuk clustering data automatic dependent surveillance-broadcast," *Explora Informatika*, pp. 30-35, 2018.

[23] A. D. Munthe, "Penerapan clustering time series untuk menggerombolkan provinsi di Indonesia berdasarkan nilai produksi padi," *Jurnal Litbang Sukowati*, vol. 2, no. 2, pp. 1-11, 2019.

[24] Suprihatin, Y. R. W. Utami, D. Nugroho, "K-Means clustering untuk pemetaan daerah rawan demam berdarah," *Jurnal TIKomSIN*, vol. 7, no. 1, pp. 8-16, 20019.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Case-Base Reasoning (CBR) and Density Based Spatial Clustering Application with Noise (DBSCAN)-based Indexing in Medical Expert Systems

Herdiesel Santoso[1], Aina Musdholifah[2]
[1]Information Systems Study Program
Sekolah Tinggi Manajemen Informatika dan Komputer El Rahma
Yogyakarta
herdiesel.santoso@stmikelrahma.ac.id
[2]Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta

Abstract-Case-based Reasoning (CBR) has been widely applied in the medical expert systems. CBR has computational time constraints if there are too many old cases on the case base. Cluster analysis can be used as an indexing method to speed up searching in the case retrieval process. This paper propose retrieval method using Density Based Spatial Clustering Application with Noise (DBSCAN) for indexing and cosine similarity for the relevant cluster searching process. Three medical test data, that are malnutrition disease data, heart disease data and thyroid disease data, are used to measure the performance of the proposed method. Comparative tests conducted between DBSCAN and Self-organizing maps (SOM) for the indexing method, as well as between Manhattan distance similarity, Euclidean distance similarity and Minkowski distance similarity for calculating the similarity of cases. The result of testing on malnutrition and heart disease data shows that CBR with cluster-indexing has better accuracy and shorter processing time than non-indexing CBR. In the case of thyroid disease, CBR with cluster-indexing has a better average retrieval time, but the accuracy of non-indexing CBR is better than cluster indexing CBR. Compared to SOM algorithm, DBSCAN algorithm produces better accuracy and faster process to perform clustering and retrieval. Meanwhile, of the three methods of similarity, the Minkowski distance method produces the highest accuracy at the threshold ≥ 90.

Keywords: case-base reasoning; clustering; dbscan; indexing; som.

1. Introduction

Expert system is a part of artificial intelligence that has been developed widely to help diagnose of diseases. The method commonly used in expert systems is rule-based reasoning, or case-based reasoning [1]. Case-based reasoning (CBR) methods have been widely applied in the medical field [2] - [6], due to the ability of CBR to work like an expert by retrieval of previous cases to solve new cases according to the given diagnosis [7]. The more old cases stored in the case base, the CBR system will be smarter in finding solutions for a given case. Problems with computation time and memory space requirements become a challenge especially when too many old cases exist on the case base. That is because the system must calculate the value of the similarity of new cases with all the old cases on the case base. A solution that can be used to shorter computational time is by finding solution that does not need to involve all data on the case base, but sufficient with some of the closest cases, so that the indexing process is needed [8].

Research focusing on the indexing process in CBR has been carried out with various methods, such as Fuzzy algorithm [9], back propagation classification algorithm [10], K-means clustering algorithm [11], and Local Triangular Kernel-Based Clustering (LTKC) algorithm [12]. K-means algorithm needs data of number of clusters that will be formed, because the assumption of the number of clusters determined at the beginning does not necessarily produce an optimal cluster. This method also has a low tolerance for data that contains noise and outliers. The back propagation and LTKC training process require quite long time because they have to try the training parameters one by one to get the best cluster. Clustering can group data sets that are not labeled into several data clusters based on similarity and dissimilarity [13]. Basically these

algorithms work by grouping cases based on the specified features. When the retrieval process is carried out on the CBR, searching for similarity values can be conducted to cases that have the same index as new cases. Clustering algorithm can describe the patterns and tendencies contained in data groups. Each group represented by the value of the center of the cluster (cluster centroid). Cluster center enables measurement of similarity between new data and all cluster centers so it can determine the most similar data groups.

The proposed clustering method uses Self-Organizing Maps (SOM) compared to Density Based Spatial Clustering Application with Noise (DBSCAN). SOM is an artificial neural network-based learning algorithm that is good in exploration and visualization of high-dimensional data [14]. The training process on the SOM algorithm does not require supervision, the SOM network will learn without having a target in advance [15]. This is different from some artificial neural network methods such as back propagation which requires a target during the learning process. Density-based clustering methods such as DBSCAN have the characteristics of clusters with high density surrounded by clusters that have with low density. DBSCAN has advantages such as: being able to handle large amounts of data in short time, having tolerance to data containing noise and outliers, being able to recognize irregular shapes, being able to handle high dimensional data, and unnecessary to know the number of clusters to be formed [16] [17].

Each clustering algorithm requires testing to determine the quality of the clustering results. The validation of the results of clustering in this study was performed by evaluating the results of the clustering algorithm based on the structure that has been determined in the data set using Davies-Bouldin index and Silhouette index [18]. The process of looking of similarity between new cases and old cases in this study uses the nearest neighbor retrieval technique, by calculating the value of similarity or closeness between new cases and old cases. Three methods were used and compared, that are manhattan distance similarity, euclidean distance similarity and minkowski distance similarity.

