CLUSTERING AREA COVID-19 INDONESIA WITH K-MEANS (CASE STUDY: KAGGLE DATASET)

Heti Mulyani[1], Ricak Agus Setiawan[2], Musawarman[3,] Annisa Romadloni[4]

[1,2,3]Politeknik Enjinering Indorama, Indonesia
[4] Politeknik Negeri Cilacap
[1]Heti.mulyani@pei.ac.id, [2]ricak.agus@pei.ac.id, [3]musawarman@pei.ac.id

***Abstract*-- The spread of the coronavirus in Indonesia is quite fast. The spread of Covid 19 is almost evenly distributed in all provinces in Indonesia. Some areas even have a fairly high mortality rate. Therefore, it is necessary to group regions to find out which areas have the highest to lowest Covid cases so that the appropriate response process can be carried out. In addition, data visualization is also needed that provides information on COVID-19 data for each province. In this study, the data were grouped using the K-Means Clustering method. The dataset used is the Indonesian Covid-19 dataset from Kaggle. The criteria for each province's covid cluster are the number of cases and deaths. The Clustering process uses the Python programming language. From the results of this study, it can be seen that there are 3 groups of covid. The first group consists of 30 provinces with several cases below 200,000 and a number of deaths below 6000. The second group contains two provinces that have the highest number of cases, namely above 600,000, but the number of deaths is less than group 3, which is 15000. In group 3 there are 2 provinces where the number of cases is below 500,000 but the death rate is above 30,000.**

Keywords*: *Covid 19, Province, Clustering, K-Means.

I. INTRODUCTION

Covid-19 is a pandemic that has occurred in almost all countries in the world, including Indonesia. The Covid-19 pandemic almost evenly hits all provinces in Indonesia. Based on data from the WHO website [1], on July 27, 2022, there were 6483 cases in 24 hours. WHO also recorded that the total number of cases from January 3, 2020, to July 26, 2022, was 6,178,873 people, with a total death of 156,929 people. The data shows that the number of COVID-19 cases is still quite high; therefore it is necessary to group the provinces that still have a lot of cases of COVID-19. One technique that can be used to group data is the existing techniques in data mining. Data mining is the process of analyzing data from different perspectives and summing it up into important pieces of information that can be used to increase profits, reduce costs, or even both [2]. One technique is clustering with k-means.

Cluster data is a technique for grouping data based on similarities. Clusters are collections of data that are similar to each other and different from documents in other clusters. This method classifies data into groups, thus data with the same characteristics are included in the same group, and data with different characteristics are grouped into others. Objects in the cluster have similar characteristics to one another [3]. One of the widely used clustering techniques is K-Means. This method is widely used because it is simple and easy to implement. Data grouping in this study was carried out on case data taken from COVID-19 data from Provinces in Indonesia on the Kaggle website [4]. The total data that has not been cleaned is 21,760 data. The data is then processed and partitioned into 34 provinces in Indonesia and then grouped based on the similarity of the data. The process of grouping data is carried out using the Python programming language. Python is a programming language with simple scripts and widespread applications, and most importantly it is open-source [5]. Meanwhile, the tableau application is utilized to display data visualization. A tableau is a tool that can analyze/describe a collection of data to be presented in an attractive form [6].

With this COVID-19 data grouping, we can see which provinces are classified as vulnerable, moderate, and safe, so that the government can anticipate the prevention process, especially for areas that are classified as vulnerable groups. Which provinces have the highest death rates and the highest number of cases can also be seen in the study as well as visualization of COVID-19 data in the form of graphs and maps.

II. METHOD

The CRISP-DM model is the research methodology used to create a data mining model in this study. This methodology consists of 6 stages, namely: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Fig. 1 shows the CRISP-DM method.

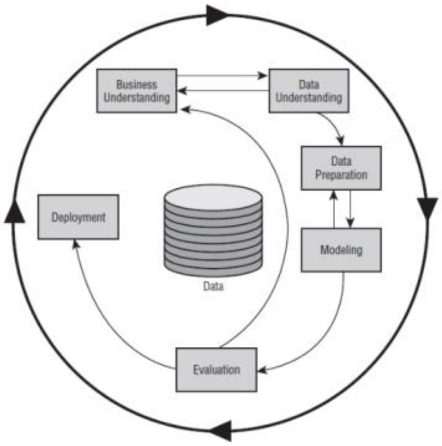


Fig. 1. Cycle CRISP-DM

A detailed explanation of the CRISP-DM flow can be described as follows:

A. Business Understanding

The business understanding stage is the stage of understanding the process and objectives from a business point of view, interpreting the knowledge in the form of defining problems in data mining, and then determining plans and strategies to achieve the data mining goals.

