Jurnal Matematika MANTIK
Volume 5, Nomor 2, October 2019, pp. 123-134

ISSN: 2527-3159 (*print*) 2527-3167 (*online*)

Application of *Expectation-Maximization* (EM) Algorithm in Grouping Popularity Tourism Objects in Malang Raya Based on Indicator of Many Visitors

Nur Atikah
Universitas Negeri Malang, *nur.atikah.fmipa@um.ac.id*

doi: https://doi.org/10.15642/mantik.2019.5.2.123-134

Abstract: Wilayah Metropolitan Malang adalah salah satu daerah di Jawa Timur yang merupakan tujuan wisata terkemuka di Indonesia dengan Kota Wisata Batu sebagai pusatnya. Mengingat perkembangan pariwisata di Malang, perlu dilakukan pengelompokan popularitas objek wisata sehingga dapat dijadikan referensi untuk pembuatan kebijakan oleh departemen pariwisata dan manajemen pariwisata. Pada artikel ini, pengelompokan dianalisis dengan menggunakan metode pengelompokan algoritma Expectation Maximation (EM). Data yang digunakan adalah data sekunder yang diperoleh dari data BPS, yaitu data banyak pengunjung wisata di Malang Raya. Hasil pengelompokan popularitas objek wisata unggulan di Malang didasarkan pada indikator jumlah pengunjung dibagi menjadi lima kelompok, ada Kelompok 1: Selecta; Grup 2: Balekambang, Pemandian Wendit dan Wisata Oleh-Oleh Brawijaya; Grup3: Museum Angkut, Coban Rondo, Museum Satwa, Taman Jatim, BNS, Petik Apel “Makmur Abadi dan Agro Kebun Teh Wonosari; Grup 4: Kusuma Agro Wisata, Kampoeng Kidz, Air Panas Cangar, Eco Green Park, Taman Hiburan Predator, Wana Wisata Coban Rais, Gunung Banyak, T-Shirt Mahajaya & Oleh-oleh, Ngliyep dan Bendungan Selorejo; Grup 5: Vihara "Dammadhipa Arama", Arung Jeram "Kaliwatu", Arung Jeram, Wana Wisata Coban Talun, Pemandian Tirta Nirwana, Pemandian Air Panas Alam Songgoriti, Wahana Air Rafting, Sahabat Air Rafting, Petik Apel Mandiri, Batu Agro Apel, Kampung Wisata.

Kata kunci: Algoritma EM, objek wisata, Malang Raya

***Abstract**: Malang Metropolitan Area is one of the areas in East Java which is a leading tourism destination in Indonesia with Batu Tourism City (Kota Wisata Batu) as the center. Considering the development of tourism in Malang, it is necessary to do a grouping of the popularity of tourism objects so that it can be used as a reference for making policy by the tourism department and tourism management. In this article, the grouping is analyzed by using the method of grouping the Expectation Maximation (EM) algorithm. The data used is secondary data obtained from BPS data, namely data of many tourism visitors in Malang Raya. The results of the grouping the popularity of leading tourism objects in Malang are based on indicators of the number of visitors divided into five groups, there are Group 1: Selecta; Group 2: Balekambang, Pemandian Wendit and Wisata Oleh-Oleh Brawijaya; Group3: Museum Angkut, Coban Rondo, Museum Satwa, Jatim Park, BNS, Petik Apel “Makmur Abadi and Agro Kebun Teh Wonosari; Group 4: Kusuma Agro Wisata, Kampoeng Kidz, Air Panas Cangar, Eco Green Park, Predator Fun Park, Wana Wisata Coban Rais, Gunung Banyak, Mahajaya T-Shirt & Oleh-oleh, Ngliyep and Bendungan Selorejo; Group 5: Vihara “Dammadhipa Arama”, Rafting “Kaliwatu”, Batu Rafting, Wana Wisata Coban Talun, Pemandian Tirta Nirwana, Pemandian Air Panas Alam Songgoriti, Wonderland Waterpark, Sahabat Air Rafting, Petik Apel Mandiri, Batu Agro Apel, Kampung Wisata.*

***Keywords:** EM algorithm, tourism object, Malang Raya*

.