2. Method

a. Knowledge Acquisition

This study used case data of medical record of patients with severe malnutrition at RSUP Dr. Sardjito Yogyakarta [3]. The malnutrition disease data consists of 90 data sets divided into 70 data as training data and 20 data as test data. The second case data is the medical record of patients with heart disease in the Medical Record Installation of RSUP Dr. Sardjito Yogyakarta [6]. The heart disease case data consists of 135 data sets divided into 115 data as training data and 20 data as test data. The third data is the diagnosis data on suspected thyroid disease from the Garvan Institute. The thyroid disease case data consists of 1428 data sets divided into 1000 data as training data and 428 data as test data.

b. Case Representation

The case representation used the frame model. Cases are represented as collections of features that characterize cases and solutions for handling these cases. Weighting of features is important to determine the level of significance of the feature to the disease. The weighting of each feature for each case is performed by an expert. If there are new cases, the weighting of disease features is divided into two categories, that are No and Yes. The value for each category is 0 for no symptoms and 1 for symptoms. After the old cases in the case base are clustered, the old case data is represented again by adding new knowledge derived from cluster center. Table 1 is a representation of cases of malnutrition in children under five who added new knowledge derived from the value of the cluster center.

Table 1. Representation of cases of malnutrition after clustering.

No	Case	Information
A	**Indication**	
1	G003	Rounded and swollen face
2	G009	Xylophone ribs
3	G019	Edema
4	G021	Very thin
B	**Patient data**	
1	Age	35 month
C	**Disease**	
1	P003	Marasmus-Kwashiorkor
D	**Indexing**	
1	Cluster	1

c. Indexing

The indexing method in this system used clustering method, i.e Density Based Spatial Clustering Application with Noise (DBSCAN) compared to Self-Organizing Maps (SOM). DBSCAN or SOM is used to group old case data into groups based on similarity and dissimilarity, so in each group contains similar data.

1) Data Normalization

The data normalization used the Min Max Normalization method. Normalization features include age, TSH, T3, TT4, and T4U since they have significant vulnerability. Min Max Normalization requires Minimum and Maximum age features. For example the age feature of malnutrition cases is a minimum value of 0 months and a maximum value of 60 months, and the age feature of a heart case minimum value is 0 years and the maximum value is 100 years. Equation 1 is the Min Max Normalization formula.

$$v' = \frac{v - \min}{\max - \min} \quad (1)$$

2) Selft Organizing Map (SOM)

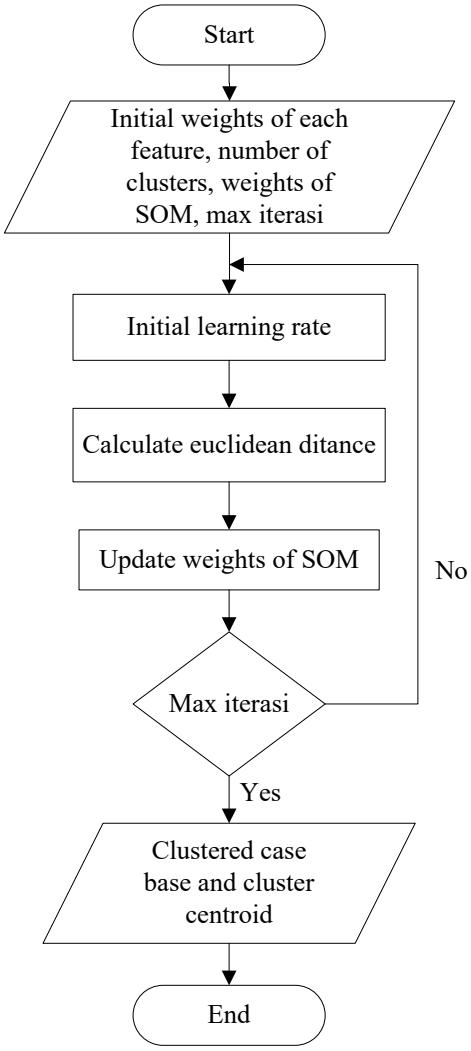


Figure 1. Clustering design using SOM algorithm.

Self-Organizing Map (SOM) algorithm or often referred as Kohonen Artificial Neural Network is one of the topology of Unsupervised Artificial Neural Network (Unsupervised ANN) in which the training process does not require supervision (target output). The clustering design using the SOM method is shown in the flowchart of Figure 1 [15]. Explanation of the flowchart diagram of Figure 1 is as follows:

a) Initializing the weights of each feature in the case base (ix) as input from SOM, number of clusters (k), initial weight (wi), and maximum iteration as SOM parameters.
b) Determine the learning rate (η) and decrease learning rate (α).
c) For each case base (xi) calculate the euclidean distance (Dj) to all initial weights of SOM (wij) using equation (2). After knowing the euclidean distance to each weight, look for the index that has the smallest value.

$$D_j = \sum_i^n \left(w_{ij} - x_i\right) \quad (2)$$

d) Each wij weight within the radius of Dj neighborhood, the weight is updated by equation (3).

$$w_{ij}(new) = w_{ij}(old) + \alpha(x_i - w_{ij}(old)) \quad (3)$$

e) Update the learning rate every 1 iteration with equation (4).

$$\eta(new) = \eta(old) \times \alpha \quad (4)$$

f) As long as the maximum number of iterations has not been reached, repeat steps c through e.
g) Output clustering using the SOM method is a clustered case database and new weights are used as cluster center values.

