ISSN No:-2456-2165

Identification of Characteristics of Covid-19 Infection Using the K-Means Clustering Method

Rina Fitriana[1]
[1]Industrial Engineering Department, Faculty of Industrial Engineering, Universitas Trisakti
[1]Jakarta, Indonesia

Yanto[2], Isdaryanto Iskandar[3]
[2,3]Faculty of Engineering, Atma Jaya Catholic University, Jakarta,Indonesia

Abstract:- 2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. Early on, many of the patients in the outbreak in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring. At this time, it's unclear how easily or sustainably this virus is spreading between people. The purpose of this study is to identify the type of data on the COVID-19 outbreak. Based on the outbreak of COVID-19 in the several area around the first identified cases, datasets for the infection based on several criteria have been made. The criteria of datasets include: reporting date; location; country; gender; and age. It evaluates how the data going to be grouped into several similar characteristics, so the report for this new viruses can be identified. Within those criteria, the data going to be analyzed with the clustering method which is specifically the k-means clustering. The k-means will group the data based on the similarity between each data for the purpose of visualizing the COVID-19 undefined data. The results obtained from the Kaggle study were data on the COVID-19 virus infection. In designing data mining, it uses the K-means clustering method. The results of k- Means clustering data mining consist of 39.97% cluster 1, 20.04% cluster 2, and 39.97% cluster 3 with Microsoft Excel, 47% cluster 1 and 53% cluster in WEKA, 32.74% cluster 1, 39.62% cluster 2, 27.64% cluster 3 at KNIME.

Keywords:- *COVID-19, Data Mining, Clustering, K-means.*

I. INTRODUCTION

According to Tan, 2006 clustering is a process to group data into several clusters or groups so that the data in one cluster has the maximum level of similarity and the data between clusters has the minimum similarity.[1]

K-Means Clustering is a data analysis method or Data Mining method that performs the modeling process without supervision (unsupervised) and is one method that performs data grouping with a partition system. The K-Means Clustering method tries to group the existing data into several groups, where the data in one group has the same characteristics as each other and has different characteristics from the data in other groups.[1]

In other words, the K-Means Clustering method aims to minimize the objective function set in the clustering process by minimizing variations between data in one cluster and maximizing variation with data in other clusters.

This is supported by previous research where data mining with the K Means Clustering method has been used to improve product quality in bakery factories [2]. Research using the K Means Clustering method has been used to analyze Covid data [3]. Research in customer management and quality management for the dairy agro-industry by applying data mining and OLAP Cube [4]. Research by applying data mining and OLAP Cube as a tool to obtain useful and accurate information in decision making in the marketing department at the bakery [5]. Research in dairy agroindustry made models and rule formulations from the acceptance of revolving cattle credit with classification techniques with decision tree algorithms from Data Mining [6]. The purpose of clustering is to determine the intrinsic clustering in an unlabeled data set depending on some measure of similarity, for example Euclidean distance [7]. The clustering methods used in this study are K-Means and K-Medoids because both methods include partitioning-based clustering which divides large datasets into K numbers (defined by the user) groups by using distance as a measure of similarity, where each group represents a cluster [8]. K-Means clustering is a clustering algorithm that is often used. Beginning by determining a point K as the initial center (centroid). Furthermore, all points are directed to the nearest centroid based on the short distance. After the cluster is formed, the centroid is performed on each cluster. Then repeat this step iteratively until there is no change in the centroid [9]. Forest patches were identified based on classification and hierarchical merging of a LiDAR-derived Canopy Height Model in a tropical rainforest in Sumatra, Indonesia. Attributes of the identified patches were used as inputs for k-medoids clustering [10]. The selection of K-Means in this study was based on several advantages compared to other methods, including its simple use, high level of accuracy, and can be applied to many things. This research studies using the K-Means method include simulation of electric power systems climate grouping [11].

II. RESEARCH METHODOLOGY

The method used in this research is part of clustering. Clustering is a method of grouping data.