How to cite: *N. Atikah, “Application of Expectation-Maximization (EM) Algorithm in Grouping Popularity Tourism Objects in Malang Raya Based on Indicator of Many Visitors”, mantik, vol. 5, no. 2, pp. 123-134, October 2019.*

1. INTRODUCTION

Nowadays, Malang is transformed into the second-largest city in East Java after Surabaya, which is only 90 kilometres away. This has become one of the causes for Malang City to grow as a trade and service area, education city and also as a city that has a creative economy industry. Malang City continues to morph into a Metropolitan City that has large magnets. Malang is no longer just an Apple City and a cool area at the foothill of Bromo and Semeru. Geographically, Malang City is very close even surrounded by Malang Regency and Batu City, then known as Malang Raya. Malang is a cool Metropolitan city. Therefore, it is reasonable if the Metropolitan Malang Raya area is a leading tourist destination in Indonesia with Batu Tourism City as the center. Malang Raya Area is the second largest metropolitan in East Java after Gerbang Kertosusila.

The Regency and City of Culture and Tourism Office (Disbudpar) revealed that the number of tourist visits in 2017 in Malang Regency reached 6.5 million people and Malang City reached 4 million people. While based on data from the Batu City Culture and Tourism Office, the number of tourists was recorded at 4.7 million people. The high level of tourist visits to Malang, which is a motivation for the Malang City Government, Malang Regency, and Batu City continue to concentrate on boosting tourism to increase Locally-Generated Revenue or PAD (Pendapatan Asli Daerah). Various existing potentials are packed again more interesting to be able to boost the economy of the community [1].

In building a good tourism industry and developing existing tourism that is better in quality and can provide many positive influences for the development of economic conditions, a specific strategy is needed to achieve it. Various important factors need to be seen and implemented in order to achieve a targeted and sustainable development and development plans, such as careful planning, effective strategies and objectives, revamping tourism objects, facilities, services to tourism promotion or marketing are important factors in supporting tourism development.

Tourist visitor data on each tourist attraction is the main and potential input for developing tourism. Of course, not all attractions are developed because of limited government funds. Thus, it is very necessary to be informed about tourism objects that are a priority of development. Of course, tourism grouping is needed, which will make it easier for managers to repair facilities and infrastructure that can increase the number of tourists. Statistically, grouping tourist objects can be done using one method in statistics, namely clustering. Some studies on clustering include Silvi grouping HIVAIDS indicators in Indonesia with Centroid Linkage and K-means Clustering methods that contain data outliers [2]. In addition to Centroid Linkage and K-Means Clustering, another method of grouping is using the Expectation-Maximization (EM) algorithm. The EM algorithm is an algorithm that functions to find the estimated Maximum Likelihood value of the parameters in a probabilistic model [3]. The advantage of the EM algorithm is that it can solve statistical problems such as estimating parameters for a combination of functions and parameters from incomplete data [4]. In this algorithm, there are two things that are used interchangeably, namely E-step that calculates the expectation value of likelihood including latent variables as if they exist, and M-step calculates the estimated value of ML from parameters by maximizing the expected value of likelihood which found in E-step.

The renewal of this paper is the application of the Expectation-Maximization (EM) method to find out the popularity of tourist objects in Malang Raya based on many visitors in 2018. Some previous studies using grouping with the EM algorithm include Darwianto & Sirait discussing implementation and clustering Expectation-Maximizationon algorithm analysis at the final assignment of Telkom University. The research aims to facilitate users in finding information on a large enough document, namely by grouping or categorizing documents according to the similarity of documents [5]. Soeyapto & Johari examines the application of data mining for data on the number of vehicles using the Expectation-Maximization (EM) algorithm in the Palembang City Dispenda [6]. The study aims to provide clearer information for the Dispenda party and simplify the analysis of the increase in the number of vehicle data by looking at the grouping of the number of vehicles in an area [7].

2. LITERATURE REVIEW

2.1 Definition of Multivariate Analysis

Multivariate analysis is one of the statistical techniques applied to understand data structures in high dimensions. Where are the intended variables these are interrelated with each other? According to Santoso, multivariate analysis can be defined simply as a method of processing variables in large quantities to find their influence on an object simultaneously [8]. Multivariate statistical analysis is a statistical method that allows conducting research on more than two variables simultaneously. Where there is at least one dependent variable and more than one independent variable, and there is a correlation between one variable and another.