B. Data Understanding

This stage begins with collecting data, describing data, evaluating data quality, and conducting data exploration.

C. Data Preparation

Several things will be done including data cleaning, data selection, records and attributes, and also data transformation to be used as input in the modeling stage.

D. Modelling

At this stage, it directly involves Machine Learning for the determination of data mining techniques, data mining tools, and data mining algorithms. The technique used in this research is clustering using K-Means. K-Means is a method of grouping by calculating the distance between data. In this algorithm, the grouping technique is based on the similarity of data that does not have any reference (unsupervised) [7]. The steps for the K-Means method can be described as: (1) Determine the number of K-Clusters, (2) Determine the center point (Centroid) for each cluster, (3) Calculate the distance between the object's data point to the central data point using the Euclidian formula *dEuclidian(x,y)*=$\sum_{k=0}(x\mathrm{i} - \mathrm{yi})^2$, (4) Grouping of object data to determine cluster members based on the minimum distance, (5) Repeat step 2 until the value in each cluster does not change place.

E. Evaluation

The evaluation stage uses the Silhouette Coefficient. The Silhouette Coefficient is used to see the quality and strength of the cluster, and how well an object is placed in a cluster. This method is a combination of cohesion and separation methods. The stages of calculating the Silhouette Coefficient are: (1) Calculate the average distance from a document for example i with all other documents that are in one cluster

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

where j is another document in cluster A, and d(i,j) is the distance between document i and j, (2) Calculate the average distance from document i to all documents in other clusters, and take the smallest value

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

where d(i,C) is the average distance of document i with all objects in cluster C where A≠C.

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

(3) The value of the Silhouette Coefficient is:

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

F. Deployment

The knowledge and information that has been obtained are disseminated at this stage by making journals using the elicited model.

III. Result and Discussion

In this section, the results and discussion of clusters with K-Means will be explained following the CRISP-DM stages. Here are the results and discussion for each section:

A. Business Understanding

COVID-19 which occurred in early 2020 continues to increase in every province comprising new cases, death cases, and recovered cases. Since then, the COVID-19 pandemic has become a concern for the government, especially for its prevention. However, the government needs to know which areas should receive priority response. The government needs to know the provinces most prone to COVID-19 in Indonesia to make policies; therefore they can be prioritized during the COVID-19 mitigation process. The priority will be given to areas that have a high mortality rate; therefore a clustering method is needed to group regions based on the number of cases and the number of deaths using the K-Means clustering algorithm.

B. Data Understanding

The data used for Indonesia's COVID-19 data management was taken from the official Kaggle.com website. The data is 31821 rows and 38 columns. The Indonesian covid-19 dataset is in the form of a time series starting from January 3, 2020 to September 16, 2022. The raw dataset from Kaggle can be seen in Table 1. Exploration of Covid data by date is shown in Fig. 2. It can be seen that the highest cases of Covid-19 occurred in July 2021 and February 2022.

C. Data Preparation

The COVID-19 dataset obtained from Kaggle is still in the form of a time series per date for each province. In addition, provincial data is still in the form of raw written ones, so the data must be cleaned first. The data preparation step is carried out by calling Kaggle raw data into google colab as shown in Fig. 3.

