

ANALYZING LECTURER PERFORMANCE FACTORS FROM COURSE EVALUATION SURVEYS USING K-MEANS CLUSTERING AND C4.5 CLASSIFICATION

Ariel Alfaro

Program Studi Kecerdasan Buatan dan Robotika
Sekolah Vokasi, Universitas Pakuan

ariel.alfaro@unpak.ac.id

ABSTRACT

One of the quality of education of a college can be seen from the quality of the performance of the lecturers in the higher education Tridharma namely education, research, and development and community service. This study aims to analyze the lecturer performance factors based on the course evaluation survey on the Tridharma of higher education in the implementation of academic evaluations and decision making for lecturers in the Computer Science study program, Pakuan University. The research method applies a combination of two data mining methods, namely k-means clustering and C4.5 classification used in the assessment of the performance of lecturers, especially in the process of education and college teaching which includes pedagogical, professional, personality and social competencies. The results of the K-means clustering mining process were assessed by learning, namely 5 sufficient lecturer clusters, 16 good cluster lecturers and 14 excellent cluster lecturers. C 4.5 classification is used to see the connectedness of factors such as learning, publication, education, PKM, support. This study shows that publication criteria are the most influential factors in the performance assessment of lecturers. Testing the level of accuracy using the K-fold Cross Validation method with 5-fold Cross Validation is 80.00% and 7-fold Cross Validation which is 82.86%.

***Keywords-** Clustering, Classification, Data Mining, Lecturer Performance Assessment.*

1. INTRODUCTION

Higher education has a big role in national development. Based on Law No. 12 of 2012 on Higher Education, higher education is obliged to develop innovative, creative and competitive academics through the implementation of Tridharma. One of the quality of education in a tertiary institution can be seen from the quality of teacher performance in carrying out their duties. In the university the scholarship is still conducting a List of Employee Work Evaluation Sheets (DP3) and BKD, while the Assessment is only limited to the evaluation of the work of lecturers for DP3, and BKD is only shown to lecturers who already have certificates. With the existence of a system called sisdata where the lecturer fills out the lecturer Tridharma activities, it will be used as a reference to assess the lecturer performance factor. Diversity of performance can be done by several factors, both internal and external factors.

Academic and teaching evaluations to determine the performance of lecturers in teaching can be done with feedback or course evaluation surveys to get the value of student satisfaction with the performance of lecturers, especially in the education and teaching process of higher education that includes pedagogical, professional, personal and social competencies as sources for clustering process. The results of the course evaluation survey become parameters for evaluating lecturers 'performance in teaching and will look at the connectedness factors in lecturers' activities using the k-means clustering method and C4.5 to see the factors that influence the achievement of lecturer assignments using the C4 classification method. 5 There are several studies

that have been carried out by many previous researchers who are concerned with the performance of lecturers using clustering and classification methods, such as the application of data mining for lecturers' performance using k-means clustering case studies at AMIK Mahaputra Riau which aims to implement data mining techniques to facilitate the process of determining lecturer performance on students in teaching and learning process by looking at criteria such as teaching and personal including pedagogical competence, professional competency, personality competency, social competency[1]. Modeling of teacher's performance index assessment using the fuzzy simple additive weighting method. elaboration, lecturer activities [2]. Conducting research predicting student satisfaction with lecturers' performance at TEDC Bandung Polytechnic by testing the decision tree (C4.5) method aimed at getting more accurate results to predict student satisfaction with the best lecturer performance and getting the highest level of accuracy [3].

So to overcome the problems and previous studies show that the assessment of lecturers' performance by using data mining techniques by using hybrid mining can be done by clustering using the k-means method and in the classification with the C45 method. Interrelation and connectedness in aspects that trigger values so that decisions will be able to provide decision-making for institutions in evaluating lecturers' performance and problems will be easily resolved.

The purpose of this study is the application of Hybrid mining using the K-Means and C45 methods for the analysis of lecturers' performance factors based on the course evaluation survey on Tridharma of tertiary

institutions.

The scope needed in this research is in the form of online questionnaire data for Computer Science students and Computer Science lecturer data. This study analyzes the factors to look for lecturers' performance evaluations from the courses evaluation

2. METHOD

Research Methodology The method applied in this is using data mining or also called Knowledge Discovery and Data Mining (KDD) [4]. As a series of processes, data mining can be divided into several stages which are shown in Figure 1. These stages are interactive, the user is directly involved or through the mediation of a knowledge base.