3) Density Based Spatial Clustering Application with Noise (DBSCAN)

Density Based Spatial Clustering Application with Noise (DBSCAN) is one of the density-based clustering algorithms. The DBSCAN algorithm works by expanding high density regions into clusters and placing irregular clusters in the spatial database as noise. The clustering design using the DBSCAN method is shown in the flow chart of Figure 2 [16]. DBSCAN has 2 parameters, that are Eps or ε psilon (maximum radius of the neighborhood) and MinPts (minimum number of points in the Eps-neighborhood of a point).

Explanation of the flow diagram of Figure 2 is as follows:

a) Initializing the weights of each feature in the case base as DBSCAN input, the maximum radius of the neighborhood (Eps) and the minimum number of points in the Eps-neighborhood of a point (MinPts) as a DBSCAN parameter.
b) Specify one data as a random starting point (p).
c) For each case data in the case base, calculate the value of ε psilon or all distances that are density reachable to p using equation (5).

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \quad (5)$$

d) If the amount of case data that meets ε psilon is more than MinPts, then p is a core point and one cluster is formed.
e) If there is no case data that is density reachable to p or the amount of case data that meets Eps is less than MinPts, then p is Noise.
f) Repeat steps c through e until all cases of case data base are processed.
g) Calculate the cluster center value (cluster centroid) using the average value for each cluster group.
h) The output of the case database is clustered and the average value is used as the cluster center value.

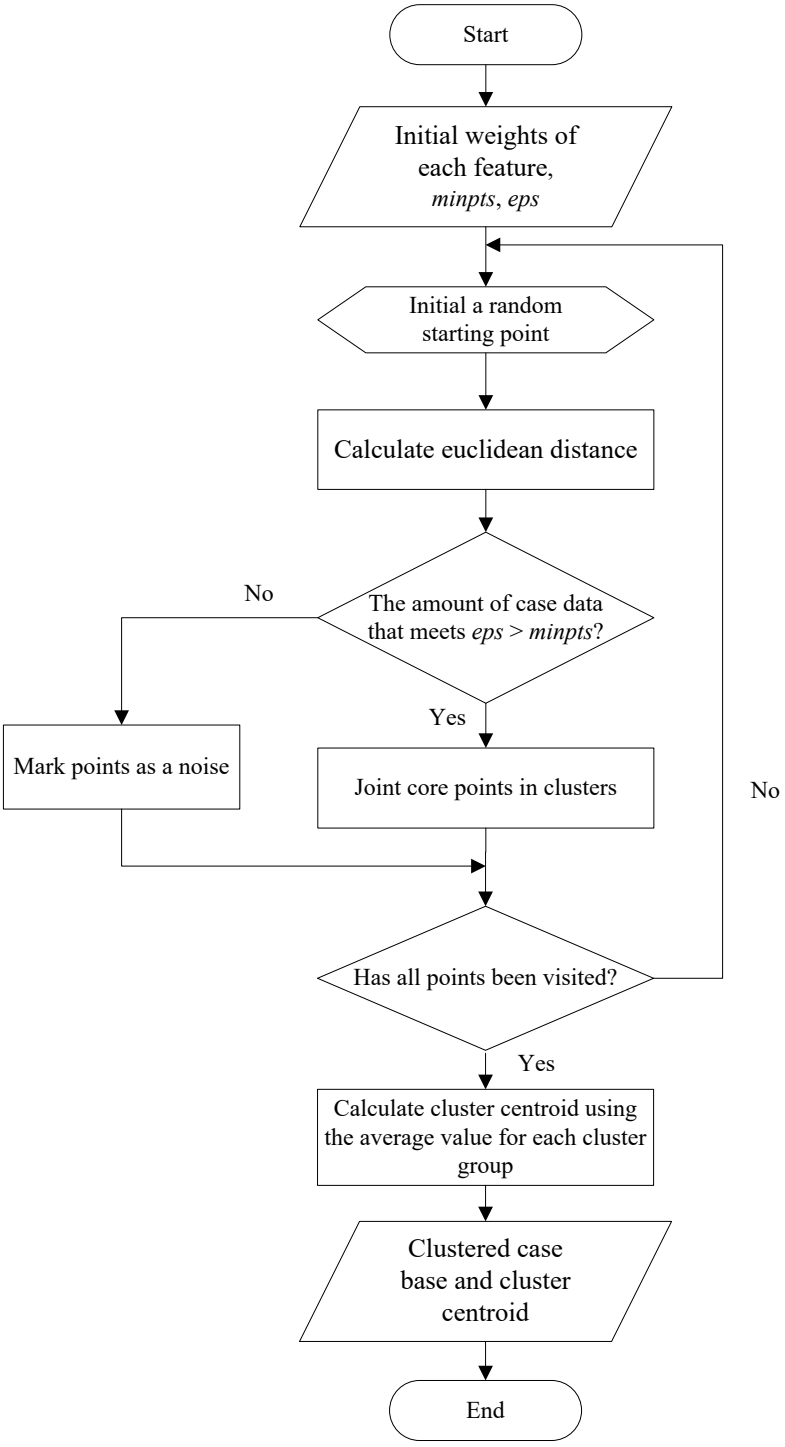


Figure 2. The design of clustering with DBSCAN algorithm.

c. Cluster Evaluation

The evaluation methods used in this system are the silhouette index and the Davies-Bouldin index methods. These methods are used to test the quality of the results of clustering. These methods are cluster validation methods that combines cohesion and separation methods. To calculate the value of silhouette index and Davies-Bouldin index, the distance between data is acquired by using the euclidean distance formula.

