International Journal of Grid and Distributed Computing
Vol. 11, No. 8 (2018), pp.87-98
http://dx.doi.org/10.14257/ijgdc.2018.11.8.08

Classification of Natural Disaster Prone Areas in Indonesia using K-Means

Bambang Supriyadi[1*], Agus Perdana Windarto[2], Triyuni Soemartono and Mungad[4]

[1]*Institut Pemerintahan Dalam Negeri, Jatinangor, Indonesia*
[2]*STIKOM Tunas Bangsa, Pematangsiantar, Indonesia*
[3]*Universitas Prof Dr Moestopo (Beragama), Jakarta, Indonesia*
[4]*Universitas Negeri Yogyakarta, Indonesia*
**bambang.supriyadi@ipdn.ac.id*

Abstract

Disaster caused by both nature and human factors has resulted in the occurrence of human casualties, environmental damage, property loss, and psychological impact. The study aims to classify disaster prone areas in Indonesia using K-means clustering method implemented in rapid miner tools. The data are collected from the Central Bureau of Statistics about the number of villages that considered as natural disaster-prone by province in Indonesia in years 2008-2014. The sample data are 34 provinces in Indonesia with 3 natural disasters commonly happen i.e. namely: Flood, Earthquake and Landslide. The final outcomes of the study were: (1) 4 provinces classified as High with cluster center 1363.333 (flood), 528.25 (earthquake) and 949.583 (landslide); 14 provinces classified as Medium with cluster center 142.619 (flood), 96.071 (earthquake) and 72.048 (landslide); and 16 provinces classified as Low with cluster center 507.396 (flood), 57.604 (earthquake) and 177.479 (landslide). This work can further provide input to the Indonesia government through mapping of disaster prone areas especially 4 provinces with very high natural disasters such as Aceh, West Java, Central Java and East Java.

Keywords: *Data mining; K-Means; Disaster; Clustering; Rapid miner*

1. Introduction

According to Law no. 24 of 2007 on Disaster Management on the Republic of Indonesia, disaster is an event or series of events that threaten and disrupt the life and livelihood of the community caused by both natural factors and / or non-internal factors and human factors resulting in the occurrence of human lives, environmental damage, property loss, and psychological effects [1]. Disaster can be divided into three i.e. natural disasters, non-natural disasters and social disaster. The Indonesia are among countries located in the area Ring of Fire (See Figure 1).

Received (February 27, 2018), Review Result (May 20, 2018), Accepted (May 28, 2018)

ISSN: 2005-4262 IJGDC

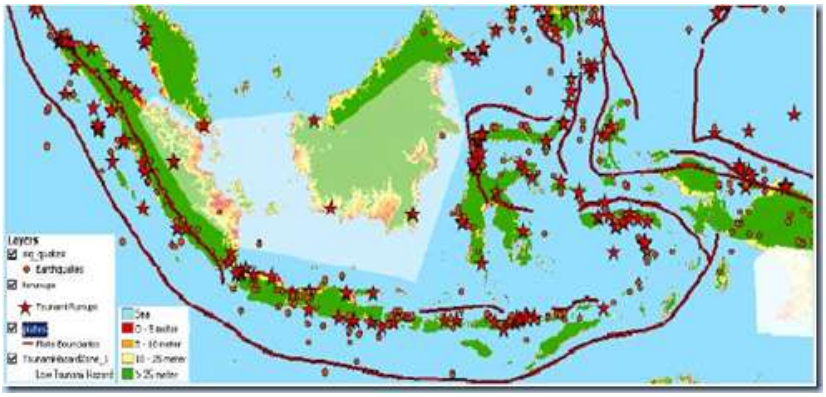

Figure 1. Indonesia Map in Ring of Fire Area [1]

The data from the Indonesia National Disaster Management Agency noted that 148.4 million people live in areas prone to earthquakes, 5 million live in areas prone to tsunami, 1.2 million live in areas prone to volcanic eruptions, 63.7 million live in areas prone to flooding, and 40.9 million live in areas prone to landslides. Based on data sources from the Central Bureau of Statistics (https://www.bps.go.id) the author intends to solve the problem of the number of villages that experienced natural disaster prone by province in Indonesia by using one computer science techniques that produce a support system decision. The sample of research data is the recapitulation of the number of villages that experienced natural disaster prone by province in Indonesia year 2008-2014. Criteria of natural disasters in question, namely: Flood, Earthquake and Landslide.

Today's data mining approach evolves to address various problems concerning data processing (See Figure 2). Many researchers [2, 3, 4, 5] use data mining techniques to solve the problem.

