Journal of Intelligent Decision Support System (IDSS) 5 (3) (2022) 89-98
Published by: IOCSCIENCE
Journal of Intelligent Decision Support System (IDSS)
Journal homepage: www.idss.iocspublisher.org

Application of rapidminer for clustering aids cases by province using data mining k-means clustering

Widya Surya Ningsih[1], Eko Haryanto[2]
[1,2]Computer System, Saintek Faculty, Universitas Pembangunan Panca Budi Medan, Indonesia

Article Info

Article history:

Received Aug 9, 2022
Revised Aug 30, 2022
Accepted Sep 20, 2022

Keywords:

AIDS Disease
Clustering
K-means
Maining Data

ABSTRACT

Acquired Immunodeficiency Syndrome or Acquired Immune Deficiency Syndrome (AIDS shortened) is a combination of symptoms and diseases caused by the HIV virus's damage to the human immune system. This study examines the WEKA Application for K-means Clustering Data Mining in Grouping AIDS Cases by Province. The increasing number of AIDS patients in Indonesia is a matter that never escapes the government's notice. People are concerned about the spread of the AIDS virus due to the persistently rising death rate. Documents supplied by the Social Security Administering Body describing the number of villages/subdistricts with health facilities were mined for data and study. This research utilizes data from the years 2008-2011 for a total of 34 provinces. There are two assessment criteria: 1) the average number of AIDS cases and 2) the average number of AIDS-related deaths controlled by three clusters: high cluster level (C1), medium cluster level (C2), and low cluster level (C3) (C3). So that the C1 cluster evaluation for AIDS cases is based on four provinces, Papua, DKI Jakarta, West Java, and East Java, nine provinces for the C2 cluster, and twenty provinces for the C3 cluster. This information can be sent to provinces who are concerned about the number of AIDS cases.

Corresponding Author:

Widya Surya Ningsih,
Computer System, Saintek Faculty,
Universitas Pembangunan Panca Budi Medan,
4, Jl. Gatot Subroto No.km, Simpang Tj., Kec. Medan Sunggal, Kota Medan, Sumatera Utara 20122
Email: rindiantika0048@gmail.com

1. Introduction

Acquired Immunodeficiency Syndrome or Acquired Immune Deficiency Syndrome (AIDS shortened) is a combination of symptoms and diseases caused by the HIV virus's damage to the human immune system. According to the data and information department of the Indonesian Ministry of Health, there were 35 million HIV-positive individuals globally in 2013, including 16 million women and 3.2 million children younger than 15 years old. In 2013, there were 2.1 million new HIV infections, including 1.9 million adults and 240,000 children younger than 15 years old. AIDS has caused 1.5 million deaths, including 1.3 million adults and 190,000 children younger than 15 years old. In 1987, HIV AIDS was first identified in the province of Bali, Indonesia. Currently, AIDS has spread to 386 districts/cities in all Indonesian provinces. From 1987 through September 2014, the following number of new AIDS cases was reported:

Figure 1. Number Of Hiv Cases In Indonesia

Figure 1 demonstrates an upward trend in the number of HIV cases from year to year since the virus was first identified (in 1987). On the other side, the number of AIDS cases tended to increase slowly, and by 2012, the number of AIDS cases had begun to fall. Since 1987 till September 2014, a total of 150,296 individuals have been living with HIV.

Several solutions, including cluster analysis, are available for the aforementioned difficulties. Cluster analysis is a multivariate approach whose primary purpose is to classify items according to their features. Currently, cluster analysis is utilized extensively in numerous domains, as documented in numerous publications and journals [1]. The clustering approach emphasizes the iterative search for cluster centers, where the cluster center is selected based on the shortest distance between each data point and the cluster center. This study drew its information from records issued by the Central Statistics Agency. In this instance, the researcher brought up the issue of classifying the number of AIDS cases by province using clustering. The variables utilized in the clustering procedure are 1) and 2) The average number of new cases of AIDS and 2). The average number of AIDS-related fatalities. The cluster's results can be used as input by the government so that provinces in the highest cluster (C1) receive more attention. C1: Patients with high AIDS instances; C2: Patients with intermediate AIDS cases; and C3: Patients with low AIDS cases.