1) Silhoutte index

Silhoutte index was used to measure the quality and strength of a cluster, how well an object is placed in a cluster. The step of calculating the silhoutte index value starts with calculating the average distance from object i to all objects in a cluster. The calculation will produce an average value called ***ai***. Next, calculate the average distance from object ***i*** to objects in other clusters. Of all the average distances, take the smallest value, the value

is called bi. Next, calculate the silhoutte index using equation (6) [18].

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (6)$$

Where ***s (i)*** is a Silhouette index value, ***a (i)*** is the average distance between point ***i*** and all points in ***A*** (the cluster where point ***i*** is located), ***b (i)*** is the average distance between point ***i*** to all points in clusters other than *A*. The silhoutte index value can vary between -1 to 1. The clustering result is good if the silhoutte index value is positive (*ai* <*bi*) and ai approaches 0, so that the maximum silhoutte index value is 1.

2) Davies-Bouldin index (DB index)

Davies-Bouldin Index has characteristics in validating clusters based on the calculation of quantity and derived features of the datas et. DB index value is calculated using equation (7) [18].

$$DB = \frac{1}{c}\sum_{c=1}^{c} Max_{i \neq j}\left\{\frac{d(x_i) + d(x_j)}{d(c_i, c_j)}\right\} \quad (7)$$

Where ***DB*** is Davies-bouldin value, ***c*** is the number of clusters, ***d (xi)*** and ***d (xj)*** case data in clusters ***i*** and clusters ***j, d (ci, cj)*** is the distance between clusters ***ci*** and ***cj***. The smaller value of Davies Bouldin Index shows that the cluster configuration scheme is optimal and the cluster quality is getting better.

d. Retrieve and Reuse

CBR systems built with cluster-indexing can provide additional knowledge derived from previous cases. This knowledge is acquired from cluster center values generated from cluster analysis and added as a representation on a case base. After the case is represented by adding knowledge to the cluster center value, the case is then stored in a database. Figure 3 shows the architecture of the CBR system architecture with cluster-indexing.

If there are new cases, the system initializes the symptoms experienced by the patient and represents them as new cases. The system will search for the most relevant clusters by calculating the similarity of symptoms of new cases to the cluster center values. Similarity calculation is performed by comparing the euclidean distance between new cases with the cluster center value using the Cosine Coefficient method. After obtaining an index or cluster that is relevant to the new case, then a calculation is performed to find the similarity value between the new case and the cases in the case base that are in the same cluster.

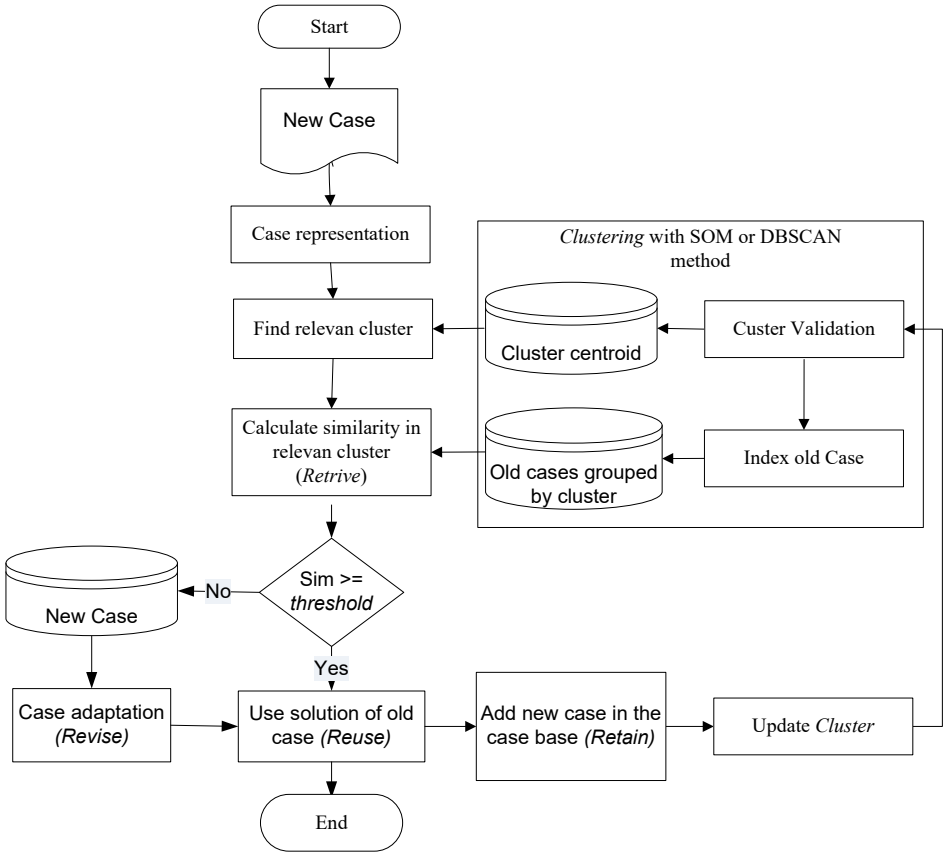


Figure 3. CBR system architecture design with cluster-indexing.

