THE MODELING OF "*MUSTAHIQ*" DATA USING K-MEANS CLUSTERING ALGORITHM AND BIG DATA ANALYSIS (CASE STUDY: LAZ)

Nurhayati Buslim[1], Rayi Pradono Iswara[2], Fajar Agustian[3]

[1,2,3]Informatics Department, Science and Technology Faculty
Syarif Hidayatullah State Islamic University
Jl. Ir. H. Juanda No. 95, Ciputat, Tangerang Selatan, Banten
nurhayati@uinjkt.ac.id, rayi.pradonoiswara@uinjkt.ac.id, fajar.agustian17@mhs.uinjkt.ac.id

ABSTRACT

There are a lot of *Mustahiq* data in LAZ (*Lembaga Amil Zakat*) which is spread in many locations today. Each LAZ has *Mustahiq* data that is different in type from other LAZ. There are differences in *Mustahiq* data types so that data that is so large cannot be used together even though the purpose of the data is the same to determine *Mustahiq* data. And to find out whether the *Mustahiq* data is still up to date (renewable), of course it will be very difficult due to the types of data types that are not uniform or different, long time span, and the large amount of data. To give *zakat* to *Mustahiq* certainly requires speed of information. So, in giving *zakat* to *Mustahiq*, LAZ will find it difficult to monitor the progress of the *Mustahiq*. It is possible that a *Mustahiq* will change his condition to become a *Muzaki*. This is the reason for the researcher to take this theme in order to help the existing LAZ to make it easier to cluster *Mustahiq* data. Furthermore, the data already in the cluster can be used by LAZ managers to develop the organization. This can also be a reference for determining the *zakat* recipient cluster to those who are entitled later. The research is "Modeling using K-Means Algorithm and Big Data analysis in determine *Mustahiq* data ". We got data *Mustahiq* with random sample from online and offline survey. Online data survey with Google form and Offline Data survey we got from BAZNAS (National *Amil Zakat* Agency) in Indonesia and another *zakat* agency (LAZ) in Jakarta. We conducted by combining data to analyzed using Big Data and K-Means Algorithm. K-Means algorithm is an algorithm for cluster n objects based on attributes into k partitions according to criteria that will be determined from large and diverse *Mustahiq* data. This research focuses on modeling that applies K-Means Algorithms and Big Data Analysis. The first we made tools for grouping simulation test data. We do several experimental and simulation scenarios to find a model in mapping *Mustahiq* data to developed best model for processing the data. The results of this study are displayed in tabular and graphical form, namely the proposed *Mustahiq* data processing model at *Zakat* Agency (LAZ). The simulation result from a total of 1109 correspondents, 300 correspondents are included in the *Mustahiq* cluster and 809 correspondents are included in the Non-*Mustahiq* cluster and have an accuracy rate of 83.40%. That means accuracy of the system modeling able to determine data *Mustahiq*. Result filtering based on Gender is “Male” accuracy 83.93%, based on Age is ”30-39” accuracy 71,03%, based on Job is “PNS” accuracy 83.39%, based on Education is “S1” accuracy 83.79%. The advantaged of research expected to be able to determine quickly whether the person meets the criteria as a *mustahik* or *Muzaki* for LAZ (*Amil Zakat* Agency). The result of modeling is K-Means clustering algorithm application program can be used if UIN Syarif Hidayatullah Jakarta want to develop LAZ (*Amil Zakat* Agency) too.

Keywords: *Big Data, Clustering, K-Means Algorithm, BAZNAS (National Amil Zakat Agency), LAZ (Amil Zakat Agency)*

I. INTRODUCTION

Big data can be defined as a collection of data whose size exceeds the capabilities of any database software tool to retrieve, store, organize and analyze. These data sets are generally generated via the internet, mobile devices, sensor networks, enterprise systems and organizations [1]. Big Data is not only focused on volume, velocity and variety, it is also focused on Big Data. The results of big data can be structured, unstructured and semi-structured [2].

Big data can be defined as a collection of data whose size exceeds the capabilities of any database software tool to retrieve, store, organize and analyze. These data sets are generally generated via the internet, mobile devices, sensor networks, enterprise systems and organizations [1].

Big Data technology is a high volume, high speed and complex management of information assets that helps companies manage data cost effectively and drives information processing innovations for decision making and increase knowledge or insight. Big Data guarantees data processing solutions with new and existing variants to provide tangible benefits for businesses [1].

The meaning of the word "*zakat*" is to grow, develop, flourish or increase. The discussion about *zakat* is in the Holy Qur’an which can be used as a reference. There is [3] [4] [5] "Allah destroys interest and gives increase for charities. And Allah does not like every sinning disbeliever" (Surah Al Baqarah (2) verse: 276); "Take Sadaqah (alms) from their wealth in order to purify them and sanctify them with it, and invoke Allah for them. Verily! Your invocations are a source of security for them, and Allah is All-Hearer, All-Knower" (Surah At-Taubah (9) verse: 103). The distribution of *zakat* funds is one way to create equal distribution of income and reduce the gap between the poor and the rich, so as to create a prosperous life as aspired by Islam.

Zakat is one of the pillars of Islam, and is worship. In this study, the researcher only discusses the institution, while the *zakat* rules will use a reference in the Islamic study book about the existing *zakat* chapter.

In discussing *Zakat* on LAZ (*Amil Zakat* Agency), it will be determined based on existing criteria based on these references [6]. The development of LAZ (*Amil Zakat* Agency) is currently in accordance with the provisions of Law of the Republic of Indonesia Number 23 of 2011 concerning *zakat* management. The criteria according to the Big Indonesian Dictionary (KBBI) is defined as a measure that is the basis for an assessment or determination of something, a measure that is the basis for an assessment or determination. Knowing these criteria will make it easier to determine the variables to be tested through simulations and experiments.

Zakat funds have a fairly good role in empowering *Mustahiq* especially in economic aspects [7]. There are a lot of *Mustahiq* data in LAZ (*Amil Zakat* Agency) which is spread in many locations now. To find out the *Mustahiq* data is still up to date, of course it will be very difficult, because types of data are different between LAZ. Amount of data need time for process. Giving *zakat* to *Mustahiq* certainly requires the speed of this information in giving *zakat* to *Mustahiq*. LAZ will find it difficult to monitor the development of the *Mustahiq*. This is the reason for the researcher to take this theme in order to help the existing LAZ to facilitate the cluster of *Mustahiq* data. Furthermore, the data that has been cluster can be used by LAZ managers to develop the organization. It can also be a reference for determine *zakat* recipients to those who are entitled later.

K-Means Clustering is, K is intended as a constant number of clusters desired, Means in this case means the value of an average of a data group which in this case is defined as a cluster, so K-Means Clustering is a method of analyzing data by grouping data by partitioning. The K-Means method has a way of working, namely the data is made into several groups. The same data is made into the same group. Different data are grouped into a different data set [8].

The description above makes the K-Means clustering algorithm method chosen by the author to be used to manage *Mustahiq* data.The author already made research before [9], so we does not do any more testing with other methods because of the time limitation of the research.

The limitations of the problem in this study are:

1. Using the simulation method by performing data clustering with the K-Means Algorithm method,

2. Make data analysis with Big Data technology using Hadoop frame work; used to build this system is Hive, Java,
3. Comparing the performance of the analysis results using simulations of several scenarios.

The description above becomes the background in this research. So, the researcher proposes the research topic “How to make modeling in determine *Mustahiq* data in LAZ using K-Means Clustering Algorithm and Big Data analysis?". The purpose of this research is to find a model that can be used for updating *Mustahiq* data. It expected to determine what factors influence the determination of the *Mustahiq*. It used the K-Means clustering algorithm method and big data analysis with simulations and experiments. The results are very useful for the development of knowledge in the IT field and implentation in *zakat* as fundrising in Islamic. It can quickly determine the *Mustahiq*. We calculated accuration of each factors influentnce the *Mustahiq*. The system calculated it, so user can choose the factor influence the *Mustahiq* in modeling. The result can used as a decision support system (DSS) in LAZ as organization and for future It can be used as model if UIN Syarif Hidayatullah Jakarta want to develop *zakat* institution too.

Based on related work of this research, the first is namely “Big Data Technology for Comparative Study of K-Means and Fuzzy C-Means Algorithms Performance” [10]. It discusses Big data is technology that can manage very large amounts of data, in very fast time to allow real-time analysis and reactions. Several clustering methods which are used to group data are Fuzzy C-Means (FCM) and K-Means Clustering. K-Means Clustering algorithm is a method of partitioning existing data into two or more group. This research goal was to compare the performance of K-Means and Fuzzy C-Means algorithms in clustering data using big data technology. In this research, Hadoop and Hive were chosen the big data technology.

The second related work is namely “Big Data Analysis Using Hadoop Framework and Machine Learning as Decision Support System (DSS) (Case Study: Knowledge of Islam Mindset)” [9]. It discusses data is a popular term which use to visualize exponential grow and un-structural and structural data storage. Therefore, we need to analyses big data accurately in real time to make better accurate result. One of the ways to do it is by using HDFS (Hadoop File Distributed File System). Another one, the big data processing can be done by using machine learning. Machine learning performs data processing based on science and engineer curiosity. The development at UIN Syarif Hidayatullah Jakarta and its diverse students and their mindset about Islam are also diverse. We need a Technical to process data to get corrected information about that.