Data mining is also a technique used for large-scale data processing; consequently, data mining plays a vital role in a variety of fields, including business, finance, meteorology, science, and technology. Data mining can alternatively be regarded as a sequence of actions designed to discover intriguing patterns among vast quantities of data, which are then saved in databases, data warehouses, or information storage. Several sciences, including data analysis, signal processing, neural networks, and pattern recognition, enable data mining approaches [1].

Clustering analysis is the process of breaking a set of data into many groups whose data similarity in one group is larger than the data similarity between groups. The promise of clustering is that it may be used to establish the data's structure, which can then be applied to a wide range of applications, including classification, image processing, and pattern recognition [2].

K-means is a data partitioning technique that divides data into multiple groups. The distance between the data and the cluster can be minimized using this approach. Using this technique in the clustering process is mostly determined by the data acquired and the conclusions to be made at the end of the procedure [3]. As a result, the following rules apply when using the kmeans algorithm [4]: (a)What is the minimum number of clusters that must be included? (b)It only contains numeric properties. Essentially, the K-means algorithm only uses a portion of the total number of components retrieved to determine the initial cluster center, which is chosen at random from the data population. The K-means algorithm will next evaluate each of the data population's components and assign them to one of the predetermined cluster centers based on the minimum distance between them and each cluster center. Furthermore, until all data components are classified into each cluster and a new cluster is formed, the position of the cluster's center will be reassessed [5].

2. Method

2.1 Method Of Collecting Data

In the data collection activities for this research, the literature study collection method was used which in this method the activities carried out were studying, searching and collecting data related to this research. The data used in grouping AIDS cases was obtained from documents describing the number of AIDS cases by province by the Central Statistics Agency in 2018 – 2011.

2.2 Data Analysis Method

Data analysis is the process of systematically searching and compiling data obtained from interviews, documentation, personal documents, observations, field notes, photographic images, and other sources by organizing the data into categories, breaking them down into units, synthesizing, organizing into patterns, selecting which ones are important and will be studied, and drawing conclusions so that they can be understood by themselves and others. Qualitative research is an analysis that is carried out by grouping data to establish a pattern of items being investigated and comparing the concepts that exist in the source in order to write this study using qualitative data analysis.

2.3 Study Of Literature

Scientific activities that are carried out in order to find solutions to a single problem, with the ultimate goal of making a theoretical or practical contribution to the advancement of the subject of science in question. The data processing of AIDS cases is included in the literature review.

3. Result and Discussion

In conducting clustering, the data obtained will be calculated first based on the number of AIDS Cases by Province. The average result is based on 2 assessment criteria, namely 1). Average number of AIDS cases by province and 2). The average number of cases of death of AIDS sufferers is shown in the following table:

Table 1
Data on the Number of AIDS Cases by Province

Province	Number of Cases				Died			
	2008	2009	2010	2011	2008	2009	2010	2011
Aceh	22	36	53	63	8	9	12	14
North Sumatra	670	485	507	509	135	93	94	94
West Sumatra	184	293	410	428	61	75	99	104
Riau	364	371	477	477	116	117	132	132
Jambi	106	165	268	291	30	50	62	66
South Sumatra	124	219	219	219	29	38	38	38
Bengkulu	33	85	131	137	10	18	29	30
Lampung	143	144	144	144	42	42	42	42
Bangka Belitung Islands	67	117	120	120	4	18	18	18
Riau islands	271	333	374	390	114	130	143	148
DKI Jakarta	2 727	2 811	3 995	3 997	440	425	576	577
West Java	2 603	3 233	3 728	3 809	503	588	665	678
Central Java	409	669	944	1 336	171	238	289	370
In Yogyakarta	129	247	505	673	46	70	108	134
East Java	2 525	3 133	3 771	3 775	575	680	779	779
Banten	71	275	401	403	12	51	67	68
Bali	869	1 506	1 747	1 747	145	275	311	311
West Nusa Tenggara	114	107	142	206	47	56	69	110
East Nusa Tenggara	110	138	242	385	23	25	36	50
West Kalimantan	730	730	1 125	1 125	110	103	138	138
Central Kalimantan	7	15	57	69	2	2	4	4