The threshold value of similarity are 0.7, 0.8, and 0.9 which means that if the highest similarity is greater than the threshold and close to 1, so this indicates that the new case has the exact same resemblance to the old case then the solution from the source case will be given to the user (reuse). If the similarity value decreases or is below the threshold, then the case will be stored in the database as a revision case which later the case under the threshold will be adjusted from the solution of the previous cases by the expert (revise). The new case is then saved to the case base by considering the cluster center value to become new knowledge (retain).

1) Determination of the Closest Cluster

During the process of finding a solution for a case, the CBR system will search for clusters that are most relevant to the new case by calculating the similarity of the symptoms of the old case with the cluster center value. Similarity calculation performed by comparing distances using the cosine coefficient method [19]. If given 2 vectors ***X*** and ***Y***, then the similarity value can be found by equation (8):

$$\text{Cos}\ (\text{X Y}) = \frac{\langle X, Y \rangle}{\|X\| \bullet \|Y\|} \tag{8}$$

where ***"‹ ›"*** denotes the multiplication of vectors ***X*** and ***Y***, and ***"| X || Y |"*** denotes the norm for each vector. For vectors with non-negative elements, the cosine similarity value always lies between 0 and 1, where 1 indicates the two vectors are really the same, and 0 indicates the opposite.

The retrieval process used the nearest neighbor method. Nearest neighbor works by calculating the value of similarity, that is, the closeness between new cases and old cases based on matching weights of a number of existing features. There are two types of similarity measurements that are local similarity and global similarity [6]. Local similarity is a measurement of proximity at the feature level, whereas global similarity is a measurement of proximity at the object level (case).

Local similarity used in this study can be divided into two types, which are numerical and symbolic. The features included in the symbolic type are the symptom features and risk factors, while the numerical features are the sex and age features. Numerical data is calculated using equation (9)

$$f(S_i, T_i) = 1 - \frac{|S_i, T_i|}{|f_{\max} - f_{\min}|} \tag{9}$$

Note: ***f (Si, Ti)*** is the similarity of the ***i-feature*** of the old case or source case (***S***) with the new case or target case (***T***), Si is the value of the ***i-feature*** of the old case (source case), ***Ti*** is the ***i-feature*** value of the new case (target case), ***fmax*** is the maximum value of the ***i-feature*** on the case base and fmin is the minimum value of the ***i-feature*** on the case base. Meanwhile, symbolic data will be calculated using equation (10).

$$f(S_i, T_i) = \begin{cases} 0, if S_i \neq T_i \\ 1, if S_i = T_i \end{cases} \tag{10}$$

Note: ***f (Si, Ti)*** is the ***i-th*** feature similarity of the ***S*** (source) and ***T*** (target) cases, ***Si*** is the ***i-th*** value feature of the old (source) case and ***Ti*** is the ***i-th*** value feature of the new case (target).

Global similarity was used to calculate the similarity between new cases and cases on the case base. The methods to calculate global similarity in this study are Manhattan distance similarity in equation (11), euclidean distance similarity in equation (12), and minkowski distance similarity in equation (12) [20].

$$Sim(X, Y) = \frac{\sum_{i=1}^{n} f(S_i, T_i) * w_i}{\sum_{i=i}^{n} w_i} \tag{11}$$

$$Sim(X, Y) = \left(\frac{\sum_{i=1}^{n} w_i^r \times |f(S_i, T_i)|^r}{\sum_{i=i}^{n} w_i^r} \right)^{1/r} \tag{12}$$

Note: ***Sim (Si, Ti)*** is the value of similarity between the old case (***S***) and the new case (***T***), ***fi (Si, Ti)*** is the similarity of the ***i-th*** feature of the old case and the new case, the similarity of the ***i-th*** feature of the source case and target case, ***n*** is the number of features in each case, ***i*** is the individual feature, between ***1 s / dn***, ***wi*** is the weight given to the ***i-th*** feature, and ***r*** is the minkowski factor (positive integer). The value of ***r*** is equal to 2 for euclidean distance and equal to 3 for minkowski distance similarity.

e. CBR System Testing

Testing is performed by applying new cases, which are 20 data as test data for cases of malnutrition and heart disease and 428 data as test data for thyroid disease cases. The results of the system are then compared with the data contained in the medical record data. System accuracy is calculated by comparing the number of correct diagnosis with the amount of test data. The accuracy in this study is acquired by comparing the number of correct decision results and the amount of test data in accordance with equation (13).

$$\text{Accuracy} = \frac{\sum_{i=i}^{n} k_i}{n} \times 100\% \tag{13}$$

Note: ***ki*** is the ***i-th*** decision (***ki*** is 1 if the decision is right and 0 is if the decision is wrong), ***n*** is the amount of test data.

3. Results and Discussion

a. Case Base Clustering Process

The process of clustering of old cases on a case base used the SOM and DBSCAN clustering algorithms. The SOM method requires three parameters, which are

number of clusters, maximum iteration, and learning rate. While the DBSCAN method requires two parameters, that are minimum points and epsilon. The parameter value is optimal if it produces the minimum Davies-Bouldin index value and the highest silhoutte coefficient and accuracy. The optimal parameter determination process carried out by clustering each case base data set using several combinations of parameters. Then each combination of these parameters is used to calculate the accuracy of the CBR retrieval process. Table 2 shows the SOM parameters and Table 3 shows the DBSCAN parameters.