As with other statistical analysis, the multivariate analysis also has types of data or scale data. The scale of the data used is of two kinds, namely metric data and non-metric data. Metric data is data that is numerical or contains numbers and mathematical calculations can be carried out in it. Metric data are also called numerical data or quantitative data. In this case, there are two kinds of metrics data, namely interval data and ratio data. While non-metric data is non-numeric data or also called qualitative data or categorical data. There are two types of non-metric data, namely nominal data and ordinal data.

2.2 Definition of Cluster Analysis

Cluster analysis is a multivariate technique whose purpose is to get object groupings by arranging objects into groups in such a way that in a group has maximum similarity (Rencher, 2002) [9]. In general, there are two types of methods used for clustering, namely hierarchical and non-hierarchical methods.

Hierarchical *cluster* analysis is done by grouping two or more objects that have the closest and the same thing so that the levels between groups become clearly visible. In hierarchical cluster analysis, the grouping results are displayed in the form of dendograms, while in non-hierarchical cluster analysis clustering begins by first determining the cluster to be formed. One of these methods can be used in this study, which is a non-hierarchical method, namely the analysis of the *Expectation-Maximization* (EM) algorithm. There are several things that need to be emphasized in the EM algorithm, namely, *Maximum Likelihood Estimation* (MLE), *Gaussian* Distribution, and *Expectation-Maximization* (EM).

2.3 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) was introduced by R. A Fisher in 1912. MLE is usually used to estimate parameter values that a function has, such as mean, variance, and so on. Bain and Engelhardt define MLE as follows [10]:

For example, X_1, X_2, X_3, ..., X_n is a random sample of a population with density $f\ (Xi\ ;\ \theta\ )$. Then the likelihood function is defined as a joint density function after data $\{x_1, x_2,\ \cdots, x_n\}$ is obtained. So that the shared density function is seen as a function of the parameters $\{\theta_1, \theta_2,\ \cdots, \theta_n\}$ and expressed by:

$$L(\theta_{1,}\theta_{2,} \dots, \theta_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

Thus, what MLE wants is to maximize the value of the likelihood function to obtain an estimator value from $\{\theta_1, \theta_2,\ \cdots, \theta_n\}$. This is very reasonable to understand because it is in accordance with determining the parameter estimator value that has the greatest chance. The steps to determine the estimator are in accordance with determining the optimal value in the differential calculation.

If this *likelihood* function is differentiated, then the possible *likelihood* estimator is $\hat{\theta}$ such that

$$\frac{\partial L(\hat{\theta})}{\partial \hat{\theta}} = 0$$

To prove that $\hat{\theta}$ really maximizes the *likelihood* function $L(\hat{\theta})$ it must be shown that:

$$\frac{\partial^2 L(\hat{\theta})}{\partial^2 \hat{\theta}} < 0$$

In many cases (especially if the value of the likelihood function is very large) where differentiation is used, it will be easier to work on the logarithm of $L(\hat{\theta})$ which is log $L(\hat{\theta})$. Of course, to determine the optimization of *likelihood* function remains the same as determining the maximum function of the logarithmic likelihood function, this is possible because the monotonous logarithm function rises at (0, ∞) which means that $L(\hat{\theta})$ has the same extreme, so to determine the *maximum likelihood estimator* from $\hat{\theta}$ as follows:

a. Determine the *likelihood* function

$$L(\theta_{1,}\theta_{2,} \dots, \theta_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

b. Forms a log-*likelihood* $l\ = \log L\big(\hat{\theta}\big)$

c. Determine the derivative of $l\ = \log L\big(\hat{\theta}\big)$ against $\hat{\theta}$

$$\frac{\partial \log\big[L\big(\hat{\theta}\big)\big]}{\partial \hat{\theta}} = 0$$

The completion of point 3 is the maximum *likelihood* estimator for $\hat{\theta}$.

d. Determine the second derivative of $l\ = \log L\big(\hat{\theta}\big)$ against $\hat{\theta}$. If $\frac{\partial^2 L(\hat{\theta})}{\partial^2 \hat{\theta}} < 0$, it will prove that $\hat{\theta}$ really maximizes the *likelihood* function.