In the third related work is namely “A Comparison of Unsupervised Learning Techniques for Encrypted Traffic Identification”. The content about the K-Means and MOGA algorithms get the best results and also cluster the data to be very small [2].

In the fourth related work is namely “Despite advances in technology, a study revealed that, "the main media used to obtain *zakat* information is word of mouth”. The result of this study can be useful for the development of LAZ (*Amil Zakat* Agency) to evaluate the efficiency of *zakat* management in meeting the needs of *Mustahiq* (*zakat* recipients) [11].

In the fifth related work is ”*Pengembangan Algoritma* Unsupervised Learning Technique *Pada* Big Data *Analisis di Media Sosial sebagai Media Promosi* Online *Bagi Masyarakat*”. It discusses how to use big data and K-Means Clustering Algorithm for developed data online and offline as promotion in social media. Disadvantaged it research object data is not for used difference type of data [12].

Based on related work above in this research we used data for different type because data collected from a few LAZ to determined *Mustahiq* data, so we created how to make modeling in determine *Mustahiq* data in LAZ using K-Means Clustering Algorithm and Big Data Analysis?

Big data can be defined as a collection of data whose size exceeds the capabilities of any database software tool to retrieve, store, organize and analyze. These data sets are generally generated via the internet, mobile devices, sensor networks, enterprise systems and organizations [13]. Big Data is not only focused on volume, velocity and variety, it is also focused on Big Data. The results of Big Data can be structured, unstructured and semi-structured [14].

1.1 K-Means

K-Means is the algorithm most often used for document clustering purposes. The main principle of K-Means is to compile the x prototype or center of mass (centroid) from a set of n-dimensional data [15]. Before applying the K-Means algorithm process, the data will be preprocessed first. The K-Means algorithm is included in partitioning clustering which separates the data into k separate sub-regions. The K-Means algorithm is used very often because of its convenience and ability to cluster large data and outliers in a very fast time.

According to [16], K-Means algorithm is based ondecomposition, most widely used in data mining field. The concept is use K as a parameter, divide n object into K clusters, to create relatively high similarity in the cluster, relatively low similarity between clusters. And minimize the total distance between the values in each cluster to the cluster center. The cluster center of each cluster is the mean value of the cluster. The calculation of similarity is done by mean value of the cluster objects. The distance between the objects iscalculated by using Euclidean distance. The closer the distance, bigger the similarity of two objects, and viceversa. the K-Means algorithm is quite effective to be applied in the process of grouping characteristics of research objects.

According to [17] K-Means is the most well-known clustering method and is widely used in various fields because of its very simple form, easy to implement, has the ability to cluster large data, is able to handle data outliers, and the time complexity is linear O (nKT) where n is the number of documents, K is the number of clusters, and T is the number of iterations. K-Means is a partitioning method of clustering that separates data into different groups. By iteratively partitioning, K-Means is able to minimize the average distance of each data to the cluster.

K-Means clustering is a method of unsupervised learning that aims to partition the review into group K where each review belongs to the group that has the closest average value [2]. The K-Means algorithm basically works in 2 processes, namely the process of detecting the location of the center of the cluster and the process of searching for members of each cluster. The clustering process begins by identifying the data to be clustered, C_{ij} (i = 1, ..., n; j = 1, ..., m) where n is the amount of data to be clustered and m is the number of variables. At the beginning of the iteration, the center of each cluster is assigned arbitrarily (it's up to the researcher), C_{kj} (k = 1, ..., k; j = 1, ..., m). Then the distance between each data set to each cluster center is calculated. To calculate the distance of the I_{th} data (x_i) at the center of the k-th cluster (c_k), named (d_{ik}), the Euclidean function can be used. A data will be a member of the k-th cluster if the distance of the data to the center of the k cluster is of the smallest value when compared to the distance to the other cluster centers. Furthermore, the calculation process uses the basic process of the K-Means algorithm, similar by previous research [7] [9] [10].

1.2 Big Data

Big Data technology is a high volume, high speed and complex management of information assets that helps companies manage data cost effectively and drive information processing innovations for decision making and increase knowledge or insight [9]. Big Data guarantees data processing solutions with new and existing variants to provide tangible benefits for businesses [9].

There are 3 initial characteristics or dimensions in Big Data, namely 3V: Volume, Variety and Velocity. IBM describes the characteristics of Big Data as follows:

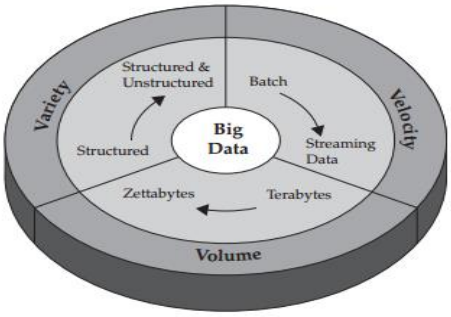


Figure 1. Big data characteristic [7]

However, in its development, Value and Veracity are added, so that it is now known to have the basic character of 5V. The following describes the five characteristics.

Volume: This means that a collection of data in a very large amount and volume some times unstructure; Variety: This means that it contains various types of files, both structured and unstructured; Velocity: This means that data can be accessed at a very fast speed so that it can be used immediately; Value: this how valuable or meaningful the data is; and Veracity (Honesty): This refers to how accurate and reliable a data is.

1.3 Diagram UML (Unified Modelling Language)

UML (Unified Modeling Language) is a language based on graphics or images for visualizing, specifying, building and documenting an object-oriented software development system [7].

1. Use Case Diagram
   Use Case Diagrams describe the activities that can be performed by the system from the view point of the user (system user) as a user (external observer) and relate to scenarios that can be performed by the user. Use case diagram is a set of actors, use case and communication between actors and usecase.
2. Activity Diagram
   The Activity Diagram illustrates the relationship of use case and responsibilities from the class. Activity diagram focuses on the behavior of an operation and shows the flow of control to complete certain processes, such as activities to receive payments. The control flow starts from the initial state or initial state and ends with another state called the end state
3. Class Diagram
   Class diagrams are the essence of object modeling. Class describes a group that has the same state and behavior. Class is a blueprint of objects in an object-oriented system. For example Volkswagen, Toyota and Ford can be grouped into classes named cars. Classes can represent both viewable and abstract concepts
4. Sequence Diagram
   Sequence diagrams describe the interactions between objects in and around the system and are used to describe scenarios, this type of diagram shows the stages that should occur to produce a use case
5. Object Diagram
   Object diagram describes the structure of the system in terms of naming objects and the way objects in the system. In the object diagram, it must be ensured that all classes defined in the class diagram must be used as objects, because otherwise the class definition cannot be accounted for.

1.4 Java Programming Language

Java was first launched as a general purpose programming language with the advantage that it can be run in a web browser as an applet. Since its inception, Java makers have planted their vision into Java for small embedded customer devices (TV, telephone, radio, etc.) so that they can communicate with each other [17].

II. METHOD

2.1 Collecting Data Method

This study uses two data collection techniques. First, primary data is data collection by distributing questionnaires online and offline to respondents. Data online got by Google form as tools. Data offline got by Interview and survey to BAZNAS and LAZ. The second, secondary data is data collection that is carried out through literature studies and related research.

2.2 Simulation Method

There are various types of life cycle which can be used for modeling and simulation studies. The steps in the simulation method are as follows [2]:

1. Problem Formulation
2. Conceptual Model
3. Collection of Input/Output Data
4. Modelling Phase
5. Simulation Phase
6. Verification,Validation, and Experimentation
7. Output Analysis Phase, we describe in chapter III

All process the steps in the simulation method are as follows:

2.2.1 Problem Formulation

In this study, the authors formulated a problem, namely the number of aspects that influence a method of making a *Mustahiq* selection model. So that a decision-making system is needed to be able to determine whether to enter the *Mustahiq* or non-*Mustahiq* clusters. In this case, the best solution is to use Hadoop and the K-Means Clustering algorithm in order to process large data and categorize the data appropriately.

2.2.2 Conceptual Model

Conceptual modeling describes the concept of the system as a whole (overall solution), starting from the initial input, the process, to the output produced by the system. The conceptual model with Hadoop and Machine Learning, namely K-Means Clustering to be implemented in the development system. The starting from connecting the system with Hadoop to storing questionnaire data. The next the system connection with Hive to create and save tables on Hadoop. After connecting, the GUI is created. Input to the "KMean Application" program is in the form of questionnaire data in a .CSV file format that can be imported into the application and stored in the Hive table into Hadoop. The concept of using K-Means Clustering is to be used on the system when analyzing questionnaire data, after which the data is analyzed with the K-Means algorithm.

2.2.3 Collection of Input/Output Data

Data source is needed to develop a system and performing simulations. Data is carrying out the modeling process (modeling). In testing analysis, it will measure how precisely the modeling is made, so that it can process the data sources obtained to become useful outputs. Data can be obtained through various sources depending on the system being created. In this study, the data sources to be used were taken from online questionnaire data and offline questionnaires. The data obtained is then processed and analyzed so that it can become information displayed in the form of data tables and graphics. This information is in the form of online and offline cluster values in the questions answered by respondents in the questionnaire.