The research carried out will be made into 2 different flowcharts related to the research methodology used to approach groups each problem, namely the data processing. Flowchart of Methodology can be seen in Figure 1

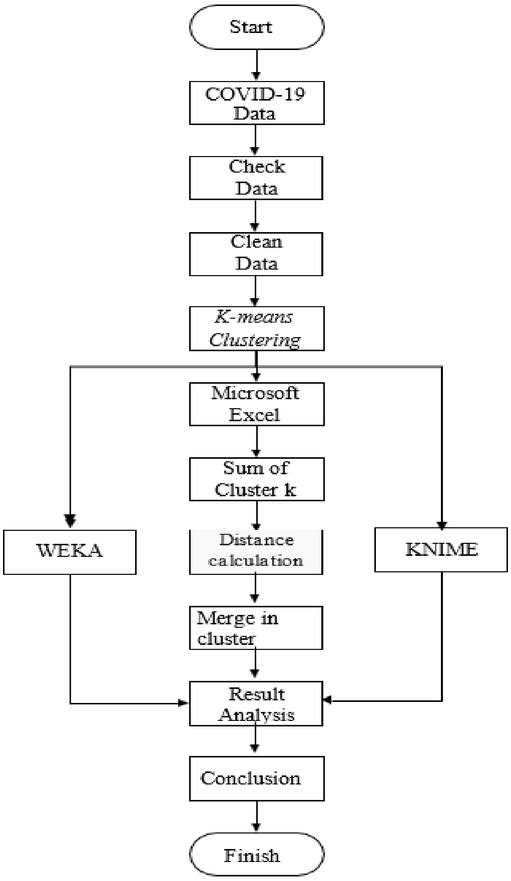


Fig 1:- Flowchart Data Processing

III. RESULT AND ANALYSIS

Based on the results of the evaluation, data on COVID-19 is obtained as follows.[12]

Reporting Date	Location	Country	Gender	Age
1/20/2020	Shenzhen, Guangdong	China	male	66
1/20/2020	Shanghai	China	female	56
1/21/2020	Zhejiang	China	male	46
1/21/2020	Tianjin	China	female	60
1/21/2020	Tianjin	China	male	58
1/21/2020	Chongqing	China	female	44
1/21/2020	Sichuan	China	male	34
1/21/2020	Beijing	China	male	37
1/21/2020	Beijing	China	male	39
1/21/2020	Beijing	China	male	56
1/21/2020	Beijing	China	female	18
1/21/2020	Beijing	China	female	32
1/21/2020	Shandong	China	male	37
1/21/2020	Yunnan	China	male	51
1/22/2020	Sichuan	China	male	57
1/22/2020	Jiangxi	China	male	56
1/22/2020	Jiangxi	China	male	50
1/22/2020	Macau	China	female	52
1/22/2020	Liaoning	China	male	33
1/22/2020	Liaoning	China	male	40
1/22/2020	Fujian	China	male	70
1/22/2020	Guizhou	China	male	51
...	...	...	...	...
...	...	...	...	...
2/25/2020	Bern	Switzerland	male	70 jhsjka

Table 1:- Covid-19 Data

The COVID-19 virus data is divided into five criteria after the cleaning process is carried out. These five criteria include reporting date, location, country, gender, and age. This data is considered sufficient and complete without any data that has empty attributes. Thus, the process can be continued in the analysis using WEKA and KNIME.

A. Microsoft Excel

The purpose of this research is how to group the cases of the spread of the COVID-19 virus infection in the area around the main spread of Wuhan, China.

The number of data that will be tested using Microsoft Excel is 843 COVID-19 infection data. This data can be seen in table 2.

