Implementation of K-means Clustering on SIPP-KLING Dashboard Applications

Fatona Fadilla Rohma, Iklima Ermis Ismail, Yoyok S Waluyo
Program Studi Teknik Informatika
Jurusan Teknik Informatika dan Komputer
Politeknik Negeri Jakarta
fatonafadilla@gmail.com, iklimaermis.ismail@tik.pnj.ac.id, yoyok_sw@yahoo.com

***Abstract* - This research focuses on grouping health house (rumah_sehat) data into five clusters, namely Very Unhealthy, Unhealthy, Not Healthy Yet, Healthy, Very Healthy. There were 17 criteria as input parameters for K-Means calculation. These research aims to grouping 8969 houses into the clusters. The results of these clustering can help decision maker (government) to analyze which parts of the houses whose need the attention more, or which areas that lower than healthy standards. The test result shows that there were 3308 Very Healthy, 2496 Healthy, 792 Not healthy Yet, 1706 Unhealthy, and 667 Very Unhealthy houses. The accuracy of this method was found 87.05% using confusion matrix, with precision of 95.64% and 75, 81%, and recall of 83.82% and 92, 98%. Based on ROC the level of diagnostic value accuracy of 87.05% includes good clustering.**

***Keywords* : k-means, healthy_home, clustering, data mining**

I. INTRODUCTION

SIPP-KLING is an environmental health mapping profile information system for UPT Puskesmas Limo. This application aims to facilitate service, health, coaching and even assistance to the community to implement PHBS (Clean and Healthy Life Behavior) in the surrounding environment. In the current SIPP-KLING system, there were 2 categories of final results, namely Healthy and Unhealthy which was obtained from the final total value. Because it only has 2 categories, the analysis obtained from each village is not specific. So that supervision and evaluation cannot be carried out optimally because the gap in the same region is too wide, hence the houses in the same region (categorie) have different health problems and require different handling. With a lack of analysis it also has an impact on the funds that will be given to each region. The amount of funds channeled will not reach the areas that need it most.

Based on the condition above, there is a need to improve the grouping of the SIPP-KLING datasets to get more specific groups. This research use K-Means Clustering to grouping 8969 houses into the cluster, to extracts the information and important patterns of interest in the SIPP-KLING application. K-Means is a most widely used and well studied method in data mining [1]. Refer to [2], the clustering analysis is useful to draw meaningful information or drawing interesting patterns form data sets and used in many fields like bioinformatics, pattern recognition, image processing, data mining, marketing, economics, etc., to get the hidden knowledge.

II. K-MEANS CLUSTERING

Clustering is a process of grouping data objects into disjointed group called clusters, so that the data in the same cluster are similar and different to other cluster [3]. The main aim of clustering is to offer a combination of imilar objects. Between classification and clustering are confuse to be different, but in classification objects is assigned in predefined classes while in clustering classes is created [1].

The K-Means algorithm uses the process repeatedly to get the cluster database. It takes the desired number of initial clusters as input and produces the number of end clusters as output. If the algorithm is needed to generate cluster K then there will be an initial K and a final K. The K-Means method will randomly select the k pattern as the starting point of the centroid. The number of iterations to reach the centroid cluster will be affected by the prospective random initial centroid cluster where the position of the new centroid does not change. The K value chosen as the initial center will

be calculated using the Euclidean Distance formula, which is to find the closest distance between the centroid point and the data / object. Data that has a short distance or closest to the centroid will form a cluster [1].

In its completion, the K-Means algorithm will produce the centroid point which is the purpose of the K-Means algorithm. After the iteration of K-Means stops, each object in the dataset becomes a member of a cluster. The cluster value is determined by looking for all objects to find clusters with the closest distance to the object. The K-means algorithm will group data items in a dataset into a cluster based on the closest distance [4]. Following are the steps of K-means algorithm[5] :

INPUT : Number of desired clusters K Data Object D={d1,d2,..dn}

Step:

- randomly elevate K data objects (as initial centers) from data set D.
- Repeat;
- Calculate the distance of each data object d_i ($1 \leq i \leq n$) from all k clusters C j ($1 \leq j \leq k$) and then assign data object d_i to the nearest cluster.
- For each cluster j ($1 \leq j \leq k$)
- Recalculate the cluster center until no change in the center of clusters.

OUTPUT : A set of K clusters

III. DESIGN AND REALIZATION

A. *Design*

Dashboard SIPP-KLING (Profile Information System Mapping Environmental Health) is a web-based system developed by applying K-means clustering methods in data processing to become information that will be easy to learn knowledge. The actors involved in this application are the admin and cadre, the admin manages the SIPP-KLING web dashboard while the cadres are actors who collect the data through the citizen. This system serves admin to control the activities of data processing, and facilitate the admin in making reports and making decisions.