2.4 Expectation *Maximation (EM) algorithm*

The *Expectation-Maximization* (EM) algorithm was first introduced by Dempster, Laird, and Rubin in 1977. According to Kusrini & Lutfi, the *Expectation-Maximization* (EM) algorithm is often used to find the *Maximum Likelihood* (ML) estimation of the parameters in a probabilistic model, where the model also depends on unknown latent variables [11]. In this algorithm, there are two things that are used interchangeably, namely E-step that calculates the expectation value of *likelihood* including *latent* variables as if they exist, and M-step calculates the estimated value of ML from parameters by maximizing the expected value of *likelihood* found in E-step.

The *Expectation-Maximization* (EM) algorithm has better properties than other approaches or methods. Some of the advantages of the EM Algorithm compared to other approaches include [12]:

a. The EM algorithm is more numerically stable, wherein each iteration the log-*likelihood* rises.
b. Under general conditions, EM algorithms converge to a reliable value. That is, by starting an arbitrary value $\theta^{(0)}$ it will almost always converge to a local *maximizer*, except wrong in taking the initial value $\theta^{(0)}$.
c. The EM algorithm tends to be easy to implement because it relies on calculating *complete data*.
d. The EM algorithms are easily programmed because they do not involve either integrals or derivatives of likelihood.
e. The EM algorithm only takes up a little *hard disk* space and memory on the computer because it doesn't use a matrix or its inverse in each iteration.
f. The analysis is easier than other methods.
g. Taking into account the increase in monotonous likelihood in iterations, it is easy to monitor convergence and program errors.
h. Can be used to estimate the value of *missing data*.

Although there are many advantages of the EM algorithm, there are weaknesses of the EM Algorithm, including:

a. It does not provide a procedure for generating covariance matrix estimates from parameter estimators.
b. The EM algorithm can converge slowly, that is if there is too much *incomplete information*.
c. The EM algorithm does not guarantee that it will converge to a maximum global value if there are multiple maxima.

The EM algorithm is a process that is divided into two steps, there are:

1) *Expectation* Step (E-Step)
   Search for expectation values to the likelihood function based on observed variables.
2) *Maximization* Step (M-Step) MLE
   A search of parameters by maximizing *likelihood* expectations generated from E-Step.

   Searching for parameters generated from the M-step will be used again for the next E-step, and this step will be repeated until it gives a convergent value and is an estimator of a parameter.

Suppose there is a sample of n items where $n_1$of the item is observed while $n_2 = n - n_1$ items are not observed. The observed items are denoted by $X' = (X_1, X_2, \dots, X_n)$ and unobserved items are denoted $' = (Z_1, Z_2, \dots, Z_n)$).. Assume X_{is} and Z_{js} are mutually independent and identically distributed variables (*independent and identically distribution*) with a probability density function $f(x|\theta)$, where $\theta \in \Omega$. Assume X_{is} and Z_{js} are mutually independent. Denote a function of the density of the combined probability of X with $g(x|\theta)$. Then $h(x, z| \theta)$ for the combined probability density function for observed and unobserved data. Whereas $k(z| \theta, x)$ represents the conditional probability density function notation of the missing data to provide observed data. Then it can be obtained

$$k(z| \theta, x) = \frac{h(x, z| \theta)}{g(x|\theta)} = 0$$

The observed *Likelihood* function of the data is

$$L(\theta|x) = g(x|\theta)$$

Then the likelihood function for complete data is defined by

$$L^c(\theta|x, z) = h(x, z|\theta)$$

Our goal is to maximize the *likelihood* function $L(\theta|x)$ by using the complete *likelihood* function $L^c(\theta|x, z)$ in the process. Use the equation $k(z| \theta, x)$, obtained

$$\log L(\theta|x) = \int \log L(\theta|x) \,.\, k(z|\theta_0, x) dz$$
$$\log L(\theta|x) = \int \log g(\theta|x) \,.\, k(z|\theta_0, x) dz$$
$$\log L(\theta|x) = \int [\log h(x, z|\theta) - \log k(z|\theta_0, x)] \,.\, k(z|\theta_0, x) dz$$
$$\log L(\theta|x) = \int \log h(x, z|\theta) k(z|\theta_0, x) dz - \int \log k\ (z|\theta, x). k(z|\theta, x)\, dz$$
$$\log L(\theta|x) = E_{\theta_0}[logL^c(\theta|x, z)|\theta_0, x] - E_{\theta_0}[logk(Z|\theta, x)|\theta_0, x]$$

Where expectations are taken under the conditional probability density function of $k(z|\theta_0, x)$. Then define the first part on the right side of the function above.

$$Q(\theta|\theta_0, x) = E_{\theta_0}[logL^c(\theta|x, z)|\theta_0, x]$$

The expectation defined by the Q function is called E-Step from the EM algorithm. Denoted $\hat{\theta}^{(0)}$ estimation initials of θ, based on the observed *likelihood* function. *Then* $\hat{\theta}^{(1)}$ becomes the argument that maximizes $Q(\theta|\theta_0, x)$. This is the first step to estimate θ then we define the EM algorithm as follows.

