IOP Conference Series: Materials Science and Engineering

Optimization of K Value at the K-NN algorithm in clustering using the expectation maximization algorithm

To cite this article: Zulkarnain Lubis *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **725** 012133

IOP Conf. Series: Materials Science and Engineering **725** (2020) 012133 doi:10.1088/1757-899X/725/1/012133

Optimization of K Value at the K-NN algorithm in clustering using the expectation maximization algorithm

Zulkarnain Lubis[1,*], Poltak Sihombing[2], Herman Mawengkang[2]

[1]Graduate School of Computer Science
[2]Department of Information Technology, Faculty of Computer Science and Information Technology, University of North Sumatra, Medan, Indonesia

zulkarnainlb@gmail.com

Abstract. Data is the most important thing in a study. The quality of the results of the research will be directly proportional to the quality of the data that will be used in the research is concerned. One of the problems that exist in the data set is the absence of a value in the data for a particular attribute or better known as the missing data. One method that is often used by researchers is the k-nearst Neighbor (KNN). However, this method has several drawbacks, one of which is the selection of appropriate values of k not to degrade the performance of the classification. In the process of calculating the parameters k KNN there that can affect the accuracy of the classification results. To use more than one parameter k then used by majority voting to determine the classification results. If the parameter k in KNN classification used 1 then the result was very tight because it will use the nearest neighbor to the results of the classification. Conversely, if the value of the parameter k used KNN is great then the classification results will blur.This research will optimize the parameters k in the UN tax cluster using the algorithm*expectation Maximation*(EM). The results of the research in the form of clustering information by using the number of clusters k value optimization and the number of clusters without using the optimization of the value k. Then analysis the results after getting data already clustered. Results from the study showed that k obtained from the optimization algorithm can improve the results of the cluster where the 66% error can be reduced to 64%, yet very close to the best result of the measurement accuracy is tested.

1. Introduction

One way to classify the data is by using Clustering. Data grouping or clustering is a method used to classify into groups or clusters based on similarity, so that related data is placed in the same cluster. There are several clustering algorithms known ie partitional (Expectation-maximization, K-Means) and hierarchical (Centroid Linkage, Single Linkage), overlapping (Fuzzy C-Means) and hybrid. The algorithm can overcome the arbitrary grouping are partitional algorithm. Where in partitional algorithm, a document can be a member of a group or cluster to a process but the subsequent processes such documents can be moved to another cluster. One partitional algorithm that can group documents that have not been labeled is Expectation-Maximization, the algorithm used to find the value of Maximum Likelihood estimation of parameters in a probabilistic model. The characteristics of this algorithm is able to classify the data that has not been labeled or unlabeled data and also the results of the classification will always convergence. This algorithm has two phases: phase and phase Expectation Maximization.

In Expectation step (E-step) use the EM algorithm - Cluster for classifying data based on the model parameters. While on the Maximization step (M-step) will be done peng updates of the model parameters by using Multiple Linear Regression. Phase E-step and M-step is continued until the probability of each cluster achieve convergence. Before performing the necessary process of grouping data pre-processing, namely cleansing, tokenizing, parsing. Labeling of a cluster is done by finding the most actual label appears on a Cluster, and then adopt the label as the label Cluster. With the implementation of the EM algorithm - Cluster in the process of budget clusterisasi it can classify and determine the appropriate number of clusters,

2. Stages of Data Mining

Data mining is actually a part of the process of Knowledge Discovery in Databases (KDD), not as a technology intact and independent. Data mining is an important part of steps in the process of KDD primarily concerned with the extraction and calculation of the data patterns are analyzed, as shown by Figure 1 below:

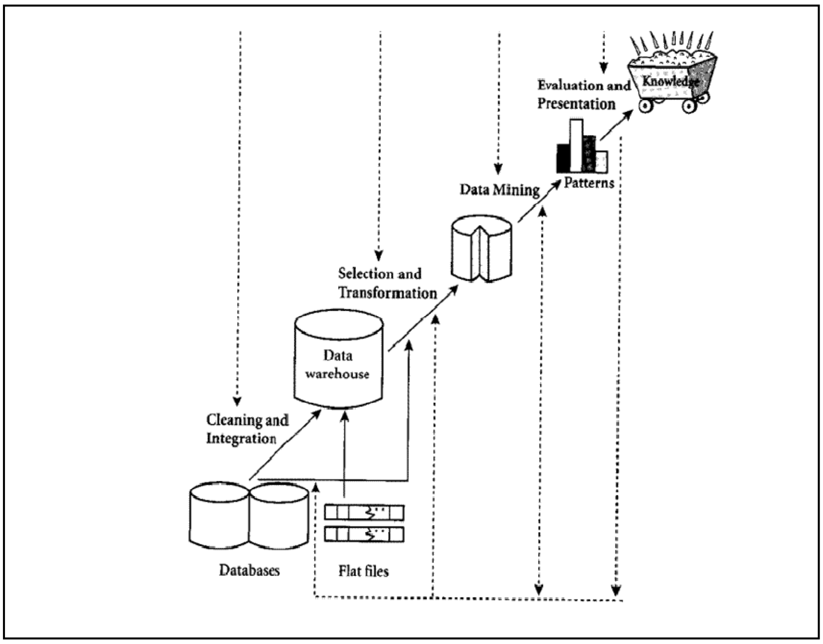

Figure 1. Stages in the process of knowledge discovery

a. *Data cleaning*
   To eliminate the data noise (irrelevant data / dealing directly with the ultimate goal of data mining process, eg data mining that aims to analyze the results of the sale, then the data in the collection as "employee name", "age", and so on can -ignore) and inconsistent.
b. *Data integration*
   To combine multiple data sources.
c. *Data selection*
   To retrieve the appropriate data for analysis.
d. *Data transformation*
   To transform data into a form more suitable for mining. Data mining is the most important process in which a particular method is applied to generate the data pattern.

IOP Conf. Series: Materials Science and Engineering **725** (2020) 012133 doi:10.1088/1757-899X/725/1/012133

e. *Pattern evaluation*
To identify whether interenting patterns obtained is sufficient to represent knowledge based on specific calculations.
f. *Knowledge presentation*
To present the knowledge that has been obtained from the user.

2.1 Method of KNN (K-Nearest Neighbor)

The working principle of the K-Nearest Neighbor (KNN) is seeking the shortest distance between the data to be evaluated by K neighbors (neighbor) closest to the training data. This technique is included in the nonparametric classification groups. Here we do not pay attention to the distribution of the data to be grouped. This technique is very simple and easy to implement. Similar to clustering techniques, we classify a new data based on the distance the new data into multiple data / neighbor (neighbor) nearby.