South Kalimantan	22	27	27	28	5	5	5	5
East Kalimantan	11	11	11	12	10	10	10	10
North Sulawesi	117	173	173	557	44	62	62	125
Central Sulawesi	8	12	12	12	4	6	6	6
South Sulawesi	143	143	591	995	62	62	62	167
Southeast Sulawesi	9	20	22	26	1	4	5	5
Gorontalo	3	3	3	3	1	1	1	1
West Sulawesi	-	-		0	-	-		0
Maluku	186	192	192	195	68	70	70	70
North Maluku	7	10	17	17	7	8	8	8
West Papua	58	58	58	397	19	19	19	152
Papua	2 294	2 681	3 665	3 938	353	358	580	602

After that, the data is compiled, and the average value of each criterion is calculated, as shown in table 2 below:

Table 2
Data on Average Number of AIDS Cases by Province

Province	Number of Cases				Average	Died				Average
	2008	2009	2010	2011		2008	2009	2010	2011	
Aceh	22	36	53	63	43,5	8	9	12	14	10,75
North Sumatra	670	485	507	509	542,75	135	93	94	94	104
West Sumatra	184	293	410	428	328,75	61	75	99	104	84,75
Riau	364	371	477	477	422,25	116	117	132	132	124,25
Jambi	106	165	268	291	207,5	30	50	62	66	52
South Sumatra	124	219	219	219	195,25	29	38	38	38	35,75
Bengkulu	33	85	131	137	96,5	10	18	29	30	21,75
Lampung	143	144	144	144	143,75	42	42	42	42	42
Kep. Bangka	67	117	120	120	106	4	18	18	18	14,5
Riau Islands	271	333	374	390	342	114	130	143	148	133,75
DKI Jakarta	2727	2811	3995	3997	3383	440	425	576	577	504,5
West Java	2 603	3233	3 728	3 809	3233	503	588	665	678	608,5
Central Java	409	669	944	1 336	674	171	238	289	370	267
In Yogyakarta	129	247	505	673	388,5	46	70	108	134	89,5
East Java	2 525	3 133	3771	3 775	3771	575	680	779	779	703,25
Banten	71	275	401	403	287,5	12	51	67	68	49,5
Bali	869	1 506	1 747	1 747	869	145	275	311	311	260,5
NTB	114	107	142	206	142,25	47	56	69	110	70,5
NTT	110	138	242	385	218,75	23	25	36	50	33,5
West Kalimantan	730	730	1 125	1 125	730	110	103	138	138	122,25
Central Kalimantan	7	15	57	69	37	2	2	4	4	3
South Kalimantan	22	27	27	28	26	5	5	5	5	5
East Kalimantan	11	11	11	12	11,25	10	10	10	10	10
North Sulawesi	117	173	173	557	255	44	62	62	125	73,25
Sul.Central	8	12	12	12	11	4	6	6	6	5,5
South Sulawesi	143	143	591	995	468	62	62	62	167	88,25
Central Sulawesi	9	20	22	26	19,25	1	4	5	5	3,75
Gorontalo	3	3	3	3	3	1	1	1	1	1

West Sulawesi	-	-		0	0	-	-		0	0
Maluku	186	192	192	195	191,25	68	70	70	70	69,5
North Maluku	7	10	17	17	12,75	7	8	8	8	7,75
West Papua	58	58	58	397	142,75	19	19	19	152	52,25
Papua	2 294	2 681	3665	3 938	3665	353	358	580	602	473,25
Indonesia	15 136	18 442	24131	26 483	24131	3 197	3708	4 539	5 056	3708

The value of each variable will be determined after the data has been amassed and the average value has been determined. The data will then be clustered using the K-means technique, which divides the data into three groups.

