Algorithm Model of K-means for Poor Households Classifying

Panggih Pawenang*, Nova El Maidah **
* BPS-Statistics of Banyumas Regency, Badan Pusat Statistik, Jl Warga Bhakti No 5 Purwokerto, Jawa Tengah 53114
**Information System Study Programe, Jln. Kalimantan 37 Kampus Bumi Tegalboto, Jember 68121
*panggih@bps.go.id, **nova.pssi@unej.ac.id

ABSTRACT
Poverty is a complex problem experienced by majority of world's nations especially in the developing countries such as Indonesia. Poverty countermeasure has become the main country development program as one of the indicators of development success is the decrease of poverty rate. Various programs either in national or regional scale have been made as the government's attempt to reduce the poverty rate. However, there have been a lot of news reports covering the government's programs which are less precise and even not on the target. Thus, it is essential to have a method which can be implemented to help the planning process of those poverty alleviation programs. This study aims at explaining the formation of K-means algorithm model for classifying poor households; taking a study case in Banyumas Regency, Central Java. The result of this study was K-means algorithm model which has been adapted to the poverty concept from the Statistics Indonesia (BPS) as well as the factors affecting the household poverty. The information obtained from each cluster formed is household characteristics, estimation of household number, and estimation of population experiencing poverty.

Keyword: K-means, partition clustering method, artificial intelligence, algorithm, poverty

1. Introduction

The object of grouping carried out in this study used Statistics Indonesia's data from Banyumas district. selection of data usage from Banyumas District because Banyumas District has the characteristics of urban residents and rural residents.

Banyumas Regency is one of 35 regencies/cities in Central Java Province which have a large population. The projection of the population of SP2010 for 2016 shows the population in Banyumas Regency was 1,650,625 people consisting of 824,717 male souls and 825,908 female souls. The administrative area of Banyumas Regency consists of 331 villages with 187 of them having rural status while the rest are urban. The number of households in Banyumas Regency in 2016 is projected at 451 211 with an average number of household members of 3.7 people [1]. Banyumas Regency is ranked third from 35 regencies/cities in Central Java Province in terms of population. The population growth rate of Banyumas Regency in 2016 is based on the SP2010 projection of 0.90 [2]. Figure 1 shows the rate of population growth in Banyumas Regency from 2010-2016.

Along with increasing population, the potential problems that arise also vary. One problem that arises is poverty. The measurement of poverty in Indonesia is carried out periodically by Statistics Indonesia since 1984 through SUSENAS (Survei Sosial Ekonomi Nasioanl, the Indonesian translation of National Social Economic Survey). The calculation of poverty is done by using the concept of the household capability approach to meeting basic needs approaches [3]. The ability to fulfill the basic needs in question is the ability from the economic side to fulfill basic food and non-food needs measured from the expenditure side. The condition of poverty in Banyumas Regency in 2016 was 17.23% with the per capita poverty line of Rp.344,514. 00. Viewed from the percentage, the number of poor people in Banyumas Regency is estimated at around 283.900 people. Figure 2 shows the poverty rate of the Banyumas District is higher then the poverty rate of Central Java Province and figure 3 shows the community poverty line of the Banyumas District for 2010-2016.

Planning for poverty reduction programs cannot be separated from the availability of accurate data and proper data analysis [4]. Many poverty reduction programs that have been carried out by the government both on a national scale or such as the launch of the Indonesian Health Card (Indonesian translation of Kartu Indonesia Sehat or KIS), Indonesia Smart Card (Indonesian translation of Kartu Indonesia Pintar or KIP), distribution of rice for the poor (RASKIN), People's Business Credit (Indonesian translation of Kredit Usaha Rakyat or KUR), Family Hope Program (Indonesian translation of Program Keluarga Harapan or PKH) or on a regional scale Healthy Banyumas Card (Indonesian translation of Kartu Banyumas Sehat or KBS), prosperous

rice distribution (RASTRA) and others that aim to reduce poverty. Good data analysis will have an impact on policy making so that the program formed becomes as targeted and effective in achieving the goals.