Reporting date	Location	Country	Gender	Age
1/20/2020	Shenzhen, Guangdong	China	male	66
1/20/2020	Shanghai	China	female	56
1/21/2020	Zhejiang	China	male	46
1/21/2020	Tianjin	China	female	60
1/21/2020	Tianjin	China	male	58
1/21/2020	Chongqing	China	female	44
1/21/2020	Sichuan	China	male	34
1/21/2020	Beijing	China	male	37
1/21/2020	Beijing	China	male	39
1/21/2020	Beijing	China	male	56
.	.	.	.	.
2/25/2020	Bern	Switzerland	male	70

Table 2:- Covid-19 Data

The data contained in Table 3 cannot be directly processed because there are different types of data. Therefore, the solution is to convert it to nominal data. However, the age data no longer needs to be converted because it already meets the requirements in the form of nominal data.

Atribut	Description	Nominal
Reporting Date	1/20/2020 1/21/2020	1 2
Location	Shenzhen, Guangdong Shanghai Zheijiang Tianjin Chongqing Shicuan Beijing	1 2 3 4 5 6 7
Country	China	1
Gender	Male Female	1 2

Table 3:- CONVERTION CRITERIA

After making the conversion criteria, the data can be directly converted and the results of the process copied

Reporting Date	Location	Country	Gender	Age
1	1	1	1	66
1	2	1	2	56
2	3	1	1	46
2	4	1	2	60
2	4	1	1	58
2	5	1	2	44
2	6	1	1	34
2	7	1	1	37
2	7	1	1	39
2	7	1	1	56
.	.	.	.	.
47	32	19	2	70

Table 4:- Data Convertion

The difference in distance or the magnitude of the numbers that are far enough can make it difficult in the grouping process. One of the solutions used to reduce the number of variables between variables is to normalize the numbers in the height and weight variables using equation 1.

$$n\text{normalized value} = \frac{(\text{initial value}-min\ \text{value})}{(max\ \text{value}-\text{min value})} \quad (1)$$

The steps for the normalization process are:

a. Finding the max and min values for the reporting date variable.

Max value (Xmax) = 2

Min value(Xmin) = 1

b. Calculating the normalized value using equation 1

$$X11 = \frac{(Xreportingdate1 - Xmin)}{X\ max\quad min}$$

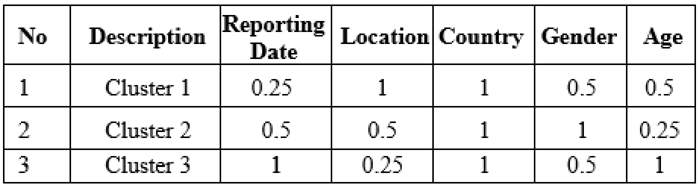
$$= \frac{(1-1)}{(2-1)} = 0$$

The same calculation is carried out until the 10th data. The results of the normalization of the data variables can be seen in table 5.

Reporting Date	Location	Country	Gender	Age
0	0	1	0	1
0	0.166666667	1	1	0.688
1	0.333333333	1	0	0.375
1	0.5	1	1	0.813
1	0.5	1	0	0.75
1	0.666666667	1	1	0.313
1	0.833333333	1	0	0
1	1	1	0	0.094
1	1	1	0	0.156
1	1	1	0	0.688
.	.	.	.	.
0.3	0.24	0.44	0	0.71

Table 5:- Normalization Data

After the number of clusters is determined, the next step is to determine the initial cluster center value for each cluster in each variable.



No	Description	Reporting Date	Location	Country	Gender	Age
1	Cluster 1	0.25	1	1	0.5	0.5
2	Cluster 2	0.5	0.5	1	1	0.25
3	Cluster 3	1	0.25	1	0.5	1

Table 6:- Initial Cluster Centre

The examples of calculating the distance to the 1st data in each cluster are:

$$(x_1, c_1) = \sqrt{(rd_1 - rd_{c1})^2 + \cdots + (age_1 - age_{c1})^2}$$

$$= \sqrt{(0-0.25)^2 + \cdots + (1-0.5)^2}$$

$$= 1.25$$

The same equations and calculations are applied to 10 data to get the distance of each data in each cluster as in table 7.