Table 1. Input data in hadoop

No.	Data Questionnaire	Data Type
1	Name	String
2	Sex	String
3	Age	String
4	Education level	String
5	Home ownership status	String
6	Salary	String
7	Debts and receivables	String
8	Days of outcome	String
9	Smoke	String
10	School Fees / Month	String
11	Electricity Cost/Month	String
12	Q1-Q15	String

2.2.4 Modeling Phase

The modeling phase is the stage of making a test scenario carried out on the system in accordance with predetermined variables. Scenarios are based on comparing the simulation output results by running the system. The first modeling phase made five UML diagram modeling used, namely use case diagrams, class diagrams, object diagrams, sequence diagrams, and activity diagrams. Then the second do the K-Means construction modeling and do the coding on the "KMean Application" programs. The K-Means algorithm process and formula followed the similar research in [9] [10].

- Construction of the K-Means Algorithm

In this modeling phase, the manually calculated K-Means Algorithm construction is carried out. Manual data is obtained by comparing the total number of *Mustahiq* and the total number of Non-*Mustahiq* then the largest value is taken. If the total number of *Mustahiq* more then the total number of Non-*Mustahiq* then it is included in the *Mustahiq* cluster, and otherwise Non-*Mustahiq* cluster.

The manual calculation process is as follows.

1. K-Means Algorithm Clustering Process:
   At this stage, the main process will be carried out, namely the segmentation of value data accessed from the database, namely a K-Means algorithm clustering method. The following is a flowchart diagram of the K-Means algorithm with the assumption that the Input parameter is the number of data sets of n data and the number of initializations of K = 2 centroid according to the desired cluster. The K-Means construction can be explained by several steps that are followed by the K-Means clustering algorithm which contains the following parts:
   a. N data: N data set to be processed where N data

consists of N attributes (Total of A Value, Total of B Value) which means N data has 2 attributes.

b. K centroid: Initialization of the data cluster center is as much as K where the initial centers are used as the number of classes to be created. Centroid is obtained randomly from N existing data sets.

c. Euclidian Distance: is the distance obtained from the calculation between all N data and K centroid which will obtain the level of closeness to the class closest to the data population. The Euclidian distance to indicate the existence of equations between each clusters with a minimum distance and has a higher equation.

$$D_{ik} = \sqrt{\sum_{j}^{m}(Cij - Ckj)^2} \quad (1)$$

C_{ij}: First Data Point
C_{kj}: Second Data Point
$D_{ik:}$ Euclidean distance is the distance between the data at *x* point and *y* point by math calculations

d. Data grouping: after a number of data populations find closeness to one of the existing centroids, the data population automatically enters the class that has the centroid in question.

e. Update new centroid: each class that was created will update the new centroid. This is done by calculating the average value of each class. If it does not meet the optimal results of the Euclidian distance measurement process, it is looping again.

f. Iteration of limit: if the clustering process is not optimal but it has met the maximum iteration limit, the process is terminated.

The following is an example of the K-Means algorithm function that we use: The total data, namely 1109 correspondents, 10 correspondents were taken as an example to be used for the construction of the K-Means algorithm manually. The experiment was carried out using the following parameters: Number of clusters: 2; Amount of data: 10; Number of attributes: 2.

Table 2. List of correspondent

No.	Name	Score *Mustahiq*	Score Non-*Mustahiq*
1	RESP 00001	925	759
2	RESP 00001	1250	325
3	RESP 00002	1075	550
4	RESP 00003	825	675
5	RESP 00002	1150	1225
6	RESP 00004	800	1000
7	RESP 00005	1000	575
8	RESL 00001	1125	525
9	RESL 00003	1025	600
10	RESL 00004	1000	575

From the list of correspondent data do simulation with iteration 1 to iteration 3, because the clustering results have reached stable and convergent. The result showed bellow:

Table 3. Result calculation manual of k-means clustering

1	2	3	4	5	6	7	8	9	10	
750	0	581.49	651.92	292.62	1101.14	1081.95	325.96	302.08	901.73	C1
277.84	727.04	524.86	212.10	454.60	382.16	232.12	445.52	452.16	167.36	C2

2.2.5 Simulation Phase

In the simulation phase, a simulation process is carried out on the "KMean Application" program which is connected to Hadoop and uses the K-Means Algorithm. The simulation is run from the login stage, data storage, and data analysis. The factors in the simulation process can be seen in Table 4.

Table 4. Variable and factor in simulation

Variable / Parameter Simulation	Data Filtering
Factor 1	Filter data based on Gender
Factor 2	Filter data based on Age
Factor 3	Filter data based on Job

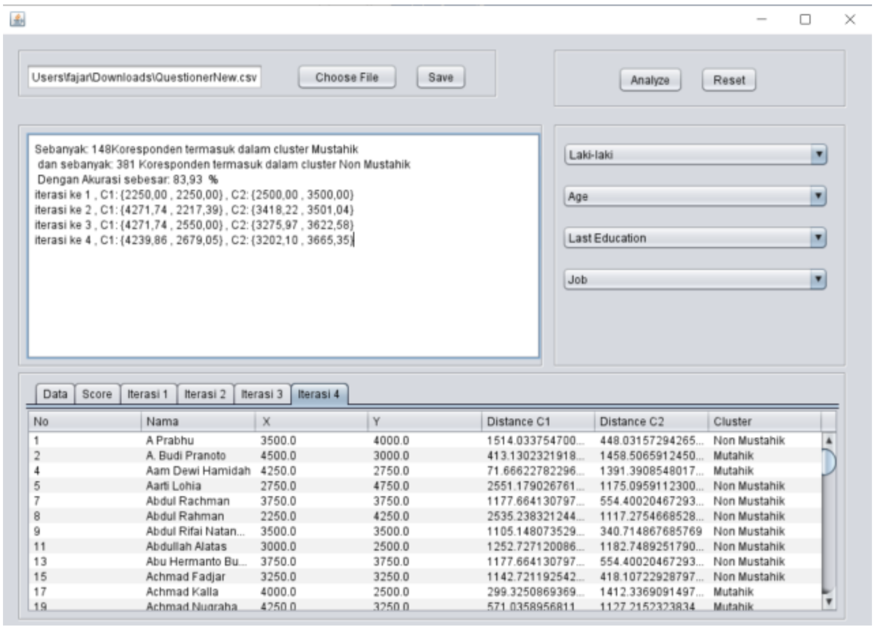
The variable used is the data filter on the analysis results of the "KMean Application" program. The data filter in question is that the data is sorted based on certain conditions and the results of data analysis are also different for each filter.

III. RESULTS AND DISCUSSION

3.1 Simulation Result

The scenario simulation process starts by running Hadoop and Hive on the Linux Terminal. Then open the program ”Kmean Applications”. The initial display when running will display menus which can be seen in the following figure.

The first factor is the data filtered by gender. The results are shown in Figure 2.



Users\fajar\Downloads\QuestionerNew.csv

```
Sebanyak: 148Koresponden termasuk dalam cluster Mustahik
 dan sebanyak: 381 Koresponden termasuk dalam cluster Non Mustahik
 Dengan Akurasi sebesar: 83,93  %
iterasi ke 1 , C1: {2250,00 , 2250,00} , C2: {2500,00 , 3500,00}
iterasi ke 2 , C1: {4271,74 , 2217,39} , C2: {3418,22 , 3501,04}
iterasi ke 3 , C1: {4271,74 , 2550,00} , C2: {3275,97 , 3622,58}
iterasi ke 4 , C1: {4239,86 , 2679,05} , C2: {3202,10 , 3665,35}
```