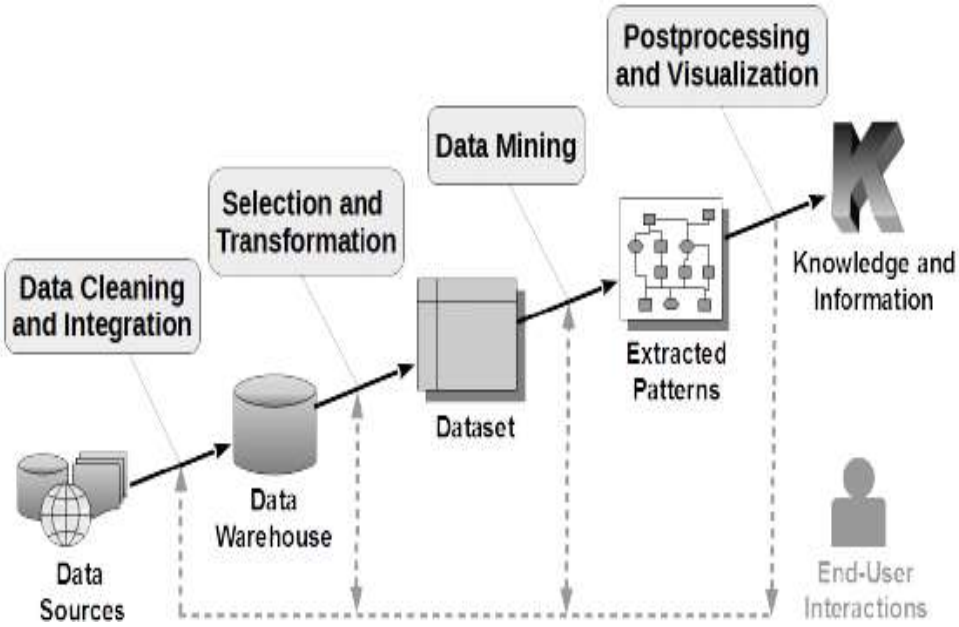

Figure 2. Data Mining Process

Data mining has several techniques, including classification, estimation, prediction and clustering. Clustering Technique [6] is a grouping of records, observing or observing and forming classes of objects that have similarities. Cluster is a collection of records that have similarities with each other and have an unlike the records in the cluster [7, 8, 9].

The purpose of this research is to classify provinces prone to natural disasters into several clusters. So that from the existing cluster can be the first step for the government and the community about disaster prone areas in Indonesia by province. This is necessary in order to provide an "early warning system" for the community regarding sites considered to be at high risk for disasters and locations that are safe from disasters. Thus, it is expected that the information can be made appropriate steps for spatial planning to improve the environment and minimize the effects of disasters effectively.

The rest of this paper is organized as follow. Section 2 presents theoretical background on artificial intelligence. Section 3 present the proposed method for prediction. Section 4 presents obtainde results and following by discussion. Finally, Section 5 concludes this work.

2. Data Mining

Data mining is an interdisciplinary subfield of computer science. It is the use of automatic data analysis techniques to uncover previously undetected relationship among data items [10]. It allows users to analyze data from various angles and dimensions, classified it and precise the relation recognized [11]. The K-means algorithm involves randomly selecting K initial centroids or mean where K is a user defined number of desired clusters. For each of the object the distance is calculated between center points and data points and with minimum distance data, cluster is generated. This data points are far from another cluster or group [8]. In order to measure the distance between objects and means different K-means clustering techniques can be used. Most popular distant metric that used is Euclidean Distance. Euclidean distance is represented as the square root of addition of squared differences between corresponding dimensions of object and the mean or cluster centroid. Euclidean distance is the most common distance metric which is most commonly used when dealing with multi- dimensional [10]. Basically the selection of initial means is up to the developer of clustering system what he/she wants. But this selection of initial means is independent of K-means clustering, because these initial means are inputs of K- means algorithm. In some cases, it is preferred to select initial means randomly from the given dataset while some others prefer to produce initial points randomly. As known that selection of initial means affects both the execution time of the algorithm and also the success of K-means algorithm [10].

2.1. Clustering

Clustering is the essential aspect of data mining. It is the technique of grouping of data in different groups by of their similarity measures. It means data items in the same group that are called cluster are more similar to each other than to those in other groups. Clustering is an unsupervised learning [4]. Clustering techniques mainly used two algorithms: Hierarchical algorithm and Partition algorithm. In the hierarchical algorithm, the dataset is divided into smaller subset in a hierarchical manner whereas in partition algorithm dataset is partitioned into the desired number of sets in a single step. K-means clustering is most popular partition algorithm. It uses in many application for producing the accurate result because of its simplicity of implementation [4].

