Mapping of the Reading Literacy Activity Index in East Java Province, Indonesia: an Unsupervised Learning Approach

H A Azies[1*], A F D Rositawati[2]

[1] Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS), Jalan Raya ITS Sukolilo 60111, Surabaya, Indonesia
[2] Research and Development Agency, Ministry of Home Affairs, Jalan Kramat Raya 10430, Jakarta, Indonesia

*Corresponding author's e-mail: harunalazies@gmail.com, harunazies.206003@mhs.its.ac.id

Abstract. One of the educational problems that must be faced by East Java Province is the low reading culture of the community. The level of reading culture can be indicated by the Reading Literacy Activity Index (Alibaca Index). Alibaca Index of East Java is only 33.19 which value is included in the low category. So, this research uses the indicators that compose the Alibaca Index to classify regencies/cities in East Java Province. The analysis process carried out in this research uses one of the unsupervised learning algorithms, namely the K-Means algorithm. Analysis using the K-Means algorithm for grouping regencies/cities in East Java Province based on the indicators that compose the Alibaca index gives the results that the regencies/cities of East Java Province are divided into 3 clusters based on the optimal number of clusters according to the result of the elbow and silhouette method. Cluster 1 consists of 20 regencies and cities, cluster 2 consists of 10 regencies, and cluster 3 consists of 8 cities. Each cluster has different characteristics, cluster 1 is the cluster with the lowest skill dimension, while the cluster 2 area is an area that dominates the access dimension, alternative dimension, and cultural dimension, meanwhile, the third cluster does not have dominance in these 3 dimensions, which means that cluster 3 is the government's priority for improving reading activities, so the result of the analysis can help the government to develop strategic policies to achieve educational equity, especially concerning literacy levels based on the characteristics of each regency/city in East Java Province.

1. Introduction

Education gives a big contribution to the success of a nation. The level of educational attainment can be shown through the literacy level in the community. Literacy has a high correlation with educational attainment. The higher the literacy level, the better the educational attainment. One of educational attainment in Indonesia is the eradication of illiteracy with significant results. This is evidenced by the increasing Literacy Rate in Indonesia as indicated by figure 1. The diagram shows that the Literacy Rate in Indonesia increasing continuously from year to year. This information shows that more Indonesians are literate, meaning that the government has succeeded in increasing the eradication of illiteracy so the education access can be expanded. However, the success of the government in eradicating illiteracy and expanding access to education has not been followed by the success in cultivating a reading culture in the community.

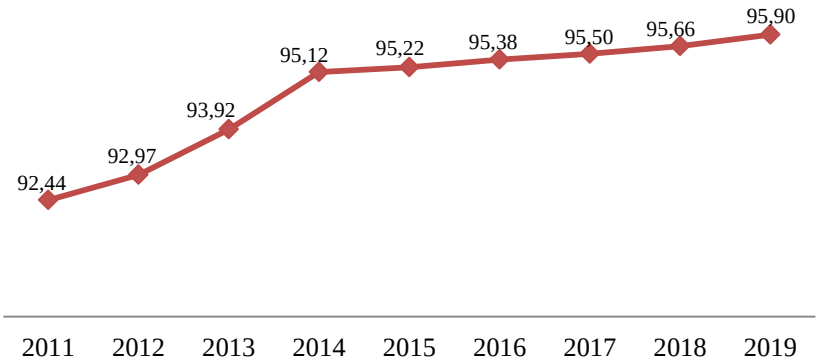

Figure 1. Literacy Rate in Indonesia (Source: bps.go.id, 2011 – 2019).

Various surveys show unsatisfactory results. Survey of Programme for International Student Assessment (PISA) in 2015, placed Indonesia at 64th of 72 countries. During the period 2012 – 2015, the PISA score for reading in Indonesia only increased 1 point from 396 to 397. The test results show that the ability to understand the reading materials, especially text documents, in Indonesian children aged 9-14 years is in the bottom ten [1]. The same issue also happened in East Java Province, where the Literacy Rate in East Java Province was already high but the reading culture in the community was still low. The Literacy Rate of each regency/city in East Java in 2019 is shown in figure 2.

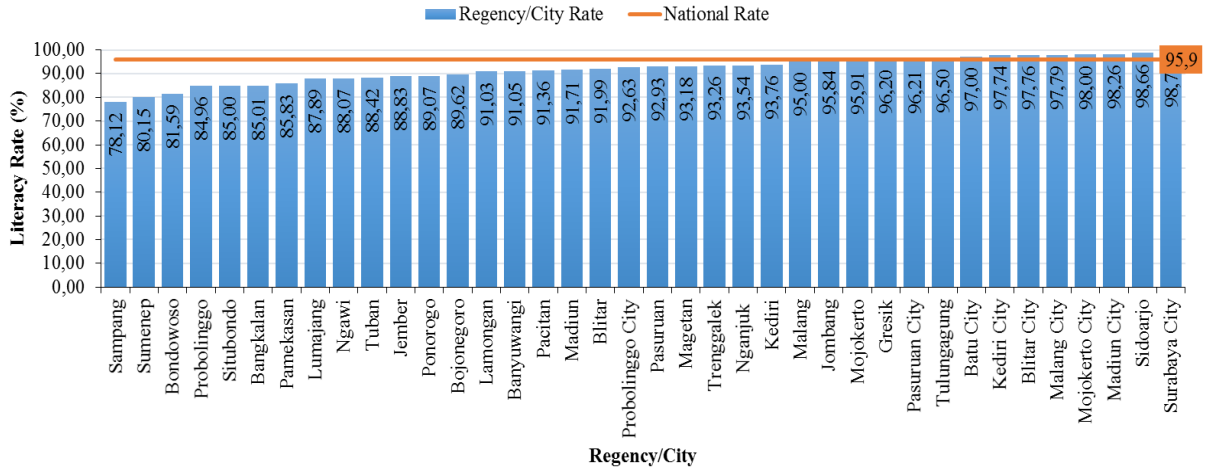

Figure 2. Literacy Rate in Each Regency/city in East Java Province, Indonesia

(Source: jatim.bps.go.id, 2019).

