# Analysis of K-Means and K-Medoids's Performance Using Big Data Technology

[(1)]Nurhayati, [(2)] Nadika Sigit Sinatrya, [(3)]Luh Kesuma Wardhani, [(4)]Busman
[(1)(2)(3)] Department of Informatics, Syarif Hidayatullah State Islamic University Jakarta
[(4)] Department of Management, School of Economics Gotong Royong Jakarta
nurhayati@uinjkt.ac.id, nadikatrya@gmail.com, luhkesuma@uinjkt.ac.id, busman.sjam@gmail.com

***Abstract- This research's goal is to find out the better performance algorithm between K-Means and K-Medoids algorithm. The performance of both algorithm are compared by testing data using Java-based application, Hadoop, and Hive. Comparison was conducted in terms of accuracy, execution time and time complexity of the algorithm. In terms of accuracy, K-Medoids is better than K-Means with an average accuracy of 63.24%, while K-Means is 52.11%. In terms of execution time, K-Medoids also has better performance with average speed of 3.1 ms, while K-Means is 3.45 ms. In terms of time complexity algorithms, both algorithms have the result of O ($n^2$). K-Medoids has better performance than K-Means, which K-Medoids has an average value of 310.157, while K-Means has greater value than K-Medoids of 377,886. So the K-Medoids algorithm is superior to K-Means in terms of accuracy, execution time and time complexity.***

***Keywords— Analysis, K-Means, K-Medoids, Big Data Technology, Java***

## I. INTRODUCTION

The rapid growth of data accumulation has created a rich state of data but minimal information [1]. The desired information can not be obtained easily due to large volumes of data. So it takes a method to get knowledge that is not visible in the data but it is potential to use is the method of data mining [2]. This technology called big data. Big data allow us to manage, collect and analyize data in various format, large amount, and grow rapidly.

The large amount of data were collected from any resource, then analyze it to find answer. Using big data technology, user can reduce cost and time, optimized and developed new product and has smart decision making. One of technology that help user to gain knowledge among large set of data is called data mining.

Data Mining is a process of extracting information from a very large set of data through the use of algorithms and withdrawal techniques in the field of statistics, machine learning and database management systems [3].

One technique known in data mining is clustering techniques [4]. Clustering is a method of grouping data. Clustering is the process of partitioning a set of data objects into subsets called clusters. Objects within the cluster have similar characteristics between each other and are different from other clusters. Clustering is widely used in various applications such as business intelligence, image pattern recognition, web search, biology, and security [5].

According to reference [6], machine learning (ML) is a variant of the artificial intelligence system that allows computers to learn without programmed explicitly. In general, ML jobs that are often used are to classify one problem into several groups. In everyday life, objects can be easily identified by humans, but may not necessarily be explained specifically. This is where the role of Machine Learning in recognizing, identifying, or predicting certain data by studying historical data. With ML, the model is created either directly or indirectly, by extracting knowledge from experts or from data that has not even been known to do with the way to learn it with certain algorithms [7].

Machine learning has two techniques: supervised learning and unsupervised learning. The practical majority of machine learning uses supervised learning [8]. Supervised learning is one type of machine learning algorithm that uses a known dataset (training dataset) to make predictions. Unsupervised learning is one type of machine learning algorithm used to draw conclusions from datasets consisting of labeled response data input. The most common unsupervised learning method is cluster analysis, which is used in data analysis to look for hidden patterns or groupings in data [9].

Several algorithms used by unsupervised learning method are K-Means and K-Medoids. The K-Means algorithm is a well-known partitioning method for clustering [10]. K-Means is a non-hierarchy data clustering method that attempts to partition existing data into one or more clusters or groups so that data that have the same characteristics are grouped into the same cluster and data that have different characteristics are grouped into the other group [11].

While K-Medoids or Partitioning around Medoids (PAM) according to [12] is a clustering algorithm similar to K-Means. The difference between these two algorithms are the K-Medoids or PAM algorithm using the object as the representative (medoid) as the center of the cluster for each cluster, while K-Means uses the mean value as the center of the cluster [13].

Several past studies have compared the performance of K-Means and K-Medoids algorithms. In 2017, [14] analyzed the performance of K-Means and K-Medoids algorithms using UCI Machine Learning Repository which is a collection of databases that are often used by researchers in the field of machine learning, then implemented with the Java programming language. Both

K-Means and K-Medoids are scored based on their approach to large datasets. In the study, the performance of K-Medoids algorithm is superior to K-Means algorithm in terms of accuracy with accuracy of 92%, while K-Means has an accuracy of 88.7%.