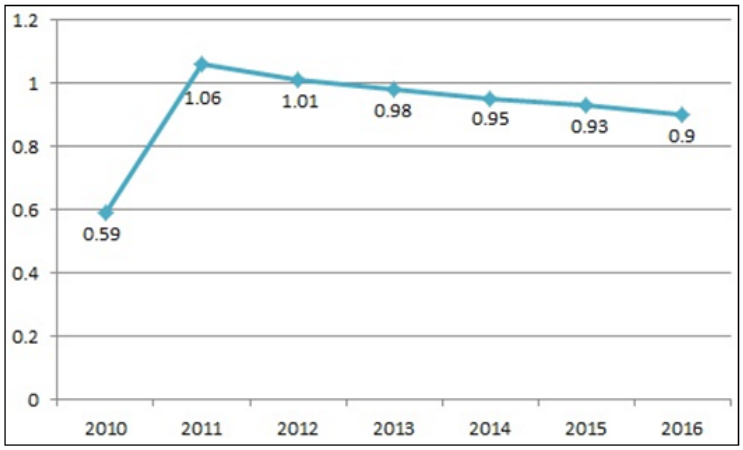

Figure 1. Graph of regency growth rates for 2010-2016 [2]

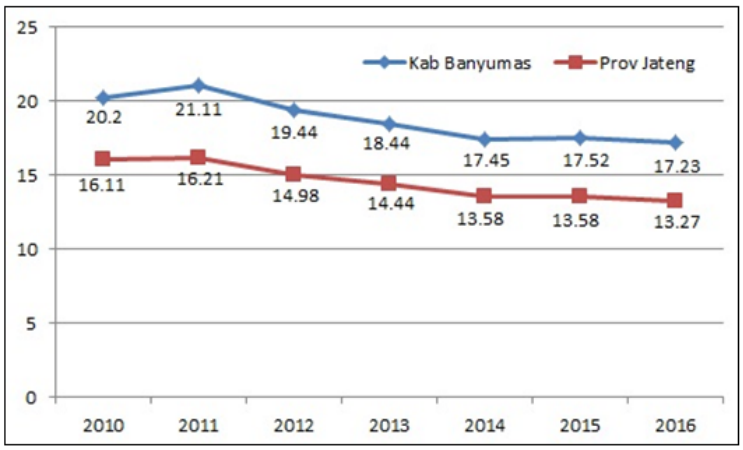

Figure 2. The poverty rate of Banyumas Regency 2010-2016 [3]

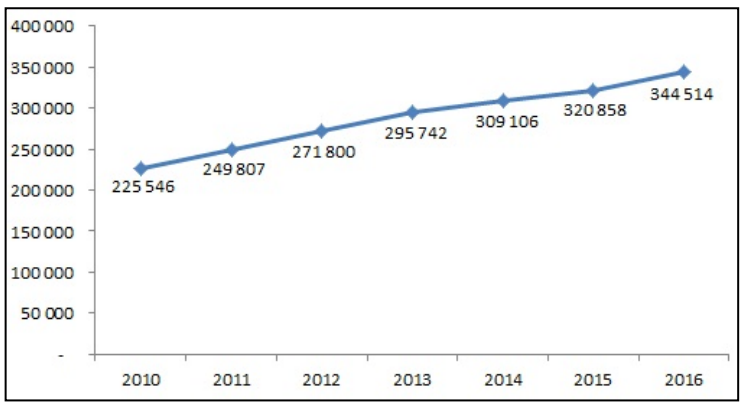

Figure 3. The poverty rate of Banyumas Regency 2010-2016 [3]

Data analysis was carried out with the aim of knowing more about the characteristics of a problem so that these problems can be classified or grouped into several more specific parts with the aim of facilitating the settlement process. Many methods can be used in data analysis ranging from simple statistical analysis to in-depth statistical analysis or even using artificial intelligence methods. One of the artificial intelligence methods that can be used in clustering is the algorithm K-means algorithm falls into the category of partition clustering method.