Laki-laki

Age

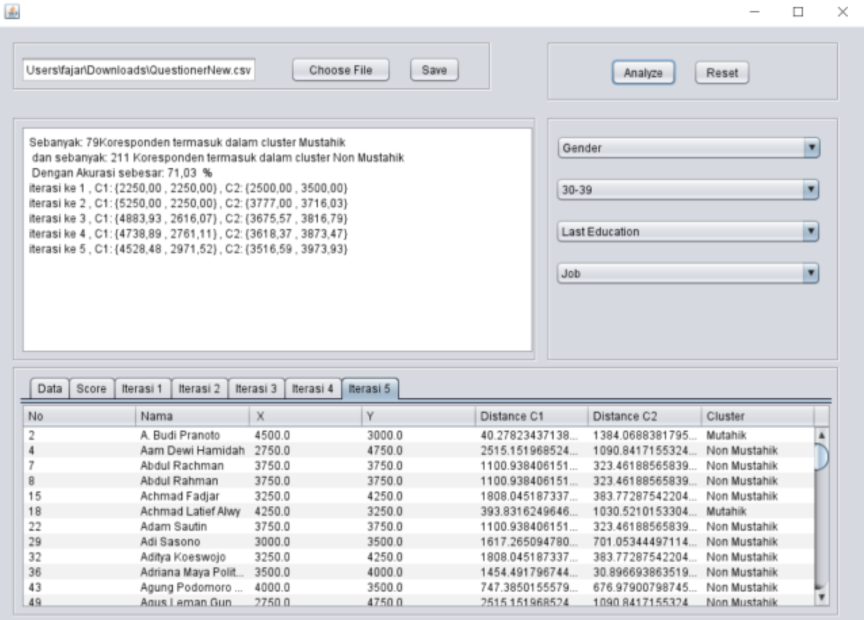
Last Education

Job

No	Nama	X	Y	Distance C1	Distance C2	Cluster
1	A Prabhu	3500.0	4000.0	1514.033754700...	448.03157294265...	Non Mustahik
2	A. Budi Pranoto	4500.0	3000.0	413.1302321918...	1458.5065912450...	Mutahik
4	Aam Dewi Hamidah	4250.0	2750.0	71.66622782296...	1391.3908548017...	Mutahik
5	Aarti Lohia	2750.0	4750.0	2551.179026761...	1175.0959112300...	Non Mustahik
7	Abdul Rachman	3750.0	3750.0	1177.664130797...	554.40020467293...	Non Mustahik
8	Abdul Rahman	2250.0	4250.0	2535.238321244...	1117.2754668528...	Non Mustahik
9	Abdul Rifai Natan...	3500.0	3500.0	1105.148073529...	340.714867685769	Non Mustahik
11	Abdullah Alatas	3000.0	2500.0	1252.727120086...	1182.7489251790...	Non Mustahik
13	Abu Hermanto Bu...	3750.0	3750.0	1177.664130797...	554.40020467293...	Non Mustahik
15	Achmad Fadjar	3250.0	3250.0	1142.721192542...	418.10722928797...	Non Mustahik
17	Achmad Kalla	4000.0	2500.0	299.3250869369...	1412.3369091497...	Mutahik
19	Achmad Nugraha	4250.0	3250.0	571.0358956811	1127.2152323834	Mutahik

Figure 2. Screenshot of analysis result filter based on male gender “male”

In the figure above, the results of the analysis are filtered by gender, namely male. The result is that male correspondents are more dominant in the *Mustahiq* cluster. Where there are 148 correspondents who belong to the *Mustahiq* cluster and 381 correspondents are included in the non-*Mustahiq* cluster. And has an accuracy rate of 83.93%.



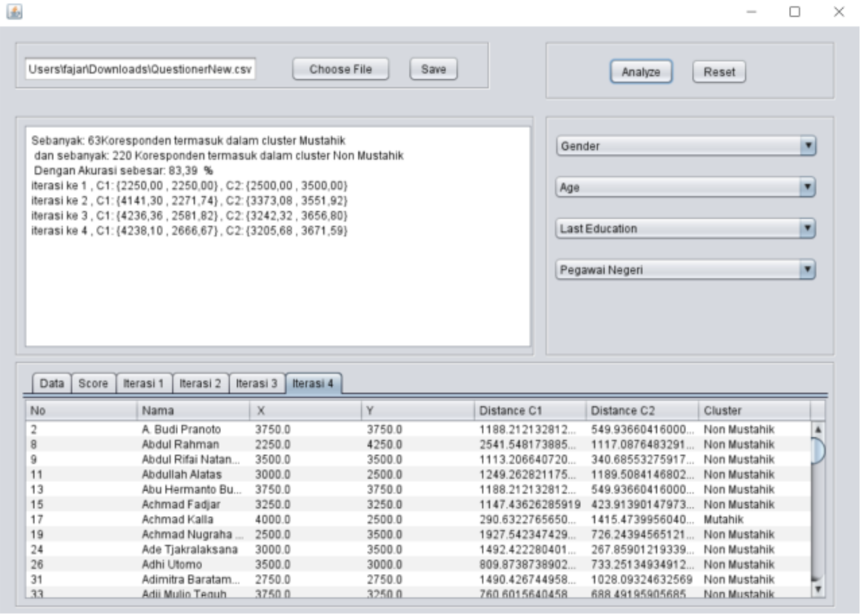
```
Sebanyak: 79Koresponden termasuk dalam cluster Mustahik
 dan sebanyak: 211 Koresponden termasuk dalam cluster Non Mustahik
 Dengan Akurasi sebesar: 71,03  %
iterasi ke 1 , C1: {2250,00 , 2250,00} , C2: {2500,00 , 3500,00}
iterasi ke 2 , C1: {5250,00 , 2250,00} , C2: {3777,00 , 3716,03}
iterasi ke 3 , C1: {4883,93 , 2616,07} , C2: {3675,57 , 3816,79}
iterasi ke 4 , C1: {4738,89 , 2761,11} , C2: {3618,37 , 3873,47}
iterasi ke 5 , C1: {4528,48 , 2971,52} , C2: {3516,59 , 3973,93}
```

Gender
30-39
Last Education
Job

No	Nama	X	Y	Distance C1	Distance C2	Cluster
2	A. Budi Pranoto	4500.0	3000.0	40.27823437138...	1384.0688381795...	Mutahik
4	Aam Dewi Hamidah	2750.0	4750.0	2515.151968524...	1090.8417155324...	Non Mustahik
7	Abdul Rachman	3750.0	3750.0	1100.938406151...	323.46188565839...	Non Mustahik
8	Abdul Rahman	3750.0	3750.0	1100.938406151...	323.46188565839...	Non Mustahik
15	Achmad Fadjar	3250.0	4250.0	1808.045187337...	383.77287542204...	Non Mustahik
18	Achmad Latief Alwy	4250.0	3250.0	393.8316249646...	1030.5210153304...	Mutahik
22	Adam Sautin	3750.0	3750.0	1100.938406151...	323.46188565839...	Non Mustahik
29	Adi Sasono	3000.0	3500.0	1617.265094780...	701.05344497114...	Non Mustahik
32	Aditya Koeswojo	3250.0	4250.0	1808.045187337...	383.77287542204...	Non Mustahik
36	Adriana Maya Polit...	3500.0	4000.0	1454.491796744...	30.896693863519...	Non Mustahik
43	Agung Podomoro ...	4000.0	3500.0	747.3850155579...	676.97900798745...	Non Mustahik
49	Agus Leman Gun	2750.0	4750.0	2515.151968524	1090.8417155324	Non Mustahik

Figure 3. Screenshoot of result filter based age (30-39 years)

In the Figure 3, the analysis results are based on the Ages of 30 to 39 years. The result is that the correspondence of the Non-*Mustahiq* cluster is more dominant than the *Mustahiq* cluster, namely 79 correspondents are included in the *Mustahiq* cluster and 211 correspondents are included in the Non-*Mustahiq* cluster. And has an accuracy rate of 71.03%.



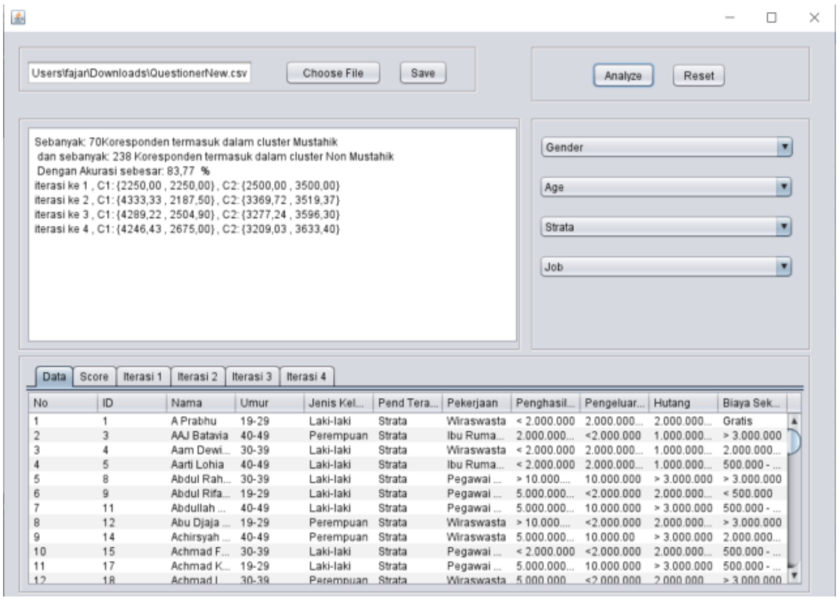
```
Sebanyak: 63Koresponden termasuk dalam cluster Mustahik
 dan sebanyak: 220 Koresponden termasuk dalam cluster Non Mustahik
 Dengan Akurasi sebesar: 83,39  %
iterasi ke 1 , C1: {2250,00 , 2250,00} , C2: {2500,00 , 3500,00}
iterasi ke 2 , C1: {4141,30 , 2271,74} , C2: {3373,08 , 3551,92}
iterasi ke 3 , C1: {4236,36 , 2581,82} , C2: {3242,32 , 3656,80}
iterasi ke 4 , C1: {4238,10 , 2666,67} , C2: {3205,68 , 3671,59}
```