KNN algorithm purpose is to classify the new objects based on attributes and sample training. Clasifier not use any model to be matched and only based on memory. Given query point, will find a number of objects or k (training points) closest to the query point. Classification using the voting majority among the classification of k objects. KNN classification algorithm uses adjacency as the predicted value of the new query instance. Algorithm KNN method is simple, operates on the shortest distance from the query instance to the training sample to determine its KNN.

K best value for this algorithm depends on the data. In general, a high k value will reduce the effect of noise on klsifikasi, but draw the line between each classification is becoming increasingly blurred. Nice k value can be selected by optimization of parameters, for example by using cross-validation. The special case where the classification is based on the training data diprekdisikan closest (in other words, k = 1) is called Nearest Neighbor algorithm.
excess KNN *(K-Nearest Neighbor):*
1. Resilient to training data that has a lot of noise.
2. Effective if training data is huge.

The weakness of KNN *(K-Nearest Neighbor):*
1. KNN need to determine the value of the parameter k (the number of nearest neighbors).
2. *Training* based on distance is not clear on what kind of distance that must be used.
3. Which attributes should be used to get the best results.
4. The computational cost is high because the necessary calculation of the distance of each query instance in the whole training sample.

2.2 KNN algorithm

1. Determine the parameter K
2. Calculate the distance between the data to be evaluated with all the training
3. Sort range formed (ascending)
4. Determine the shortest distance to the order of K
5. Pair the corresponding class
6. Find the number of classes from the nearest neighbor and set the class as a class data to be evaluated

KNN formula:

$$d_i = \sqrt{\sum_{i=1}^{p} \left(x_{2i} - x_{1i}\right)^2} \qquad \text{... 1}$$

Information:

x_1 = Sample Data
x_2 = Data Test / Testing
i = Variable Data
d = Distance formed
p = Dimension Data

Below is a flowchart of the method KNN:

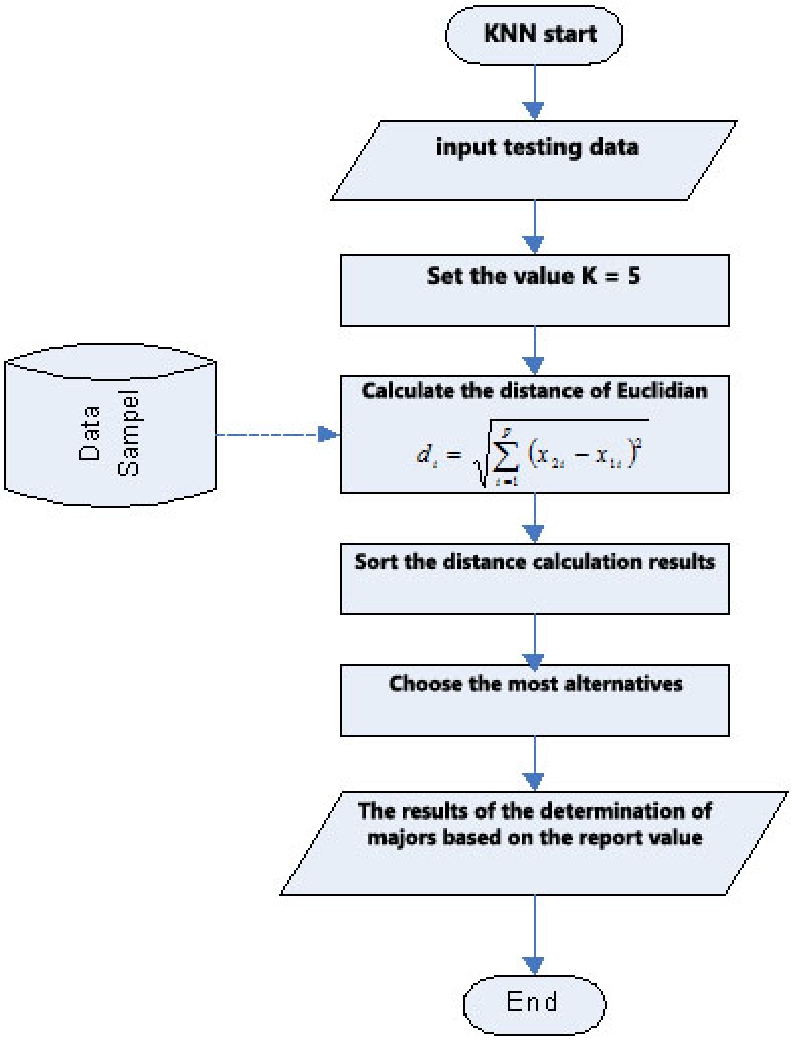

Figure 2. Flowchart of KNN Method

3. Methodology

Broadly speaking, the stages in this study is illustrated in Figure 3.

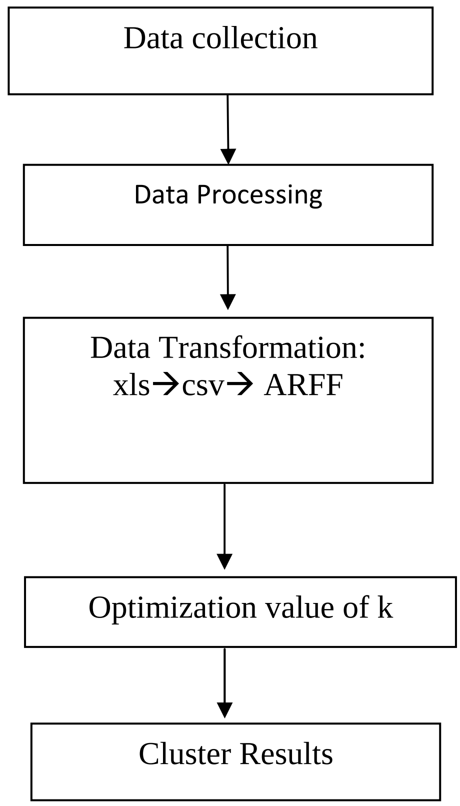

Figure 3. Block diagram of the stages of research

Figure 3 above is a research methodology that will be done by the author. The research methodology aims to outline all the activities carried out during the course of the study. From the picture above, it is known that there are three stages to be done to resolve the case at this research that includes: data collection, pre-process the data, data transformation, optimization of the value of k and cluster results. The preparation process includes three main things:

3.1 Data Selection

Select the data that will be used in the data mining process. In the process of the election is done also attributes that are tailored to the data mining process. In this study, the data used is in the form of data-ready, meaning that the data obtained has been the form of the target data. At this stage the problem to be faced is noisy data and missing values. Data cleaning process pelu done to clean data from duplicate data, the data is inconsistent, or typographical errors. So the data that has been through this process are ready to be processed in data mining. In this study, the data used is data that has been consistent, so that the data cleansing process is only performed on any data missing value.