3.1 Centroid Data

In the application of the K-means algorithm, the midpoint or centroid value is obtained from the data obtained provided that the desired clusterization is 3, the cluster determination is divided into three parts, namely high clusters (C1), medium clusters (C2) and low-level clusters (C3). . then the value of the midpoint or centroid also has 3 points. Determination of the cluster point is done by taking the largest value (maximum) for the high cluster (C1), the average value (average) for the medium cluster (C2) and the smallest value (minimum) for the low cluster (C3). The point value can be seen in the following table:

Table 3. Initial Data Centroid (Iteration 1)

	Centroid	
Max (C1)	3771	703,25
Average (C2)	635,3636	125
Min (C3)	0	0

3.2 Clustering Data

The data obtained can be clustered into three clusters by utilizing the centroid. Clustering is done by calculating the shortest distance between each processed data point. In iteration 1, categories for the three clusters were derived from data on the number of villages/kelurahan with health services per province. DKI Jakarta, West Java, East Java, and Papua are clusters of patients with high AIDS cases (C1). Clusters of persons with moderate AIDS cases (C2) are found in nine provinces, whereas clusters of people with low AIDS cases (C3) are found in twenty more. The following table and picture describe the method of finding the smallest distance, grouping data in iteration 1, and top clustering:

Table 4
Calculation of Cluster Center Distance Iteration 1

Province	**X**	**Y**	**iteration**			
			C1	**C2**	**C3**	**Shortest Distance**
Aceh	43,50	10,75	3791,28	602,79	44,81	44,81
North Sumatra	542,75	104,00	3283,40	94,96	552,62	94,96
West Sumatra	328,75	84,75	3497,37	309,24	339,50	309,24
Riau	422,25	124,25	3398,44	213,11	440,15	213,11
Jambi	207,50	52,00	3622,52	434,05	213,92	213,92
South Sumatra	195,25	35,75	3637,52	449,07	198,50	198,50
Bengkulu	96,50	21,75	3737,16	548,67	98,92	98,92
Lampung	143,75	42,00	3687,03	498,57	149,76	149,76

Kep. Bangka	106,00	14,50	3729,16	540,77	106,99	106,99
Riau islands	342,00	133,75	3475,97	293,49	367,22	293,49
DKI Jakarta	3382,50	504,50	436,39	2773,23	3419,92	436,39
West Java	3233,00	608,50	546,28	2642,25	3289,77	546,28
Central Java	674,00	267,00	3127,57	147,16	724,96	147,16
In Yogyakarta	388,50	89,50	3437,73	249,40	398,68	249,40
East Java	3771,00	703,25	0,00	3188,51	3836,01	0,00
Banten	287,50	49,50	3544,31	355,96	291,73	291,73
Bali	869,00	260,50	2935,58	270,09	907,21	270,09
NTB	142,25	70,50	3683,50	496,12	158,76	158,76
NTT	218,75	33,50	3614,84	426,54	221,30	221,30
West Kalimantan	730,00	122,25	3096,00	94,68	740,17	94,68
Central Kalimantan	37,00	3,00	3799,09	610,67	37,12	37,12
South Kalimantan	26,00	5,00	3809,54	621,07	26,48	26,48
East Kalimantan	11,25	10,00	3823,13	634,62	15,05	15,05
North Sulawesi	255,00	73,25	3572,00	383,87	265,31	265,31
Central Sulawesi	11,00	5,50	3824,19	635,70	12,30	12,30
South Sulawesi	468,00	88,25	3359,77	171,35	476,25	171,35
Southeast Sulawesi	19,25	3,75	3816,40	627,93	19,61	19,61
Gorontalo	3,00	1,00	3832,88	644,41	3,16	3,16
West Sulawesi	0,00	0,00	3836,01	647,54	0,00	0,00
Maluku	191,25	69,50	3635,42	447,57	203,49	203,49
North Maluku	12,75	7,75	3822,06	633,56	14,92	14,92
West Papua	142,75	52,25	3686,19	497,96	152,01	152,01
Papua	3665,00	473,25	253,25	3049,59	3695,43	253,25