While in 2011, [15] also performed an analysis on the performance of K-Means and K-Medoids algorithms. In this research, artificial data were generated and grouped using K-Means and K-Medoids. He brought up 3 classes with each class having 120 objects that have 2 variables. The result is K-Means has an advantage in terms of execution time than K-Medoids.

In [16] an analysis was conducted using the K-Means algorithm. Analysis was done using an application with Java programming language that was connected with Hadoop. The answer data from the .csv formatted questionnaire is stored into Hadoop and analyzed in the application. Then the result of K-Means clustering is compared with its manual value to get its accuracy value. The study yielded accuracy on the overall data analysis of 75.08%.

Based on previous studies and the above description, we were interested in doing research on analysis of the performance of K-Means and K-Medoids algorithm with big data technology. Data which was taken from [16] were testing using of K-Means and K-Medoids algorithm to observe those two algorithm in term of accuracy, execution time and complexity.

This paper is structured as follows. Section II describes some previous work that related to this paper. Section III describes the method in doing research. Section IV describes the result of this research. Finally section V presents our discussion and conclusions.

## II. RELATED WORK

Several past studies have compared the performance of K-Means and K-Medoids algorithms. The first related work is "*Comparative Analysis of K-Means and K-Medoids Algorithm on IRIS Data*" by Kalpit G. Soni. The author analyzed the performance of K-Means and K-Medoids algorithms using UCI Machine Learning Repository which is a collection of databases that are often used by researchers in the field of machine learning, then implemented with the Java programming language. Both K-Means and K-Medoids are scored based on their approach to large datasets. In the study, the performance of K-Medoids algorithm is superior to K-Means algorithm in terms of accuracy with accuracy of 92%, while K-Means has an accuracy of 88.7%. [14].

Aishwarya Batra in her research entitled "*Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms*" also performed an analysis on the performance of K-Means and K-Medoids algorithms. In this research, artificial data were generated and grouped using K-Means and K-Medoids. He brought up 3 classes with each class having 120 objects that have 2 variables. The result is K-Means has an advantage in terms of execution time than K-Medoids [15].

In Nurhayati's research (2017) entitled "Big Data Analysis Using Hadoop Framework and Machine Learning for Islam Mindset Monitoring student and lecturer as Decision Support System (DSS)", She performed an analysis using the K-Means algorithm. Analysis is done using an application with Java programming language that is connected with Hadoop. The answer data from the .csv formatted questionnaire is stored into Hadoop and analyzed in the application. Then the result of K-Means clustering is compared with its manual value to get its accuracy value. The study yielded accuracy on the overall data analysis of 75.08% [16].

## III. METHOD

There are several methods used in this research. Data used was taken from [16], The authors used the previous questionnaire data, thus the data came from 1236 respondents (Students and Lecturers of state Islamic university of Jakarta).

The method used in this research is the simulation method, the following are the steps of the simulation method.

1. *Problem Formulation,* identify problem from previous research resulta.
2. *Conceptual Model,* discusses this entire research.
3. *Collection of I/O Data,* questionnaires data used as input applications to produce the output of tables, accuracy, graphics, and algorithm complexity.
4. *Modeling Phase,* modeling K-Means and K-Medoids algorithms by calculating multiple sample data, accuracy, and Big O manually.
5. *Simulation Phase,* perform simulation features on K-Means K-Medoids application.
6. *Conclusion (verification, validation, and experimentation),* verify, validate, experiment to make sure the app matches the concept.
7. *Output Analysis Phase,* performs an analysis of the output based on filtering data scenarios in terms of accuracy, execution time, and Big O.

As mentioned earlier, in modelling phase K-Means and K-Medoids algorithm was used to analyze large data set. K-Means Clustering is a method of analyzing data or data mining methods that perform unsupervised data modelling. K-Means is a method of grouping existing data into groups, where data in one group have the same characteristics with each other and have different data from other groups [17].

Meanwhile, K-Medoids is a classical partitioning technique for clustering data. K-Medoids can be defined as cluster objects, which averages the difference for all objects in a minimal cluster that is the most central point of the given data. The K-Medoids algorithm has the advantage of overcoming weaknesses in the K-Means algorithm which is sensitive to noise and outlier, where objects with large values that allow to deviate from from

the data distribution. Another advantages is the result of the clustering process is not dependent on the incoming sequence of the dataset [12].