Figure 2 shows that the Literacy Rate in each regency/city in East Java has reached a number above 80 except for one regency, is Sampang Regency which has a Literacy rate of 78.12, but the figure of 78.12 can be said that Literacy Rate in Sampang Regency is also high. This means that many residents of East Java Province were literate, but unfortunately, it is not followed by a high reading culture. The level of reading culture can be indicated by the Reading Literacy Activity Index (Alibaca Index). Reading Literacy Activity Index (Alibaca Index) of East Java Province only 33.19 which is in the low category, and if the province in Indonesia is sorted from the one with the highest index to the lowest, then the 33.19 figure put East Java province at rank 26 of 34 provinces in Indonesia, or in other words, East Java Province occupies the 9th lowest position among other provinces [2].

The problem of low the Reading Literacy Activity Index (Alibaca Index) of East Java is one of the problems that must be completed by all residents of East Java Province, especially the Governor of East Java, who makes education one of the points contained in Nawa Bhakti Satya. Nawa Bhakti Satya

was the nine main programs in the governor campaign promise during the East Java gubernatorial election last year. So, it is necessary to do research that could help the government in making decisions to increase the Alibaca Index of East Java Province. So, one of the points Nawa Bhakti Satya namely East Java Smart can be realized.

Therefore, this research was conducted with the purpose to help the government in realizing the East Java Smart by clustering regencies/cities in East Java Province based on the indicators that compose the Alibaca Index. The clustering method is used because the educational attainment in each regency/city in East Java Province is different which can be caused by the gap in educational resources and facilities between regions, community groups, and socioeconomic levels. So, it is necessary to know which regencies/cities have the same characteristics and are in the same cluster. Thus, the education policies taken can be adjusted to the characteristics of educational attainment in each regency/city, and the policies can be equated to regencies/cities that are in the same cluster and have the same characteristics. So, in this case, the cluster method can be useful for determining which areas have indicators of the Alibaca index that is already high, and which areas have is still low and must be prioritized by the government.

One of the previous research that provided benefits from using the Cluster method was conducted by Soemartini and Supartini, that research aimed to find out which areas in West Java were prioritized to get assistance from the Government, and the results of clustering with the K-Means method shows that regencies/cities in the second cluster require more assistance than regencies/cities in the first cluster [3]. Therefore, the cluster method used in this research will also be able to provide benefits to find out which areas in East Java should be prioritized by the government in increasing the community's reading activities. Other research related to K-Means in the field of education also has been carried out, including Liu in 2017 used the K-Prototypes clustering method related to distance education [4]. Shovon in 2012 by comparing k-means clustering and decision trees related to student academic achievement [5], with the same subject Rani et al. in 2021 using K-Means and FP Growth [6]. Meanwhile, a new development related to K-means by Sinaga in 2020 [7] and by Wu in 2020 [8]. Based on several previous researchers, no research was found that groups the province of East Java with any of the unsupervised learning approaches, namely K-means in the case study of the Reading Literacy Activity Index (Alibaca Index). Therefore, this study will be the first study to explore the results of the Alibaca cluster of indices in East Java using the K-Means method.

2. Literature Review

This section will explain Reading Literacy Activities Index (Alibaca Index), and previous research that is relevant to this research.

2.1. Reading Literacy Activity Index (Alibaca Index)

The problem of low reading culture is one of the important issues in understanding the low level of literacy of the Indonesian people, including the inhabitants of East Java. It is believed that people who have a high reading culture also have a high level of literacy. However, to encourage people to have a high reading culture, several prerequisites are needed. The efforts to improve people's interest in reading must start with the effort to improve people’s ability to read. The ability to read is a prerequisite for accessing reading. After having the ability to read, then further is developing the habit of reading. Efforts to develop the habit of reading is cannot be done without the availability of materials for reading and other supporting facilities. So, the culture of reading does not grow by itself but requires several components, there are: (1) the ability to read, (2) the availability of materials reading, and (3) the development of reading habits. Without one of the three components, will be difficult to build a reading culture [9].

Another reference mentions the same thing that the culture of reading does not grow by itself, there are four dimensions that can influence the occurrence of literacy activities. The fourth dimensions are among others [10]:

- The skill dimension is the first requirement for someone to be able to access and read literacy resources.

- The Access dimension is supporting resources where the community can take advantage of literacy resources, such as libraries or bookstores.
- The alternative dimension is a variety of information and entertainment technology options. Alternative here can be interpreted as another option provided by electronic and digital devices in accessing literacy resources.
- The cultural dimension is ideas, values, norms, and meanings formed by families, communities, and the wider environment that also influence literacy behavior. In this case, culture is interpreted as an effort to form literacy habits.