2.2. Terms K-Means

a. *Cluster*: A cluster is an ordered list of objects, which have some common characteristics. The objects belong to an interval [*a*, *b*] and specifically in our case is [0. 1].

b. *Distance between Two Clusters:* The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed.

c. *Similarity:* A similarity measure SIMILAR (*Di*, *Dj*) can be used to represent the similarity between the documents. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement.

d. *Average Similarity:* If the similarity measure is computed for all pairs of documents (Di, *Dj*) except when *i=j*, an average value AVERAGE SIMILARITY is obtainable. Specifically, AVERAGE SIMILARITY = CONSTANT SIMILAR (*Di*, *Dj*), where *i*=1. 2.…,*n* and *j*=1. 2.…,*n* and *i* < > *j*.

e. *Threshold :* The lowest possible input value of similarity required to join two objects in one cluster.

f. *Similarity Matrix:* Similarity between objects calculated by the function SIMILAR (*Di, Dj)*, represented in the form of a matrix is called a similarity matrix.

g. *Dissimilarity Coefficient:* The dissimilarity coefficient of two clusters is defined to be the distance between them. The smaller the value of dissimilarity coefficient, the more similar two clusters are.

h. *Cluster Seed:* First document or object of a cluster is defined as the initiator of that cluster *i.e.*, every incoming object's similarity is compared with the initiator. The initiator is called the cluster seed [11].

2.3. A Centroid-Based Clustering Technique

According to the basic K-mean clustering algorithm, clusters are fully dependent on the selection of the initial clusters centroids. K data elements are selected as initial centers; then distances of all data elements are calculated by Euclidean distance formula. Data elements having less distance to centroids are moved to the appropriate cluster. The process is continued until no more changes occur in clusters *k*-1 [7][12]. This partitioning clustering is most popular and fundamental technique [12]. It is vastly used clustering technique which requires user specified parameters like number of clusters k, cluster initialisation and cluster metric [13]. First it needs to define initial clusters which makes subsets (or groups) of nearest points (from centroid) inside the data set and these subsets (or groups) called clusters [12]. Secondly, it finds means value for each cluster and define new centroid to allocate data points to this new centroid and this iterative process will goes on until centroid [14] does not changes. The simplest algorithm for the traditional *K*-means [12] is as follows;

a. Input: $D = \{d_1, d_2, d_3 ..., d_n\}$ // set of n numbers of data points *K* // The number of desire Clusters

b. Output: A set of *k* clusters

- Select k points as initial centroids.
- Repeat
- From *K* clusters by assigning each data point to its nearest centroid.
- Recomputed the centroid for each cluster until centroid does not change [13-22].

2.3.1. Important Equations

Distance measure will determine how the similarity of two elements is calculated and it will influence the shape of the clusters. The Euclidean distance is given by:

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2} \quad (1)$$

where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k-th attributes (components) or data objects p and q.

2.3.2. Improved Technique

Part1: Determine initial centroids

Step1.1: Input Dataset

Step1.2: Check the Each attributes of the Records

Step1.3: Find the mean value for the given Dataset.

Step1.4: Find the distance for each data point from mean value using Equation (Equ).

IF

The Distance between the mean value is minimum then it will be stored in Then Divide datasets into *k* cluster points don't needs to move to other clusters.

ESLE

Recalculate distance for each data point from mean value using Equation (1) until divide datasets into k cluster

Part2: Assigning data points to nearest centroids

Step2.1: Calculate Distance from each data point to centroids and assign data points to its nearest centroid to form clusters and stored values for each data.

Step2.2: Calculate new centroids for these clusters.

Step2.3: Calculate distance from all centroids to each data point for all data points.

IF

The Distance stored previously is equal to or less then Distance stored in Step2.1 Then Those Data points don't needs to move to other clusters.

ESLE

From the distance calculated assign data point to its nearest centroid by comparing distance from different centroids. Step2.5: Calculate centroids for these new clusters again.

Until

The convergence criterion met.

OUTPUT

A Set of K clusters[10]

3. Experimental Design

In this study the method used is the method of data mining as follows. 1). Data collection stage, 2). Data processing stage, 3). Clustering stage and 4). Stage Analysis [4]. Stages in the method are further described as follows:

3.1. Data Collection Stage

Implementation of data mining clustering the number of villages that experienced natural disaster prone by province in Indonesia with tools RapidMiner, required relevant data. Sources of research data obtained from the Central Bureau of Statistics. Data used in

2008-2014 which consists of 3 criteria, namely: a. Flood, b. Earthquake and c. Landslide. Data will be processed in 3 clusters of (C1) Cluster Disaster High, (C2) Cluster Disaster Medium and (C3) Cluster Disaster Low.