B. *Implementation of K-Means Algorithm*

This stage will explain the steps to operate the K-Means algorithm manually:

1. Determination of the number of clusters is five (k = 5) to be made, the determination of clusters is based on conceptual observations by experts grouped into five, namely, Healthy, Unhealthy, Unhealthy, Very Healthy, Very Unhealthy. The amount of data used is 8969 SIPP-KLING data and 17 attributes from the criteria for determining the Dinkes Health House. The criteria are taken based on interview process with dr. Tiur. The criteria are initialized by r (r1 up to r17) namely: ceiling, wall, floor, the bed room window, the family room window, ventilation, kitchen exhaust fan, lighting, toilet, safe drinking water, sewer system, rubbish management, opening the bed room window, opening the family room window, keep the house and the yard clean, throw the feces of the baby and toddler into the toilet, and throw the garbage in its place.

2. Determination of the centroid of each cluster can be seen in TABLE 1.

TABLE 1. The Centroid Value of Clusters

	Criteria																
	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17
C1	62	62	62	31	31	62	62	62	100	100	100	75	88	88	88	88	88
C2	62	62	62	31	31	62	62	62	75	75	75	75	88	88	88	88	88
C3	62	62	62	31	31	62	62	62	50	50	50	50	88	88	88	88	88
C4	31	31	31	31	31	31	31	31	25	25	25	25	44	44	44	44	44
C5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

3. Calculate the distance between data and the centroid.

   Measuring the distance between the data and the centroid used Euclidian Distance formula (eq 1)

   $$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

   Then the distance matrix will be obtained, namely C1, C2, C3, C4 and C5 as follows:

   Data distance of cluster **1** is:

   - **d (X1, C1)** = $\sqrt{}$ (62-62) ^{2} + (62-62) ^{2} + (62-62) ^{2} + (0-31) ^{2} + (0-31) ^{2} + (0-62) ^{2} + (31-62) ^{2} + (31-62) ^{2} + (100-100) ^{2} + (100-100) ^{2} + (50-100) ^{2} + (50-75) ^{2} + (0-88) ^{2} + (88-88) ^{2} + (88-88) ^{2} + (88-88) ^{2} + (44-88) ^{2} = **143.153763485**

- **d (X2, C2)** = $\sqrt{(62-62)^2 + (62-62)^2 + (62-62)^2 + (31-31)^2 + (31-31)^2 + (31-62)^2 + (62-62)^2 + (62-62)^2 + (100-100)^2 + (75-100)^2 + (100-100)^2 + (75-100)^2 + (88-88)^2 + (88-88)^2 + (88-88)^2 + (88-88)^2 + (88-88)^2}$ = **39.8246155035**
- **d (X3, C3)** = $\sqrt{(62-62)^2 + (62-62)^2 + (62-62)^2 + (0-31)^2 + (31-31)^2 + (31-62)^2 + (31-62)^2 + (31-62)^2 + (100-100)^2 + (100-100)^2 + (100-100)^2 + (25-100)^2 + (0-88)^2 + (44-88)^2 + (88-88)^2 + (88-88)^2 + (88-88)^2}$ = **126.585939188**
- **d (X4, C4)** = $\sqrt{(62-62)^2 + (62-62)^2 + (62-62)^2 + (31-31)^2 + (31-31)^2 + (31-62)^2 + (31-62)^2 + (62-62)^2 + (100-100)^2 + (100-100)^2 + (100-100)^2 + (50-100)^2 + (88-88)^2 + (88-88)^2 + (88-88)^2 + (88-88)^2 + (88-88)^2}$ = **50.4678115238**
- **d (X5, C5)** = $\sqrt{(62-62)^2 + (62-62)^2 + (31-62)^2 + (31-31)^2 + (31-31)^2 + (31-62)^2 + (31-62)^2 + (31-62)^2 + (100-100)^2 + (100-100)^2 + (50-100)^2 + (50-100)^2 + (44-88)^2 + (44-88)^2 + (88-88)^2 + (88-88)^2 + (44-88)^2}$ = **113.035392687**

The calculation is done to another data distance for clusters 2, 3, 4 and 5.

4. Then grouping the data according to its cluster, which is the data that has the shortest distance. TABLE 2 shows the result of 1st iteration calculation.

TABLE 2. The result of 1st Iteration

Iteration 1				
C1	**C2**	**C3**	**C4**	**C5**
143.153763485	140.953893171	11.484819124332	158.461351754931	252.234018324254
39.8246155035	47.021271782	84.917607125967	175.314004004244	303.435330836737
126.585939188	133.787144375	147.135991518051	168.709810028937	266.775186252395
50.4678115238	66.4981202742	97.066987178958	176.136310850432	300.732439221312
113.035392687	110.236110236	121.045446010992	135.255314128503	244.438131231606

5. The process returns again step 2.
   For the next repetition (1st repetition to completion), the new centroid is calculated by calculating the average value of the data in each cluster, like shown in TABLE 3.