Data	Jarak C1	Jarak C2	Jarak C3
1	1.25	1.436140662	1.145643924
2	1.020833333	0.743315116	1.163873144
3	1.128082198	1.137278672	0.804716996
4	1.077105496	0.752599661	0.589623821
5	1.060660172	1.224744871	0.612372436
6	0.979166667	0.530739133	0.946713981

7	1.044163674	1.193151755	1.261062163
8	0.988705751	1.234671642	1.278197584
9	0.964709315	1.22832775	1.234671642
10	0.920682491	1.300540753	0.954021095
.	.	.	.
843	1.089862377	1.277810628	1.066677083

Table 7:- Data Distance In Each Cluster

After each data is calculated the distance for each cluster, the next step is to group the data according to its cluster. The cluster group of data is taken from the shortest distance of the data to a cluster.

Data	C1	C2	C3
1			✓
2		✓	
3			✓
4			✓
5			✓
6		✓	
7	✓		
8	✓		
9	✓		
10	✓		
.	.	.	.
843			✓

Table 8:- Cluster Placement

The results of grouping 843 data can be seen in the table 9.

Data	**Cluster**
1	3
2	2
3	3
4	3
5	3
6	2
7	1
8	1
9	1
10	1
.	.
843	3

Table 9:- Cluster Distribution

After the data is divided into each cluster, then the amount of data from each cluster can be calculated. The results of clustering can be seen in table 10.

Cluster 1	Cluster 2	Cluster 3
337	169	337

Table 10:- Cluster Result

Based on the results of the division of each cluster, the data are grouped into three clusters. The comparisons obtained in the three clusters contained in the data are 39.97% : 20.04% : 39.97%.

B. WEKA

Based on the data regarding the COVID-19 virus in Table 1, the dataset must be converted to .csv form first so that the data can be processed into a method using the WEKA software.

Attribute checking is done with the aim of knowing the distribution of data. Can be seen in Figure 2.

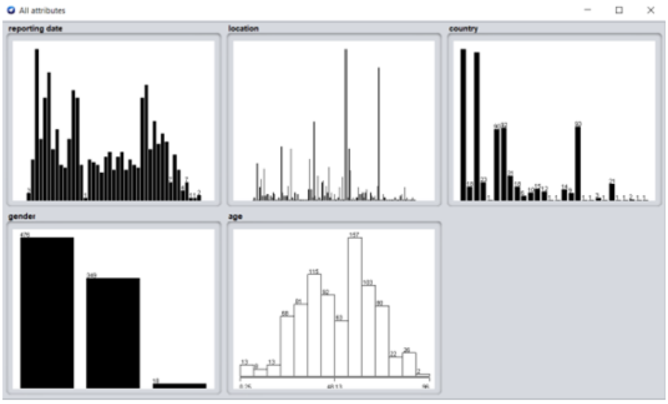
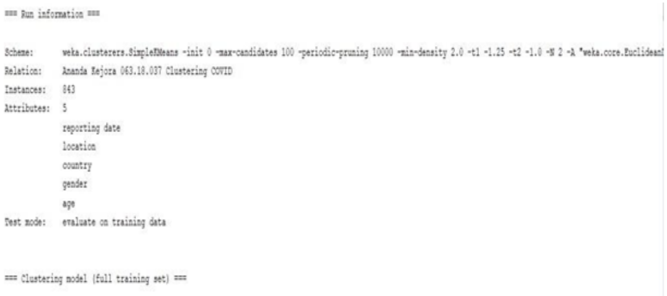
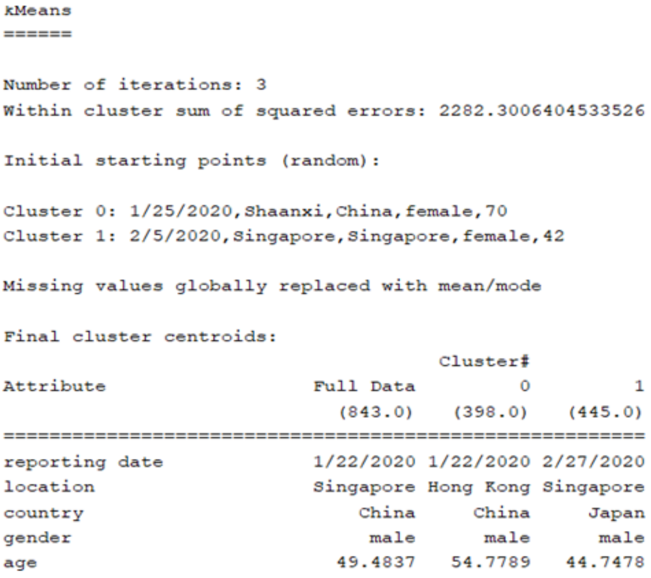