Denoted $\hat{\theta}^{(m)}$ in estimating step *m*. Then to estimate steps to (*m*+1):

a. *Expectation* Step (E-Step)

$$Q\left(\theta\middle|\hat{\theta}^{(m)}, x\right) = E_{\hat{\theta}^{(m)}}\left[logL^c(\theta|x, z)\middle|\hat{\theta}^{(m)}, x\right]$$

Where expectations are taken from the conditional probability density function $k\left(z\middle|\hat{\theta}^{(m)}, m\right)$

b. *Maximization* Step (M-Step)

$$\hat{\theta}^{(m+1)} = Arg \max Q\left(\theta\middle|\hat{\theta}^{(m)}, x\right)$$

Where,

$$Q(\hat{\theta}^{(m+1)}|\hat{\theta}^{(m)},x) \geq Q(\hat{\theta}^{(m)}\big|\hat{\theta}^{(m)},x\big)$$

The following are some of the advantages of using the EM algorithm [13]:

a. The EM algorithm is quite stable and easy to make the program.
b. In general, the EM algorithm has reliable convergence, means that it always converges almost to its local maximum point.
c. Requires a small storage capacity on the computer.
d. Can be used to estimate the value of lost data, because, in the EM algorithm, there is a process of distributing incomplete data to complete data based on the conditional opportunity value.

3. RESEARCH METHODS

The approach in this research proposal is quantitative research. The quantitative approach aims to test the theory, build facts, show relationships between variables, provide statistical descriptions, estimate and forecast results. The type of this research used is the descriptive quantitative study of existing problems, modelling in the input and output systems in WEKA Data Mining applications, solving problems and interpreting them. This study aims to classify the level of popularity of tourist attractions in Malang Raya based on indicators of many tourist attractions using the Expectation-Maximization (EM) algorithm. The steps to be carried out in this study are described as follows:

a. Input data into the WEKA Data Mining application 3.8.
b. It is using *clustering* methods that are used to determine the value of the cluster to be processed. The method used is by grouping data that has been inputted from the data.
c. Calculate the value of the *centroid cluster*. Determine the *cluster* value first if, after clustering, there is still data that changes then it is repeated again to the *cluster* iteration process.
d. Displays *clustering* grouping, aims to show the iterative process and *class* in the *cluster* with the number of *records* in the *store name.*

4. RESULTS AND DISCUSSION

In this study, the author will determine *clustering* into two groups, three groups, four groups and five groups. Grouping with two groups consists of groups of tourist objects with high popularity and tourist groups with low popularity. Grouping with three groups consists of groups of tourist objects with high levels of popularity, groups of tourist objects with moderate levels of popularity and groups of tourist objects with low levels of popularity. Grouping with four groups consists of a group of tourist objects with a very high level of popularity, a group of tourist objects with a high level of popularity, a group of tourist objects with low popularity and tourist groups with very low levels of popularity. Grouping with five groups consists of groups of tourist objects with a very high level of popularity, groups of tourist objects with a high level of popularity, groups of tourist objects with moderate levels of popularity, groups of tourists with low popularity and tourist groups with a high level of popularity very low.

Analysis with the *Expectation-Maximization* (EM) algorithm is calculated using the help of Weka *software* 3.8. The algorithm begins with the expectation step, namely initializing the initial value then iterating so that it reaches a convergent value. The results of the grouping are as follows:

Grouping with 2 Groups

Based on the analysis with the EM algorithm by using the help of Weka 3.8 s*oftware*, it is known that the iteration step carried out with two groups is 27 times so that the results obtained are group 1 which is included in the group of attractions with high popularity consisting of 22 attractions. Group 2 which is included in the group of tourist objects with a low level of popularity consists of 18 attractions. The division of group members is as follows:

Group 1 : Kusuma Agro Wisata, Selecta, BNS, Museum Satwa, Jatim Park, Petik Apel "Makmur Abadi", Air Panas Cangar, Museum Angkut, Eco Green Park, Predator Fun Park, Wana Wisata Coban Rais, Pemandian Tirta Nirwana, Wana Wisata Coban Talun, Gunung Banyak, Mahajaya T-Shirt & Oleh-oleh, Wisata Oleh-Oleh Brawijaya, Agro Kebun Teh Wonosari, Bendungan Selorejo Balekambang, Ngliyep, Pemandian Wendit and Coban Rondo.