Four dimensions that were described above had a role important and related to each other in supporting the literacy activities, so the absence of one dimension will affect the function of the other dimensions. For example, reading skills and abilities will influence how access to reading materials can be optimally utilized. Likewise, the ability to access various alternative information technology will affect how the technology is used to access information. So, the Reading Literacy Activity Index (Alibaca Index) is composed of four dimensions, that is Skill Dimension, Access Dimension, Alternative Dimension, and Cultural Dimension. Each dimension is considered as a factor that jointly supports the occurrence of reading literacy activities [1].

2.2. Previous Research

Many previous kinds of research using the K-Means Clustering method have been carried out, one of them is research to Clustering the Illiteracy Rate in Indonesia. The purpose of that research is to cluster the provinces in Indonesia by the level of the Illiteracy Rate, so that can be prioritized to areas with a high level of Illiteracy Rate. The result of that research gives information that the K-Means Clustering method can classify the provinces in Indonesia based on the level of the Illiteracy Rate. So the result can help the Government for taking strategic policy in achieving educational equity in Indonesia [11]. Another research by the method of K-Means Clustering has also been done with the title of Cluster Analysis with K-Means Methods to cluster the regencies/cities in Maluku Province using the Human Development Index (HDI) indicators. That research gives information about the benefits of the K-Means algorithm to clustering the regencies/cities in Maluku Province based on the same characteristics of the area in terms of the four indicators of the HDI. The four indicators are Life Expectancy, Per Capita Expenditure, Literacy Rate, and Average Duration of School. Two of them are measures in the field of education, that is the Literacy Rate and Average Duration of School [12]. In addition, research related to clustering related to participation rates in the Indonesian province was carried out using the K-Means method with Rapidminer. The analysis gives the result that the provinces in Indonesia are clustered into three clusters, with DI Yogyakarta as the highest cluster and North Kalimantan as the lowest cluster, while the remaining 32 provinces form clusters [13].

In addition, the research carried out by Monica et al comparing several unsupervised learning methods including K-Means, K-medoids, and Self-organizing map (SOM), the results of this study indicate that the K-means method is the most optimal method based on the silhouette coefficient value, compared to the other two unsupervised learning methods [14]. Based on the two studies previously mentioned, it can be seen that the K-Means Clustering method able to cluster observations or regions based on the indicators of education like Literacy Rate, Illiteracy Rate, and Average Duration of School. So, in this research, regencies/cities in East Java Province will be grouped based on one measure of education, namely the Reading Literacy Activity Index (Alibaca Index). Research on the Reading Literacy Activity Index (Alibaca Index) has never been done before because this index is a new measure in the field of education that was just published in 2019 by the Ministry of Education and Culture.

3. Material and Method

This section will explain the data source and research variable. Besides that, it also explains the method that is used in this research, which is Cluster analysis.

3.1. Data Source and Research Variable

The data in this research were obtained as a secondary from the publication of the Central Statistics Agency of East Java Province in 2019. The research data used is an indicator of the Reading Literacy Activity Index (Alibaca Index) that consists of four dimensions, which are the Skills Dimension, the Access Dimension, the Alternative Dimension, and the Cultural Dimension. The indicators for each dimension used are described in Table 1.

Table 1. Indicators for Each Dimension of the Alibaca Index.

Dimension	Indicator	The Name of Indicator
Skill	X1	Literacy Rate
	X2	Average Duration of School
Access	X3	Number of Schools
	X4	Number of Villages According to Availability of Community Reading Gardens
Alternative	X5	Number of Villages According to Base Transceiver Station (BTS)
	X6	Number of Villages According to Cellular Phone Signal Presence
	X7	Number of Villages According to the Presence of GSM or CDMA Internet Signal
Cultural	X8	Number of Villages According to Availability of Illiteracy Eradication Activities
	X9	Number of Villages According to Availability of Al-Quran Education Park
	X10	Number of Villages According to Television and Radio Programs Accepted by Residents
	X11	Number of Villages According to Availability of Educational Activities Package A/B/C
	X12	Number of Villages According to Availability of Playgroups

- The Skill Dimension explains the level of community skills in accessing reading materials, this dimension is described by two indicators: (1) Literacy Rate, and (2) Average Duration of School.
- The Access dimension explains the availability of literacy resources both at school and in the community, this dimension is described by two indicators: (1) Number of Schools, and (2) Number of Villages According to Availability of Community Reading Gardens.
- The Alternative dimension explains the options or possibilities provided by electronic and digital devices in accessing information both at school and in the community. To be able to access information electronically and digitally, a signal or internet network must be available. So, the indicators that can describe the Alternative dimension are (1) Number of Villages According to Base Transceiver Station (BTS), (2) Number of Villages According to Cellular Phone Signal Presence, and (3) Number of Villages According to the Presence of GSM or CDMA Internet Signal.
- The Cultural dimension explains the extent to which people's habits or behavior in accessing literacy materials. The availability of activities or programs in the community is expected to improve people's habits or behavior in accessing literacy materials. So, in this case, the cultural dimension can be described through indicators (1) Number of Villages According to Availability of Illiteracy Eradication Activities, (2) Number of Villages According to Availability of Al-Quran Education Park, (3) Number of Villages According to Television and Radio Programs Accepted by Residents, (4) Number of Villages According to Availability of

Educational Activities Package A/B/C, and (5) Number of Villages According to Availability of Playgroups.