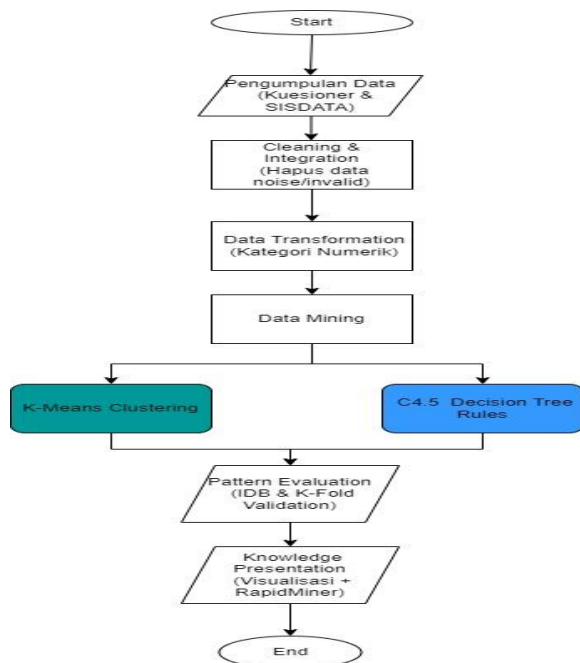


Figure 1. Data Mining Stages

2.1. Cleaning And Integration Data

Data cleaning is the process of removing noise and inconsistent or irrelevant data. [5]. Cleansing the questionnaire data was obtained from PUTIK in the academic and student division section of the urak see unpak.ac.id from the online questionnaire menu and for the data on the activities of Tridharma lecturers taken from the lecturers' data. Removing inconsistent data in the student satisfaction index data or called the online questionnaire obtained from the Communication Information Technology Center obtained questionnaire data on computer science as many as 7000 data and lecturer activities from 35 data collected data will be calculated based on student satisfaction index to be obtained clustering and classified.

2.2. Data Transformation

Data transformation according to [5] in the database is often not all used, therefore only data that is suitable for analysis will be retrieved from the database. Then the transformation of lecturer questionnaire data with 200

survey based on the three Tridharma of the college, and other lecturer activity data related to the research. The competencies that are assessed are Pedagogic, Professional, Personality and Social, and lecturer activity data from the SISDATA Application and implementation using the RapidMiner application.

respondents based on the subject taught data is changed or combined into a format suitable for processing in data mining. That is the result of each answer the question has been determined the weight rating on each question answer that is A = 5, B = 4, C = 3, D = 2, E = 1. The following is a description of the data that has been selected with a categorical data type, but in the process determine the value of learning using clustering the need for numeric data types. For this reason, weights are made for each satisfaction index based on aspects of the competency of the three-Tridharma of higher education in order to become numerical values in categorical data types for the conversion of categorical data into numerical data.

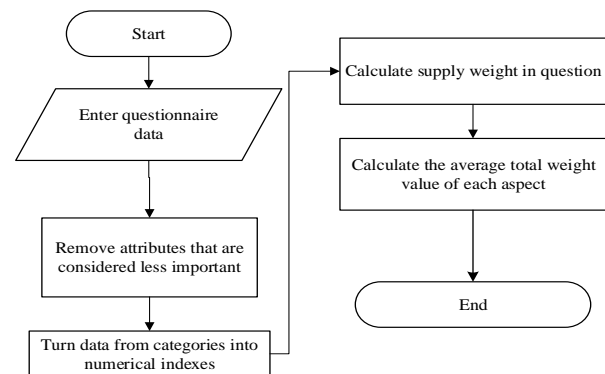


Figure 2. Data transformation flowchat

2.3. Procces Mining

Is a major process when the method is applied to find valuable and hidden knowledge from data [6]. hybrid mining combining two opposing elements but still maintaining the character of these elements. is a combination of two or more data mining techniques or methods that aim to minimize their shortcomings and optimize the data collected to be used as material in the decision making process [9]. To analyze the connectedness of a factor in lecturer performance with the lecturer Tridharma activity. the mining process of this research data uses the K-Means and C45 methods. This research will conduct a grouping of data on the accumulation of questionnaire values by clustering using the k-means method.