TABLE 3. Average Centroid Value in 1st Repeat

	The Criteria																
	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17
C1	59.87	61.48	61.70	27.75	30.54	54.05	49.57	57.77	99.03	91.86	97.57	61.03	65.78	67.43	85.97	79.80	81.92
C2	30.17	54.55	51.29	59.38	97.13	96.54	50.71	60.16	66.49	66.89	86.78	84.09	76.96	30.17	54.55	51.29	59.38
C3	30.72	54.50	47.95	55.87	87.21	84.74	47.68	16.23	55.75	60.31	86.83	81.26	60.48	30.72	54.50	47.95	55.87
C4	26.495	32.00	27.07	45.23	74.33	74.56	50.37	29.09	26.14	33.25	77.20	45.38	28.85	26.49	32.00	27.07	45.23
C5	0	0	0	0	30	30	30	5	0	0	44	17.6	26.4	0	0	0	0

If the new centroid is different from the previous centroid, then the process continues the next step. However, if the new centroid is calculated the same as the previous centroid, then the clustering process is complete.

In this calculation, the processes stop at 12th iteration, like shown in TABLE 4 and TABLE 5. After getting the cluster label for each data, the average value is searched by adding up all members of each cluster and dividing the number of members.

TABLE 4. Data Grouping in the 12th Repetition

Iteration 11				
C1	**C2**	**C3**	**C4**	**C5**
131.9010703186250	120.9553208498430	114.7087233018670	85.1339968684470	123.3096352165050
39.077105049157 2	70.469321117856 9	112.7033587137170	109.2693158563240	138.9591651270630
108.2886032123300	114.6481471926570	133.8441119688580	47.838639555203 3	126.9050507430600
44.577674052077 6	70.209642990202 1	102.7415759492260	98.551401356583 8	134.6284680869400
97.925942945531 5	82.628011956220 2	77.290287574655 2	65.025382327006 4	97.259405221802 3

TABLE 5. Centroid Average Value in 12th Repetition

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17
C1	60.77	61.76	61.88	30.613	30.83	58.04	55.31	60.03	99.25	90.49	97.83	63.77	78.12	74.14	86.64	80.39	86.59
C2	60.83	61.62	61.78	30.46	30.86	59.17	57.91	61.25	98.22	96.79	47.30	56.90	72.90	69.60	87.27	86.35	87.40
C3	48.80	57.94	59.32	29.50	30.56	49.19	40.33	58.17	91.99	94.53	56.22	34.05	74.52	72.12	85.31	84.98	2.963
C4	52.91	58.13	59.09	14.37	27.07	35.06	25.67	44.86	92.71	91.46	74.42	47.68	14.52	36.56	83.20	83.33	77.43
C5	54.20	59.87	58.85	25.28	29.76	43.84	39.06	54.82	71.5	69.71	55.5	31.53	39.66	44.88	79.07	17.16	28.34

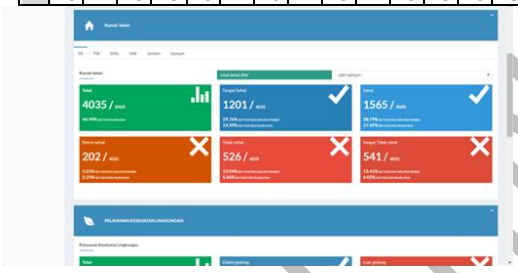

Figure 1 Main Page

Figure 2 Analysis Page

Because the centroid doesn't change (same as the previous centroid), the clustering process is complete, then the data will be grouped based on the closest distance to the cluster, from 8969 K-Means data managed to group 3308 into Very Healthy categories, 2496 Healthy categories, 792 Unhealthy categories, 1706 Unhealthy categories, and 667 Very Unhealthy categories.

C. *System Implementation*

Figure 1 is the implementation of the main page who displays the SIPP-KLING dashboard. On this page the system displays the total amount of data based on the status category such as whether the place is healthy or not, feasible or not, and many or few deviations. The grouped data is the result of the K-Means process. Figure 2 is an implementation of the analysis page for each village, analysis of the location is carried out in each attribute to see which attributes most influence the village. The highest scale on the graph is the scale that most influences and becomes a warning for a village. On the page there are several buttons, the first button to send the status of the database to the database, the second button to send the distance value of the attribute to the database.

D. *Testing Data Accuracy*

The accuracy of the algorithm is tested using Confusion Matrix [6]. The testing aims to determine the performance of the K-Means Clustering algorithm in classifying the data into a predetermined cluster.