3.2. Cluster Analysis

One of the statistical methods used for clustering of observations is cluster analysis [15]. Cluster analysis is a data mining method that is used to search for data and then group them based on the similarity between one data and others. So, observations in the same group have relatively homogeneous than observations in the different groups [16]. Cluster analysis consists of two types, that is hierarchical and non-hierarchical clustering. In hierarchical clustering, it is not yet known how many groups will be formed. While in the non-hierarchical clustering, it is already known how many groups will be formed. This research will use one of the non-hierarchical methods, that is K-Means Clustering.

3.2.1 K-Means Clustering. K-means clustering is one of the non-hierarchical methods that partition the observation into one cluster or more, so the observation that has the same characteristics will be grouped into one group, and the observation that has the different characteristics will be grouped into another group [17]. K-Means Clustering grouped the data into the k groups based on the centroid of each group [18]. K-Means method is the clustering algorithm based on the distance that divides the data into the clusters and the algorithm is only working on the numeric attributes [19]. Grouping the data by the K-Means method follow these algorithms below [20]:

- Define the number of groups
- Allocate the data to the group by random
- Calculate the center of the group or the centroid for each group. The centroid in each group is calculated from the average of all data for each feature. If M represents the number of data in a group, i represents the i-th feature in a group, and p represents the data dimensions, then the centroid of the i-th feature is calculated using equation (1).

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \tag{1}$$

  calculations using equation (1) are performed as many as the data dimensions which is represented by p. So, the calculations are performed from $i = 1$ until $i = p$.
- Calculate Euclidean to know the distance of the data to the centroid. Once the distance of the data to the centroid is known, then each data is allocated to the closest centroid. Euclidean can be calculated using equation (2).

$$d = [(x_1 - x_2)^2 + (y_1 - y_2)^2]^{\frac{1}{2}} \tag{2}$$

  compare the distance of the data to the centroid of each existing group, then reallocate the data into each group. Data will be reallocated to the group that has centroid which closest to the data. According to [1] this allocation can be determined using equation (3).

$$a_{i1} = \begin{cases} 1, & d = min\{D(x_i, c_1)\} \\ 0, & \text{other} \end{cases} \tag{3}$$

  a_{i1} is the x_i point membership value to the c_1 centroid, d represents the closest distance from the data x_i to K groups, and c_1 is the 1st centroid. The membership value of the data in the group and the distance are then used to determine the objective function used in the K-Means method. According to [20] the objective function can be calculated using equation (4).

$$J = \sum_{i=1}^{n}\sum_{i=1}^{k} a_{ic}\, D(x_i, c_1)^2 \tag{4}$$

  where n represents the number of data, while the number of the group is represented by k, a_{i1} is the value membership of a point x_i to a group c_1 that followed. a has a value of 0 or 1. The value $a_{i1} = 1$ if the data is a member of a group. If not, then the value $a_{i1} = 0$.
- Go back to the 3rd step if there are still data that move to other groups or if there is a change in centroid position.

3.2.2 The Elbow Method. It is important to decide the optimal number of clusters. One method that can show the optimal number of clusters is The Elbow method. This method works by considering the percentage of comparative results among the number of clusters that will shape an elbow at a point. The elbow method will choose the value of the cluster and then add the value of the cluster to be applied as a model of data in deciding the optimal number of clusters. Here are the steps in defining the value of *K* in the K-means using the Elbow method [21]:

- Determine the initial *K* value
- Increase the *K* value
- Do the calculation using equation (5) to get the Sum of Square Error (SSE) from each *K* value

$$SSE = \sum_{K=1}^{k} \sum_{x_i e S_k} \|x_i - c_k\|_2^2 \quad (5)$$

- Note the SSE result of the *K* value that dropped drastically
- Determine the *K* value in the form of an angle.

3.2.3 Silhouette Coefficient Method. With this method, it will be known the power and quality of a cluster and can show how optimally an observation is entered into a certain cluster. Here are the steps to calculate the Silhouette Coefficient [21]:

- Get the average distance between one observation with all other observations that are in the same cluster by doing a calculation using equation (6).

$$a(i) = \frac{1}{|A|-1} \sum j \, \epsilon_{A, j \neq i} d(i,j) \quad (6)$$

  equation (6) above gives an example to calculate the average distance between observation *i* with observation *j*. Both observation *i* and *j* are in cluster *A*. While, $d(i,j)$ indicates the distance from observation *i* to observation *j*.
- Get the average distance between observation *i* with all other observations that are in the different clusters by doing a calculation using equation (7), then select the smallest average distance value.

$$d(i,C) = \frac{1}{|A|} \sum j \, \epsilon \, C \, d(i,j) \quad (7)$$

  where *C* is another cluster that is different from cluster *A*, and the average distance between observation *i* with all observations in another cluster *C* is indicated by $d(i,C)$, and value $b(i)$ is calculated using equation (8).

$$b(i) = \min C \neq A \, d(i,j) \quad (8)$$

- The last step is doing calculations using equation (9) to get the Silhouette Coefficient.

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (9)$$

4. Result and Discussion

In this section, we will discuss the results of the analysis consisting of an overview of the literacy conditions, the results of the K-Means clustering analysis, and the mapping of the clustering results based on the indicators of the reading literacy activity index in the province of East Java.