K-Means Clustering Is a non-hierarchical data clustering method that groups data in the form of one or more clusters / groups. Data that has the same characteristics are grouped in one cluster / group and data that have different characteristics according to [7]. The steps to clustering using the K-Means method are as follows:

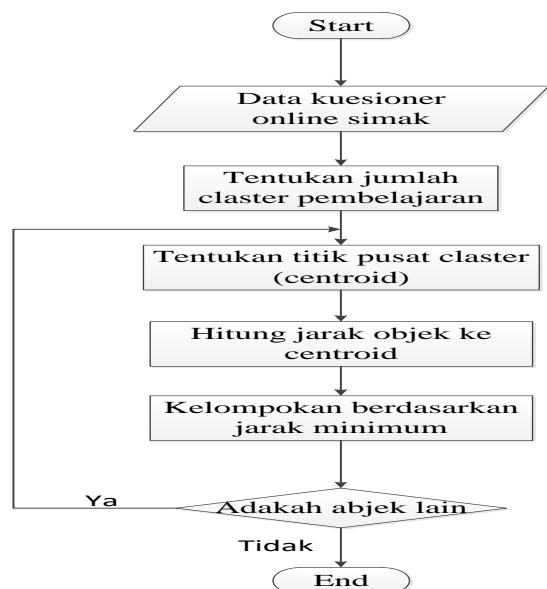


Figure 3. K-means flow chart

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^n (x_j - c_j)^2} \dots\dots\dots(1)$$

Euclidean Distance formula description:

d = Distance

j = Amount of data

c = Centroid

x = Data

C4.5 algorithm and decision tree are two inseparable models, because to build a decision tree, C4.5 algorithm is needed. In the late 1970s until the early 1980s, J. Ross Quinlan, a researcher in the field of machine learning, developed a decision tree model called ID3 (Iterative Dichotomiser), even though this project had actually been made before by EB Hunt, J. Marin, and PT Stone. Then Quinlan made an algorithm from the development of ID3 called C4.5 based on supervised learning. There are several steps in making a decision tree with the C4.5 algorithm [8], namely:

1. Prepare training data. Training data is biased from historical data that has happened before and has been grouped into certain classes.
2. Determine the roots of the tree. the root will be taken from the selected atribut by calculating the Gain value of each attribute, the highest Gain value which will be the first root. Before calculating the Gain value of the attribute, first calculate the entropy value, namely:

$$Entropy(S) = - \sum_{i=1}^n p_i * \log_2 p_i \dots\dots\dots(2)$$

Information:

S: Case set

A: Attribute

n: Number of partitions S

pi: Proportion of Si to S

3. Then calculate the Gain value with the information gain method:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \dots\dots\dots(3)$$

Information:

S: case set

A: attribute

n: number of attribute attributes A

| Si |: number of cases on the i-th partition

| S |: number of cases in S

4. Repeat step 2 until all tuples are partitioned.

5. The decision tree partition process will stop when: a. All tuples in node N have the same class. b. There are no attributes in tuples that are partitioned anymore. c. There are no tuples in the empty branch.

2.4. Pattern Evaluation

To identify interesting patterns into knowledge based that are found. In this stage the results of data mining techniques in the form of distinctive patterns and prediction models are evaluated to assess whether the existing hypotheses have indeed been reached [6]. If it turns out the results obtained do not fit the hypothesis there are several alternatives that can be taken such as making feedback to improve the data mining process, or accept this result as an unexpected result that might be useful.

2.5. Knowledge Presentation

Knowledge presentation is a visualization and presentation of knowledge about the methods used to obtain the knowledge obtained by users. The last stage of the data mining clustering process is in the form of a web-based system. Classification is how to formulate a decision or action from the results of the analysis obtained. In this presentation, the results containing rules are implemented in RapidMiner.

3. RESULTS AND DISCUSSION

3.1. Implementation of the K-Means Mining Process

Following this with the data used from the results of the questionnaire 7,000 data and carried out the process of transforming the data to get the results of the questionnaire values obtained by the teacher as many as 35 data. Academic and teaching evaluation to determine lecturer performance in teaching which includes pedagogical, professional, personal and social competencies as a source for grouping processes. The following table 1 data for the k-means process.

Table 1. Calculation data for the K-means process.