The results of clustering with the SOM method depend on the initial weight given and the number of neurons in the output layer. Meanwhile, in DBSCAN the greater the ***minPts*** value, the more noise will be, this also affects the quality of the cluster. Therefore, determining the ε***psilon*** and ***minPts*** values at the beginning of the clustering process is very important. The quality of clustering results for the SOM and DBSCAN methods can be seen from the Davies-Bouldin index value, Silhoutte index and accuracy. The smaller the Davies-Bouldin index value, shows that the cluster parameters are optimal and the better the cluster quality. Meanwhile, for the Silhoutte index value getting closer to 1 shows that each case data is in the right cluster and there is no overlapping classes. Accuracy is determined by comparing the system diagnosis results and the actual diagnosis without applying a threshold value. The accuracy values of each trial are compared and the highest value for each data set is searched on the case base. The highest accuracy is used as the optimal clustering parameter.

Table 2. Optimal SOM parameters.

SOM attribute	Malnutrition Case Data	Heart Case Data	Thyroid Case Data
Amount of Clusters	3	5	5
Iteration	50	50	500
Learning Rate	0.1 – 0.2	0.4 – 0.5	0.7 – 0.8
Silhoutte index	0.378	0.279	0.303
DB index	0.812	0.324	0.587
Time (s)	0.439	1.167	11.48
Accuracy	100%	100%	87.15%

Table 3. Optimal DBSCAN parameters.

DBSCAN attribute	Malnutrition Case Data	Heart Case Data	Thyroid Case Data
Epsilon	1	13	0.5
MinPoints	3	3	10
Amount of Clusters	4	4	11
Amount of Noise	2	6	163
Silhoutte index	0.365	0.268	0.688
DB Index	0.888	0.462	0.420
Time (s)	0.124	0.282	8.37
Accuracy	100%	100%	90.89%

b. System Capability Analysis

The process of analyzing the ability of the system is divided into three scenarios. The first scenario is the diagnosis of the system using CBR non-indexing, the second scenario is the diagnosis of the CBR system with indexing using the SOM algorithm and the third scenario is the diagnosis of the CBR system with indexing using the DBSCAN algorithm. The searching process of relevant clusters with CBR cluster-indexing used the cosine similarity method and the similarity calculation process for all three scenarios used the Manhattan distance similarity method, euclidean distance similarity and minkowski distance similarity. Testing is performed by applying new cases, which are 20 data as test data for cases of malnutrition and heart disease and 428 data as test data for thyroid disease cases. Then the amount of the correct data is calculated, and the accuracy is determined according to the threshold and the average retrieval time for each similarity method. Based on the 3 testing scenarios, there are differences in the results of each scenario, as seen in Table 4 for cases of malnutrition, Table 5 for cases of heart disease and table 6 for cases of thyroid disease.

Table 4. Comparison of system capability in CBR non-indexing and CBR cluster-indexing for cases of malnutrition.

Scenario	Method	Threshold	Manhattan Distance	Euclidean Distance	Minkowski Distance
Scenario 1	CBR non-indexing	≥70	18 (90%)	18 (90%)	17 (85%)
		≥80	17 (85%)	18 (90%)	17 (85%)
		≥90	9 (45%)	18 (90%)	17 (85%)
Average retrieve time (seconds)			0.02598	0.02792	0.02925
Scenario 2	CBR SOM indexing	≥70	20 (100%)	20 (100%)	20 (100%)
		≥80	18 (90%)	20 (100%)	20 (100%)
		≥90	9 (45%)	20 (100%)	20 (100%)
Average retrieve time (seconds)			0.02269	0.02323	0.02425
Scenario 3	CBR indexing DBSCAN	≥70	20 (100%)	20 (100%)	20 (100%)
		≥80	18 (90%)	20 (100%)	20 (100%)
		≥90	9 (45%)	20 (100%)	20 (100%)
Average retrieve time (seconds)			0.02245	0.02288	0.02305

Table 5. Comparison of system capabilities in CBR non-indexing and CBR cluster-indexing for cardiac case data.

Scenario	Method	Threshold	Manhattan Distance	Euclidean Distance	Minkowski Distance
Scenario 1	CBR non-indexing	≥70	16 (80%)	20 (100%)	19 (95%)
		≥80	13 (65%)	20 (100%)	19 (95%)
		≥90	6 (30%)	12 (60%)	19 (95%)
Average retrieve time (seconds)			0.0535	0.0565	0.0469
Scenario 2	CBR SOM indexing	≥70	17 (85%)	20 (100%)	19 (95%)
		≥80	13 (65%)	20 (100%)	19 (95%)
		≥90	6 (30%)	12 (60%)	19 (95%)
Average retrieve time (seconds)			0.0417	0.0423	0.0424
Scenario 3	CBR indexing DBSCAN	≥70	17 (85%)	20 (100%)	20 (100%)
		≥80	13 (65%)	20 (100%)	20 (100%)
		≥90	6 (30%)	12 (60%)	19 (95%)
Average retrieve time (seconds)			0.0411	0.0418	0.0421

Table 6. Comparison of system capabilities in CBR non-indexing and CBR cluster-indexing for thyroid case data.