3.2. Data Processing Stage

Before the data goes into the process, it is first clustered according to predetermined criteria by taking the average number of villages that experienced natural disaster prone by province in Indonesia from 2008-2014. The clustered data are then processed using K-Means of 3 clusters.

3.3. Clustering Stage

In determining the cluster based on the data already available, it takes a flowchart to facilitate in determining the flow of calculation as a groove to find the results of the application of the cluster to the data to be processed [4].

3.4. Stage Analysis

At this stage, data analysis of the average number of villages that experienced natural disaster prone by province in Indonesia will be processed with the RapidMinner tool. RapidMiner is a machine learning environment for data mining, text mining and predictive analysis [4]. In the previous stage, has been determined in 3 clusters, namely: (C1) Cluster Disaster High, (C2) Cluster Disaster Medium and (C3) Cluster Disaster Low.

4. Results and Discussion

In grouping, the data obtained will be calculated in advance based on the number of villages that experienced natural disaster prone by province in Indonesia in 2008-2014. Sum result based on 3 assessment criteria, namely: a. Flood, b. Earthquake and c. Landslide, as shown in Table 1.

Table 1. Data of Villages that Experienced Natural Disaster Prone in Indonesia, 2008-2014

Provinsi	Data of villages that experienced natural disaster prone in Indonesia								
	Flood			Earthquake			Landslide		
	2008	2011	2014	2008	2011	2014	2008	2011	2014
Aceh	1991	1463	1649	745	260	1228	310	227	273
North Sumatra	682	649	807	196	34	191	470	407	569
West Sumatra	243	315	306	634	496	78	205	244	225
Riau	479	328	512	2	0	0	24	23	24
Jambi	402	357	518	250	132	40	51	27	58
South Sumatra	328	499	745	31	36	2	136	147	145
Bengkulu	145	215	213	776	24	56	88	114	151
Lampung	251	432	508	15	7	5	58	82	82
Bangka Islands	20	16	58	2	0	0	1	0	4
Riau Islands	33	24	51	0	0	0	10	11	13
DKI Jakarta	178	53	151	0	0	0	1	1	0
West Java	1162	989	1193	68	2169	412	1610	1477	1578
Central Java	1367	1266	1273	905	116	129	1254	1410	1222
DI Yogyakarta	52	89	76	410	8	27	61	78	77
East Java	1419	1370	1218	90	10	207	696	673	665
Banten	535	401	531	15	41	19	127	140	150
Bali	33	71	58	27	4	0	105	162	150
West Nusa Tenggara	199	282	286	183	166	68	28	59	46
East Nusa Tenggara	612	557	445	21	14	97	621	565	581
West Kalimantan	394	740	616	0	0	0	35	67	65
Central Kalimantan	451	316	534	0	0	0	10	9	23

South Kalimantan	533	591	623	0	0	0	40	44	40
East Kalimantan	478	463	409	0	17	4	113	71	55
North Kalimantan	0	0	140	0	0	13	0	0	40
North Sulawesi	375	336	353	186	174	102	303	294	308
Central Sulawesi	583	565	731	40	144	158	178	143	205
South Sulawesi	801	746	728	16	20	22	364	278	280
Southeast Sulawesi	276	351	702	15	8	175	55	51	123
Gorontalo	276	307	323	12	60	99	54	57	73
West Sulawesi	181	221	159	36	24	8	159	220	157
Maluku	119	122	233	60	13	43	48	68	122
North Maluku	132	155	285	128	51	143	34	23	52
West Papua	50	32	88	30	196	160	18	13	54
Papua	363	411	308	38	157	341	291	336	251

(Source : Central Bureau of Statistics of Indonesia)

From Table 1. the data is then accumulated based on 3 criteria, namely a. Flood, b. Earthquake and c. Landslide. The process of data accumulation is done by finding the average value of each province based on predetermined criteria. The results of complete calculation can be seen in the following Table 2.