N O	NAMA DOSEN	PEDAD OGIK	PROFESI ONAL	KEPRIBAD IAN	SOSI AL
1	ASA	4,00	3,89	3,88	3,86
2	APP	4,01	4,02	4,01	4,03
3	AI	4,13	4,17	4,20	4,09
4	AC	4,26	4,24	4,03	4,28
5	AQ	4,16	4,16	4,17	4,24
35	YY	4,27	4,25	4,26	4,35

This research will be conducted on the accumulation of questionnaire data that has been processed in the previous stage to assess Tridharma namely education and learning by initializing a cluster center randomly or randomly. This stage is the number of clusters is 3, namely decisions C1 (Fair), C2 (Good) and C3 (Very Good) and will be calculated the distance of each cluster from one data where the cluster will calculate the distance of all data to

each center of mass using equation (1). Then do the same thing for P5 to P35 and also do it on Centroid 2 to Centroid 3 which produces Iteration 1, After that do the same thing calculate P1 to Centroid 3 to determine 2nd iteration and so on iteration until it reaches the same membership value that make the counting process stop and if the membership results are different then proceed in step 3 until the membership value is the same.

NO	NAMA DOSEN	ITERASI 1			ITERASI 2			ITERASI 3		
		C1	C2	C3	C1	C2	C3	C1	C2	C3
1	ASA	1			1			1		
2	APP	1			1			1		
3	AI		1			1			1	
4	AC			1			1			1
5	AO		1			1			1	
6	AM			1			1			1
7	AS		1			1			1	
8	BHS			1			1			1
9	DA	1			1			1		
10	DKU		1			1			1	
11	DS		1			1			1	
12	EK		1			1			1	
13	ETT			1			1			1
14	FDW		1			1			1	
15	HTS			1			1			1
16	HFA		1			1			1	
17	HBS			1			1			1
18	HT		1			1			1	
19	IW	1			1			1		
20	IA			1			1			1
21	IM		1			1			1	
22	KS		1			1			1	
23	LKS			1			1			1
24	MSI		1			1			1	
25	MIS			1			1			1
26	MYT		1			1			1	
27	PH		1			1			1	
28	SA			1			1			1
29	SH			1			1			1
30	SS			1			1			1
31	SM		1			1			1	
32	TPN		1			1			1	
33	TAZ	1			1			1		
34	VIS		1			1			1	
35	YY			1			1			1

Figure 4. Results of manual learning clustering calculations

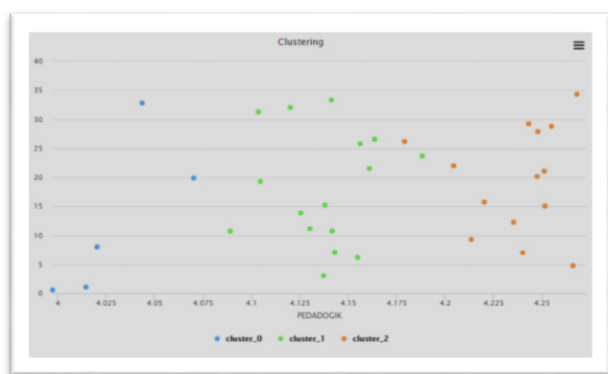


Figure 5. Rapidminer clustering learning outcomes

Clustering from various Pedagogic aspects, namely the percentage of dominant attribute values, in the 1st and 2nd iteration the cluster position does not change, the iteration is stopped and the final result will be obtained, namely: The results obtained that 5 lecturers included in the category for cluster C1 with each attribute average value, namely Pedagogic of 4,026, Professional of 3,992, Personality of 3,988, Socially of 4,026. Then 16 lecturers for the C2 cluster with each value from the average attribute, namely Pedagogic for 4,136, Professional for 4,133, Personality for 4,137, Social for 4,159. Furthermore, 14 lecturers in the C3 cluster category with each attribute average value, namely Pedagogic of 4,236, Professional of 4,232, Personality of 4,226, Social of 4,252.

3.2. Implementasi Proses Mining C45

The results in the calculation of clustering which have been processed in the previous stage to assess Tridharma

namely education and learning. After getting the results of learning will look for connectedness to determine Tridharma factor analysis for assessing lecturer performance, will get a decision rule. At this stage, the authors test and design the system using RapidMiner data mining software. The following is an explanation of the implementation of the C4.5 process with rapidminer.

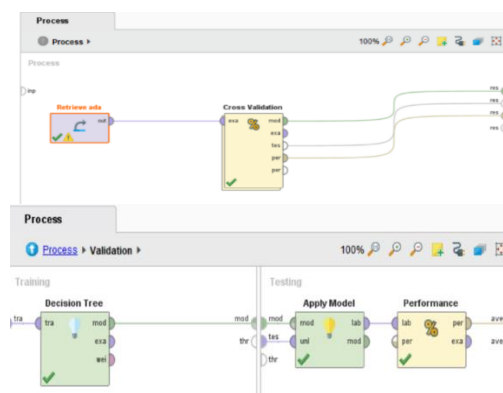


Figure 6. Implementation of the cross validation model on C45.