4.1. Overview of Literacy Conditions in East Java Province

The problem of low reading culture is one of the important issues in understanding the low level of literacy of the Indonesian people, including the inhabitants of East Java Province. It is believed that people who have a high reading culture also have a high level of literacy.

However, to encourage people to have a high reading culture, several prerequisites are needed. There are four dimensions that can influence the occurrence of literacy activities that were described in the previous section. Through the Reading Literacy Activity Index (Alibaca Index), those dimensions become a starting point for further research of the dynamics and development of community literacy in East Java Province.

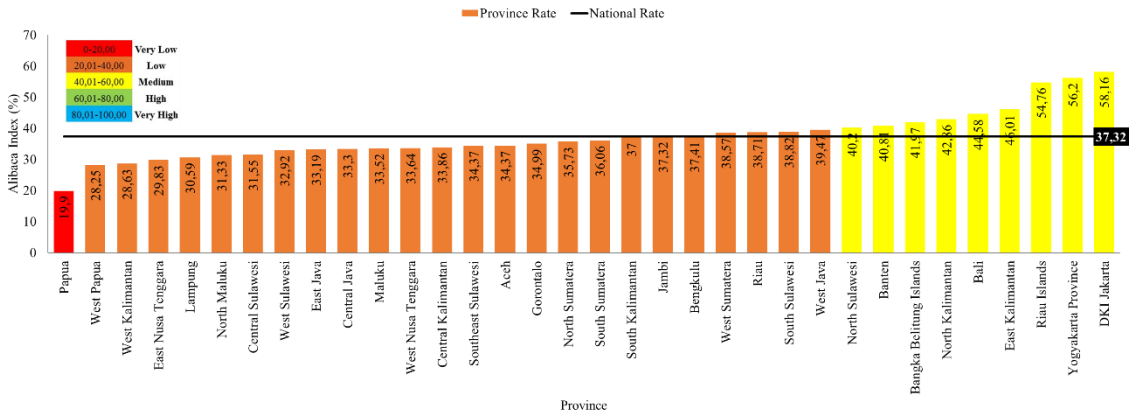

Figure 3. Provincial Alibaca Index by Rank from High to Low (Source: Ministry of Education and Culture, 2019).

Data from the Ministry of Education and Culture in 2019 shows that the Alibaca index of East Java Province is in the category of low, which is 33.19 (Figure 3), this value is also lower than the Alibaca index of National (Figure 3). The Alibaca index of East Java Province is composed of four dimensions, that is the Skill Dimension of 71.69, the Access Dimension of 15.99, the Alternative Dimension of 43.54, and the Cultural Dimension of 24.32 (Figure 4). The access dimension is the dimension with the lowest value, even the East Java province is included in the 5 provinces with the lowest value of the access dimension at the national level.

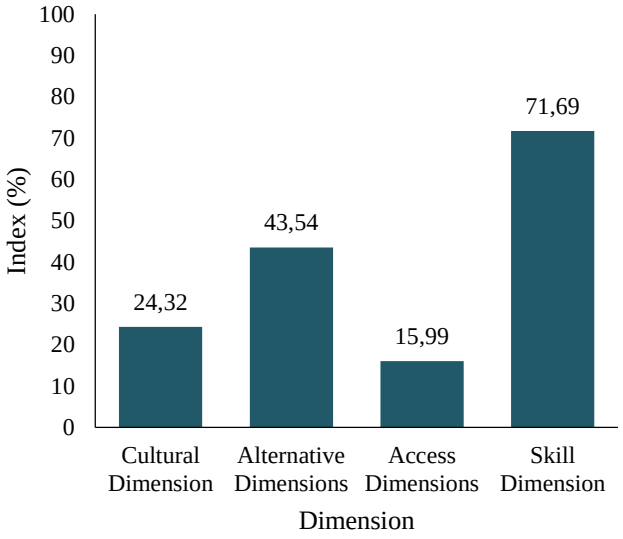

Figure 4. Alibaca Index of East Java by Dimension (Source: Ministry of Education and Culture, 2019).

4.2. K-Means Clustering Algorithm Analysis of The Reading Literacy Activity Index Indicator in East Java Province

Cluster analysis is an analytical method for classifying observation objects into several clusters (groups) based on the characteristics that these objects have. Between observations in a cluster are homogeneous while between clusters are mutually heterogeneous. The method used in this research is the K-means clustering algorithm which is one of the popular algorithms of unsupervised machine

learning. In this research, certain methods are used to determine the number of clusters, those are the elbow method and the silhouette coefficient method.

Based on figure 5, it can be seen that the line has a fracture that forms a bend at k = 3. Thus, using this method, the optimal cluster is obtained when it is at k = 3. Then the silhouette method is used, this method is to determine the quality of the cluster, which is indicated by the average value of the silhouette coefficient being the maximum. The following is the result of optimizing clusters with the Silhouette method.