IOP Conf. Series: Materials Science and Engineering **725** (2020) 012133 doi:10.1088/1757-899X/725/1/012133

Table 1. The data in excel format

	A	B	C	D	E	F	G	H	I
1									
2	SPPT	TARGET	SPPT Y Bayar	Tercapai	SPPT Belum Bayar	Target Tak Capai	Penerimaan	Pencapaian	Kategori
3	997	111.687.118	450	90.128.985	547	21.558.133	80,70	TERCAPAI	NAIK
4	673	58.896.882	316	49.427.433	357	9.469.449	83,92	TERCAPAI	TURUN
5	2.108	106.493.314	935	72.531.398	1.173	33.961.916	68,11	TERCAPAI	NAIK
6	1.400	57.434.795	535	25.445.940	865	31.988.855	44,30	TIDAK TERCAPAI	TURUN
7	498	32.945.502	233	18.709.761	265	14.235.741	56,79	TIDAK TERCAPAI	TURUN
8	473	52.273.697	175	34.690.376	298	17.583.321	66,36	TERCAPAI	TURUN
9	942	30.941.684	524	17.561.782	418	13.379.902	56,76	TIDAK TERCAPAI	TURUN
10	718	25.902.318	390	13.759.250	328	12.143.068	53,12	TIDAK TERCAPAI	NAIK
11	385	26.294.309	169	10.269.476	216	16.024.833	39,06	TIDAK TERCAPAI	TURUN
12	336	10.101.119	173	5.799.489	163	4.301.630	57,41	TIDAK TERCAPAI	TURUN
13	991	70.296.473	414	49.007.088	577	21.289.385	69,71	TERCAPAI	TURUN
14	941	156.416.754	489	138.097.411	452	18.319.343	88,29	TERCAPAI	TURUN
15	1.111	147.920.197	588	106.572.381	523	41.347.816	72,05	TERCAPAI	NAIK
16	2.903	144.972.498	1.420	79.203.686	1.483	65.768.812	54,63	TIDAK TERCAPAI	NAIK
17	288	44.674.689	169	21.339.751	119	23.334.938	47,77	TIDAK TERCAPAI	TURUN
18	2.725	329.456.622	1.537	204.438.210	1.188	125.018.412	62,05	TERCAPAI	NAIK
19	1	3.166.668	1	3.166.668	0	0	100,00	TERCAPAI	TETAP
20	344	23.859.036	223	16.708.075	121	7.150.961	70,03	TERCAPAI	TURUN
21	1.196	43.499.255	780	31.458.923	416	12.040.332	72,32	TERCAPAI	TURUN
22	1.482	51.070.436	387	16.191.750	1.095	34.878.686	31,70	TIDAK TERCAPAI	TURUN
23	1.658	70.966.662	856	42.877.842	802	28.088.820	60,42	TERCAPAI	TURUN
24	507	22.847.146	207	10.553.302	300	12.293.844	46,19	TIDAK TERCAPAI	NAIK
25	1.974	117.421.573	530	57.231.401	1.444	60.190.172	48,74	TIDAK TERCAPAI	NAIK
26	1.459	93.092.242	538	36.377.251	921	56.714.991	39,08	TIDAK TERCAPAI	TURUN
27	437	23.395.822	196	12.469.996	241	10.925.826	53,30	TIDAK TERCAPAI	NAIK
28	873	24.245.854	574	17.610.111	299	6.635.743	72,63	TERCAPAI	NAIK
29	1.433	42.586.720	820	26.287.982	613	16.298.738	61,73	TERCAPAI	NAIK
30	725	20.110.320	381	9.477.328	344	10.632.992	47,13	TIDAK TERCAPAI	TURUN
31	1.367	40.779.994	862	26.035.516	505	14.744.478	63,84	TERCAPAI	NAIK
32	2.469	63.490.805	1.879	49.900.323	590	13.590.482	78,59	TERCAPAI	TURUN
33	693	61.402.741	399	47.953.448	294	13.449.293	78,10	TERCAPAI	NAIK
34	874	33.368.374	443	17.998.100	431	15.370.274	53,94	TIDAK TERCAPAI	NAIK
35	1.099	111.948.767	328	53.875.448	771	58.073.319	48,13	TIDAK TERCAPAI	NAIK
36	1.543	91.571.016	795	54.217.988	748	37.353.028	59,21	TIDAK TERCAPAI	NAIK
37	1.506	34.395.422	857	19.946.262	649	14.449.160	57,99	TIDAK TERCAPAI	NAIK
38	5.664	219.955.560	2.577	108.554.531	3.087	111.401.029	49,35	TIDAK TERCAPAI	TURUN
39	1.358	110.809.551	785	72.628.801	573	38.180.750	65,54	TERCAPAI	NAIK

Sheet1 Sheet2 Sheet3

3.2 Data transformation

This study procedures carried out as in figure 3.1, namely, the data obtained from the database of the UN tax revenue Deli Serdang. Data will be modified. Data in the form of Excel 2016 spreadsheet files (.xls) as input to the Weka open source software. Before the data is transformed into ARFF, the data is converted first into the .csv format. Weka transform data from .csv be ARFF. The result of the transformation is preliminary data that will be used for optimization prosesn with k values. The results of the data transformation xls, csv, ARFF can be seen in the picture below.

IOP Conf. Series: Materials Science and Engineering **725** (2020) 012133 doi:10.1088/1757-899X/725/1/012133