The concept of the K-means algorithm is to form groups of data by calculating the closest distance between data points as needed [5]. The algorithm has been applied in several studies such as the efficiency of

consumer segmentation [6], prediction of student academic achievement [7], pattern recognition and data mining [8]. One of the advantages of algorithms that fall into the category of artificial intelligence is flexibility and speed in the problem solving process. Referring to the strengths of the K-means algorithm, this study focused on the application of the k-means algorithm for the classification of poor households so as to provide an overview of the conditions and characteristics of poor households that are expected to help plan poverty alleviation programs.

2. Research Method

2.1. Researched Object

This study focuses on the formation of the K-Means algorithm model for the classification of poor households. The formation of the model was carried out by paying attention to the concept of poverty used by BPS and the concept of the K-Means algorithm. Statistics Indonesia conducts poverty calculations using an approach to the ability of households to fulfill basic needs. The ability to fulfill the basic needs in question is the ability from the economic side to fulfill basic food and non-food needs measured from the expenditure side. Information about poverty is obtained from SUSENAS which is periodically carried out by BPS every year. This study utilizes SUSENAS data as the main reference for the classification of poor households with household case studies in Banyumas Regency, Central Java Province.

2.2. Data Analisys

In addition to BPS data, library studies are also conducted to determine the factors that are considered influential in the process of classifying poor households. When these factors have been determined, the next step is to combine these factors with the basic concept of the K-Means algorithm so that the K-Means algorithm model will be formed which can be used for the process of classification of poor households.

In 2005, BPS first conducted data collection of poor households to obtain poverty micro data through an activity called Pendataan Sosial Ekonomi (PSE 2005) and SUSENAS produces poverty indicators (GK). The results of the data collection are then updated every three years until the last activity is done in 2015. There are four classifications of poor households used by BPS in the data collection that is: (1) housekeeping is very poor (SM) = SM < 0,8 GK; (2) rumah tangga miskin (M) = 0,8 GK < M < 1 GK; (3) almost poor households (HM) = 1 GK <HM <1,2 GK; and (4) other poor vulnerable households (RML) = 1.2 GK <RML <1.6 GK.

2.3. K-Means Algorithm

Cluster is a set of data that has similarities to one another and also has inequalities with other data outside the cluster. Cluster formation can be done in many ways, one of which uses K-means algorithm. K-means algorithm is included in the category of partitioning methods in the field of clustering. The concept of K-means algorithm is to split a set of data into a K cluster that has a certain similarity. The usual formula used to measure antardata equality is euclidean distance. The steps in K-means algorithm are as follows:

(1) Initialization: determine the number of clusters to be established (K) and assign a number of K data as a temporary centroid. The process of determining K can be done randomly or using heuristic methods.
(2) Calculation of the similarity between data: determine the distance between data with the temporary centroid. This is done on any existing data. Then put the data into a particular centroid based on the closest distance.
(3) Calculation of centroid: after placing the data into a particular centroid, then do the centroid value renewal.
(4) Convergent conditions: repeat steps 2 and 3 until the converging conditions are reached. Convergent conditions are conditions in which each existing data is exactly on one particular cluster and the data distance with clustered centroid cluster is the least value.

3. Result and Analysis

The poverty rate of Banyumas Regency in 2016 was 17.23%. This means there are 17.23% of the total population of Banyumas Regency (1,650,625 inhabitants), namely $\pm$ 283.900 people included in the category of poor population. Other data stated that the number of households in Banyumas Regency in that year reached $\pm$454,200 with an average number of household members as many as 3.6 people [9]. Noting the description, the number of poor households in Banyumas Regency is estimated to reach 78,800 with a poverty line position of Rp.344,514.00 per capita per month. The average length of school for residents of Banyumas Regency (age 25 years and over) in 2016 was 7.39. Its means that the average education of the adult population (25 years and above) can only be completed up to junior high school level 1 [9]. Regionally, some areas of Banyumas Regency are in the form of rural areas with the majority of employment in the trade category [9].