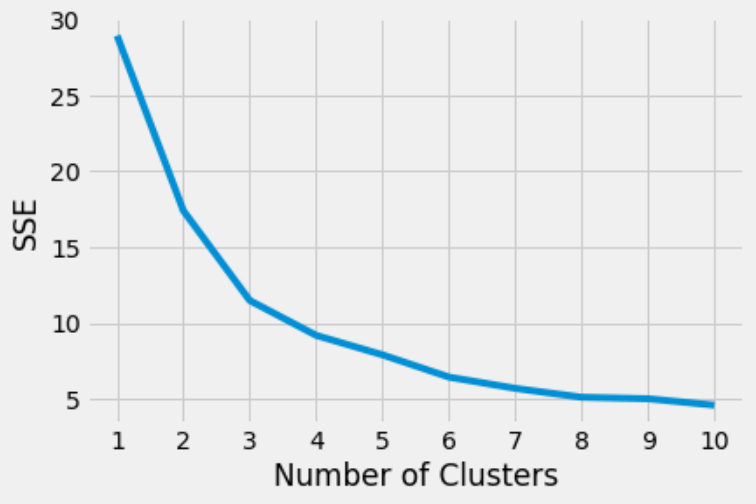

Figure 5. The results of Applying the Elbow Method to Determine the Optimum Number of Clusters.

Determination of the optimal number of clusters in K-Means clustering using the silhouette coefficient method shows that the optimal number of clusters is 3 clusters. This can be seen because the highest silhouette value [22] lies in the number of clusters of 3 clusters with an average silhouette value of 0.366 (Figure 6).

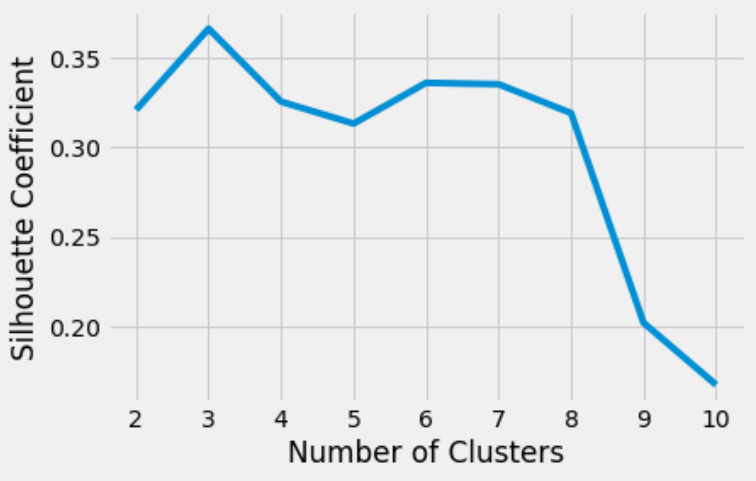

Figure 6. Results of Application of the Silhouette Method.

Based on the results of determining the optimal number of clusters using the elbow and silhouette coefficient methods, it is found that the optimal number of clusters is 3 clusters. (Figure 7) visual results of the three-cluster K-means algorithm (k=3) using python software.

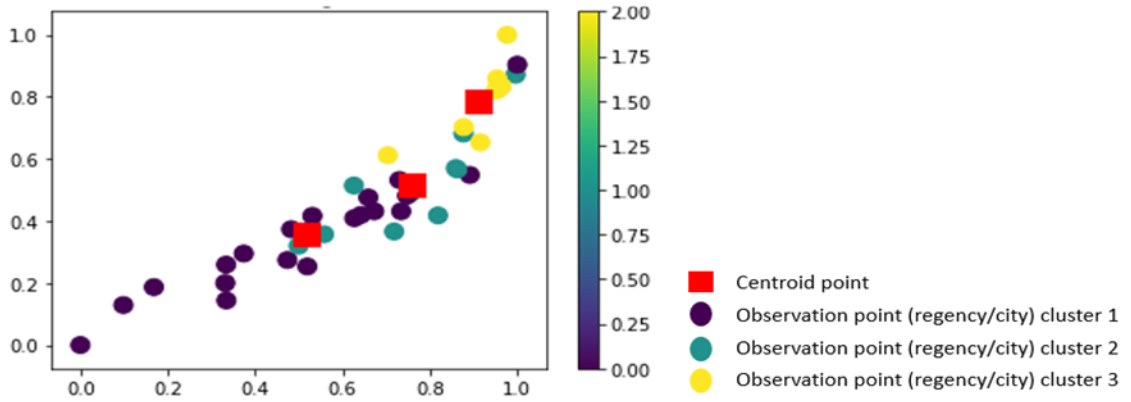

Figure 7. Clustering Results of Regencies/cities in East Java Using the K-Means Algorithm.

Based on figure 7, it can be seen that 38 regencies/cities in East Java Province were divided into 3 clusters which were formed with the composition of the members of each cluster shown in Table 2.

Table 2. The results of the Regency/City grouping are based on the cluster division.

Cluster	Members of the Cluster	Number of Regencies/Cities
Cluster 1	Bangkalan, Banyuwangi, Blitar, Bondowoso, Jember, Lumajang, Madiun, Magetan, Nganjuk, Ngawi, Pacitan, Pamekasan, Ponorogo, Probolinggo, Sampang, Situbondo, Sumenep, Surabaya City, Trenggalek, Tulungagung	20 (52,63%)
Cluster 2	Bojonegoro, Gresik, Jombang, Kediri, Lamongan, Malang, Mojokerto, Pasuruan, Sidoarjo, Tuban	10 (26,32%)
Cluster 3	Batu City, Blitar City, Kediri City, Madiun City, Malang City, Mojokerto City, Pasuruan City, Probolinggo City	8 (21,05%)

4.3. Mapping the Clustering Results of The Alibaca Index in East Java Province

After obtaining the optimal cluster results with the number of clusters of 3 clusters, the next step is to analyze the results of the formed clusters. Figure 8 is a mapping of the results of clustering using the k-means algorithm method, the visualization of the mapping using the ArcView GIS software.