Table 2. The data as a .csv

	A
1	SPPT,TARGET,SPPT Y Bayar,Tercapai,SPPT Belum Bayar,Target Tak Capai,Penerimaan,Pencapaian,Kategori
2	997,"111,687,118",450,"90,128,985",547,"21,558,133",80.70,TERCAPAI,NAIK
3	673,"58,896,882",316,"49,427,433",357,"9,469,449",83.92,TERCAPAI,TURUN
4	2,108,"106,493,314",935,"72,531,398","1,173","33,961,916",68.11,TERCAPAI,NAIK
5	1,400,"57,434,795",535,"25,445,940",865,"31,988,855",44.30,TIDAK TERCAPAI,TURUN
6	498,"32,945,502",233,"18,709,761",265,"14,235,741",56.79,TIDAK TERCAPAI,TURUN
7	473,"52,273,697",175,"34,690,376",298,"17,583,321",66.36,TERCAPAI,TURUN
8	942,"30,941,684",524,"17,561,782",418,"13,379,902",56.76,TIDAK TERCAPAI,TURUN
9	718,"25,902,318",390,"13,759,250",328,"12,143,068",53.12,TIDAK TERCAPAI,NAIK
10	385,"26,294,309",169,"10,269,476",216,"16,024,833",39.06,TIDAK TERCAPAI,TURUN
11	336,"10,101,119",173,"5,799,489",163,"4,301,630",57.41,TIDAK TERCAPAI,TURUN
12	991,"70,296,473",414,"49,007,088",577,"21,289,385",69.71,TERCAPAI,TURUN
13	941,"156,416,754",489,"138,097,411",452,"18,319,343",88.29,TERCAPAI,TURUN
14	1,111,"147,920,197",588,"106,572,381",523,"41,347,816",72.05,TERCAPAI,NAIK
15	2,903,"144,972,498","1,420","79,203,686","1,483","65,768,812",54.63,TIDAK TERCAPAI,NAIK
16	288,"44,674,689",169,"21,339,751",119,"23,334,938",47.77,TIDAK TERCAPAI,TURUN
17	2,725,"329,456,622","1,537","204,438,210","1,188","125,018,412",62.05,TERCAPAI,NAIK
18	1,"3,166,668",1,"3,166,668",0,0,100.00,TERCAPAI,TETAP
19	344,"23,859,036",223,"16,708,075",121,"7,150,961",70.03,TERCAPAI,TURUN
20	1,196,"43,499,255",780,"31,458,923",416,"12,040,332",72.32,TERCAPAI,TURUN
21	1,482,"51,070,436",387,"16,191,750","1,095","34,878,686",31.70,TIDAK TERCAPAI,TURUN
22	1,658,"70,966,662",856,"42,877,842",802,"28,088,820",60.42,TERCAPAI,TURUN
23	507,"22,847,146",207,"10,553,302",300,"12,293,844",46.19,TIDAK TERCAPAI,NAIK
24	1,974,"117,421,573",530,"57,231,401","1,444","60,190,172",48.74,TIDAK TERCAPAI,NAIK
25	1,459,"93,092,242",538,"36,377,251",921,"56,714,991",39.08,TIDAK TERCAPAI,TURUN
26	437,"23,395,822",196,"12,469,996",241,"10,925,826",53.30,TIDAK TERCAPAI,NAIK
27	873,"24,245,854",574,"17,610,111",299,"6,635,743",72.63,TERCAPAI,NAIK
28	1,433,"42,586,720",820,"26,287,982",613,"16,298,738",61.73,TERCAPAI,NAIK
29	725,"20,110,320",381,"9,477,328",344,"10,632,992",47.13,TIDAK TERCAPAI,TURUN
30	1,367,"40,779,994",862,"26,035,516",505,"14,744,478",63.84,TERCAPAI,NAIK
31	2,469,"63,490,805","1,879","49,900,323",590,"13,590,482",78.59,TERCAPAI,TURUN
32	693,"61,402,741",399,"47,953,448",294,"13,449,293",78.10,TERCAPAI,NAIK
33	874,"33,368,374",443,"17,998,100",431,"15,370,274",53.94,TIDAK TERCAPAI,NAIK
34	1,099,"111,948,767",328,"53,875,448",771,"58,073,319",48.13,TIDAK TERCAPAI,NAIK
35	1,543,"91,571,016",795,"54,217,988",748,"37,353,028",59.21,TIDAK TERCAPAI,NAIK
36	1,506,"34,395,422",857,"19,946,262",649,"14,449,160",57.99,TIDAK TERCAPAI,NAIK
37	5,664,"219,955,560","2,577","108,554,531","3,087","111,401,029",49.35,TIDAK TERCAPAI,TURUN
38	1,358,"110,809,551",785,"72,628,801",573,"38,180,750",65.54,TERCAPAI,NAIK
39	1,387,"96,585,740",578,"50,338,175",809,"46,247,565",52.12,TIDAK TERCAPAI,NAIK

pajak

Table 3. The data in ARFF format

pajak.arff

Relation: pajak

No.	1: SPPT Nominal	2: TARGET Nominal	3: SPPT Y Bayar Nominal	4: Tercapai Nominal	5: SPPT Belum Bayar Nominal	6: Target Tak Capai Nominal	7: Penerimaan Numeric	8: Pencapaian Nominal	9: **Kategori** Nominal
1	997.0	111,687...	450.0	90,128,...	547.0	21,558,133	80.7	TERCAPAI	NAIK
2	673.0	58,896,...	316.0	49,427,...	357.0	9,469,449	83.92	TERCAPAI	TURUN
3	2,108	106,493...	935.0	72,531,...	1,173	33,961,916	68.11	TERCAPAI	NAIK
4	1,400	57,434,...	535.0	25,445,...	865.0	31,988,855	44.3	TIDAK TER...	TURUN
5	498.0	32,945,...	233.0	18,709,...	265.0	14,235,741	56.79	TIDAK TER...	TURUN
6	473.0	52,273,...	175.0	34,690,...	298.0	17,583,321	66.36	TERCAPAI	TURUN
7	942.0	30,941,...	524.0	17,561,...	418.0	13,379,902	56.76	TIDAK TER...	TURUN
8	718.0	25,902,...	390.0	13,759,...	328.0	12,143,068	53.12	TIDAK TER...	NAIK
9	385.0	26,294,...	169.0	10,269,...	216.0	16,024,833	39.06	TIDAK TER...	TURUN
10	336.0	10,101,...	173.0	5,799,489	163.0	4,301,630	57.41	TIDAK TER...	TURUN
11	991.0	70,296,...	414.0	49,007,...	577.0	21,289,385	69.71	TERCAPAI	TURUN
12	941.0	156,416...	489.0	138,097...	452.0	18,319,343	88.29	TERCAPAI	TURUN
13	1,111	147,920...	588.0	106,572...	523.0	41,347,816	72.05	TERCAPAI	NAIK
14	2,903	144,972...	1,420	79,203,...	1,483	65,768,812	54.63	TIDAK TER...	NAIK
15	288.0	44,674,...	169.0	21,339,...	119.0	23,334,938	47.77	TIDAK TER...	TURUN
16	2,725	329,456...	1,537	204,438...	1,188	125,018,412	62.05	TERCAPAI	NAIK
17	1.0	3,166,668	1.0	3,166,668	0.0	0.0	100.0	TERCAPAI	TETAP
18	344.0	23,859,...	223.0	16,708,...	121.0	7,150,961	70.03	TERCAPAI	TURUN
19	1,196	43,499,...	780.0	31,458,...	416.0	12,040,332	72.32	TERCAPAI	TURUN
20	1,482	51,070,...	387.0	16,191,...	1,095	34,878,686	31.7	TIDAK TER...	TURUN
21	1,658	70,966,...	856.0	42,877,...	802.0	28,088,820	60.42	TERCAPAI	TURUN
22	507.0	22,847,...	207.0	10,553,...	300.0	12,293,844	46.19	TIDAK TER...	NAIK
23	1,974	117,421...	530.0	57,231,...	1,444	60,190,172	48.74	TIDAK TER...	NAIK
24	1,459	93,092,...	538.0	36,377,...	921.0	56,714,991	39.08	TIDAK TER...	TURUN
25	437.0	23,395,...	196.0	12,469,...	241.0	10,925,826	53.3	TIDAK TER...	NAIK
26	873.0	24,245,...	574.0	17,610,...	299.0	6,635,743	72.63	TERCAPAI	NAIK
27	1,433	42,586,...	820.0	26,287,...	613.0	16,298,738	61.73	TERCAPAI	NAIK
28	725.0	20,110,...	381.0	9,477,328	344.0	10,632,992	47.13	TIDAK TER...	TURUN
29	1,367	40,779,...	862.0	26,035,...	505.0	14,744,478	63.84	TERCAPAI	NAIK
30	2,469	63,490,...	1,879	49,900,...	590.0	13,590,482	78.59	TERCAPAI	TURUN
31	693.0	61,402,...	399.0	47,953,...	294.0	13,449,293	78.1	TERCAPAI	NAIK
32	874.0	33,368,...	443.0	17,998,...	431.0	15,370,274	53.94	TIDAK TER...	NAIK
33	1,099	111,948...	328.0	53,875,...	771.0	58,073,319	48.13	TIDAK TER...	NAIK
34	1,543	91,571,...	795.0	54,217,...	748.0	37,353,028	59.21	TIDAK TER...	NAIK
35	1,506	34,395,...	857.0	19,946,...	649.0	14,449,160	57.99	TIDAK TER...	NAIK
36	5,664	219,955...	2,577	108,554...	3,087	111,401,029	49.35	TIDAK TER...	TURUN
37	1,358	110,809...	785.0	72,628,...	573.0	38,180,750	65.54	TERCAPAI	NAIK
38	1,387	96,585,...	578.0	50,338,...	809.0	46,247,565	52.12	TIDAK TER...	NAIK
39	622.0	21,117,...	449.0	16,045,...	173.0	5,071,953	75.98	TERCAPAI	NAIK