In Figure 6 explains the lecturer activity data associated with cross validation operators in which there is a process model of the C4.5 algorithm testing data by applying the model and performance on the rapidminer, where the operator and the parameters used will affect the accuracy and model of the decision root. The following picture is the implementation of the cross validation model on C45.

Table 2. Results of Entropy Calculation and Information Gain C45

		JUMLAH KASUS (S)	CUKUP	BAIK	AMAT BAIK	ENTROPY	GAIN
TOTAL		35	23	7	5	1,2635	
PEMBELAJARAN	AMAT BAIK	14	7	3	4	1,4926	
	BAIK	16	12	3	1	1,0141	
	CUKUP	5	4	1	0	0	
JAD							0,6512
	AMAT BAIK	14	3	6	5	1,5306	
	BAIK	14	13	1	0	0,0000	
	CUKUP	7	7	0	0	0	
PENDIDIKAN							0,1557
	AMAT BAIK	6	1	3	2	1,4591	
	BAIK	29	22	4	3	1,0351	
	CUKUP						
PUBLIKASI							0,9666
	AMAT BAIK	6	0	2	4	0	
	BAIK	8	2	5	1	1,0351	
	CUKUP	21	21	0	0	0	
PKM							0,3157
	AMAT BAIK	2	0	0	2	0	
	BAIK	2	0	2	0	0	
	CUKUP	31	23	5	3	1,0701	
PENUNJANG							0,2638
	AMAT BAIK	7	1	3	3	1,4488	
	BAIK	4	2	1	1	1,5000	
	CUKUP	24	20	3	1	0,7852	

The results in table 2 will be in the form of a decision tree which is then interpreted in a decision rule that can be used to predict factors for evaluating lecturer performance. the data used by the lecturer student activity data by classification using the C45 algorithm, is used to make a decision tree, the first step is to count the number of classes for the results of AMAT GOOD, GOOD, and ENOUGH. Then calculate the entropy value of all classes divided by the attributes of Learning, JAD, Education, Publication, PKM and Support. Based on table 2 the entropy value is obtained to calculate the attribute value in determining as root in the decision tree. The results obtained that the gain value of each attribute is Learning 0.2028, JAD 0.6512, Education 0.1557, Publication

0.9666, PKM 0.3157, Supporting 0.2638 and PUBLICATION get the highest gain information 0.9666, then this attribute is used as the root node (root node). Following are the results of the decision tree using Rapidminer.



Figure 7. Decision tree using Rapidminer

Based on Figure 7 we can explain the rules produced by the decision tree, there are several rules, this can be seen in accordance with Decree of Pakuan University Chancellor Number 19./KEP/REK/IV/2018. that the indicators Publication, Learning, Education, Support. Influence on factors to determine the evaluation of lecturer performance with the Publication indicator has the biggest influence factor. Based on Figure 8 we can explain the rules produced by the decision tree, there are several rules, this can be seen in accordance with Decree of Pakuan University Chancellor Number 19./KEP/REK/IV/2018. that the indicators Publication, Learning, Education, Support. Influence on factors to determine the evaluation of lecturer performance with the Publication indicator has the biggest influence factor.

3.3. Evaluation of Clustering Validation Results

In this study the results of grouping using the K-Means IDB algorithm are used to validate data on each Cluster so that outliers are placed in the correct Cluster the more IDB values indicate the most optimal cluster scheme [11]. In the validation trial phase with the number of 3 clusters, 4 clusters and 5 test clusters. In the number of clusters 4 with the largest value of 0.802 and in cluster 5 the lowest value with IDB is 0.600.

Table 3. Clustering Validation Test

Jumlah Cluster	IDB
C3	0.679
C4	0.802
C5	0.600

3.4. Evaluation of K-fold Cross Validation Results

In this study the data used are lecturer activities with 35 data. According to Fu [10] k-Fold Cross Validation repeats k-times to divide a set of samples randomly into k mutually independent subsets, each repetition leaving one subset for testing and a subset others for training. In the k-times research used 5 and 7 with a total of 35 data. The results of this test are seen as a whole in the lecturer activity data for evaluating lecturer performance. Based on Figure 8, it is obtained that for 5-Fold Cross Validation with an accuracy of 80.00% and 7-Fold Cross Validation of 82.86%. This shows that the method used has a fairly large degree of accuracy better.