Table 2. Accumulated Data of Villages that Experienced Natural Disaster Prone in Indonesia

Provinsi	Flood	Earthquake	Landslide
Aceh	1701.00	744.33	270.00
North Sumatra	712.67	140.33	482.00
West Sumatra	288.00	402.67	224.67
Riau	439.67	0.67	23.67
Jambi	425.67	140.67	45.33
South Sumatra	524.00	23.00	142.67
Bengkulu	191.00	285.33	117.67
Lampung	397.00	9.00	74.00
Bangka Islands	31.33	0.67	1.67
Riau Islands	36.00	0.00	11.33
DKI Jakarta	127.33	0.00	0.67
West Java	1114.67	883.00	1555.00
Central Java	1302.00	383.33	1295.33
DI Yogyakarta	72.33	148.33	72.00
East Java	1335.67	102.33	678.00
Banten	489.00	25.00	139.00
Bali	54.00	10.33	139.00
West Nusa Tenggara	255.67	139.00	44.33
East Nusa Tenggara	538.00	44.00	589.00
West Kalimantan	583.33	0.00	55.67
Central Kalimantan	433.67	0.00	14.00
South Kalimantan	582.33	0.00	41.33
East Kalimantan	450.00	7.00	79.67
North Kalimantan	46.67	4.33	13.33
North Sulawesi	354.67	154.00	301.67
Central Sulawesi	626.33	114.00	175.33
South Sulawesi	758.33	19.33	307.33
Southeast Sulawesi	443.00	66.00	76.33
Gorontalo	302.00	57.00	61.33
West Sulawesi	187.00	22.67	178.67
Maluku	158.00	38.67	79.33
North Maluku	190.67	107.33	36.33
West Papua	56.67	128.67	28.33
Papua	360.67	178.67	292.67

The average Accumulated Data of villages that experienced natural disaster prone in Indonesia in Table 2 will be processed into clustering by applying the K-Means algorithm using RapidMinner has been determined in 3 clusters, namely: (C1) Cluster Disaster High, (C2) Cluster Disaster Medium and (C3) Cluster Disaster Low.

4.1. Centroid Data

In the application of K-means algorithm using tools RapidMiner, the centroid value is derived from the grouping of data performed. In this case, the researchers conducted 3 clusters, namely: (C1) Cluster Disaster High, (C2) Cluster Disaster Medium and (C3) Cluster Disaster Low. The cluster point determination is done by taking the highest value for the Cluster Disaster High, the average value for the Cluster Disaster Medium and the smallest value for the Cluster Disaster Low. The calculation process can be seen below.

a. Cluster Disaster High MAX (value on criterion 1 (Flood)
MAX (value on criterion 2 (Earthquake)
MAX (value on criterion 3 (Landslide)

b. Cluster Disaster Medium AVERAGE (average value on criterion 1 (Flood)
AVERAGE (average value on criterion 2 (Earthquake)
AVERAGE (average value on criterion 3 (Landslide)

c. Cluster Disaster LowMIN (value on criterion 1 (Flood)
MIN (value on criterion 2 (Earthquake)
MIN (value on criterion 3 (Landslide)

The centroid data for each cluster can be seen in the Table 3 below:

Table 3. Centroid Initial Data

Data Cluster	Flood	Earthquake	Landslide
Cluster Disaster High	1701.00	883.00	1555.00
Cluster Disaster Medium	457.89	128.81	224.90
Cluster Disaster Low	31.33	0.00	0.67

4.2. K-Means Clustering Using RapidMiner

In natural disaster data grouping, the writer uses K-means algorithm to group data based on attribute at cluster central distance from a set of data as in Table 3. Iteration is the process of execution on K-Means to group data based on cluster center distance value. The cluster central distance value will continue to change according to the number of iterations performed. The K-Means process will continue to iterate until the data grouping equals the previous iteration data grouping. In other words, the process will continue iterating until the data in the last iteration is the same as the previous iteration.