3.3 Optimization Rated K

k-Nearest Neightbor (KNN) is a method using supervised algorithms where the results of the new query instance is classified based on the majority of categories on KNN. The purpose of this algorithm is to classify a new object attributes and training Based on the sample. Classifier does not use any model to be matched and only based on memory. Given query point, will find a number of objects or K (training points) closest to the query point.

3.4 Expectation Maximization Clustering

Expectation maximization algorithm is an algorithm unsupservised learning that has the ability to perform searches darisekumpulan knowledge of data that do not have labels or targets a particular class, by seeingthe value of any instances distributed into the Gaussian distribution, more tepatnyaadalah Gaussian mixture, then do iterations ascending to seek the highest likelihood value for each instance (see proximity to each cluster instances). Expectation Maximization algorithm (EM algorithm) is an algorithm that utilizes the mixture of Gaussian mixture.

Basically EM algorithm consists of two steps, ie, expectation and maximization. Calculating expektasi to a likelihood probability value, then the second step of fixing the value of the probability of the stretcher by changing parameters on Gaussian mixture so as to achieve maximum likelihood. There some things that need to be emphasized in the EM algorithm Algorithm namely:

1. Maximum Likelihood Estimation (MLE)
2. Mixtures of Gaussians
3. Estimation-Maximization (EM)

But the EM algorithm using Gaussian mixture or words of a Gaussian lainlebih used or seeking mixture of yangdidapatkan distribution. EM Algorithm has the task of finding each Gaussian yangterdapat on Gaussian mixture distribution and develop each Gaussian yangditemukan at the optimum condition (so the model is more fit) that's called maximization, and the clustering process.

3.5. Interpretation / Evaluation

At this stage of the evaluation and interpretation of the patterns obtained based on the results of clustering data using EM-cluster method. If the results obtained are not appropriate, then the process would be repeated to the stage of the clustering process data. Knowledge of this stage is the final part of the KDD process where possible to investigate whether a pattern or information found in conflict with the facts. Pattern information generated from the data mining process should be presented in a form easily understood by the parties concerned.

4. Result and Discussion

Furthermore, from the data of the parameter with a k-nn algorithm, the data in the pull to get the Weka application cluster also using two parts of the cluster with no parameters and cluster k-nn-nn with parameter k. Both parts are in the cluster by using an algorithm *expectation Maximization* (EM). This algorithm is already available in Weka and can be directly used. The output from these two different parts and will be compared. For the results of the cluster with no parameters can be viewed as in Table 4.

IOP Conf. Series: Materials Science and Engineering **725** (2020) 012133 doi:10.1088/1757-899X/725/1/012133