Gender
Age
Last Education
Pegawai Negeri

No	Nama	X	Y	Distance C1	Distance C2	Cluster
2	A. Budi Pranoto	3750.0	3750.0	1188.212132812...	549.93660416000...	Non Mustahik
8	Abdul Rahman	2250.0	4250.0	2541.548173885...	1117.0876483291...	Non Mustahik
9	Abdul Rifai Natan...	3500.0	3500.0	1113.206640720...	340.68553275917...	Non Mustahik
11	Abdullah Alatas	3000.0	2500.0	1249.262821175...	1189.5084146802...	Non Mustahik
13	Abu Hermanto Bu...	3750.0	3750.0	1188.212132812...	549.93660416000...	Non Mustahik
15	Achmad Fadjar	3250.0	3250.0	1147.43626285919	423.91390147973...	Non Mustahik
17	Achmad Kalla	4000.0	2500.0	290.6322765650...	1415.4739956040...	Mutahik
19	Achmad Nugraha ...	2500.0	3500.0	1927.542347429...	726.24394565121...	Non Mustahik
24	Ade Tjakralaksana	3000.0	3500.0	1492.422280401...	267.85901219339...	Non Mustahik
26	Adhi Utomo	3500.0	3000.0	809.8738738902...	733.25134934912...	Non Mustahik
31	Adimitra Baratam...	2750.0	2750.0	1490.426744958...	1028.09324632569	Non Mustahik
33	Adii Mulio Teguh	3750.0	3250.0	760.6015640458	688.49195905685	Non Mustahik

Figure 4. Screenshot of result filter based on job "PNS"

Figure 4 above, the filter results based on the selected Job, namely Civil Servants (PNS). The result is that more correspondents are included in the Non-*Mustahiq* cluster. Where there are 63 correspondents who enter the *Mustahiq* cluster, while 220 correspondents enter the Non-*Mustahiq* cluster. And has an accuracy rate of 83.39%.



Users\fajar\Downloads\QuestionerNew.csv

```
Sebanyak: 70Koresponden termasuk dalam cluster Mustahik
 dan sebanyak: 238 Koresponden termasuk dalam cluster Non Mustahik
 Dengan Akurasi sebesar: 83,77  %
iterasi ke 1 , C1: {2250,00 , 2250,00} , C2: {2500,00 , 3500,00}
iterasi ke 2 , C1: {4333,33 , 2187,50} , C2: {3369,72 , 3519,37}
iterasi ke 3 , C1: {4289,22 , 2504,90} , C2: {3277,24 , 3596,30}
iterasi ke 4 , C1: {4246,43 , 2675,00} , C2: {3209,03 , 3633,40}
```

No	ID	Nama	Umur	Jenis Kel...	Pend Tera...	Pekerjaan	Penghasil...	Pengeluar...	Hutang	Biaya Sek...
1	1	A Prabhu	19-29	Laki-laki	Strata	Wiraswasta	< 2.000.000	2.000.000...	2.000.000...	Gratis
2	3	AAJ Batavia	40-49	Perempuan	Strata	Ibu Ruma...	2.000.000...	<2.000.000	1.000.000...	> 3.000.000
3	4	Aam Dewi...	30-39	Laki-laki	Strata	Wiraswasta	< 2.000.000	2.000.000...	1.000.000...	2.000.000...
4	5	Aarti Lohia	40-49	Laki-laki	Strata	Ibu Ruma...	< 2.000.000	2.000.000...	1.000.000...	500.000 - ...
5	8	Abdul Rah...	30-39	Laki-laki	Strata	Pegawai ...	> 10.000....	10.000.000	> 3.000.000	> 3.000.000
6	9	Abdul Rifa...	19-29	Laki-laki	Strata	Pegawai ...	5.000.000...	<2.000.000	2.000.000...	< 500.000
7	11	Abdullah ...	40-49	Laki-laki	Strata	Pegawai ...	5.000.000...	10.000.000	> 3.000.000	500.000 - ...
8	12	Abu Djaja ...	19-29	Perempuan	Strata	Wiraswasta	> 10.000....	<2.000.000	2.000.000...	> 3.000.000
9	14	Achirsyah ...	40-49	Perempuan	Strata	Wiraswasta	5.000.000...	10.000.00	> 3.000.000	2.000.000...
10	15	Achmad F...	30-39	Laki-laki	Strata	Pegawai ...	< 2.000.000	<2.000.000	2.000.000...	500.000 - ...
11	17	Achmad K...	19-29	Laki-laki	Strata	Pegawai ...	5.000.000...	10.000.000	> 3.000.000	500.000 - ...
12	18	Achmad I	30-39	Perempuan	Strata	Wiraswasta	5.000.000	<2.000.000	2.000.000	> 3.000.000

Figure 5. Screenshot of result filter based on education “S1”

Figure 5 above, the result was that more correspondents entered the Non-*Mustahiq* cluster. In this education ’S1” based filter, as many as 70 correspondents entered the *Mustahiq* cluster and 238 correspondents entered the Non-*Mustahiq* cluster. The accuracy rate of 83.76%.

3.2 Output Analysis Phase

Output by calculating the accuracy of the KMean application with the formula described in the previous chapter. The results of the calculation of the accuracy of the "KMean Application" program can be seen in the following table:

Table 5. Accuracy of the kmean application

Data Filtering	Manual		Kean Application		Accuracy		
	Mustahiq	Non-*Mustahiq*	*Mustahiq*	Non-*Mustahiq*	T	F	Percentage
Gender “Male”	233	296	148	381	444	85	83.93%
Age “30-39	129	161	79	211	206	84	71.03%
Job “PNS”	110	173	63	220	236	47	83.39%
Education “S1”	120	188	70	238	258	50	83.76%

The Table 5 above is the result of manual the comparison of data calculations and the program "ApplicationsKMean". The correspondence that is included in the *Mustahiq* cluster and the Non-*Mustahiq* cluster, and the accuracy is true or false, the accuracy has an average percentage of 80.05%. So that the results of this accuracy can guarantee that the calculation results of the K-Means algorithm in the "KMean Application" program are in accordance with the desired target when doing concept modeling and simulation modeling.

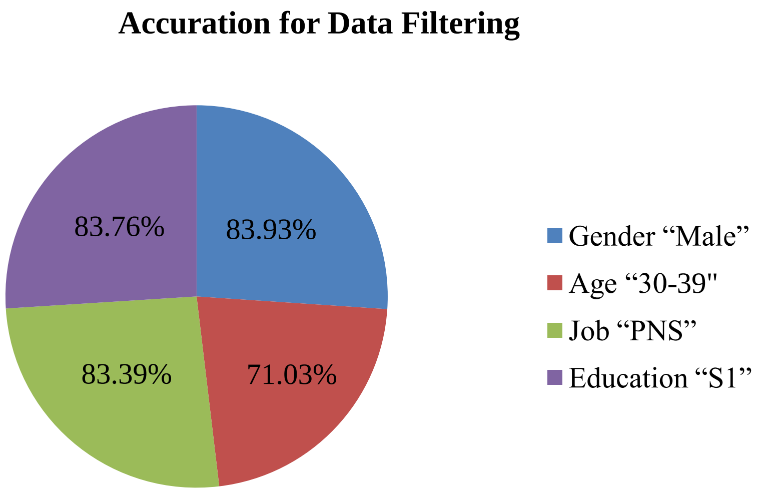
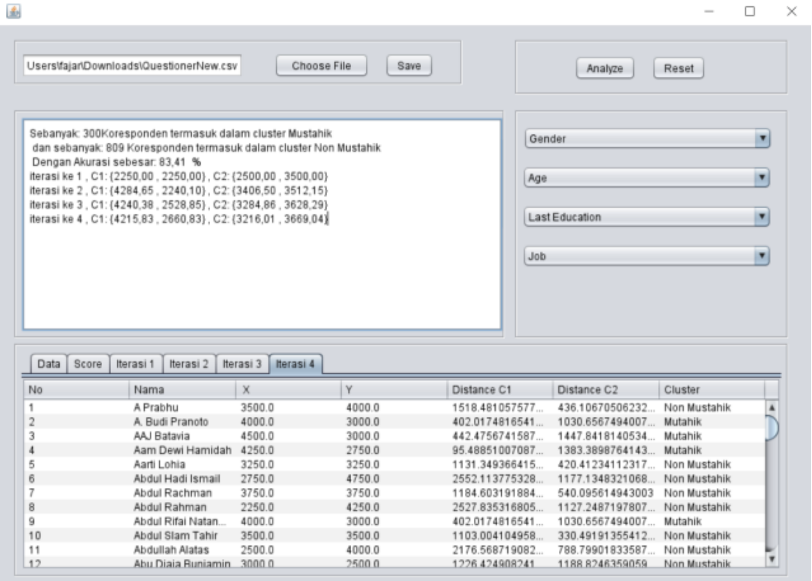


Figure 6. Graph results accuracy by data filtering

Figure 6 shown output accuracy by calculating the of the K-Means on data filtering based on gender,age, job,and education. It refer to Table 5 for calculation. It proven accuracy can guarantee that results can used determined *Mustahiq* data in concept modeling and simulation of the KMeans algorithm in the "KMean Application" program. It are in accordance with the desired target when doing simulation and modeling scenario.



```
Sebanyak: 300Koresponden termasuk dalam cluster Mustahik
 dan sebanyak: 809 Koresponden termasuk dalam cluster Non Mustahik
 Dengan Akurasi sebesar: 83,41  %
iterasi ke 1 , C1: (2250,00 , 2250,00) , C2: (2500,00 , 3500,00)
iterasi ke 2 , C1: (4284,65 , 2240,10) , C2: (3406,50 , 3512,15)
iterasi ke 3 , C1: (4240,38 , 2528,85) , C2: (3284,86 , 3628,29)
iterasi ke 4 , C1: (4215,83 , 2660,83) , C2: (3216,01 , 3669,04)
```