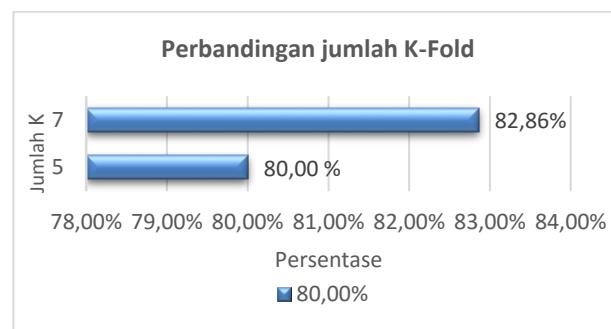


Figure 8. Accuracy results using k-Fold Cross Validation.

4. CONCLUSION

The application of Hybrid mining to analyze lecturer performance factors based on the course evaluation survey using the K-Means and C45 methods. There are 3 results of the decision from the clustering process the results of the decision were obtained from the questionnaire data of student satisfaction index for lecturers which included aspects of Tridharma competencies namely Pedagogic, Professional, Personality, Social Wherein that 5 lecturers included in the C1 category. Furthermore 16 lecturers were included in the C2 category. C3 The factors evaluating the performance of lecturers for classification are Learning, Education, Publication, PKM, Supporting. Based on the results of testing using the C45 method it was found that the most influential factor was Publication with a decision root value of 0.9666. The results of testing with k-fold cross validation on the system for evaluating lecturer performance obtained an accuracy of 5-fold cross validation of 80.00% while 7-fold cross validation of 82.86%.

REFERENSI

- [1] Darmanta Sukrianto, 2016. Penerapan Data Mining Untuk Kinerja Dosen Menggunakan K-means Clustering (Studi Kasus Di Amik Mahaputra Riau). Jurnal PI-Cache. 5(1): 54- 63.
- [2] Muhamad, M., Febri, T., Kasmi, Leni, A. 2017. Pembuatan model penilaian indeks kinerja dosen menggunakan metode fuzzy simple additive weighting. Seminar Nasional Teknologi Informasi dan Multimedia, STMIK AMIKOM Yogyakarta: 4-7 Februari 2017. ISSN : 2302-3805.
- [3] Ade Yuliana and D. B. Pratomo. 2017. Algoritma Decision Tree (C4.5) untuk memprediksi kepuasan mahasiswa Terhadap kinerja dosen politeknik TEDC Bandung. Seminar Nasional Inovasi Teknologi, UN PGRI Kediri: 22 Februari 2017. e-ISSN : 2549-7952.
- [4] Han, J. & Kamber, M. 2006, "Data Mining Concept and Techniques Second Edition ". Morgan Kauffman, San Francisco.
- [5] Tosida et al. 2015. Visualization model of small and medium enterprises (SMEs) Telematics services potentiality map in Indonesia. Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Pakuan, Bogor.

- [6] Budanis, D. M, Nofi, S. 2015. Aplikasi Data Mining Untuk Menghasilkan Pola Kelulusan Siswa Dengan Metode Naïve Bayes. Jurnal Ilmiah NERO. 1(3): 182-189.
- [7] Johan Oscar Ong. 2013. Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University. Jurnal Ilmiah Teknik Industri. 12(1): 1412- 6869.
- [8] Kusriani and E. T. Luthfi. 2009. Algoritma Data Mining. Yogyakarta : Andi.
- [9] Sung, H. H., & Seong H. J. 2010. A *Hybrid Data Mining Method for the Medical Classification of Chest Pain*. International Journal of Computer and Information Engineering. 4(1):99-104.
- [10] Fu L. 1994. *Neural Network In Computer Intelligence*. Singapura : McGraw Hill.
- [11] Budi, D. S. Bain, K. K., Adam, M. 2015. Pengelompokan Tingkat Kesehatan Masyarakat Menggunakan Shelf Organizing Maps Dengan Cluster Validation Idb dan I-Dunn Seminar Nasional Aplikasi Teknologi Informasi (SNATi) 2015 Yogyakarta, 6 Juni 2015
- [12] DIKTI KEMENTERIAN RISET. 2019. Pedoman Operasional Penilaian Angka Kredit Kenaikan Jabatan Akademik/Pangkat Dosen Direktorat Jenderal Sumber Daya IPTEK dan Dikti Kementrian Riset, Teknologi dan Pendidikan Tinggi. Jakarta.