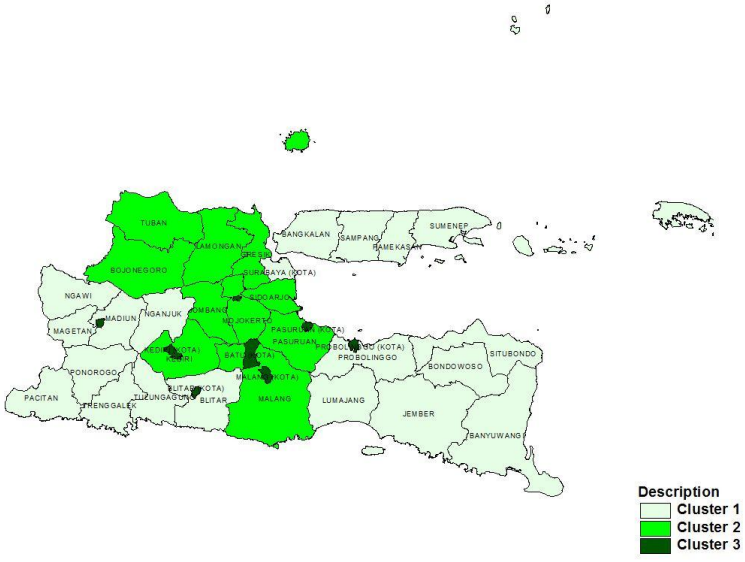

Figure 8. Mapping the Clustering Results of The Alibaca Index.

Based on figure 8, it can be seen that 38 regencies/cities in East Java Province were grouped into 3 clusters according to the results presented in Table 2, while the descriptive results of each cluster are as follows.

Based on Table 3, it can be seen that cluster 3 is the cluster with the highest skill dimension compared to the other clusters, with the literacy rate (X1) of 96.92 percent and the average duration of school (X2) of the people in Cluster 3 area is 9.6 or 10 years, which means that the average of people in cluster 3, cluster 2, and cluster 1 were not able to have received formal education for 12 years, or that the average of East Java's people just attends school up maximum to the junior high school level. Meanwhile, the area of cluster 2 is the area that dominates the access dimension, the alternative dimension, and the cultural dimension, because the value of each indicator of those 3 dimensions in cluster 2 has the highest value compared to the other clusters.

Table 3. Tabulation of Clustering Mapping Results.

Dimension	Indicator	Cluster 1	Cluster 2	Cluster 3
Skill Dimension	X1	88.79	93.74	96.92*)
	X2	6.78	7.84	9.68*)
Access Dimensions	X3	603.55	648.50*)	106.38
	X4	64.40	91.50*)	22.38
Alternative Dimensions	X5	111.75	163.60*)	23.88
	X6	229.55	364.20*)	32.00
	X7	89.15	195.80*)	29.13
	X8	63.85	68.60*)	4.38
Cultural Dimension	X9	174.75	349.30*)	31.00
	X10	1495.20	2365.80*)	220.63
	X11	48.05	60.20*)	6.38
	X12	93.40	253.50*)	23.25

Note: *) Variable dominance in the cluster

Each cluster has different characteristics, cluster 1 is the cluster with the lowest skill dimension, while the cluster 2 area is an area that dominates the access dimension, alternative dimension, and cultural dimension, meanwhile, the third cluster does not have dominance in these 3 dimensions, and when compared to other clusters, cluster 3 has the lowest score for all indicators in all three dimensions**,** which means that cluster 3 is the government's priority for improving reading activities. Based on these clusters, the recommendations formulated in this research are as follows.

a. Skill Dimension

Priority for improvement in cluster 1 for the Literacy Level indicator and the Average Duration of School for all clusters, because the average population in cluster 3, cluster 2, and cluster 1 are not able to receive formal education for 12 years.

b. Access Dimensions

The average number of schools in cluster 3 is very low compared to other clusters, so there needs to be an even distribution of schools in this cluster, as well as the number of villages that have availability of community reading gardens, there needs to be an addition in cluster 3.

c. Alternative Dimensions

Priority for improvement in cluster 3 for the number of villages according to base transceiver station (BTS) indicator, the number of villages according to cellular phone signal presence, and the number of villages according to the presence of GSM or CDMA internet signal indicator.

d. Cultural Dimensions

Cluster 3 is an area with limitations on all indicators in the cultural dimension, therefore cluster 3 becomes a priority cluster for improvement and equity in all indicators of the cultural dimension. So

the result of the analysis can help the government to develop strategic policies to achieve educational equity, especially concerning literacy levels in East Java Province based on the characteristics of each regency/city.