Statistics Indonesia, in its publication with the title of 2016 Indonesian macro poverty calculation and analysis states that the concept used in measuring poverty in Indonesia is the concept of meeting basic needs.

Based on this concept, poverty is seen as an economic inability to fulfill basic food and non-food needs measured from the expenditure side. The minimum limit in terms of the economy to meet these needs is called the poverty line [10]. Statistics Indonesia first calculated the number and percentage of poor people in 1984 through SUSENAS with households as survey objects. A household is a person or group of people who inhabit part or all of the physical building/census and are usually in fulfilling their daily needs bound in one economic entity. The scope of the survey is a description of the main household and consumption module.

When poverty is seen as an inability of the economic terms to meet basic food and non-food needs measured in terms of expenditure, the main factor affecting the size of a household's expenditure is the number of household members. The expenditure of a household is obtained by adding up the expenditure of all household members. The average expenditure of household members is represented in a variable called per capita expenditure. A household was said to be not poor if it can economically meet its basic needs. Another factor that is considered to influence the poverty of a household is income. The size of household income is broadly approximated from the number of household members who work and the type of work. Number of household members who work have a positive correlation with household income [10]. Based on the description, it can be concluded that there are three factors that are considered to influence the poverty of a household, namely the number of household members, the value of household expenditure, and the number of household members who work.

The formation of the K-Means Algorithm model for the classification of poor households begins with modeling the household (object of research). Based on the description above, there are 3 factors that are considered to influence the poverty of a household, namely the number of household members (denoted x), the value of expenditure per capita (denoted y), and the number of household members working (denoted z) so that a household r can modeled in form $r(x_r,y_r,z_r)$.

The next step is the implementation of the model into the K-Means algorithm. The results of the implementation are as follows:

(1) Initialization

Determine the number of clusters formed by one cluster and specify a number of data as temporary centeroids. The process of determining k can be done randomly or using the heuristic method.

(2) Calculation of similarities between points.

Calculation of interdata similarity is done by calculating the distance between points with temporary centroids. This is done at every point that exists. Point $r(x_r,y_r,z_r)$ and $c(x_c,y_c,z_c)$ was the centroid. while from a cluster k, the formula used to determine the distance (d_{rc}) is:

$$d_{rc} = \sqrt{[(x_r - x_c)^2 + (y_r - y_c)^2 + (z_r - z_c)^2]} \quad (1)$$

(3) Placing a point into the cluster *k*

The process of placing a point r into the cluster k is carried out based on the calculation of the similarity between the centroids. A point r is said to enter in cluster t if drck is the minimum value of the distance point r for all centroids ($d_{rct} = min\{d_{rc1}, d_{rc2}, d_{rc3}, \ldots, d_{rct}, \ldots, d_{rck}\}$).

(4) Updated centroid.

After placing a point into a particular k cluster, then updating the centroid value.

(5) Repeat step 2 until the convergent condition is reached. Convergent conditions are conditions where each existing data is located in one particular cluster and the distance between the selected data with the centroid cluster with minimum value.

The data used for testing the k-means algorithmic model for the classification of poor households is the SUSENAS Banyumas Regency data in 2016. The survey was conducted in March 2016 and involved 953 respondents. Table I shows some of the data used in the testing process.

Table 1. Model Data Test

Respondent	Number of Household Members	per capita expenditure (in million rupiah)	working household members
1	5	3.4	1
2	1	6.1	0
3	6	3.6	3
4	5	0.8	2
...	...	...	...
953	4	0.6	0

Using data as in Table I, by determining k = 7, the results of testing the k-means algorithm model for the classification of poor households are shown in Table II and Table III. The centroid position in each cluster showed by Table II and the number of points for each cluster showed by Table III. The cluster with the highest number of points was cluster 5 (300 points) while cluster 1 is the cluster with the smallest number of points (8 points).