Table 5
Results of Grouping Iteration 1

Province	**C1**	**C2**	**C3**
Aceh			1
North Sumatra		1	
West Sumatra		1	
Riau		1	
Jambi			1
South Sumatra			1
Bengkulu			1
Lampung			1
Kep. Bangka			1
Riau islands		1	
DKI Jakarta	1		
West Java	1		
Central Java		1	

In Yogyakarta		1	
East Java	1		
Banten			1
Bali		1	
West Nusa Tenggara			1
East Nusa Tenggara			1
West Kalimantan		1	
Central Kalimantan			1
South Kalimantan			1
East Kalimantan			1
North Sulawesi			1
Central Sulawesi			1
South Sulawesi		1	
Southeast Sulawesi			1
Gorontalo			1
West Sulawesi			1
Maluku			1
North Maluku			1
West Papua			1
Papua	1		
Total	4	9	20

The K-means process will iterate until the data grouping is identical to the prior iteration's data grouping. To put it another way, the process will continue to iterate until the data in the most recent iteration is identical to that in the prior iteration. Data clusters on the number of AIDS cases were obtained by province in iteration 1. At the next iteration, the iteration process will come to an end, and the process of determining the middle or centroid value will begin. The same procedure is used after obtaining the midway or centroid value by determining the closest distance. The following table summarizes the shortest distance search method, data grouping in the last iteration, and data clustering:

Table 6.
Centroid Data Iteration 2

	Centroid	
Cluster C1	3512,88	572,38
Cluster C2	529,47	141,58
Cluster C3	107,51	28,06

Table 7. Perhitungan Jarak Pusat Cluster Iterasi 2

Province	**X**	**Y**	**Iteration 2**			**Shortest Distance**
			C1	**C2**	**C3**	
Aceh	43,50	10,75	3514,54	503,28	66,31	66,31
North Sumatra	542,75	104,00	3006,83	39,86	441,81	39,86
West Sumatra	328,75	84,75	3221,25	208,61	228,38	208,61
Riau	422,25	124,25	3122,94	108,61	329,11	108,61

Jambi	207,50	52,00	3346,09	334,20	102,81	102,81
South Sumatra	195,25	35,75	3360,74	350,58	88,07	88,07
Bengkulu	96,50	21,75	3460,46	449,25	12,69	12,69
Lampung	143,75	42,00	3410,62	398,37	38,83	38,83
Kep. Bangka	106,00	14,50	3452,25	442,13	13,65	13,65
Riau islands	342,00	133,75	3201,07	187,64	257,20	187,64
DKI Jakarta	3382,50	504,50	146,99	2876,02	3309,46	146,99
West Java	3233,00	608,50	282,20	2743,55	3178,93	282,20
Central Java	674,00	267,00	2855,25	191,36	614,82	191,36
In Yogyakarta	388,50	89,50	3161,47	150,29	287,63	150,29
East Java	3771,00	703,25	289,41	3289,83	3725,19	289,41
Banten	287,50	49,50	3267,48	258,90	181,26	181,26
Bali	869,00	260,50	2662,21	359,75	796,17	359,75
West Nusa Tenggara	142,25	70,50	3407,78	393,69	54,84	54,84
East Nusa Tenggara	218,75	33,50	3337,91	328,98	111,37	111,37
West Kalimantan	730,00	122,25	2819,04	201,46	629,57	201,46
Central Kalimantan	37,00	3,00	3522,20	511,60	74,83	74,83
South Kalimantan	26,00	5,00	3532,73	521,67	84,71	84,71
East Kalimantan	11,25	10,00	3546,50	534,67	97,94	97,94
North Sulawesi	255,00	73,25	3295,89	282,85	154,25	154,25
Central Sulawesi	11,00	5,50	3547,46	536,03	99,11	99,11
South Sulawesi	468,00	88,25	3083,12	81,38	365,48	81,38
Southeast Sulawesi	19,25	3,75	3539,60	528,51	91,55	91,55
Gorontalo	3,00	1,00	3556,08	544,92	107,96	107,96
West Sulawesi	0,00	0,00	3559,20	548,08	111,11	111,11
Maluku	191,25	69,50	3359,48	345,82	93,43	93,43
North Maluku	12,75	7,75	3545,37	533,77	96,92	96,92
West Papua	142,75	52,25	3410,03	396,91	42,74	42,74
Papua	3665,00	473,25	181,57	3153,02	3585,23	181,57