## IV. RESULT

### A. Analysis of data output

To analyze the results of accuracy on K-Means Kmedoids Application done by analysis on the results of clustering graphs first. The results of clustering graphs are the overall results of K-Means and K-Medoids, as well as the results for which data are filtered based on existing scenarios. After that the author compared both accuracy.

The overall output on the results of analysis using K-Means and K-Medoids is shown in Figures 1, 2, and 3.

Figure 1 shows the results of K-Means and K-Medoids clustering of the entire data on first iteration. In the K-Means results (left graph) has a central value of cluster C1 which is 925 on the x line (do not understand history) and 400 on the y line (understand history), and the center value of cluster C2 is 825 on the x line (do not understand history) and 525 on the line y (understand history). There are 671 respondents in the understand History cluster, and 565 respondents in the cluster Do not understand History. In the first iteration there is no changes in K-Means cluster members, so the iteration is continued.

In the K-Medoids results in Figure 1 (right graph) has a central value of cluster C1 which is 950 on the x line (do not know history) and 700 on the y line (know history), and the center of cluster C2 is 925 on the x line (do not know history) and 825 on the y line (know history). The results of K-Medoids clustering is different from the results of K-Means clustering, where there are 112 respondents in the Know History cluster, and 1124 respondents in the cluster Do not Know History. At the first iteration the total cost is 335707.72 and iteration is continued.

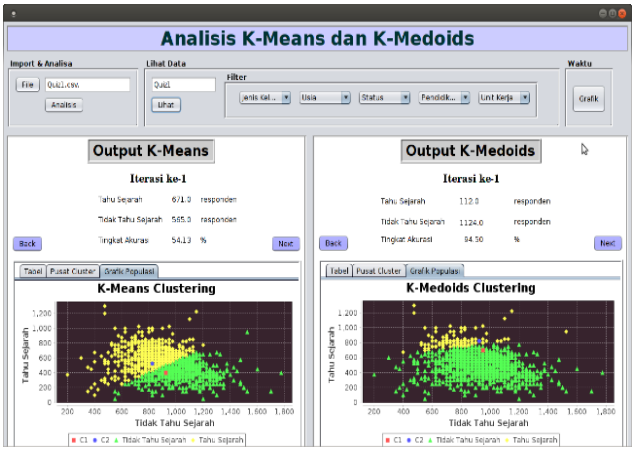


Figure 1. The overall output on the first iteration

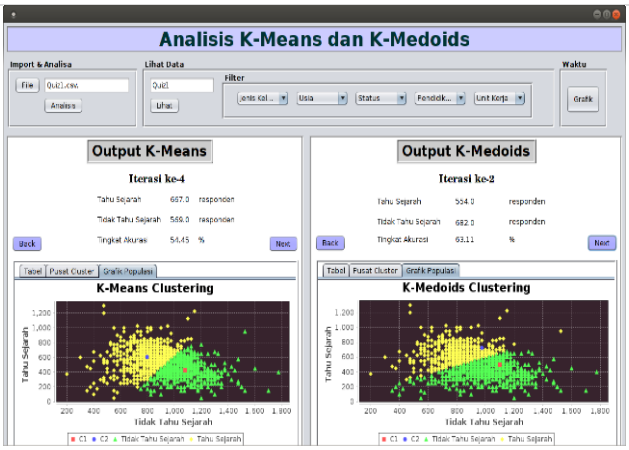


Figure 2. Overall output on the 4th iteration of K-Means and 2nd iteration K-Medoids

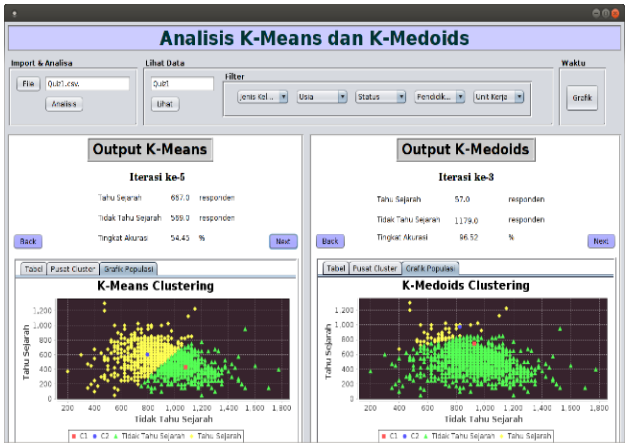


Figure 3. Overall output on the last iteration

In the 2nd to 4th iteration of K-Means clustering, cluster members are still changing. Figure 2 shows the overall data clustering results for the 4th iteration of K-Means and the 2nd iteration of K-Medoids. In the K-Means results (left graph) has a central value of cluster C1 of 1079.5856 on the x line and 429.67374 on the y line, and center cluster of C2 is 797.04785 on the x line and 605.90436 on the y line. The clustering results of K-Means are 667 respondents in the understand History cluster, and 569 respondents in the cluster Do not understand History. In the 4th iteration K-Means cluster members are still changing so the iteration will continue.