No	Nama	X	Y	Distance C1	Distance C2	Cluster
1	A Prabhu	3500.0	4000.0	1518.481057577...	436.10670506232...	Non Mustahik
2	A. Budi Pranoto	4000.0	3000.0	402.0174816541...	1030.6567494007...	Mutahik
3	AAJ Batavia	4500.0	3000.0	442.4756741587...	1447.8418140534...	Mutahik
4	Aam Dewi Hamidah	4250.0	2750.0	95.48851007087...	1383.3898764143...	Mutahik
5	Aarti Lohia	3250.0	3250.0	1131.349366415...	420.41234112317...	Non Mustahik
6	Abdul Hadi Ismail	2750.0	4750.0	2552.113775328...	1177.1348321068...	Non Mustahik
7	Abdul Rachman	3750.0	3750.0	1184.603191884...	540.095614943003	Non Mustahik
8	Abdul Rahman	2250.0	4250.0	2527.835316805...	1127.2487197807...	Non Mustahik
9	Abdul Rifai Natan...	4000.0	3000.0	402.0174816541...	1030.6567494007...	Mutahik
10	Abdul Slam Tahir	3500.0	3500.0	1103.004104958...	330.49191355412...	Non Mustahik
11	Abdullah Alatas	2500.0	4000.0	2176.568719082...	788.79901833587...	Non Mustahik
12	Abu Djaja Bunjamin	3000.0	2500.0	1226.424908241	1188.8246359059	Non Mustahik

Figure 7. Screenshot output simulation

The total entered into the *Mustahiq* cluster = 300 correspondents
The total is included in the non-*Mustahiq* cluster = 809 correspondents
Accuracy Level = 83.41%
Iteration 1, C1 = (2250.00, 2250.00) and C2 = (2250.00, 3500.00)
Iteration 2, C1 = (4284.65, 2240.10) and C2 = (3406.50, 3512.15)
Iteration 3, C1 = (4240.38, 2528.85) and C2 = (3284.86, 3628.29)
Iteration 4, C1 = (4215.83, 2660.83) and C2 = (3216.01, 36269.04)

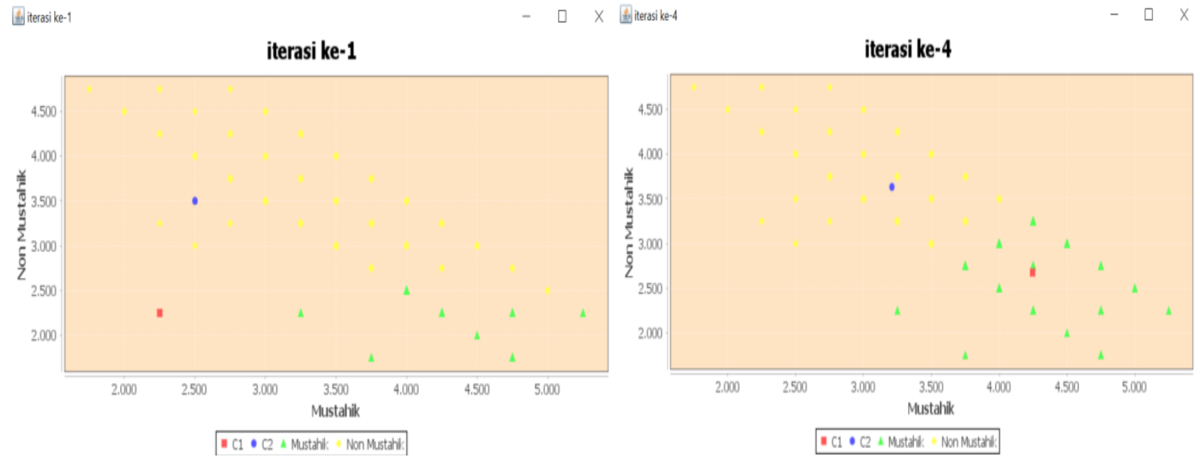


Figure 8. Graph results of simulation output

Figure 7 and 8 show screenshot and graph of the simulation output results without a data filter. The first image shows the pattern in the first iteration. Most of them are included in the Non-*Mustahiq* cluster, which is seen in the number of yellow rhombuses that approach the center of the Non-*Mustahiq* cluster (blue circle) compared to the center of the *Mustahiq* cluster (red box). However, in the second image, which shows the fourth iteration, there is an addition to the *Mustahiq* cluster which can be seen from the number of green triangles that approach the center of the *Mustahiq* cluster (red box). However, in the fourth iteration the Non-*Mustahiq* cluster (yellow rhombus) still looks more dominant. This iteration is carried out four times because in the fourth iteration, the value is stable and convergent or there is no change in value. So, it can be concluded that correspondents tend to belong to the Non-*Mustahiq* cluster. Of the 1,109 correspondents, as many as 300 correspondents entered the *Mustahiq* clusters and 809 correspondents entered the Non-*Mustahiq* cluster, and had an accuracy rate of 83.41

Table 6. Output results are based on the selected filtering data, namely "gender"

No.	Name	Score C1	Score C2	Result
1	RESL00001	1514,033755	448,0315729	Non-*Mustahiq*
2	RESL00002	413,1302322	1458,506591	*Mustahiq*
3	RESL00003	71,66622782	1391,390855	*Mustahiq*
4	RESP00003	2551,179027	1175,095911	Non-*Mustahiq*
5	RESL00004	1177,664131	554,4002047	Non-*Mustahiq*
6	RESL00005	2535,238321	1117,275467	Non-*Mustahiq*
7	RESL00006	1105,148074	340,7148677	Non-*Mustahiq*
8	RESL00007	1252,72712	1182,748925	Non-*Mustahiq*
9	RESL00008	1177,664131	554,4002047	Non-*Mustahiq*
10	RESL00009	1142,721193	418,1072293	Non-*Mustahiq*

The simulation output results are based on the selected filtering data, namely male gender.
The total entered into the *Mustahiq* cluster = 148 correspondents
The total is included in the non-*Mustahiq* cluster = 381 correspondents
Accuracy Level = 83.93%
Iteration 1, C1 = (2250.00, 2250.00) and C2 = (2250.00, 3500.00)
Iteration 2, C1 = (4271.74, 2217.39) and C2 = (3418.22, 3501.04)
Iteration 3, C1 = (4271.74, 2550.00) and C2 = (3275.97, 3622.58)
Iteration 4, C1 = (4239.86, 2679.05) and C2 = (3202.10, 3565.354)

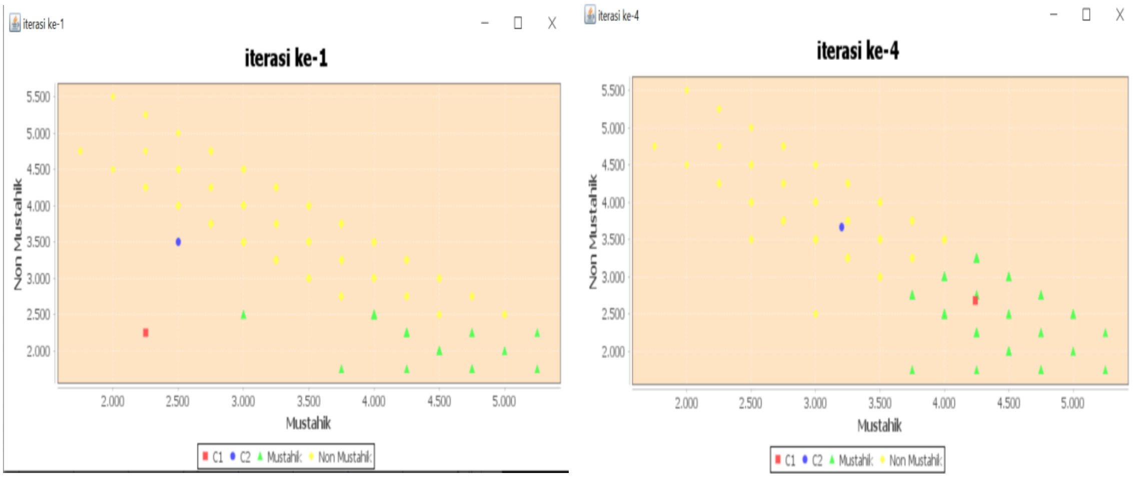


Figure 9. Factor 1 simulation output graph (based on male gender)

Figure 9 shows a graph of the output results from the simulation of the first factor, namely the data filtered based on Male Gender. The first image shows the pattern in the first iteration. Most of them are included in the Non-*Mustahiq* cluster, which is seen in the number of yellow rhombuses that approach the center of the Non-*Mustahiq* cluster (blue circle). However, in the second image, which shows the fourth iteration, there is an addition to the *Mustahiq* cluster which can be seen from the number of green triangles that approach the center of the *Mustahiq* cluster (red box). However, in the fourth iteration the Non-*Mustahiq* cluster (yellow rhombus) still looks more dominant. This iteration is carried out four times because in the fourth iteration, the value is stable and convergent or there is no change in value. So it can be concluded that male respondents tend to belong to the Non-*Mustahiq* cluster. Of the 529 correspondents, 148 correspondents entered the *Mustahiq* cluster and 381 correspondents entered the Non-*Mustahiq* cluster.