Group 2 : Vihara "Dammadhipa Arama", Rafting "Kaliwatu", Kampoeng Kidz, Batu Rafting, Pemandian Air Panas Alam Songgoriti, Wonderland Waterpark, Sahabat Air Rafting, Petik Apel Mandiri, Batu Agro Apel, Kampung Wisata Kungkuk, Desa Wisata Sumberejo, Desa Wisata Bumiaji, Mega Star Indonesia, Wisata Oleh-oleh Deduwa, Candi Jago, Sengkaling, Pemandian Dewi Sri and Candi Kidal.

Grouping with 3 Groups

Based on the analysis with the EM algorithm by using the help of Weka 3.8 *software*, it is known that the iteration step carried out with three groups is 18 times so that the results are group 1 which belongs to the group of tourist objects with a high level of popularity consisting of 5 attractions. Group 2 is included in the group of tourist objects with a popularity level that is comprised of 17 attractions. Group 3 which is included in the group of tourist objects with a low level of popularity consists of 18 attractions. The division of group members is as follows:

Group 1 : Selecta, Wisata Oleh-Oleh Brawijaya, Balekambang, Pemandian Wendit and Coban Rondo.

Group 2 : Kusuma Agro Wisata, Jatim Park, Air Panas Cangar, BNS, Petik Apel "Makmur Abadi", Museum Satwa, Eco Green Park, Museum Angkut, Predator Fun Park, Gunung Banyak, Pemandian Tirta Nirwana, Wana Wisata Coban Talun, Wana Wisata Coban Rais, Mahajaya T-Shirt & Oleh-oleh, Agro Kebun Teh Wonosari, Ngliyep and Bendungan Selorejo.

Group 3 : Vihara "Dammadhipa Arama", Rafting "Kaliwatu", Kampoeng Kidz, Batu Rafting, Pemandian Air Panas Alam Songgoriti, Wonderland Waterpark, Sahabat Air Rafting, Petik Apel Mandiri, Batu Agro Apel, Kampung Wisata Kungkuk, Desa Wisata

Sumberejo, Desa Wisata Bumiaji, Mega Star Indonesia, Wisata Oleh-oleh Deduwa, Candi Jago, Sengkaling, Pemandian Dewi Sri and Candi Kidal.

Grouping with 4 Groups

The iteration step carried out with four groups is 29 times so that the results obtained are group 1 which belongs to the group of tourist objects with a very high level of popularity consisting of 6 attractions. Group 2 which is included in the group of tourist objects with a high level of popularity consists of 5 attractions. Group 3 which is included in the group of tourist objects with low popularity consists of 12 attractions. Group 4 which is included in the group of attractions with a very low level of popularity consists of 17 attractions. The division of group members is as follows:

Group 1 : Selecta, Museum Angkut, Wisata Oleh-Oleh Brawijaya, Balekambang, Pemandian Wendit and Coban Rondo.

Group 2 : Museum Satwa, Jatim Park, BNS, Petik Apel “Makmur Abadi and Agro Kebun Teh Wonosari.

Group 3 : Kusuma Agro Wisata, Kampoeng Kidz, Air Panas Cangar, Eco Green Park, Predator Fun Park, Wana Wisata Coban Rais, Pemandian Tirta Nirwana, Wana Wisata Coban Talun, Gunung Banyak, Mahajaya T-Shirt & Oleh-oleh, Ngliyep and Bendungan Selorejo.

Group 4 : Vihara “Dammadhipa Arama”, Rafting “Kaliwatu”, Batu Rafting, Pemandian Air Panas Alam Songgoriti, Wonderland Waterpark, Sahabat Air Rafting, Petik Apel Mandiri, Batu Agro Ape, Kampung Wisata Kungkuk, Desa Wisata Sumberejo, Desa Wisata Bumiaji, Mega Star Indonesia, Wisata Oleh-oleh Deduwa, Candi Jago, Sengkaling, Pemandian Dewi Sri and Candi Kidal.