Fig 2:- All Attributes of COVID-19 Data on WEKA



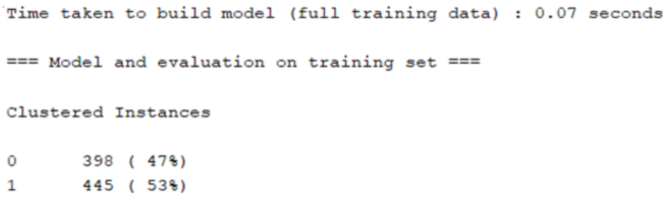
```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.Euclidean
Relation:     Ananda Kejora 063.18.037 Clustering COVID
Instances:    843
Attributes:   5
              reporting date
              location
              country
              gender
              age
Test mode:    evaluate on training data


=== Clustering model (full training set) ===
```

Fig 3:- Run Information on WEKA



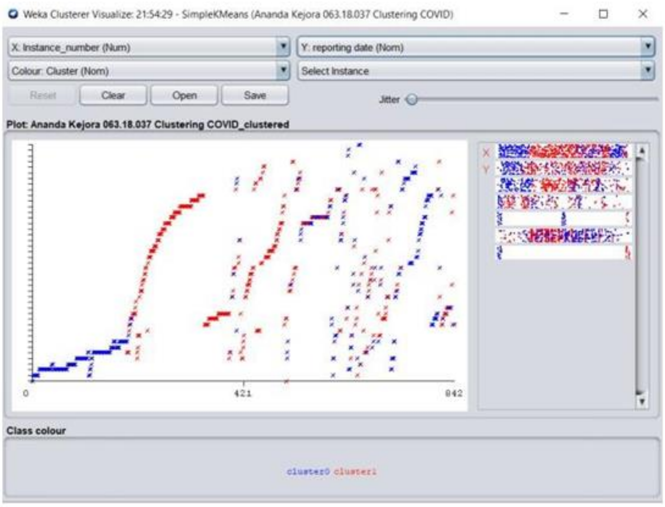
```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 2282.3006404533526

Initial starting points (random):

Cluster 0: 1/25/2020,Shaanxi,China,female,70
Cluster 1: 2/5/2020,Singapore,Singapore,female,42

Missing values globally replaced with mean/mode

Final cluster centroids:
                                    Cluster#
Attribute               Full Data          0          1
                          (843.0)    (398.0)    (445.0)
=========================================================
reporting date          1/22/2020  1/22/2020  2/27/2020
location                Singapore  Hong Kong  Singapore
country                     China      China      Japan
gender                       male       male       male
age                       49.4837    54.7789    44.7478
```

Fig 4:- K-means Clustering with WEKA



```
Time taken to build model (full training data) : 0.07 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       398 ( 47%)
1       445 ( 53%)
```

Fig 5:- Results of K-means Clustering with WEKA



Weka Clusterer Visualize: 21:54:29 - SimpleKMeans (Ananda Kejora 063.18.037 Clustering COVID)

X: Instance_number (Num)

Y: reporting date (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: Ananda Kejora 063.18.037 Clustering COVID_clustered