In the K-Medoids results in Figure 2 (right graph) has a central value of cluster C1 that is 1100 on the line x and 500 on the y line, and the center of cluster C2 is 975 on the line x and 725 on the line y. Seen inthe clustering results K-Medoids produced 554 respondents in the understand History cluster, and 682 respondents in the cluster Do not Know understand. At the 2nd iteration the total cost is 264537.53 and the difference in total cost is -71170.195. Because the total cost difference is still negative, then the iteration still continues.

Figure 3 shows the overall data clustering results for the 5th K-Means iteration and the 3rd iteration of K-Medoids. In the K-Means results (left graph) has a central value of cluster C1 that is 1080.58 on the line x and 432.24957 on the line y, and center cluster C2 is 795.3523 on the line x and 604.2354 on the y line. There are 667 respondents in the Understand History cluster, and 569 respondents in the cluster Do not Understand History. The cluster members in the 5th iteration is the same as the cluster members in the 4th iteration, hence the iteration of the K-Means algorithm is stopped and the end result is in the 5th iteration.

In the K-Medoids results in Figure 3 has a cluster center of C1 with of 925 on the x line and 750 on the y line, and the center of cluster C2 is 825 on the x line and 975 on the y line. The clustering results K-Medoids produced 57 respondents in the Understand History cluster, and 1179 respondents on the cluster Do not Understand History. At the 3rd iteration the total cost is 374093.75 and the total cost difference is 109556.23. Because the total cost difference is positive, the iteration is stopped and the result is in the 2nd iteration.

B. Results of accuracy

To analyze the clustering graph, the result of accuracy of K-Means and K-Medoids was compared. The comparison is shown in table 1 and figure 19.