Grouping with 5 Groups

In grouping with these five groups, there were no iterations made, so that the results obtained were group 1 which was included in the group of tourist objects with a very high level of popularity consisting of 1 tourist attraction. Group 2 which is included in the group of tourist objects with a high level of popularity consists of 3 attractions. Group 3 is included in the tourist attraction group with a popularity level that consists of 7 attractions. Group 4 which is included in the group of tourist objects with a low level of popularity consists of 10 attractions. Group 5 which is included in the group of attractions with a very low level of popularity consists of 19 attractions. The division of group members is as follows:

Group 1 : Selecta.

Group 2 : Balekambang, Pemandian Wendit and Wisata Oleh-Oleh Brawijaya.

Group 3 : Museum Angkut, Coban Rondo, Museum Satwa, Jatim Park, BNS, Petik Apel “Makmur Abadi and Agro Kebun Teh Wonosari.

Group 4 : Kusuma Agro Wisata, Kampoeng Kidz, Air Panas Cangar, Eco Green Park, Predator Fun Park, Wana Wisata Coban Rais, Gunung Banyak, Mahajaya T-Shirt & Oleh-oleh, Ngliyep and Bendungan Selorejo.

Group 5 : Vihara “Dammadhipa Arama”, Rafting “Kaliwatu”, Batu Rafting, Wana Wisata Coban Talun, Pemandian Tirta Nirwana, Pemandian Air Panas Alam Songgoriti, Wonderland Waterpark, Sahabat Air Rafting, Petik Apel Mandiri, Batu Agro Ape, Kampung Wisata Kungkuk, Desa Wisata Sumberejo, Desa Wisata Bumiaji, Mega Star Indonesia, Wisata Oleh-oleh Deduwa, Candi Jago, Sengkaling, Pemandian Dewi Sri and Candi Kidal.

With the grouping of 5 groups, it was found that the iteration process had stopped (no more iterations). This shows that the grouping of attractions divided into five groups is the maximum result.

Selection of the Best Model

In the Expectation-Maximization (EM) algorithm, the best model represents the best data or model is indicated by the largest log-*likelihood* value. The following is the log-likelihood value of each cluster.

Tabel 4.1 Log *Likelihood* Value

Number of *Clusters*	**Log-*Likelihood***
2	-12,80051
3	-12,66082
4	-12,5927
5	-12,09286

Based on Table 4.1, it can be seen that testing with five groups has the largest log-*likelihood* value, so the best model is the *Expectation-Maximization* (EM) algorithm testing with five groups that produce one tourist attraction with a very high level of popularity, three attractions with popularity the high, seven tourist objects with moderate popularity, ten attractions with low popularity and 19 tourist attractions with low popularity.

5. CONCLUSION AND RECOMMENDATION

The conclusions from this study are: (a) the application of the *Expectation-Maximization* (EM) algorithm in grouping the popularity of leading tourist objects in Malang Raya based on indicators of many visitors was carried out using WEKA 3.8 *software* assistance. The number of groupings used was two groups, three groups, four groups and 5 groups. Based on the log *likelihood* value, it was found that the group with 5 groups had the largest log-*likelihood* value, so the model suitable for this research was a model with 5 groups, namely a group of tourists with very high popularity, a group of tourists with high popularity, groups tourist attraction with a moderate level of popularity, a group of tourist objects with low popularity and tourist groups with very low levels of popularity, (b) the results of grouping the popularity of leading tourist objects in Malang Raya based on indicators of many visitors using the *Expectation-Maximization* (EM) algorithm are as follows:

Group 1	:	Selecta.
Group 2	:	Balekambang, Pemandian Wendit and Wisata Oleh-Oleh Brawijaya.
Group 3	:	Museum Angkut, Coban Rondo, Museum Satwa, Jatim Park, BNS, Petik Apel "Makmur Abadi and Agro Kebun Teh Wonosari.
Group 4	:	Kusuma Agro Wisata, Kampoeng Kidz, Air Panas Cangar, Eco

Green Park, Predator Fun Park, Wana Wisata Coban Rais, Gunung Banyak, Mahajaya T-Shirt & Oleh-oleh, Ngliyep and Bendungan Selorejo.