Table 2. Centroid for Each Clusters

Centroid	x	y	z
1	2.625	7.421	1.250
2	1.676	1.003	0.555
3	6.234	0.639	1.812
4	7.384	0.501	3.615
5	4.303	0.707	1.553
6	4.177	0.838	3.114
7	2.746	0.853	1.678

Table 3. Member Distribution

Clusters	Distribution	%
1	8	0.8
2	207	21.7
3	64	6.7
4	26	2.7
5	300	31.5
6	96	10.1
7	252	26.5
SUM	953	100.0

Table 4. Cluster Characteristics

Cluster	Household members avg	Per capita expenditure avg	Working household member avg	Estimation of household numbers
1	3	7.42	1.25	2 849
2	2	1.00	0.61	94 781
3	6	0.64	1.84	35 038
4	7	0.50	3.65	14 623
5	4	0.71	1.57	151 169
6	4	0.84	3.12	47 364
7	3	0.85	1.69	1 104

The k-means algorithm model for the classification of poor households is manifested in a program based on the MATLAB programming language. The matlab version used is MATLAB 2018a which is installed on notebooks with Intel core-i5 4300 processor specifications, 8GB RAM, 128GB SSD.

The output of the k-means algorithm model for the classification of poor households can also provide household characteristics for each cluster formed. Summary of the characteristics of each cluster showed by Table IV. Cluster 1 shows information on household characteristics having an average of 3 household members and 1.25 people working with an average of Rp.7,420,000.00 per person per month. The scope of expenditure was in the form of food and non-food expenditure. Food expenditure includes food and processed foods, while non-food expenses include the distribution of housing, fuel, education, health, telecommunications, etc. [4]. Because the data used are survey results, each data has a conversion value for population estimates. Based on the conversion value, the estimated number of households in cluster 1 is 2,849 households.

4. Conclusion

The results of this study are an algorithmic model that combines the K-Means algorithm concept and the Statistics Indonesia concept of poverty. The algorithm model can be used as an alternative method in classifying poor households. Information obtained from data processing uses the algorithm model in the form of household characteristics, estimated number of households, estimated number of poor people.

References

[1] Kabupaten Banyumas. Badan Pusat Statistik, 2017, *Kabupaten Banyumas Dalam Angka 2017*, BPS Kabupaten Banyumas, Purwokerto.

[2] Kabupaten Banyumas. Badan Pusat Statistik, 2017, *Statistik Daerah Kabupaten Banyumas 2017*, BPS Kabupaten Banyumas, Purwokerto.

[3] Badan Pusat Statistik, 2017, *Data dan Informasi Kemiskinan Kabupaten Kota Tahun 2016*, Badan Pusat Statistik. Jakarta.

[4] Badan Pusat Statistik, 2016, *Perhitungan dan Analisis Kemiskinan Makro Indonesia 2016*, Badan Pusat Statistik. Jakarta.

[5] Bijuraj. L. V, 2013, Clustering and its Applications, Proceeding of National Conference on New Horizons in IT – NCHIT 2013, page 169-172.

[6] Ezenkwu Chinedu Pascal. Ozuomba Simeon. Kalu constance, Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services, *International Journal of Advanced Research in Artificial Intelligence Vol 4 No 10 2015, page 40-44*.

[7] Oyelade O J et al, 2010, *Application of k-Means Clustering algorithm for prediction of Students' Academic Performance*. Internaational Journal of Computer Science and Information Security Vol 7 No 1 2010. Page 292-295.

[8] Barkha Narang et al, 2016, *Application based, advantageous K-means Clustering Algorithm in Data Mining – Review*. International Journal of Lastest Trends in Engineering and Technology Vol 7. Page 121-126.

[9] Kabupaten Banyumas. Badan Pusat Statistik, 2017, *Statistik Daerah Kabupaten Banyumas 2017*, BPS Kabupaten Banyumas, Purwokerto.

[10] Subdirektorat Statistik Kerawanan Sosial. Badan Pusat Statistik, 2016. *Data dan Informasi Kemiskinan Kabupaten/ Kota Tahun 2016*, Badan Pusat Statistik, Jakarta.