Table 8
Results of Grouping Iteration 1

Province	**C1**	**C2**	**C3**
Aceh			1
North Sumatra		1	
West Sumatra		1	
Riau		1	
Jambi			1
South Sumatra			1
Bengkulu			1
Lampung			1
Kep. Bangka			1
Riau islands		1	
DKI Jakarta	1		
West Java	1		
Central Java		1	
In Yogyakarta		1	

East Java	1		
Banten			1
Bali		1	
West Nusa Tenggara			1
East Nusa Tenggara			1
West Kalimantan		1	
Central Kalimantan			1
South Kalimantan			1
East Kalimantan			1
North Sulawesi			1
Central Sulawesi			1
South Sulawesi		1	
Southeast Sulawesi			1
Gorontalo			1
West Sulawesi			1
Maluku			1
North Maluku			1
West Papua			1
Papua	1		
total	4	9	20

3.3 Data analysis

The data grouping conducted on three clusters in iteration 2 yielded the same findings as iteration 1. Four provinces with high-level clusters for AIDS cases, namely Papua, DKI Jakarta, West Java, and East Java, nine provinces with moderate-level clusters, and 20 provinces with low-level clusters, may be identified from 34 data on the number of AIDS cases by province.

3.4 RapidMiner Tool Implementation

Researchers use K-means to organize data based on attributes in clustering. In data grouping (clustering) operations, K-Means is a relatively simple and quick technique. The fundamental idea behind this method is to create k prototypes/centroids/means out of a set of n-dimensional data. The Rapidminer application will aid in the clustering and grouping process.