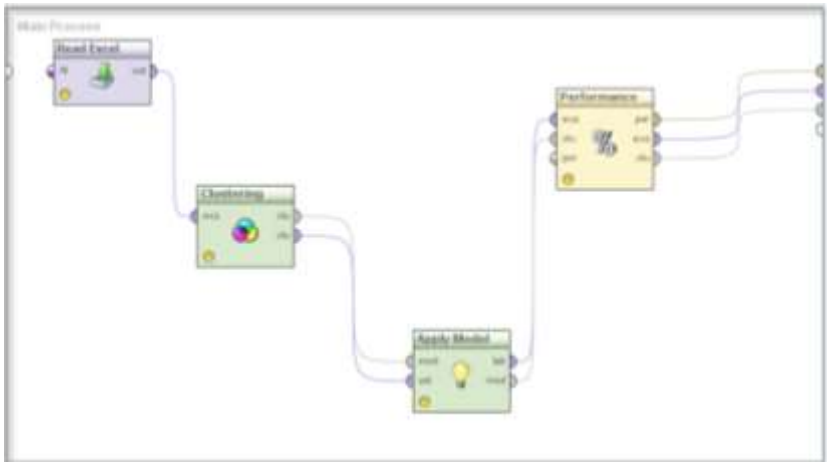

Figure 3. Design of K-Means Algorithm Usage with value K = 3

Figure 3 above describes a model first trained and read through Read Excel using excel data; information related to Read Excel is studied by the model. Then the model can be applied to other Read Excel normally for prediction. All required parameters are stored in the model object. It is mandatory that both Read Excel should have the exact same number, order, type, and role attribute. If the metadata property of Read Excel is inconsistent, this can cause serious errors. In this case the researcher uses 34 provinces data samples in the area of natural disasters using 3 attributes of assessment, namely: floods, earthquakes and landslides. In this Example Process, the 'natural disaster' data set is loaded using the Retrieve operator. The clustered model trained in Read Excel uses the K-Means operator. This model is then supplied to the Model App operators model input port. A collection of 'natural disaster' data is loaded using the Retrieve operator and is provided in the data input port unlabeled by the Applicable Model operator. Applicable Model Operators apply models trained by K-Means operators on 'natural disasters' to classify attribute values with defined clusters.

Based on the design of Figure 1. the RapidMiner tools will group the 'natural disaster' area based on the input data provided. The results of the final grouping can be seen in the picture below:

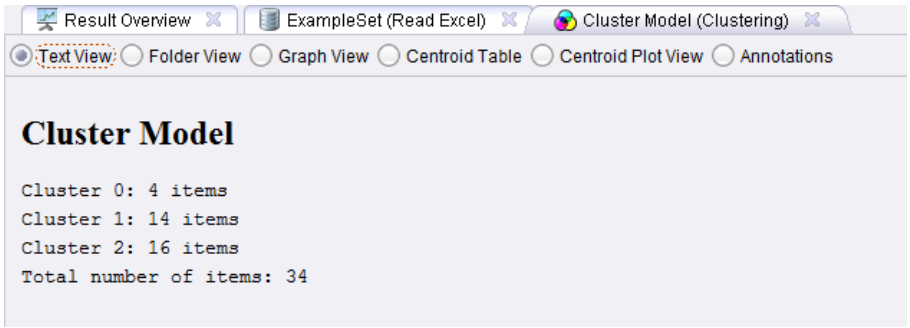

Figure 4. Results Grouping

Figure 4 above describes the classification of 'natural disaster' areas in Indonesia by province that 4 provinces in (C1) Cluster Disaster High, 16 provinces in (C2) Cluster Disaster Medium and 16 provinces in Cluster Disaster Low (C3). The 4 provinces in the 'disaster prone' areas are Aceh, West Java, Central Java and East Java.

Result Overview | ExampleSet (Read Excel) | Cluster Model (Clustering)

Text View | Folder View | Graph View | Centroid Table | Centroid Plot View | Annotations

Attribute	cluster_0	cluster_1	cluster_2
Flood	1363.333	142.619	507.396
Earthquake	528.250	96.071	57.604
Landslide	949.583	72.048	177.479

Figure 5. The final Centroid Value at the Last Iteration

Based on Figure 5 above, Cluster center distance value continues to change based on the number of iterations performed. This happens because the distance of each data to each centroids that measures the distance between the data with the cluster center using the Euclidian is always different on each iteration.