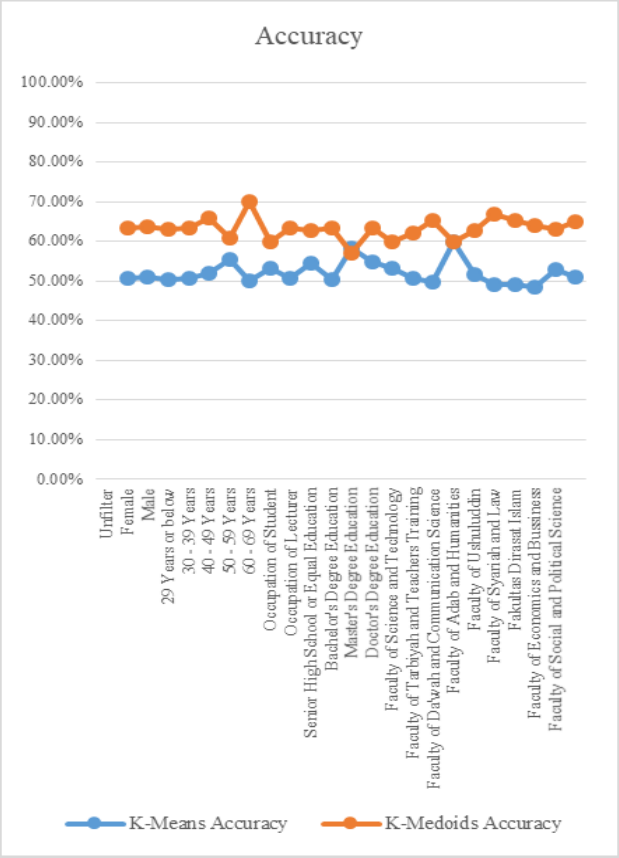


Figure 19. Graph of Accuracy

TABLE 1
COMPARISON OF ACCURACY RESULT ANALYSIS

| No. | Filtered By | K-Means | K-Medoids |
|---|---|---|---|
| 1 | Without filter | 50.73% | 63.33% |
| 2 | Gender: female | 51.08% | 63.74% |
| 3 | Gender: male | 50.39% | 62.94% |
| 4 | Age: 29 years or less | 50.62% | 63.35% |
| 5 | Age: 30 -39 years | 52.00% | 66.00% |
| 6 | Age: 40 -49 years | 55.45% | 60.91% |
| 7 | Age: 50 -59 years | 50.00% | 70.00% |
| 8 | Age: 60 -69 years | 53.33% | 60.00% |
| 9 | Employment status: Student | 50.62% | 63.35% |
| 10 | Employment status: Lecturer | 54.44% | 62.78% |
| 11 | Education: senior high school or equal | 50.53% | 63.42% |
| 12 | Education: Bachelor's degree | 58.46% | 56.92% |
| 13 | Education: Master's degree | 54.67% | 63.33% |
| 14 | Education: Doctor degree | 53.33% | 60.00% |
| 15 | Faculty of Science and Technology | 50.70% | 62.21% |
| 16 | Faculty of Science of Tarbiyah | 49.78% | 65.33% |
| 17 | Faculty of Science of Da'wah | 60.00% | 60.00% |
| 18 | Faculty of Courtesy | 51.75% | 62.66% |
| 19 | Faculty of Ushuluddin | 49.07% | 66.74% |
| 20 | Faculty of Sharia | 49.13% | 65.44% |
| 21 | Faculty of Dirasat Islam | 48.50% | 64.00% |
| 22 | Faculty of Economics and Business | 52.91% | 63.09% |
| 23 | Faculty of Social Sciences | 51.08% | 64.92% |
| **Average Accuracy** | | **52.11%** | **63.24%** |

TABLE II
RESULTS OF EXECUTION TIME

| **Average of All Tests** | | | | |
|---|---|---|---|---|
| Sample | K-Means | K-Medoids | Selisih | Faster Algorithm |
| 1 | 4.75 | 1 | 3.75 | K-Medoids |
| 2 | 2.25 | 8.5 | 6.25 | K-Means |
| 3 | 3 | 1.75 | 1.25 | K-Medoids |
| 4 | 4.5 | 1 | 3.5 | K-Medoids |
| 5 | 2.75 | 3.25 | 0.5 | K-Means |
| **Average** | **3.45** | **3.1** | **0.35** | **K-Medoids** |

C. Result of execution time

Analysis of the results of execution time is done by comparing the execution time of each algorithm by 3 times testing, and the results can be seen in table II.

D. Result of Big O

Table III shows the result of Big O analysis by each algorithm.

## V. DISCUSSION AND CONCLUSION

In this research we have conducted some experiment to gain some information about the performance of K-Means and K-Medoids Algorithm. K-Medoids Algorithm better than K-Means Algorithm in terms of accuracy, execution time and time complexity. The test was performed 5 times on both algorithms and produced different outputs because the first cluster center selection in K-Means clustering was done by randomly selecting data and cluster center selection each iteration in K-Medoids clustering was done by selecting random data. Each test had different results in each filtering data scenario and the average accuracy of 5 tests are 52.11% for K-Means and 63.24% for K-Medoids. The performance of the execution time is displayed in the form of time charts with per-milisecond units. Testing the execution time is also performed 5 times and executed simultaneously with testing accuracy. Then the average time is calculated from the five test of the execution time and K-Medoids algorithm get an average execution time of 3.1 ms, while the K-Means algorithm get the average execution time is slower by 3.45 ms. The result of time complexity is also directly proportional to the result of the execution time of the algorithm, where the K-Means algorithm is more complex than the K-Medoids algorithm with an average value of 377,886 and K-Medoids with an average grade of 301,157. Factors that affect the time complexity of an algorithm are the number of instructions and loops in the algorithm.

TABLE III
BIG O EQUATION ANALYSIS

| No. | n | K-Means (98n^2+11n-19) | K-Medoids (80n^2+40n-43) | Difference |
|---|---|---|---|---|
| 1 | 10 | 9,891 | 8357 | 1,534 |
| 2 | 20 | 39,401 | 32757 | 6,644 |
| 3 | 30 | 88,511 | 73157 | 15,354 |
| 4 | 40 | 157,221 | 129557 | 27,664 |
| 5 | 50 | 245,531 | 201957 | 43,574 |
| 6 | 60 | 353,441 | 290357 | 63,084 |
| 7 | 70 | 480,951 | 394757 | 86,194 |
| 8 | 80 | 628,061 | 515157 | 112,904 |
| 9 | 90 | 794,771 | 651557 | 143,214 |
| 10 | 100 | 981,081 | 803957 | 177,124 |
| **Avarage** | | **377,886** | **310,157** | **67,729** |