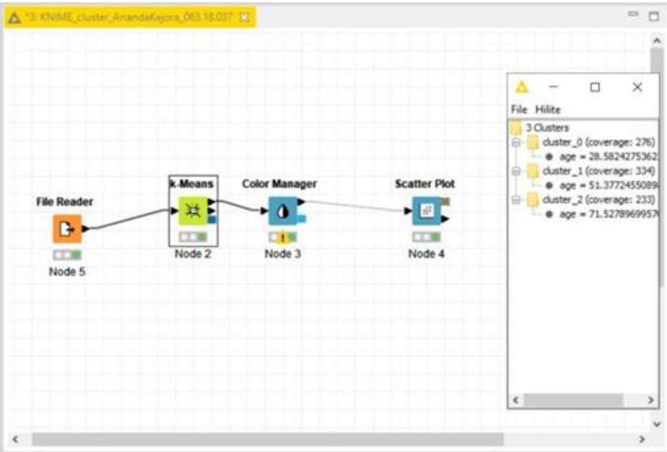
Fig 6:- Visualization of Data with WEKA

The analysis obtained is that the COVID-19 data is divided into two clusters with data sharing of 398 data in cluster 0 and 445 data in cluster 1 which can be seen in Figure 5. Visualization of the distribution data can be seen in Figure 6. Visualization of Data with WEKA.

C. KNIME

It's the same with the work at WEKA. The data regarding the COVID-19 virus in Table 1, the dataset must be converted into .csv form first so that the data can be processed into a method using the KNIME software.

The nodes needed for analysis using KNIME are file reader, k-means, color manager, and scatter plot. File reader has a function to read data so that it can be processed followed by k-means as the chosen method. Color manager and Scatter plot have functions as data visualization.



*3: KNIME_cluster_AnandaKejora_063.18.037

3 Clusters
- cluster_0 (coverage: 276)
  - age = 28.5824275362
- cluster_1 (coverage: 334)
  - age = 51.3772455089
- cluster_2 (coverage: 233)
  - age = 71.5278969957

Fig 7:- K-means Clustering on KNIME

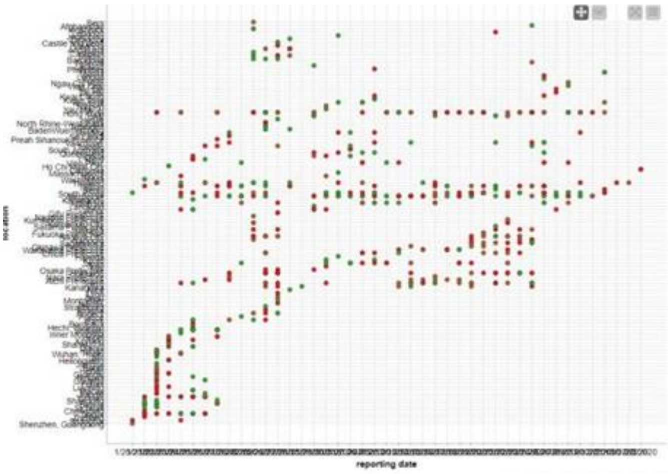
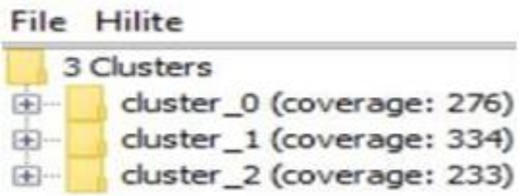


Fig 8:- Data Visualization on KNIME



File Hilite

3 Clusters
- cluster_0 (coverage: 276)
- cluster_1 (coverage: 334)
- cluster_2 (coverage: 233)

Fig 9:- Distribution of K-means Clustering on KNIME

The analysis obtained is that the COVID-19 data is divided into three clusters with data sharing of 276 data in cluster 0, 334 data in cluster 1, and 233 data in cluster 2 which can be seen in Figure 9.