Scenario	Method	Threshold	Manhattan Distance	Euclidean Distance	Minkowski Distance
Scenario 1	CBR non-indexing	≥70	392 (91.56%)	393 (91.82%)	393 (91.82%)
		≥80	392 (91.56%)	393 (91.82%)	393 (91.82%)
		≥90	385 (89.95%)	392 (91.56%)	392 (91.56%)
Average retrieve time (seconds)			0.124	0.127	0.130
Scenario 2	CBR SOM indexing	≥70	353 (82.45%)	373 (87.15%)	373 (87.15%)
		≥80	352 (82.24%)	373 (87.15%)	373 (87.15%)
		≥90	324 (75.70%)	352 (82.24%)	352 (82.24%)
Average retrieve time (seconds)			0.112	0.114	0.119
Scenario 3	CBR indexing DBSCAN	≥70	389 (90.89%)	389 (90.89%)	389 (90.89%)
		≥80	389 (90.89%)	389 (90.89%)	389 (90.89%)
		≥90	371 (86.68%)	389 (90.89%)	389 (90.89%)
Average retrieve time (seconds)			0.105	0.106	0.107

The testing results of the three scenarios shows that the best accuracy and retrieval time at the threshold ≥ 90 for malnutrition disease data, acquired using the Minkowski distance method which is implemented on the CBR with indexing using the DBSCAN method. The accuracy is 100% with an average retrieval time of 0.02305 seconds. Research [3] with the same case data, reached the best accuracy of 85% with a threshold ≥ 0.75. So in the case of malnutrition, CBR with indexing using the DBSCAN method can improve accuracy. The best accuracy and retrieval time value at threshold ≥ 80 of heart disease data acquired using the Minkowski distance method implemented on CBR with DBSCAN indexing. The accuracy is 100% with an average retrieval time of 0.0421 seconds. This accuracy is as good as research [6] which produces 100% accuracy at threshold ≥ 80. The best retrieval time for thyroid disease data at threshold ≥ 90 is acquired using CBR with DBSCAN indexing of 0.107 seconds. This value is better than research [12] with an average retrieval time of 0.3045 seconds. But for accuracy calculation, CBR non-indexing is able to guess 392 correct data from 428 test data and produce an accuracy of 91.56%. Whereas CBR with cluster-indexing is able to guess 389 data from 428 test data and produces an accuracy of 90.89% which is implemented with the DBSCAN algorithm. This accuracy is smaller than the accuracy produced by research [12] using the Minkowski distance method with an accuracy of 92.52%.

In CBR with cluster-indexing the number of clusters greatly influences the retrieval time. Because the increasing number of clusters will make the cluster size of each cluster being relatively reduced. The retrieval time of the old case matching process will also be reduced, as the number of clusters decreases. On the other hand the time to search for relevant clusters will also increase in the process of finding the cluster center along with the increasing number of clusters. The number of clusters in the SOM algorithm is determined based on the number of output neurons while the initial weighting of the initial neurons is determined randomly. In DBSCAN, the larger value of ε***psilon*** the wider scope of the cluster. While too small εpsilon will produce a large number of clusters and the distance of objects are very close each other. Likewise, too large ***minPts*** will produce a lot of noise. This will affect the accuracy of the CBR system with cluster-indexing.

Non-indexing CBR always provides the highest similarity value as a solution. The solution is required by comparing new cases with all cases on a case base. If the CBR non-indexing finds cases with the same similarity value, the cases are sorted by the earliest calculation process and the top case is taken to be a solution. The diagnosis with the highest similarity is not always the same as the diagnosis given by experts. This is because the similarity method does not consider the level of confidence in the new cases. For the next research, it is necessary to add the level of expert confidence in diagnosing the disease since the different features that exist in a particular case.

4. Conclusion

The results of clustering with SOM algorithm are depend on the initiation of the initial weight given to the cluster and the number of neurons in the output layer. Initial weight initiation in the SOM algorithm is generated randomly so it is possible to obtain different clustering results for the same parameters. Likewise with the DBSCAN algorithm, the results of clustering are depend on the value of εpsilon and minPts specified at the beginning. Therefore, a proper method is needed to determine the most appropriate parameters for the SOM and DBSCAN algorithms in order to produce the best cluster.

In the case of malnutrition and heart disease data testing, CBR with cluster-indexing has better accuracy and shorter processing time than non-indexing CBR. Whereas in the case of thyroid disease the accuracy of non-indexing

CBR is better than non-indexing CBR, even though CBR with cluter-indexing has a better average retrieval time. Cluster-indexing method with DBSCAN algorithm has a better accuracy, faster processing and retrieval time than SOM. Whereas, of the three similarity methods, the Minkowski distance method produced the highest accuracy at the threshold of ≥ 90. Further research needs to consider the level of confidence in the new case and the level of expert confidence of a case in calculating the value of similarity due to differences in features that exist in a particular case.

References

[1] P. Berka, "NEST : A Compositional Approach to Rule-Based and Case-Based Reasoning," *Adv. Artif. Intell.*, vol. 2011, 2011.

[2] N. Rumui, A. Harjoko, and A. Musdholifah, "Case-Based Reasoning for Stroke Disease Diagnosis," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 12, no. 1, pp. 33–42, 2018.

[3] Nurfalinda and N. Nikentari, "Case Based Reasoning untuk Diagnosis Penyakit Gizi Buruk pada Balita," *J. Sustain. J. Has. Penelit. dan Ind. Terap.*, vol. 06, no. 02, 2017.

[4] M. Benamina, B. Atmani, and S. Benbelkacem, "Diabetes Diagnosis by Case-Based Reasoning and Fuzzy Logic," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 3, pp. 72–80, 2018.