## REFERENCE

[1] S. Haryati, A. Sudarsono, E. Suryana, "Implementasi Data Mining Untuk Memprediksi Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," Jurnal Media Infotama, vol 11, No. 2, pp.130-138, 2015. Available : Jurnal Media Infotama : https://jurnal.unived.ac.id/index.php/ jmi/article/view/258

[2] E. Sabna and M. Muhardi, "Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi dan Hasil Belajar," Jurnal CoreIT, vol.2, no.2, 2016. Available: Jurnal CoreIT, http://ejournal.uin-suska.ac.id/index.php/coreit/article/view/2392

[3] S. Taruna, S. Hiranwal, "Enhanced Naïve Bayes Algorithm for Intrusion Detection in Data Mining, " *International Journal of Computer Science and Information Technology* (IJCSIT), vol. 4 (6), pp. 960-962, 2013

[4] S. Rahayu, D. Nugrahadi, F. Indriani, "Clustering Penentuan Potensi Kejahatan Daerah di Kota Banjarbaru dengan Metode K-Means," Jurnal Ilmiah KLIK, vol.1, no.1, 2014. Available: Jurnal Ilmiah KLIK: http://klik.unlam.ac.id/index.php/klik/article/view/7

[5] E. Irwansyah, "Clustering", 2016. [Online]. Available : https://socs.binus.ac.id/ 2017/03/09/clustering/ [Accessed : December 14, 2017]

[6] D. Suhartono, "Word Vector Representation: WORD2VEC & Glove," 2016. [Online]. Available : https://socs.binus.ac.id/2016/12/22/word-vector-representation-word2vec-glove/. [Accessed : December 14, 2017]

[7] S.A.I. Alfarozi, "Analyrical Forward Learning (AFL) Pada Jaringan Syaraf Tiruan (JST)," Thesis, Univesitas Gadjah Mada, Yogyakarta, 2016.

[8] Browlee, "Supervised Learning and Unsupervised Learning," 2016. [Online]. Available : https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ [Accessed : December 14, 2017]

[9] Anonim, Unsupervised Learning," 2016. [Online]. Available: https://www.mathworks.com/discovery/supervised-learning.html. [Accessed : December 13, 2017]

[10] J. Han, M. Kamber, and J. Pei., *Data Mining: Concepts and Techniques,* San Francisco California: Morgan Kaufmann Publishers, 2012

[11] B.M. Metisen, H.L. Sari, "Analisis Clustering menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk pada Swalayan Fadhila,"Jurnal Media Infotama, vol.11, no.2, pp. 110-118, 2015. Available : Jurnal Media Infotama : https://jurnal.unived.ac.id/index.php/jmi/article/view/258

[12] Pramesti, Dyang Falila., Furqon, M. Tanzil., Dewi, Candra. 2017. Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Peersebaran Titik Panas (Hotspot). Malang: Universitas Brawijaya. ISSN: e-ISSN: 2548-964X.

[13] Kaur, Noor K., Kaur, Usvir., & Singh, Dr. Dheerendra., 2014. K-Medoids Clustering Algorithm – A Review. [pdf] International Journal of Computer Application and Technology (IJCAT). ISSN. 2349-1841 Vol. 1, Issue 1. April 2014

[14] K.G. Soni, A. Patel, "Comparative Analysis of K-Means and K-Medoids Algorithm on IRIS Data," International Journal of Computatuons Intelligent Research, vol.13, no.5, pp. 899-906, 2017.

[15] A. Batra, *Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms*," http://www.apiit.edu.in,. [Online]. Available: http://www.apiit.edu.in/downloads/all%20chapters/CHAPTER-59.pdf. [Accessed : December 13, 2017]

[16] Nurhayati, V. Amrizal, N. Sinatrya, and T.S. Kania, *Analisa Big Data Menggunakan Hadoop Framework dan Machine Learning sebagai Pemantau Pola Pikir Mahasiswa dan Dosen UIN tentang Islam untuk Decision Support System (DSS).* Research report, Universitas Islam Negeri Syarif Hidayatullah Jakarta.

[17] F. Nasari, S. Darma, "Penerapan K-Means Clustering Pada Data Penerimaan Mahasiswa Baru (Studi Kaus: Universitas Potensi Utama)", in Proc. Seminar Nasional Teknologi informasi dan Multimedia, 2015, pp.73-78