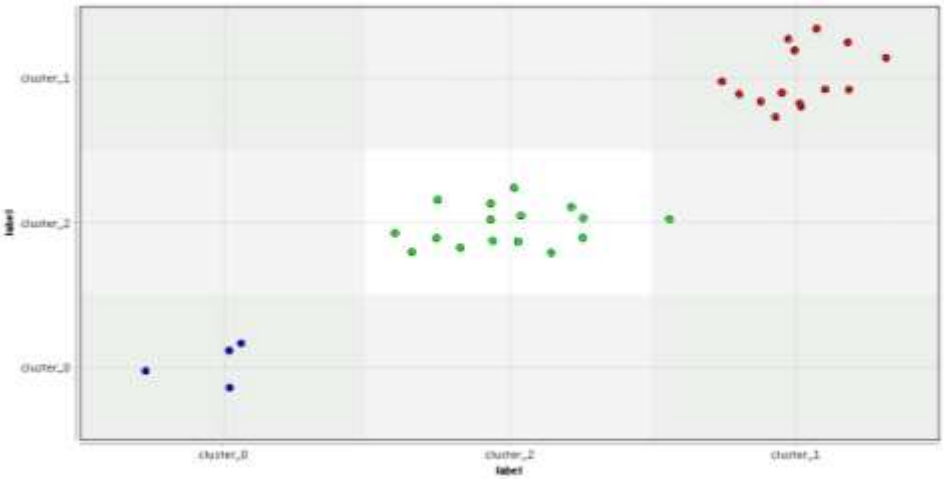

Figure 6. Results of K-Means Grouping with RapidMiner

Figure 7 above explains that the k-means algorithm has successfully classified 'natural disaster' data in Indonesia by province into three clusters. The final cluster results show by province that 4 provinces in (C1) Cluster Disaster High, 16 provinces in (C2) Cluster Disaster Medium and 16 provinces in Cluster Disaster Low (C3).

4.3. K-Means Performance Accuracy with RapidMiner

One of the operators used to measure K-Means accuracy is % Performance. Performance% is used for performance evaluation of centroid-based grouping methods. This operator provides a list of performance criteria values based on the cluster centroid. The %Performance measurement parameters are avg._within_centroid_distance and davies_bouldin. avg._within_centroid_distance: The average in cluster distance is calculated by the average distance between the centroid and all cluster examples. davies_bouldin: Algorithms that produce clusters with low intra-cluster distance (high intra-cluster similarity) and high inter-cluster spacing (low inter-cluster similarity) will have a low Davies-Bouldin index, a grouping algorithm that generates a set clusters with the smallest Davies-Bouldin index are considered the best algorithms based on the criteria. The results obtained on the Davies-Bouldin index for 'natural disaster' areas are –0.796 (See Figure 7). Based on the results of these performance, therefore it can be considered as the best algorithm based on criteria.

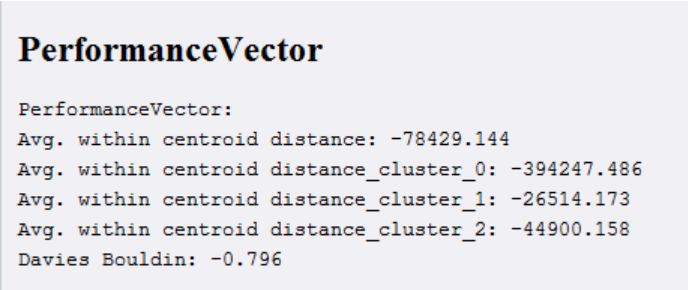

Figure 5. Result Performance Vector K-Means with RapidMiner

5. Conclusion

This paper has presented the usage of K-means clustering method implemented in rapid miner tools to classify disaster prone areas in Indonesia. In summary, to cluster areas of 'natural disaster' in Indonesia by province can be done by applying datamining clustering technique with K-Means method. Attributes used in the study are: (1) Flood; (2) Earthquakes, and (3) Landslide. The clustering used 3 clusters: (C1) Cluster Disaster High, (C2) Cluster Disaster Medium and (C3) Cluster Disaster Low. From the results, we obtained provincial areas prone to 'natural disasters' *i.e.*, (1) 4 provinces (C1) Cluster Disaster High: Aceh, Wast Java, Central Java and East Java; (2) 14 provinces in (C2) Cluster Disaster Medium: West Sumatera, Bengkulu, Bangka Islands, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, West Nusa Tenggara, Nort Kalimantan, Gorontalo, West Sulawesi, Maluku, North Maluku and West Papua; (3) 16 provinces on (C3) Cluster Disaster Low: North Sumatera, Riau, Jambi, South Sumatra, Lampung, Banten, East Nusa Tenggara, West Kalimantan, North Kalimantan, Central Sulawesi, South Sulawesi, Southeast Sulawesi and Papua; (4) In the assessment of performance accuracy with Performance model measured from 2 parameters, namely: Avg. within centroid distance and Davies Bouldin, we obtained the smallest value which means as the best algorithm based on criteria. The value of Performance is as follows: (1) Avg. within centroid distance: –78429.144; (2) Avg. within centroid distance_cluster_1: –394247.486; (3) Avg. within centroid distance_cluster_2: –26514.173; (4) Avg. within centroid distance_cluster_3: –44900.158; (5) Davies Bouldin as –0.796.

Acknowledgment

This research is supported by Institut Pemerintahan Dalam Negeri, Jatinangor, Indonesia.