[5] L. G. Vedayoko, E. Sugiharti, and M. A. Muslim, "Expert System Diagnosis of Bowel Disease Using Case Based Reasoning with Nearest Neighbor Algorithm," *Sci. J. Informatics*, vol. 4, no. 2, pp. 7–10, 2017.

[6] E. Wahyudi and S. Hartati, "Case-Based Reasoning untuk Diagnosis Penyakit Jantung," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 11, no. 1, pp. 1–10, 2017.

[7] S. Mulyana and S. Hartati, "Tinjauan Singkat Perkembangan Case – Based Reasoning," *semnasIF UPNVN Yogyakarta*, pp. 17–24, 2009.

[8] A. Sarkheyli and D. Söffker, "Case Indexing in Case-Based Reasoning by Applying Situation Operator Model as Knowledge Representation Model," *IFAC-PapersOnLine*, vol. 28, no. 1, pp. 81–86, 2015.

[9] J. Lu, D. Bai, N. Zhang, T. Yu, and X. Zhang, "Fuzzy Case-Based Reasoning System," *Appl. Sci.*, vol. 6, no. 7, p. 189, 2016.

[10] T. Rismawan and S. Hartati, "Case-Based Reasoning untuk Diagnosa Penyakit THT (Telinga Hidung dan Tenggorokan)," *Indones. J. Comput. Cybern. Syst.*, vol. 6, no. 2, pp. 67–78, 2012.

[11] S. Guo, F. Yang, Q. Lu, and X. Liu, "Combination Case-Based Reasoning and Clustering Method for Similarity Analysis of Production Manufacturing Process," *Proc. - 2015 Int. Conf. Ind. Informatics - Comput. Technol. Intell. Technol. Ind. Inf. Integr. ICIICII 2015*, pp. 97–101, 2015.

[12] D. Riyadi and A. Musdholifah, "Local Triangular Kernel-Based Clustering (LTKC) for Case Indexing on Case-Based Reasoning," *Indones. J. Comput. Cybern. Syst.*, vol. 12, no. 2, pp. 139–148, 2018.

[13] D. L. Olson, *Descriptive Data Mining*, 1st ed. Singapore: Springer Singapore, 2017.

[14] R. Popovici and R. Andonie, "Music genre classification with Self-Organizing Maps and edit distance," *Proc. Int. Jt. Conf. Neural Networks*, 2015.

[15] R. Umar, A. Fadlil, and R. R. Az Zahra, "Self Organizing Maps (SOM) untuk Pengelompokan Jurusan di SMK," *KHAZANAH Inform.*, vol. 4, no. 2, pp. 131–137, 2018.

[16] H. Shah, K. Napanda, and D. Lynette, "Density Based Clustering Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 3, no. 11, pp. 54–57, 2015.

[17] A. Musdholifah, S. Hashim, and S. Zaiton, "Cluster Analysis on High-Dimensional Data: A Comparison of Density-based Clustering Algorithms," *Aust. J. Basic …*, vol. 7, no. 2, pp. 380–389, 2013.

[18] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus External cluster validation indexes," *Int. J.*, vol. 5, no. 1, pp. 27–34, 2011.

[19] H. Seetha, M. N. Murty, and B. K. Tripathy, *Modern Technologies for Big Data Classification and Clustering*. Hershey PA: IGI Global, 2018.

[20] J. M. Merigó and M. Casanovas, "A new minkowski distance based on induced aggregation operators," *Int. J. Comput. Intell. Syst.*, vol. 4, no. 2, pp. 123–133, 2011.

Peer Reviewers

The Board of Editors greatly appreciate the participation of the following reviewers that help during the review process for the publication of Khazanah Informatika volume 6 (year 2019).

1. Ahmad Ramdani, Institut Teknologi Sumatera
2. Aris Rakhmadi, UMS
3. Azizah Fatmawati, UMS
4. Bana Handaga, UMS
5. Dedi Gunawan, UMS
6. Devi AP Putri, UMS
7. Diah Priyawati, UMS
8. Dwi Pangestuty, Universitas Muhammadiyah Kalimantan Timur
9. Endah Sudarmilah, UMS
10. Frieyadie, STMIK Nusa Mandiri Jakarta
11. Hari Prasetyo, UMS
12. Herry Sujaini, Universitas Tanjungpura
13. Indra Waspada, Universitas Diponegoro
14. Leon Abdillah, Universitas Bina Darma
15. Mardhiya Hayati, Universitas AMIKOM Yogyakarta
16. Naufal Azmi Verdikha, Universitas Muhammadiyah Kalimantan Timur
17. Puji Ramadhan
18. Ramalia Narotama Putri, Sekolah Tinggi Ilmu Komputer Pelita Indonesia
19. Rofilde Hasudungan, Universitas Muhammadiyah Kalimantan Timur
20. Rudiman, Universitas Muhammadiyah Kalimantan Timur
21. Sayekti Harits Suryawan, Universitas Muhammadiyah Kalimantan Timur
22. Sitaresmi Handani, AMIKOM Purwokerto
23. Siti Puspita Sakti, STMIK Syaikh Zainuddin NW Anjani
24. Tati Ernawati, Politeknik TEDC Bandung
25. Umi Fadlilah, UMS
26. Yogiek Indra Kurniawan, Universitas Jenderal Soedirman
27. Yuliant Sibaroni, Universitas Telkom