TABLE 6. Value of Confussion Matrix

Cofussion Classification	Value
True Positive (TP)	4865
True Negative (TN)	2943
False Positive (FP)	222
False Negative (FN)	939

a. Precision

Precision is the amount of data that is true positive (the amount of positive data that is correctly recognized as positive) is divided by the amount of data that is recognized positively. From the test of precision value of the K-means is 95.63% for Healthy House and 75.71% for Unhealthy House.

TP / FP + TP) x 100% ↔ 4865 / (4865+222) = 0.956359 (Healthy House)

TN / (FN + TN) x 100% ↔ 2943/ (2943 + 939) = 0.7571 (Unhealthy House)

Assuming that status of the K-means is very healthy and healthy compared

b. Recall

Recall is the amount of data which is true positive divided by the amount of data which is actually positive (true positive + true negative). For recall

value is 83.82% for Healthy House and 92.98% for Unhealthy House.

TP / (FN + TP) x 100% ↔ 4865 / (939 + 4865) = 0.8382 (Healthy House)

TN / (FP + TN) x 100% ↔ 2943 / (222 + 2943) = 0.9298 (Unhealthy House)

c. Accuracy

By knowing the amount of data grouped correctly it can be seen the accuracy of the prediction that is:

(TP + TN) / (TP + TN + FP + FN) x 100% ↔ (4865+2943) / (4865 + 939 + 222 + 2943) = 0.8705541309

Refer to [7], the accuracy value is 0.870, and include as good classification.

d. F-Measure

F-Measure is a value obtained from precision and recall measurements between clustered classes and actual classes. The ratio of f-measure is reversed, so the higher the fmeasure value, the smaller the difference between the same recall precision.

F-Measure = 2 / (1/recall + 1/precision) or
F(i,j) = (2 * recall(i,j) * precision(i,j)) / ((precision(i,j) + recall(i,j)) **(eq 2)**

For Healthy House :

2 / (1 / 0.9563 + 1 / 0.8382) = 0.893363789

(2 * 0.9563 * 0.8382) / (0.9563 + 0.8382) = 0.893363789

For Unhealthy House :

2 / (1 / 0.7581 + 1 / 0.9298) = 0.835216992

(2 * 0.7581 * 0.9298) / (0.7581 + 0.9298) = 0.83521

e. Data Analysis

From the above test results shows the accuracy of this method is 87, 05%, with precision of 95.64% for healthy class and 75, 81% for unhealthy class, and recall 83.82% for healthy and 92, 98 % to unhealthy. Results accuracy, precision and recall in the grouping of data Healthy House can be seen in the TABLE 7.

TABLE 7 Accuracy Result

Accuracy 87,05%	True Healthy	True Unhealthy	Class Precision
Pred. Healthy	4865	939	95,64%
Pred. Unhealthy	222	2943	75,81%
Class Recall	83,82%	92,98%	

IV. CONCLUSION

The implementation of K-Means Clustering can be classified the 8969 data into 5 categories of Healthy House namely: Very Healthy, Healthy, Not Healthy, Unhealthy and Very Unhealthy The data obtained from the clustering results of K-means can help to analyze which parts of a house should be better addressed, or which areas have lower levels of health. The K-Means Clustering accuracy test obtained an accuracy of 87% that included in the category of good clustering.

REFERENCES

[1] Arpita A and Hitesh G 2013 Global K-Means (GKM) Clustering Algorithm: A Survey *International Journal of Computer Applications (IJCA)* vol 79 no 2 p 20-24

[2] Chunfei Z and Zhiyi F 2013 An Improved K-Means Clustering Algorithm *Journal of Information & Computational Science* (JICS) 10: 1 p 193-199

[3] Unnati R Raval and Chaita J 2016 Implementing & Improvisation of K-means Clustering Algorithm *International Journal of Computer Science and Mobile Computing (IJCSMC)* vol 5 issue 5 p 191-203

[4] Bhoomi B, Prof. Nirali M, and Prof. Vimal P 2013 A survey on Efficient Enhanced K-Means Clustering Algorithm *International Journal for Scientific Research & Development (IJSRD)* vol 1 issue 9 p 1756-1758

[5] Shi N, Liu X and Guan Y 2010 Research on K-means Clustering Algorithm An Improved K-means Clustering Algorithm *Third International Symposium on Intelligent Information Technology and Security Informatics* (IITSI) p 63-67

[6] Jasmina Dj N, Alempije V, Sinisa S. I, Zeljko P and Milica T 2017 Evaluation of Classification Models in Machine Learning *Theory and Applications of Mathematics & Computer Science 7* p 39-46

[7] Gorunescu F 2011 Data Mining Concepts, Model and Techniques Berlin: Springer