Table 7. Output results are based on the selected filtering data, namely age

No.	Name	Score C1	Score C2	Result
1	RESL00001	40,27823437	1384,068838	*Mustahiq*
2	RESP00001	2515,151969	1090,841716	Non-*Mustahiq*
3	RESP00002	1100,938406	323,4618857	Non-*Mustahiq*
4	RESL00002	1100,938406	323,4618857	Non-*Mustahiq*
5	RESP00004	1808,045187	383,7728754	Non-*Mustahiq*
6	RESP00005	393,831625	1030,521015	*Mustahiq*
7	RESL00003	1100,938406	323,4618857	Non-*Mustahiq*
8	RESL00004	1617,265095	701,053445	Non-*Mustahiq*

The output simulation results are based on the selected filtering data, namely Age.
The total entered into the *Mustahiq* cluster = 79 correspondents
The total is included in the non-*Mustahiq* cluster = 211 correspondents
Accuracy Level = 71,03%
Iterasi 1, C1= (2250.00, 2250.00) dan C2 = (2250.00, 3500.00)
Iterasi 2, C1= (5250.00, 2250.00) dan C2 = (3777.00, 3716.03)
Iterasi 3, C1 = (4738.89, 2761.11) dan C2 = (3618.37, 3873.47)
Iterasi 4, C1 = (4528.48 , 2971.52) dan C2 = (3516.59 , 3973.93)

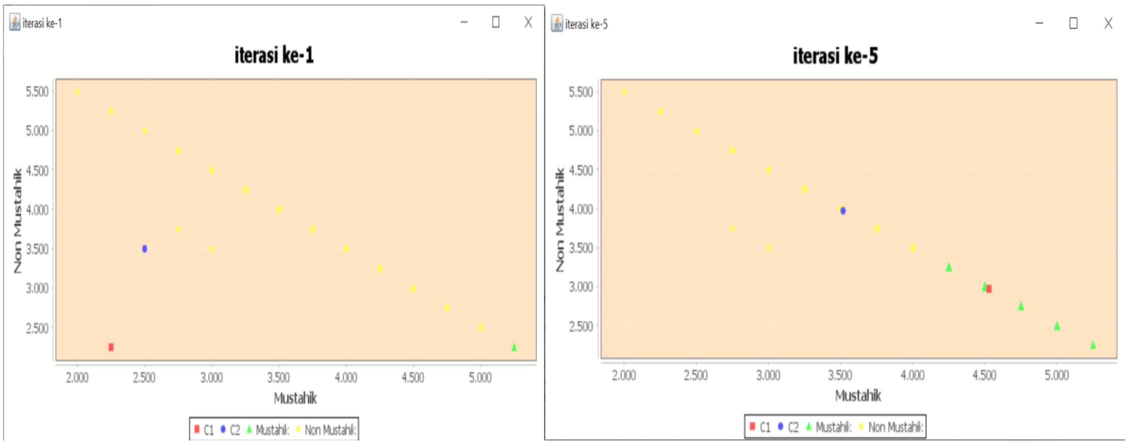


Figure 10. Factor 2 simulation output graph (based on age ”30-39 ”)

Figure 10 shows a graph of the output results from the second factor simulation filtered by age 30-39 years. The first image shows the pattern in the first iteration. Most of them are included in the Non-*Mustahiq* cluster, which is seen in the number of Non-*Mustahiq* clusters (yellow rhombus) that are approaching the center of the Non-*Mustahiq* cluster (blue circle). However, in the second picture, which shows the fifth iteration, there is an addition to the *Mustahiq* cluster as seen from the number of *Mustahiq* clusters (green triangles) that approach the center of the *Mustahiq* cluster (red box). However, in the fifth iteration, the Non-*Mustahiq* cluster still looks more dominant. This iteration is done five times because in the fifth iteration, the value is stable and convergent or there is no change in value. So it can be concluded that respondents aged 30 to 39 years tend to belong to the Non-*Mustahiq* cluster. Of the 290 correspondents, 79 correspondents entered the *Mustahiq* cluster and 211 correspondents entered the Non-*Mustahiq* cluster.

Table 8. Output results are based on the selected filtering data, namely job

No.	Name	Score C1	Score C2	Result
2	RESL00001	1188,212	549,9366	Non-*Mustahiq*
8	RESP00001	2541,548	1117,088	Non-*Mustahiq*
9	RESP00002	1113,207	340,6855	Non-*Mustahiq*
11	RESP00003	1249,263	1189,508	Non-*Mustahiq*
13	RESL00002	1188,212	549,9366	Non-*Mustahiq*
15	RESP00004	1147,436	423,9139	Non-*Mustahiq*
17	RESP00005	290,6323	1415,474	*Mustahiq*
19	RESL00003	1927,542	726,2439	Non-*Mustahiq*
24	RESL00004	1492,422	267,859	Non-*Mustahiq*
26	RESP00006	809,8739	733,2513	Non-*Mustahiq*

The total entered into the *Mustahiq* cluster = 63 correspondents
The total is included in the non-*Mustahiq* cluster = 220 correspondents
Accuracy Level = 83.39%
Iteration 1, C1 = (2250.00, 2250.00) and C2 = (2250.00, 3500.00)
Iteration 2, C1 = (4141.30, 2271.74) and C2 = (3373.08, 3551.92)
Iteration 3, C1 = (4236.36, 2581.82) and C2 = (3242.32, 3656.80)
Iteration 4, C1 = (4238.10, 2666.67) and C2 = (3205.68, 3671.59)

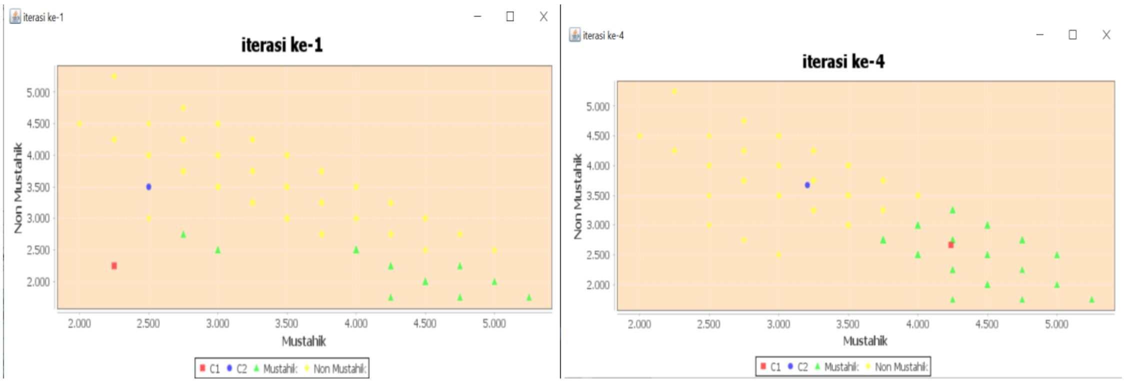


Figure 11. Factor 2 simulation output graph (based on job ‘PNS’)

Figure 11 shows a graph of the output results from the third factor simulation filtered by Civil Servant Occupation. The first image shows the pattern in the first iteration. Most of them are included in the Non-*Mustahiq* cluster, it can be seen from the number of yellow rhombuses that approach the center of the Non-*Mustahiq* cluster (blue circle). However, in the fourth iteration the *Mustahiq* cluster increases, it can be seen from the green triangle approaching the center of the *Mustahiq* cluster (red box). Iterations are carried out four times because in the fourth iteration, the value is stable and convergent or there is no change in value. So it can be concluded that from a total of 283 correspondents, 63 people entered the *Mustahiq* cluster and 220 people entered the Non-*Mustahiq* cluster.