Group 5 : Vihara "Dammadhipa Arama", Rafting "Kaliwatu", Batu Rafting, Wana Wisata Coban Talun, Pemandian Tirta Nirwana, Pemandian Air Panas Alam Songgoriti, Wonderland Waterpark, Sahabat Air Rafting, Petik Apel Mandiri, Batu Agro Ape, Kampung Wisata Kungkuk, Desa Wisata Sumberejo, Desa Wisata Bumiaji, Mega Star Indonesia, Wisata Oleh-oleh Deduwa, Candi Jago, Sengkaling, Pemandian Dewi Sri and Candi Kidal.

Based on the conclusions obtained, it appears that groups 4 and 5 are in the form of tourist objects that are related to low lands and are consumptive. Whereas group 1, group 2, and group 3 are attractions related to water and the beauty of nature in both the highlands and waters (sea, waterfall). Therefore, we need to recommend: (a) to the local government of Malang Raya to prioritize the improvement of services in tourism objects related to the waters (beaches, seas, and waterfalls) and to further explore the beauty of nature, (b) for other researchers, it is recommended to add indicators that influence the popularity of tourist objects in Malang to represent the characteristics of each tourist attraction better, and (c) the use of the EM estimation method is possible to obtain endless iterations. To minimize this occurrence, it is recommended to use other methods to group the popularity of attractions in Malang Raya.

Acknowledgement

We gratefully acknowledge to Faculty of Mathematics and Sciences of Malang State University of the year 2018.

References

[1] E. Kurniawan, "*Malang City Government Concentration on Boosting Tourism to Increase PAD*". 2017. https://malangtoday.net/malang-raya/kota-malang/dongkrak-pariwisata-untuk-tingkatkan-pad/amp/

[2] R. Silvi, "Analisis Cluster dengan Data Outlier Menggunakan Centroid Linkage dan K-Means Clustering untuk Pengelompokkan Indikator HIV/AIDS di Indonesia", *mantik*, vol. 4, no. 1, pp. 22-31, May 2018.

[3] S. Borman, "*The Expectation Maximization Algorithm: A Short Tutorial*", July 2006.

[4] T. A. Kusuma and Suparman. "Algoritma Expectation-Maximization (EM) untuk Estimasi Distribusi Mixture", *Jurnal Konvergensi*, Vol. 4, No. 2 Oktober 2014.

[5] R. E. D. Sirait, E. Darwianto, and D. D. J. Suwawi, "Implementasi dan Analisis Algoritma Clustering Expectation–Maximization (EM) pada Data Tugas Akhir Universitas Telkom", *e-Proceeding of Enginering*, Vol. 2, No. 2, Agustus 2015.

[6] I. Johari, D. Soeyapto, and Mardiani, "Penerapan data Mining untuk Data Jumlah Kendaraan Menggunakan Algoritma Expectation Maximization (EM) pada Dispenda Kota Palembang", STMIK MDP, 2015.

[7] Clustering, K. “Implementation and Analysis of Clustering Expectation - maximization (EM) Algorithms on Telkom University Final Project Data”, Vol. 2, No. 2, pp. 6711–6717, 2015.

[8] S, Santoso, “*Statistik Multivariat: Konsep dan Aplikasi dengan SPSS*”, Jakarta. Elex Media Komputindo, 2004.

[9] A. C. Rencher, “*Method of Multivariate Analysis (Second Edition)*”, New York: John Wiley and Sons, Inc. 2002.

[10] L. J. Bain and M. Engelhardt, “*Introduction to Probablity and Mathematical Statistcs*”. California: Duxbury Press. 1992.

[11] Kusrini and E. T. Luthfi, “Algoritma Data Mining”, Yogyakarta: Andi, 2009.

[12] H. Glanz, H and L. Carvalho, “An expectation-maximization algorithm for the matrix normal distribution with an application in remote sensing”. *Journal of Multivariate Analysis*, Vol. 167, pp. 31-48, September 2018, doi: 10.1016/j.jmva.2018.03.010.

[13] G. J. McLachlan and T. Krishnan, “*The EM Algorithm and Extensions*”, John Wiley & Sons, Hoboken, 2008, doi: 10.1002/9780470191613.