IV. CONCLUSION

Based on the results of k-means clustering using two different methods starting from WEKA and KNIME. Each method has initial starting points which are taken randomly.

The method using Microsoft Excel was applied to ten data samples for testing. The data is divided into three clusters, namely: cluster 1 with 337 data, cluster 2 with 169 data, and cluster 3 with 337 data. The percentage difference between these three clusters has a difference of 39.97% : 20.04% : 39.97%.

The method using WEKA divides data into two number of clusters, namely: cluster 0 totaling 398 data and cluster 1 totaling 445 data. The percentage difference between these two clusters has a difference of 47%: 53%.

The method using KNIME divides data into three clusters, namely: cluster 0 totaling 276 data, cluster 1 totaling 334 data, and cluster 2 totaling 233 data. The percentage difference between these three clusters has a difference of 32.74% : 39.62% : 27.64%.

RECOMMENDATION

Further research on the COVID-19 virus with a combination of several data mining methods can continue to be developed.

REFERENCES

[1]. Tan, Steinbach Kumar. *Introduction to Data Mining*. Pearson Education,Inc. Addison Wesley. 2006.

[2]. Fitriana R, Saragih J, dan Lutfiana N. 2017. *Model business intelligence system design of quality products by using data mining in R Bakery Company*. Proceeding of 10th . International Seminar on Industrial Engineering and Management. Belitung, Indonesia: 7-9 September 2017.

[3]. Indraputra, R. A., Fitriana, R. K-Means Clustering Data COVID-19. *Jurnal Teknik Industri 10*(3), 275–282, 2020.

[4]. Fitriana R, Eriyatno, Djatna T, Kusmuljono BS. Peran sistem intelijensia bisnis dalam manajemen pengelolaan pelanggan dan mutu untuk agroindustri susu skala usaha menengah. Jurnal Teknologi Industri Pertanian. 22 (3) : 131 – 139, 2012

[5]. Fitriana, R., Saragih, J., & Hasyati, B. A. Perancangan Model Sistem Intelijensia Bisnis Untuk Menganalisis Pemasaran Produk Roti di Pabrik Roti Menggunakan Metode Data Mining dan Cube. *Jurnal Teknologi Industri Pertanian*, *28*(1), 113–126.2018.

[6]. Fitriana R. 2013. Rancang bangun sistem intelijensia bisnis untuk agroindustri susu skala menengah di Indonesia. [Disertasi]. Bogor : Institut Pertanian Bogor. 2013

[7]. Shukur, B. K., & Alrashid, S. Z. N. (2014). Evaluation of Clustering Image Using Steady State Genetic and Hybrid K-Harmonic Clustering Algorithms. *IJCCCE*, *14*(1), 10–20.

[8]. Djouzi, K., & Beghdad-Bey, K. (2019). A Review of Clustering Algorithms for Big Data. *Proceedings - ICNAS 2019: 4th International Conference on Networking and Advanced Systems*, 1–6. https://doi.org/10.1109/ICNAS.2019.8807822

[9]. Aggarwal, C. C., & Reddy, C. K. *Data Clustering Algorithms and Applications*. CRC Press.2014.

[10]. Alexander, C., Korstjens, A. H., Usher, G., Nowak, M. G., Fredriksson, G., & Hill, R. A. (2018). LiDAR patch metrics for object-based clustering of forest types in a tropical rainforest. *International Journal of Applied Earth Observation and Geoinformation*, *73*(January), 253–261. https://doi.org/10.1016/j.jag.2018.06.020

[11]. Sadeghi, M., Naghedi, R., Behzadian, K., & Shamshirgaran, A. Customisation of green buildings assessment tools based on climatic zoning and experts judgement using K -means clustering and fuzzy AHP. *Building and Environment*, *223*(May), 109473,2022. https://doi.org/10.1016/j.buildenv.2022.109473

[12]. Novel Corona Virus 2019 Dataset, https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019- dataset?select=covid_19_data.csv