Table 4. Cluster results with the original data

hs-1.arff hs-2.arff

Relation: pajak_clustered

No.	1: Instance_number	2: SPPT	3: TARGET	4: SPPT Y Bayar	5: Tercapai	6: SPPT Belum Bayar	7: Target Tak Capai	8: Penerimaan	9: Pencapaian	10: Kategori	11: Cluster
	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Nominal	Nominal	Nominal
1	0.0	997.0	111,687...	450.0	90,128,...	547.0	21,558,133	80.7	TERCAPAI	NAIK	cluster4
2	1.0	673.0	58,896,...	316.0	49,427,...	357.0	9,469,449	83.92	TERCAPAI	TURUN	cluster4
3	2.0	2,108	106,493...	935.0	72,531,...	1,173	33,961,916	68.11	TERCAPAI	NAIK	cluster10
4	3.0	1,400	57,434,...	535.0	25,445,...	865.0	31,988,855	44.3	TIDAK TER...	TURUN	cluster2
5	4.0	498.0	32,945,...	233.0	18,709,...	265.0	14,235,741	56.79	TIDAK TER...	TURUN	cluster3
6	5.0	473.0	52,273,...	175.0	34,690,...	298.0	17,583,321	66.36	TERCAPAI	TURUN	cluster10
7	6.0	942.0	30,941,...	524.0	17,561,...	418.0	13,379,902	56.76	TIDAK TER...	TURUN	cluster3
8	7.0	718.0	25,902,...	390.0	13,759,...	328.0	12,143,068	53.12	TIDAK TER...	NAIK	cluster0
9	8.0	385.0	26,294,...	169.0	10,269,...	216.0	16,024,833	39.06	TIDAK TER...	TURUN	cluster2
10	9.0	336.0	10,101,...	173.0	5,799,489	163.0	4,301,630	57.41	TIDAK TER...	TURUN	cluster3
11	10.0	991.0	70,296,...	414.0	49,007,...	577.0	21,289,385	69.71	TERCAPAI	TURUN	cluster10
12	11.0	941.0	156,416...	489.0	138,097...	452.0	18,319,343	88.29	TERCAPAI	TURUN	cluster7
13	12.0	1,111	147,920...	588.0	106,572...	523.0	41,347,816	72.05	TERCAPAI	NAIK	cluster9
14	13.0	2,903	144,972...	1,420	79,203,...	1,483	65,768,812	54.63	TIDAK TER...	NAIK	cluster0
15	14.0	288.0	44,674,...	169.0	21,339,...	119.0	23,334,938	47.77	TIDAK TER...	TURUN	cluster6
16	15.0	2,725	329,456...	1,537	204,438...	1,188	125,018,412	62.05	TERCAPAI	NAIK	cluster10
17	16.0	1.0	3,166,668	1.0	3,166,668	0.0	0.0	100.0	TERCAPAI	TETAP	cluster1
18	17.0	344.0	23,859,...	223.0	16,708,...	121.0	7,150,961	70.03	TERCAPAI	TURUN	cluster10
19	18.0	1,196	43,499,...	780.0	31,458,...	416.0	12,040,332	72.32	TERCAPAI	TURUN	cluster9
20	19.0	1,482	51,070,...	387.0	16,191,...	1,095	34,878,686	31.7	TIDAK TER...	TURUN	cluster2
21	20.0	1,658	70,966,...	856.0	42,877,...	802.0	28,088,820	60.42	TERCAPAI	TURUN	cluster10
22	21.0	507.0	22,847,...	207.0	10,553,...	300.0	12,293,844	46.19	TIDAK TER...	NAIK	cluster6
23	22.0	1,974	117,421...	530.0	57,231,...	1,444	60,190,172	48.74	TIDAK TER...	NAIK	cluster6
24	23.0	1,459	93,092,...	538.0	36,377,...	921.0	56,714,991	39.08	TIDAK TER...	TURUN	cluster2
25	24.0	437.0	23,395,...	196.0	12,469,...	241.0	10,925,826	53.3	TIDAK TER...	NAIK	cluster0
26	25.0	873.0	24,245,...	574.0	17,610,...	299.0	6,635,743	72.63	TERCAPAI	NAIK	cluster9
27	26.0	1,433	42,586,...	820.0	26,287,...	613.0	16,298,738	61.73	TERCAPAI	NAIK	cluster10
28	27.0	725.0	20,110,...	381.0	9,477,328	344.0	10,632,992	47.13	TIDAK TER...	TURUN	cluster6
29	28.0	1,367	40,779,...	862.0	26,035,...	505.0	14,744,478	63.84	TERCAPAI	NAIK	cluster10
30	29.0	2,469	63,490,...	1,879	49,900,...	590.0	13,590,482	78.59	TERCAPAI	TURUN	cluster4
31	30.0	693.0	61,402,...	399.0	47,953,...	294.0	13,449,293	78.1	TERCAPAI	NAIK	cluster9
32	31.0	874.0	33,368,...	443.0	17,998,...	431.0	15,370,274	53.94	TIDAK TER...	NAIK	cluster0
33	32.0	1,099	111,948...	328.0	53,875,...	771.0	58,073,319	48.13	TIDAK TER...	NAIK	cluster6
34	33.0	1,543	91,571,...	795.0	54,217,...	748.0	37,353,028	59.21	TIDAK TER...	NAIK	cluster3
35	34.0	1,506	34,395,...	857.0	19,946,...	649.0	14,449,160	57.99	TIDAK TER...	NAIK	cluster3
36	35.0	5,664	219,955...	2,577	108,554...	3,087	111,401,029	49.35	TIDAK TER...	TURUN	cluster6
37	36.0	1,358	110,809...	785.0	72,628,...	573.0	38,180,750	65.54	TERCAPAI	NAIK	cluster10
38	37.0	1,387	96,585,...	578.0	50,338,...	809.0	46,247,565	52.12	TIDAK TER...	NAIK	cluster0
39	38.0	622.0	21,117,...	449.0	16,045,...	173.0	5,071,953	75.98	TERCAPAI	NAIK	cluster9
40	39.0	89.0	4,324,718	53.0	2,542,244	36.0	1,782,474	58.78	TIDAK TER...	TURUN	cluster3
41	40.0	710.0	28,904,...	357.0	11,189,...	353.0	17,714,316	38.71	TIDAK TER...	TURUN	cluster2
42	41.0	3,366	83,397,...	2,346	57,677,...	1,020	25,719,730	69.16	TERCAPAI	NAIK	cluster10
43	42.0	149.0	4,018,434	118.0	2,919,124	31.0	1,099,310	72.64	TERCAPAI	TURUN	cluster9
44	43.0	155.0	4,619,470	142.0	4,383,164	13.0	236,306	94.88	TERCAPAI	TURUN	cluster8
45	44.0	2,077	72,942,...	1,300	47,117,...	777.0	25,824,935	64.6	TERCAPAI	NAIK	cluster10
46	45.0	2,312	168,610...	1,470	130,836...	842.0	37,774,270	77.6	TERCAPAI	NAIK	cluster9
47	46.0	836.0	28,197,...	644.0	22,471,...	192.0	5,725,451	79.69	TERCAPAI	NAIK	cluster4
48	47.0	3,591	112,869...	2,468	79,228,...	1,123	33,640,598	70.2	TERCAPAI	TURUN	cluster10
49	48.0	414.0	15,922,...	390.0	13,929,...	24.0	1,992,601	87.49	TERCAPAI	NAIK	cluster7
50	49.0	2,190	89,851,...	1,471	59,326,...	719.0	30,525,407	66.03	TERCAPAI	TURUN	cluster10
51	50.0	279.0	5,403,156	279.0	5,403,156	0.0	0.0	100.0	TERCAPAI	TETAP	cluster1
52	51.0	1,313	36,361,...	816.0	24,260,...	497.0	12,100,889	66.72	TERCAPAI	NAIK	cluster10
53	52.0	1,950	56,221,...	1,317	40,624,...	633.0	15,596,724	72.26	TERCAPAI	NAIK	cluster9
54	53.0	1,133	38,172,...	663.0	24,093,...	470.0	14,079,098	63.12	TERCAPAI	NAIK	cluster10
55	54.0	428.0	9,051,527	366.0	8,110,587	62.0	940,940	89.6	TERCAPAI	TURUN	cluster5
56	55.0	337.0	12,270,...	132.0	7,034,686	205.0	5,235,339	57.33	TIDAK TER...	NAIK	cluster3

Table 5. Results of Cluster Without Parameter K-NN

No.	cluster	Total Instant
1	0	15
2	1	31
3	2	10
4	3	14
5	4	12
6	5	6
7	6	19
8	7	11
9	8	14
10	9	25
11	10	48

IOP Conf. Series: Materials Science and Engineering **725** (2020) 012133 doi:10.1088/1757-899X/725/1/012133

The second phase of testing is done with the data optimization results using the k value of K Nearest Neighbor with cluster model validation is performed on the original data. When implemented generate data as in Table 6.