5. Conclusion

East Java province is among the 10 lowest regions with the lowest literacy index in Indonesia, with three dimensions below 50 percent, including access dimension, cultural dimension, and alternative dimension. Based on the results of unsupervised learning using the K-Means algorithm method for grouping regencies/cities in East Java based on the indicators that compose the Alibaca index gives the results that the regencies/cities of East Java Province are divided into 3 clusters based on the result of determining the optimal number of clusters using the elbow and silhouette method. 52.63% of regencies and cities in East Java are areas that fall into cluster 1, while 26.32% of regencies and cities in East Java are areas that fall into cluster 2, and the remaining 21.05% of regencies and cities in East Java are areas that belong to cluster 3. Each cluster has different characteristics, cluster 1 is the cluster with the lowest skill dimension, while the cluster 2 area is an area that dominates the access dimension, alternative dimension, and cultural dimension, meanwhile, the third cluster does not have dominance in these 3 dimensions, and when compared to other clusters, cluster 3 has the lowest score for all indicators in all three dimensions, which means that cluster 3 is the government's priority for improving reading activities.

References

[1] Ministry of Education and Culture 2017 *Panduan Gerakan Literasi Nasional 2017* (Jakarta: Puslitjakdikbud)

[2] Ministry of Education and Culture 2019 *Indeks Aktivitas Literasi Membaca 34 Provinsi* (Jakarta: Puslitjakdikbud)

[3] Soemartini and Supartini E 2017 *Analisis K-Means Cluster Untuk Pengelompokan Kabupaten/Kota Di Jawa Barat Berdasarkan Indikator Masyarakat* Konferensi Nasional Penelitian Matematika dan Pembelajarannya II (KNPMP II) pp 144-154

[4] S Liu and M d'Aquin 2017 *Unsupervised learning for understanding student achievement in a distance learning setting IEEE Global Engineering Education Conference (EDUCON)* pp 1373-1377

[5] Shovon M H, and Haque M 2012 *An Approach of Improving Students Academic Performance by using k means clustering algorithm and Decision tree*. *ArXiv*. abs/1211.6340

[6] Rani L N, Defit S, and Muhammad, L.J. 2021 *Determination of Student Subjects in Higher Education Using Hybrid Data Mining Method with the K-Means Algorithm and FP Growth*. *International journal of artificial intelligence*. 5

[7] Sinaga K P. and Yang M 2020 *Unsupervised K-Means Clustering Algorithm*. *IEEE Access*. 8, 80716-80727

[8] Wu, S., Rupprecht, C., & Vedaldi, A 2020 *Unsupervised learning of probably symmetric deformable 3d objects from images in the wild*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1-10)

[9] Sutarno N S 2003 *Perpustakaan dan Masyarakat* (Jakarta: Yayasan Obor Indonesia)

[10] Miller J W and Micahel M M 2016 *World Literacy: How Countries Rank and Why It Matters* (New York: Routledge)

[11] Nengsih W and Fadly A 2017 *Klasterisasi Tingkat Buta Huruf di Indonesia Berbasis Point-Based K-Means Analysis* (Pekanbaru: Politeknik Caltex Riau)

[12] Talakua M W, Leleury Z A and Talluta A W 2017 *Analisis Cluster dengan Menggunakan Metode K-Means untuk Pengelompokan Kabupaten/Kota di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014* (Ambon: Universitas Pattimura)

[13] Windarto A P and Herawan T 2022 *K-Means Algorithm with Rapidminer in Clustering School Participation Rate in Indonesia*. In: Ab. Nasir A.F., Ibrahim A.N., Ishak I., Mat Yahya N., Zakaria M.A., P. P. Abdul Majeed A. (eds) Recent Trends in Mechatronics Towards Industry 4.0. Lecture Notes in Electrical Engineering, vol 730. Springer, Singapore.

[14] Monica M., Ayuningtiyas N.U., Al Azies H., Riefky M., Khusna H., Rahayu S.P 2021 *Unsupervised Learning Approach for Evaluating the Impact of COVID-19 on Economic Growth in Indonesia*. In: Mohamed A., Yap B.W., Zain J.M., Berry M.W. (eds) Soft Computing in Data Science. SCDS 2021. Communications in Computer and Information Science, vol 1489. Springer, Singapore. https://doi.org/10.1007/978-981-16-7334-4_5

[15] Edwards A, and Cavalli-Sforza L 1965 *A Method for Cluster Analysis*. *Biometrics, 21*(2), 362-375

[16] Likas A, Vlassis, N, and Verbeek J 2003 *The global k-means clustering algorithm*. *Pattern Recognit*. 36, 451-461

[17] Zhai C and Aggarwal C C 2012 *Mining Text Data* (New York: Springer)

[18] Witten et al 2012 *Data Mining Practical Machine Learning Tools and Technique 2nd Edition* (San Fransisco: Morgan Kaufmann)

[19] Serna, L. A., Hernández, K. A., & González, P. N 2018. *A K-Means Clustering Algorithm: Using the Chi-Square as a Distance*. In International Conference on Human Centered Computing (pp. 464-470). Springer, Cham.

[20] MacQueen J B 1967 *Some Methods for Classification and Analysis of Multivariate Observations Hierarchical Grouping to Optimize an Objective Function* (Berkeley: University of California Press)

[21] Yuan, C., & Yang, H. 2019 *Research on K-value selection method of K-means clustering algorithm*. *J*, *2*(2), 226-235

[22] Dinh DT, Fujinami T, Huynh VN 2019 *Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient*. In: Chen J., Huynh V., Nguyen GN., Tang X. (eds) Knowledge and Systems Sciences. KSS 2019. Communications in Computer and Information Science, vol 1103. Springer, Singapore