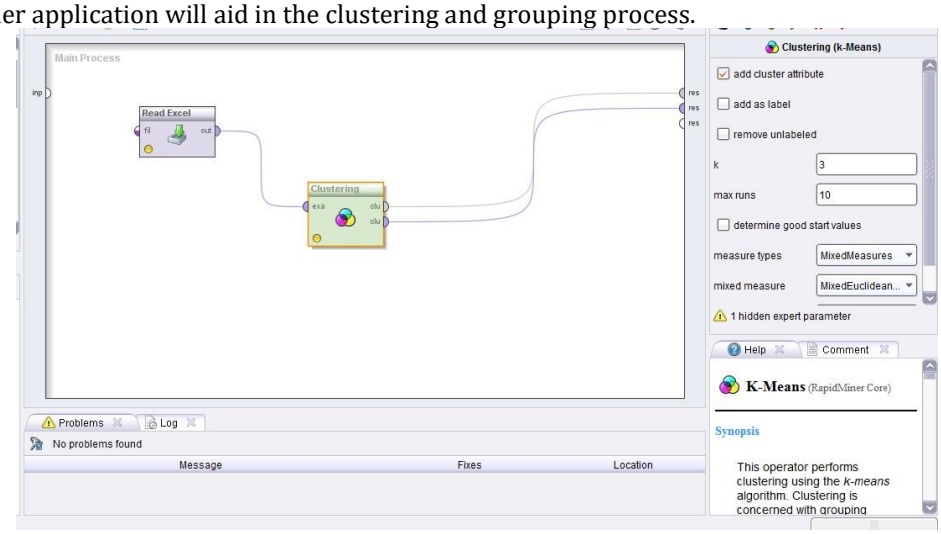

Figure 2. Using the K-Means Algorithm with a value of K = 3

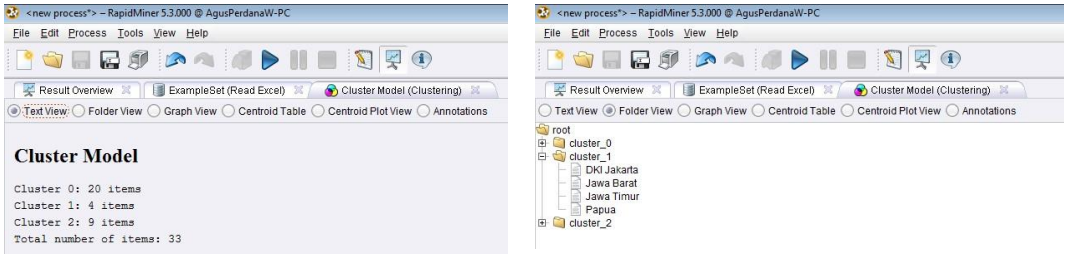

Figure 3. Grouping Results with K=3

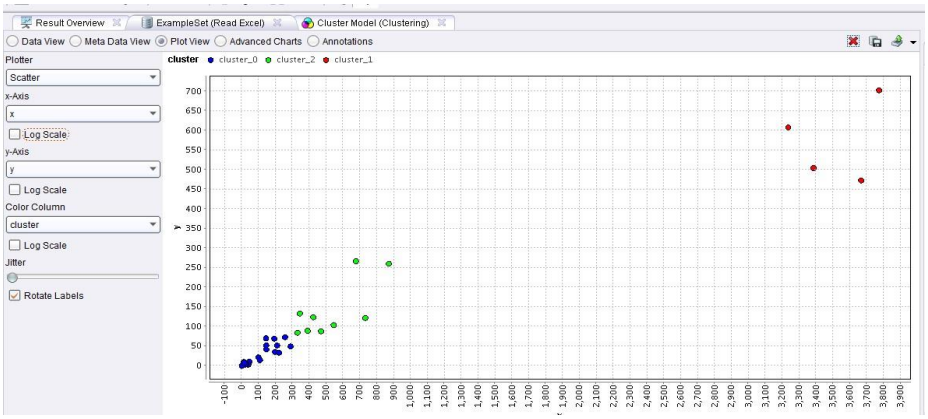

Figure 4. Rapidminer graph with K=3

The same results are achieved in cluster grouping, as shown in Figure 3, where C1 cluster comprises of four provinces: DKI Jakarta, West Java, East Java, and Papua.

4. Conclusion

The K-means clustering method can be used to assess the instances of AIDS patients by province. The information is analyzed to determine the worth of AIDS cases. Ms. Excel is used to find the centroid value in three clusters: high cluster (C1), medium cluster (C2), and low cluster (C3) (C3). As a result, an assessment is made based on the clustering of AIDS cases by province, with four provinces in cluster C1 (Papua, DKI Jakarta, West Java, and East Java), nine provinces in cluster C2, and 20 provinces in cluster C3. The findings of the study can be used to inform the government, particularly those provinces with the highest concentration of AIDS patients, so that further action can be taken.

References

[1] Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

[2] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence, 110, 104743.

[3] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295.

[4] Fränti, P., & Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. Applied Intelligence, 48(12), 4743-4759.

[5] Kuswandi, D., Surahman, E., Thaariq, Z. Z. A., & Muthmainnah, M. (2018, October). K-Means clustering of student perceptions on project-based learning model application. In 2018 4th International Conference on Education and Technology (ICET) (pp. 9-12). IEEE.