Table 6. The results of Cluster with parameter k

hs-1.arff hs-2.arff

Relation: pajak_predicted_clustered

No.	1: Instance_number	2: SPPT	3: TARGET	4: SPPT Y Bayar	5: Tercapai	6: SPPT Belum Bayar	7: Target Tak Capai	8: Penerimaan	9: Pencapaian	10: predictedKategori	11: Kategori	12: **Cluster**
	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Nominal	Nominal	Nominal	Nominal
1	0.0	4,731	140,089...	1,922	63,003,...	2,809	77,085,995	44.97	TIDAK TER...	TURUN	TURUN	cluster7
2	1.0	942.0	30,941,...	524.0	17,561,...	418.0	13,379,902	56.76	TIDAK TER...	NAIK	TURUN	cluster7
3	2.0	143.0	4,739,950	124.0	4,026,752	19.0	713,198	84.95	TERCAPAI	NAIK	TURUN	cluster8
4	3.0	725.0	20,110,...	381.0	9,477,328	344.0	10,632,992	47.13	TIDAK TER...	TURUN	TURUN	cluster7
5	4.0	2,190	89,851,...	1,471	59,326,...	719.0	30,525,407	66.03	TERCAPAI	NAIK	TURUN	cluster5
6	5.0	288.0	44,674,...	169.0	21,339,...	119.0	23,334,938	47.77	TIDAK TER...	TURUN	TURUN	cluster7
7	6.0	3,075	173,122...	1,591	84,942,...	1,484	88,179,478	49.07	TIDAK TER...	NAIK	TURUN	cluster7
8	7.0	2,469	63,490,...	1,879	49,900,...	590.0	13,590,482	78.59	TERCAPAI	NAIK	TURUN	cluster5
9	8.0	1,543	91,571,...	795.0	54,217,...	748.0	37,353,028	59.21	TIDAK TER...	NAIK	NAIK	cluster2
10	9.0	213,...	9,776,2...	126,704	6,331,9...	87,290	3,444,314,683	64.77	TERCAPAI	NAIK	NAIK	cluster4
11	10.0	521.0	17,866,...	353.0	11,994,...	168.0	5,871,261	67.14	TERCAPAI	NAIK	NAIK	cluster4
12	11.0	1,974	117,421...	530.0	57,231,...	1,444	60,190,172	48.74	TIDAK TER...	NAIK	NAIK	cluster1
13	12.0	634.0	31,770,...	259.0	16,105,...	375.0	15,665,036	50.69	TIDAK TER...	NAIK	NAIK	cluster2
14	13.0	21.0	2,596,911	19.0	2,428,466	2.0	168,445	93.51	TERCAPAI	NAIK	NAIK	cluster0
15	14.0	981.0	36,193,...	751.0	28,903,...	230.0	7,289,873	79.86	TERCAPAI	NAIK	NAIK	cluster3
16	15.0	2,725	329,456...	1,537	204,438...	1,188	125,018,412	62.05	TERCAPAI	NAIK	NAIK	cluster4
17	16.0	177.0	4,969,121	177.0	4,969,121	0.0	0.0	100.0	TERCAPAI	TETAP	NAIK	cluster6
18	17.0	414.0	15,922,...	390.0	13,929,...	24.0	1,992,601	87.49	TERCAPAI	NAIK	NAIK	cluster0
19	18.0	1,358	110,809...	785.0	72,628,...	573.0	38,180,750	65.54	TERCAPAI	NAIK	NAIK	cluster4
20	19.0	1.0	755,400	1.0	755,400	0.0	0.0	100.0	TERCAPAI	TETAP	TETAP	cluster6
21	20.0	222.0	4,624,147	222.0	4,624,147	0.0	0.0	100.0	TERCAPAI	TETAP	TETAP	cluster6
22	21.0	1,400	57,434,...	535.0	25,445,...	865.0	31,988,855	44.3	TIDAK TER...	TURUN	TURUN	cluster7
23	22.0	428.0	9,051,527	366.0	8,110,587	62.0	940,940	89.6	TERCAPAI	NAIK	TURUN	cluster8
24	23.0	2,249	148,257...	1,246	77,602,...	1,003	70,654,635	52.34	TIDAK TER...	NAIK	TURUN	cluster7
25	24.0	155.0	4,619,470	142.0	4,383,164	13.0	236,306	94.88	TERCAPAI	NAIK	TURUN	cluster8
26	25.0	589.0	37,510,...	572.0	36,421,...	17.0	1,089,038	97.1	TERCAPAI	TETAP	TURUN	cluster8
27	26.0	718.0	33,115,...	416.0	17,479,...	302.0	15,636,177	52.78	TIDAK TER...	NAIK	TURUN	cluster7
28	27.0	2,836	55,243,...	1,349	25,564,...	1,487	29,679,266	46.28	TIDAK TER...	TURUN	TURUN	cluster7
29	28.0	1,334	30,985,...	1,120	27,137,...	214.0	3,847,585	87.58	TERCAPAI	NAIK	NAIK	cluster0
30	29.0	327.0	15,852,...	281.0	13,062,...	46.0	2,790,167	82.4	TERCAPAI	NAIK	NAIK	cluster0
31	30.0	1,379	40,147,...	926.0	28,126,...	453.0	12,021,090	70.06	TERCAPAI	NAIK	NAIK	cluster3
32	31.0	556.0	31,030,...	324.0	20,404,...	232.0	10,625,664	65.76	TERCAPAI	NAIK	NAIK	cluster3
33	32.0	6,088	148,995...	4,193	99,970,...	1,895	49,024,602	67.1	TERCAPAI	NAIK	NAIK	cluster4
34	33.0	594.0	49,768,...	410.0	34,328,...	184.0	15,439,676	68.98	TERCAPAI	NAIK	NAIK	cluster4
35	34.0	876.0	23,229,...	552.0	14,428,...	324.0	8,801,052	62.11	TERCAPAI	NAIK	NAIK	cluster4
36	35.0	405.0	17,432,...	405.0	17,432,...	0.0	0.0	100.0	TERCAPAI	TETAP	NAIK	cluster6
37	36.0	337.0	12,270,...	132.0	7,034,686	205.0	5,235,339	57.33	TIDAK TER...	NAIK	NAIK	cluster2
38	37.0	635.0	16,715,...	367.0	9,730,674	268.0	6,984,362	58.22	TIDAK TER...	NAIK	NAIK	cluster2
39	38.0	573.0	44,187,...	402.0	28,474,...	171.0	15,713,208	64.44	TERCAPAI	NAIK	NAIK	cluster4
40	39.0	897.0	23,391,...	656.0	16,266,...	241.0	7,124,626	69.54	TERCAPAI	NAIK	NAIK	cluster3
41	40.0	141.0	3,360,665	141.0	3,360,665	0.0	0.0	100.0	TERCAPAI	TETAP	TETAP	cluster6
42	41.0	11.0	2,205,752	11.0	2,205,752	0.0	0.0	100.0	TERCAPAI	TETAP	TETAP	cluster6
43	42.0	1,196	43,499,...	780.0	31,458,...	416.0	12,040,332	72.32	TERCAPAI	NAIK	TURUN	cluster5
44	43.0	1,063	156,487...	645.0	133,343...	418.0	23,144,419	85.21	TERCAPAI	NAIK	TURUN	cluster8
45	44.0	991.0	70,296,...	414.0	49,007,...	577.0	21,289,385	69.71	TERCAPAI	NAIK	TURUN	cluster5
46	45.0	619.0	10,835,...	442.0	6,808,120	177.0	4,027,637	62.83	TERCAPAI	NAIK	TURUN	cluster5
47	46.0	2,194	81,285,...	841.0	41,036,...	1,353	40,249,517	50.48	TIDAK TER...	NAIK	TURUN	cluster7
48	47.0	989.0	28,633,...	486.0	13,375,...	503.0	15,258,266	46.71	TIDAK TER...	TURUN	TURUN	cluster7
49	48.0	675.0	44,992,...	357.0	31,046,...	318.0	13,945,454	69.0	TERCAPAI	NAIK	TURUN	cluster5
50	49.0	298.0	20,600,...	149.0	12,600,...	149.0	7,999,791	61.17	TERCAPAI	NAIK	NAIK	cluster4
51	50.0	1,133	38,172,...	663.0	24,093,...	470.0	14,079,098	63.12	TERCAPAI	NAIK	NAIK	cluster4
52	51.0	1,169	24,327,...	653.0	13,834,...	516.0	10,492,953	56.87	TIDAK TER...	NAIK	NAIK	cluster2
53	52.0	1,387	96,585,...	578.0	50,338,...	809.0	46,247,565	52.12	TIDAK TER...	NAIK	NAIK	cluster2
54	53.0	160.0	4,986,255	109.0	3,342,965	51.0	1,643,290	67.04	TERCAPAI	NAIK	NAIK	cluster4