Table 9. Output results are based on the selected filtering data, namely job

No.	Name	Score C1	Score C2	Result
1	RESL00001	1150,432	385,5858	Non-*Mustahiq*
3	RESP00001	407,8628	1013,325	*Mustahiq*
4	RESP00002	412,2177	1437,983	*Mustahiq*
5	RESP00003	75,08499	1365,288	*Mustahiq*
8	RESL00002	1184,089	553,389	Non-*Mustahiq*
9	RESP00005	2542,902	1140,148	Non-*Mustahiq*
11	RESL00003	1112,556	320,0904	Non-*Mustahiq*
12	RESL00004	2192,176	798,1991	Non-*Mustahiq*
14	RESP00006	1603,098	598,0886	Non-*Mustahiq*
15	RESL00005	1184,089	553,389	Non-*Mustahiq*

The total entered into the *Mustahiq* cluster = 70 correspondents
The total is included in the non-*Mustahiq* cluster = 238 correspondents
Accuracy Level = 83.77%
Iteration 1, C1 = (2250.00, 2250.00) and C2 = (2250.00, 3500.00)
Iteration 2, C1 = (43333.33, 2187.50) and C2 = (3369.72, 3519.37)
Iteration 3, C1 = (4289.22, 2504.90) and C2 = (3277.24, 3596.30)
Iteration 4, C1 = (4246.43, 2675.00) and C2 = (3209.03, 3633.40)

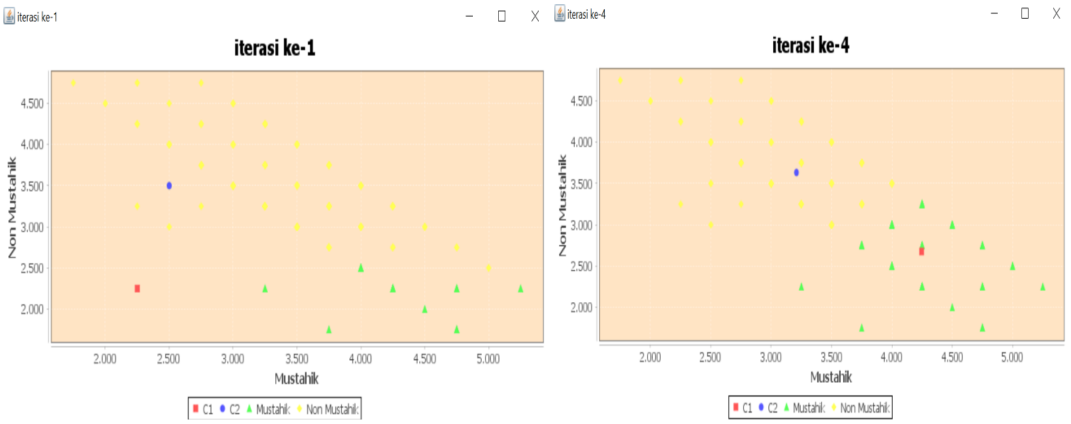


Figure 12. Factor 4 simulation output result graph (based on S1 education)

Figure 12 shows a graph of the output results from the simulation of the fourth factor filtered based on the latest education, namely S1 education. The first image shows the pattern in the first iteration. Most of them are included in the Non-*Mustahiq* cluster, which is seen in the number of yellow rhombs that approach the center of the Non-*Mustahiq* cluster (blue circle). Likewise, in the second picture, the yellow rhombus approaching the center of the Non-*Mustahiq* cluster looks more dominant. This iteration is carried out four times because in the fourth iteration, the value is stable and convergent or there is no change in value. So it can be concluded that correspondents tend to belong to the Non-*Mustahiq* cluster. Of the 308 correspondents, 70 people entered the *Mustahiq* cluster and 238 people entered the Non-*Mustahiq* cluster.

VI. CONCLUSION

Big Data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. This requires an accurate and real time Big Data analysis in order to produce more precise decisions. One way to analyze big data is using HDFS (Hadoop File Distributed File System). The analysis process in this study uses the K-Means algorithm by clustering the data about the model for determining *Mustahiq* data manual The application program "K-Means clustering Algorithm" results of this research is proven to be able to calculate the accuracy of data filtering to determine *Mustahiq* quickly.

We can summary from all graph in simulation result; The first output provides information about the amount of data in each cluster and the center point value in the first and second iterations in text form. The second output describes the correspondent data, the distance of each data to the cluster center and the cluster of each data in tabular form. The third output displays the data population pattern and cluster center in the first and second iterations in graphical form. From a total of 1109 correspondents, 300 correspondents are included in the *Mustahiq* cluster and 809 correspondents are included in the Non-*Mustahiq* cluster and have an accuracy rate of 83.40%. The first test is by doing filtering data on the male gender, where as many as 148 correspondents are included in the *Mustahiq* cluster and 381 correspondents are included in the Non-*Mustahiq* cluster and have an accuracy rate of 83.93%. The second test is by doing filtering data at the age of 30-39 years, where as many as 79 correspondents are included in the *Mustahiq* cluster and 211 correspondents are included in the Non-*Mustahiq* cluster and have an accuracy rate of 71.03%. The third test is by doing filtering data on Civil Servant Jobs (PNS), which is as many as 63 correspondents included in the *Mustahiq* cluster and 220 correspondents included in the Non-*Mustahiq* cluster and has an accuracy rate of 83.39%. The fourth test is by doing filtering data in undergraduate education (*Strata* 1), namely 70 correspondents are included in the *Mustahiq* flow cluster and 283 correspondents are included in the Non-*Mustahiq* cluster and have an accuracy rate of 83.76%.

The average accuracy of the four tests above is 83.76%. The accuracy is quite large. Based on the results of this accuracy can guarantee that the calculation results of the K-Means algorithm in the "KMean Application" program are in accordance with the desired target when doing concept modeling and simulation modeling.

Finally, application program for the modeling *Mustahiq* data with K-Means clustering Algorithm and Big Data analysis can be proposed for the model develop in LAZ *(Amil Zakat* Agency*)* to determined *Mustahiq* data. Even the application program of K-Means clustering Algoritm more quick in calculated and we can add conditional with filtering data as needed with quickly shown result. This model can be proposed to as model, if UIN Syarif Hidayatullah Jakarta want to development *zakat* agency.

The suggestions from the authors are recommend the "KMean Application" program be able to retrieve online questionnaire data in real time which is connected to the application; It is recommend that the output on the KMean application be displayed online or connected to a web application; in the next reseach we hope for collect data in the LAZ can be easier for research because data is it very difficult to collect from each LAZ in this research.

REFERENCES

[1] S. Hartati and A. Nugroho, “MongoDB: implementasi VLDB (very large database) untuk sistem basis data tersebar (distributed database),” Jurnal Teknik Informatika, 2012.

[2] B. Carlos, et al., “A comparison of unsupervised learning techniques for encrypted traffic identification. Dalhousie University,” 2009.

[3] “Holly Qur’an online,” Available: https://www.islamicfinder.org/quran/surah-al-baqara/276/?translation=english-saheeh-international&language=ms https://ayatalquran.net/ [Accessed: Jan. 22, 2021]

[4] “Holly Qur’an online,” Available: https://www.islamicfinder.org/quran/surah-at-tawba/103/?translation=english-muhammad-taqi-ud-din-al-hilali-and-muhammad-muhsin-khan [Accessed: Jan. 22, 2021]

[5] “Al Quran terjemah,” Available: https://ayatalquran.net/ [Access Jan. 22, 2021].

[6] “Panduan zakat,” Available: https://www.dompetdhuafa.org/uploads/media/PANDUAN-ZAKAT-1433-web.pdf [Accessed: Jan. 10, 2021]

[7] D. D. Prahesti. and P. P. Putri, “Learn to pronounce empowerment of small and micro enterprises through earning zakat funds,” Ilmu Dakwah: Academic Journal for Homiletic Studies, 2012, Vol. 12 no. 1, pp. 141-160. Available at: 10.15575/idajhs.v12i.190..

[8] Apriyanti, et al., “Algoritma K-Means clustering dalam pengolahan citra digital landsat,” Universitas Lambung Mangkurat, Kalimantan, 2015.

[9] Nurhayati, Busman, and V. Amrizal, “Big data analysis using hadoop framework and machine learning as decision support system (DSS) (case study: knowledge of Islam mindset),” 6th International Conference on Cyber and IT Service Management (CITSM), 2018. DOI: 10.1109/CITSM.2018.8674354

[10] Nurhayati, et al., “Big data technology for comparative study of K-Means and fuzzy C-Means algorithms performance,” 7th International Conference on Computer and Communication Engineering (ICCCE), 2018.

[11] Ahmad, et al., “Assessing the satisfaction level of zakat recipients towards zakat management,” Procedia Economics and Finance, 31, pp. 140-151, 2015.

[12] Nurhayati, Busman, and R. P. Iswara. “Pengembangan algoritma unsupervised learning technique pada big data analisis di media sosial sebagai media promosi online bagi masyarakat,” Jurnal Teknik Informatika Vol. 12 No. 1, April 2019. http://journal.uinjkt.ac.id/index.php/ti/article/view/11342/pdf. 2019.

[13] Ediyanto, et al., “Pengklasifikasian karakteristik dengan metode k-means cluster analysis,” Universitas Tanjungpura, 2013.

[14] M. I. Jordan. and T. M. Mitchel, Machine Learning: Trends, Perspectives, and Prospects. American Association for the Advancement of Science, 2015.

[15] A. Wijaya, ”Analisis algoritma k-means untuk sistem pendukung keputusan penjurusan siswa di MAN Binong Subang,” Skripsi. Bandung: Universitas Komputer Indonesia, 2010.

[16] A. M. Baswade, K. D. Joshi, and P. S. Nalwade, International Journal of

Engineering Research & Technology (IJERT). ISSN: 2278-0181, 2012.

[17] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 1967.

[18] V. Prajapati, "Big data analytics with r and Hadoop," Birmingham: Packet Publishing Ltd., 2013.

[19] "Oracle Java Technologies," Available: http://www.oracle.com

Copyright