References

[1] Pemerintah Republik Indonesia, “Undang-Undang No. 24 Tahun 2007 tentang Penanggulangan Bencana”, Jakarta: DPR RI, **(2007)**.
[2] O. J. Oyelade, O. O. Oladipupo and I. C. Obagbuwa, “Application of k Means Clustering algorithm for prediction of Students Academic Performance”, Int. J. Comput. Sci. Inf. Secur., vol. 7, no. 1, **(2010)**, pp. 292-295.
[3] S. Kumar and S. K. Rathi, “Performance Evaluation of K-Means Algorithm and Enhanced Mid-point based K-Means Algorithm on Mining Frequent Patterns”, Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 4, no. 10, **(2014)**, pp. 545-548.
[4] A. Yadav and S. Dhingra, “An Enhanced K-Means Clustering Algorithm to Remove Empty Clusters”, IJEDR, vol. 4, no. 4, **(2016)**, pp. 901-907.
[5] N. Aggarwal, K. Aggarwal and K. Gupta, “Comparative Analysis of K-means and Enhanced K-means Clustering Algorithm for Data Mining”, Int. Jouranl Sci. Eng. Res., vol. 3, no. 3, **(2012)**.
[6] T. M. Kodinariya and P. R. Makwana, “Review on determining number of Cluster in K-Means Clustering”, Int. J. Adv. Res. Comput. Sci. Manag. Stud., vol. 1, no. 6, **(2013)**, pp. 2321-7782.
[7] U. R. Raval and C. Jani, “Implementing and Improvisation of K-means Clustering,” Int. J. Comput. Sci. Mob. Comput., vol. 5, no. 5, **(2016)**, pp. 72-76.

[8] M. K. Arzoo, A. Prof, and K. Rathod, "K-Means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8 . 2." Int. Res. J. Eng. Technol., vol. 4. no. 4. pp. 2363–2368. 2017.
[9] A. P. Windarto, "Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method", Int. J. Artif. Intell. Res., vol. 1, no. 2, **(2017)**, pp. 26-33.
[10] N. Kaur, J. K. Sahiwal, N. Kaur and P.- Punjab, "Efficient K-Means Clustering Algorithm Using Ranking Method", Int. J. Adv. Res. Comput. Eng. Technol., vol. 1, no. 3, **(2012)**, pp. 85-91.
[11] K. Singh, D. Malik and N. Sharma, "Evolving limitations in K-means algorithm in data mining and their removal", IJCEM Int. J. Comput. Eng. Manag. ISSN, vol. 12, **(2011)** April, pp. 2230-7893.
[12] S. Shinde and B. Tidke, "Improved K-means Algorithm for Searching Research Papers", Int. J. Comput. Sci. Commun. Networks, vol. 4, no. 6, **(2014)**, pp. 197-202.
[13] R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", Wiley, **(1973)**.
[14] W. Gropp, E. Lusk and A. Skjellum, "Using MPI: Portable Parallel Programming with the Message Passing Interface", The MIT Press, Cambridge, MA, **(1996)**.
[15] M. Snir, S. W. Otto, S. Huss-Lederman, D. W. Walker and J. Dongarra, "MPI: The Complete Reference", The MIT Press, Cambridge, MA, **(1997)**.
[16] P. S. Pacheco, "A User_s Guide to MPI", **(2008)**.
[17] M. Snir, S. Otto, S. Huss-Lederman, D. Walker and J. Dongarra, "MPI: The Complete Reference", The MIT Press, Cambridge, Massachusetts London, England, **(2002)**.
[18] K. Stoel and A. Belkoniene, "Parallel k/h-Means Clustering for Large Data Sets", Springer-Verlag Berlin Heidelberg, Euro-Par'99, LNCS 1685, **(1999)**, pp. 1451-1454.
[19] W. Fang, K. Keung Lau, M. Lu, X. Xiao, C. Kit Lam, P. Yang Yang, B. He, Q. Luo, P. V. Sander and K. Yang, "Parallel Data Mining on Graphics Processors", Technical Report HKUSTCS0807, **(2008)** October.
[20] K. Kerdprasop and Nittaya Kerdprasop, "A lightweight method to parallel k-Means clustering", International Journal of Mathematics and Computers in Simulation, iss. 4, vol. 4, **(2010)**, pp. 144-153.
[21] A. Raghuvira Pratap, J. Rama Devi, K. Suvarna Vani and K. Nageswara Rao, "An Efficient Density based Improved KMedoids Clustering algorithm", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 2. no. 6, **(2011)**, pp. 49-54.
[22] D. Pettinger and G. Di Fatta "Space Partitioning for Scalable K-Means", Ninth International Conference on Machine Learning and Applications, IEEE, **(2010)**, pp. 319-24.