Table 7. Results of Cluster Parameters K-NN

No.	cluster	Total Instant
1	0	26
2	1	10
3	2	16
4	3	22
5	4	28
6	5	24
7	6	32
8	7	31
9	8	16

4.1. Influence Selection of Parameter Values k

In the test will be analyzed the effect of optimization parameters k value the success rate with algorithms clusterexpectation *Maximization*, The k value is the number of nearest neighbors for use as consideration in determining the number of cluster decision.

Distance parameter used to optimize the use of simulation data that euclidean distance and the Hamming distance, while the value of k used is k = 13. Based on the above data processing results, when using early data without any additional parameters obtained by the

IOP Conf. Series: Materials Science and Engineering **725** (2020) 012133 doi:10.1088/1757-899X/725/1/012133

number of clusters found and incorrect as many as 11 clusters of 66% then when using the optimization parameters obtained by the number of clusters k sebnayak 9 and can minimize incorrect cluster to 64%.

5. Conclusion

By using clustering algorithms can mengdentifikasi Cluster EM-attainment status and budget plans in the coming year. In the process of this grouping K-NN with k = 13 an algorithm and can be used for the type of data berimensi high. Determination of parameter k in K-NN algorithm can affect and improve the number of clusters in advance.

References

[1] Connolly, Thomas, C. B 2010, Database Systems: A Practical Approach to Design, Implemantation, and Management Fifth Edition: Pearson Education Inc.

[2] Cuzzocrea, Alfredo 2011, warehousing Data and Knowledge Discovery, Springer-Verlag Berlin Heidelberg, London.

[3] Hermawati, Fajar Astuti, 2013. Data Mining. ANDI Publisher: Yogyakarta.

[4] Kimball, R, Margy R, Warren T, Joy M and Bob B 2008, The Data Warehouse Lifecycle Toolkit, Wiley Publishing Inc., Canada.

[5] Indrajani 2009 Database System In Package Five In One, PT.Elex Media Komputindo, Jakarta.

[6] Ponniah, Paulraj 2010, Data warehouseing. Canada: John Wiley & Sons Inc.

[7] Tantra, Rudi 2012, Project Management Information Systems, Andi Offset, Yogyakarta.

[8] Jiawei Han and Micheline Kambar 2010, Data Mining Concepts and Techniques, Verlag Berlin.

[9] Oded Maimon and Lior Rokach, 2010, Data Mining And Knowledge Discovery Hanbook, Springer Science.

[10] Ernastuti, S. &. (2010). Graduation Prediction of Gunadarma University Students Using Algorithm and C4.5 Naive Bayes Algoritmh.

[11] Gunadi, G., Sensuse, D., I., 2012, Application of Data Mining Methods Market Basket Analysis to book the product sales data by using algorithms Apriori and Frequent Pattern Growth (FP-Growth), MKom TELEMATIKA Journal, Vol. 4, No. 1, 118-132.

[12] Gorunescu, F. (2011). Data Mining Concepts and Techniques Models. Craiova: Springer.

[13] Hastuti, K. (2012, June). ANALYSIS COMPARISON OF CLASSIFICATION OF DATA MINING ALGORITHM National V.Seminar Applied Information & Communication Technology (979 - 26 - 0255-0), 241 249.

[14] Ian H. Witten, f. E. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). (ASBurlington, Ed.) United States of America: Morgan Kaufmann.

[15] Kalyankar, Q. &. (2010). Drop Out Feature of Student Performance Data Using Decision Tree forAcademic techniques. Global Journal of Computer Science and Technology, 2-4.

[16] Mardiani, 2014, Comparison Algorithm K-Means and EM for Clusterisasi Value Based Home School Students, CITEC Journal, Vol. 1, No. 4, 316-325

[17] Oyelade, O. &. (2010). Application of Kmeans Clustering algorithm for predicting of Students AcademicPerformace. International Journal of Computer Science and Information Security, 292-295.

[18] Syaifullah. (2010). Implementation of Data Mining Algorithm Apriori Sales System, Amikom, Yogyakarta.
[19] Tahyudin, I. (2013, December). Comparing Classification Of Data Mining Algorithm to Predict the Graduation Students on Time. Information Systems International Conference (ISICO).
[20] Vercellis. (2009). Business Intelligence: Data Mining and Optimization for Decision Making Decision Making. John Wiley & Sons Inc: Southern Gate.
[21] Vrettos, K. &. (2009). Sentivity Analysis of Neural Network for Identifying the Factors for Success College Students. World Congress on Computer Science and Information Engineering. (978-0-7695-3507-4).
[22] Yanto. R, Khoiriah. R., 2015, Implementation of Data Mining with Apriori Algorithm Method in Determining Drug Purchasing Patterns, CITEC Journal, Vol. 2, No. 2, 101-113.