TABLE 1 Dataset COVID-19 Indonesia

No	Date	Location ISO Code	Location	New Cases	New Deaths	...	Growth Factor of New Deaths
0	03/01/2020	ID-JK	DKI Jakarta	2	0	...	NaN
1	03/02/2020	ID-JK	DKI Jakarta	2	0	...	1.00
2	03/02/2020	IDN	Indonesia	2	0	...	NaN
3	03/02/2020	ID-RI	Riau	1	0	...	NaN
4	03/03/2020	ID-JK	DKI Jakarta	2	0	...	1.00
...	...	...	...	...	...	...	...
31821	9/16/2022	IDN	Indonesia	2358	27	...	1.29

Source: Kaggle.com

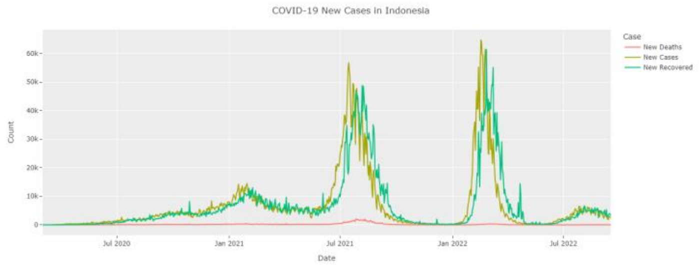
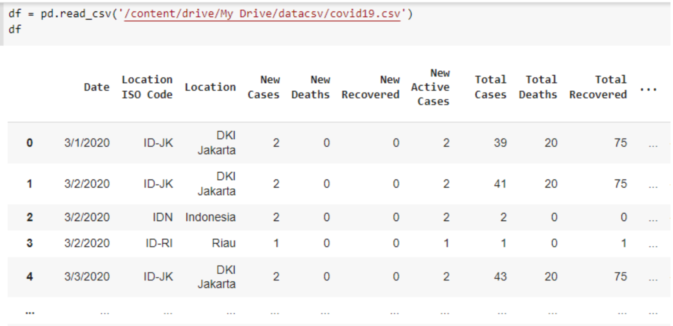


Fig. 2. Covid data chart

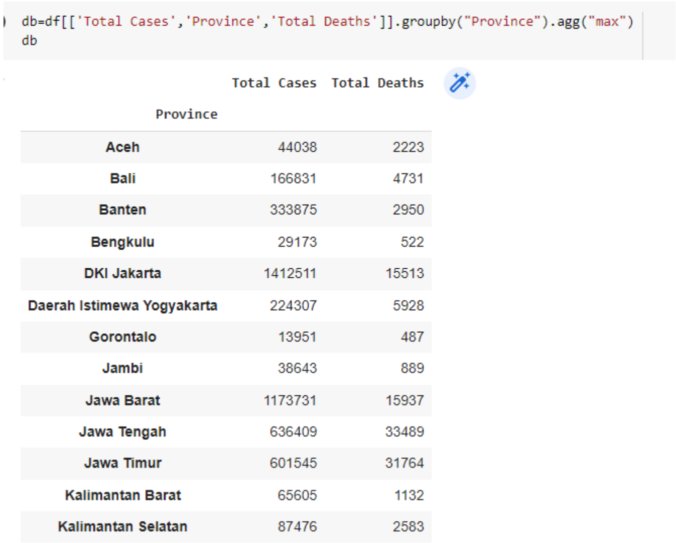


```
df = pd.read_csv('/content/drive/My Drive/datacsv/covid19.csv')
df
```

	Date	Location ISO Code	Location	New Cases	New Deaths	New Recovered	New Active Cases	Total Cases	Total Deaths	Total Recovered	...
0	3/1/2020	ID-JK	DKI Jakarta	2	0	0	2	39	20	75	...
1	3/2/2020	ID-JK	DKI Jakarta	2	0	0	2	41	20	75	...
2	3/2/2020	IDN	Indonesia	2	0	0	2	2	0	0	...
3	3/2/2020	ID-RI	Riau	1	0	0	1	1	0	1	...
4	3/3/2020	ID-JK	DKI Jakarta	2	0	0	2	43	20	75	...
...	...	...	...	...	...	...	...	...	...	...	...

Fig. 3. The calling process of the COVID-19 dataset on google colab

The next step is to call the required attributes in the data mining process for the cluster. The required attributes in this study are province, Total cases, and Total death. In addition, the data needs to be arranged per province by using the group by command since the data in Fig. 1 is still in the form of a time series. Fig. 4 shows the command and results of the run program grouping in python.



```
db=df[['Total Cases','Province','Total Deaths']].groupby("Province").agg("max")
db
```

```
                            Total Cases  Total Deaths
              Province
                  Aceh            44038          2223
                  Bali           166831          4731
                Banten           333875          2950
              Bengkulu            29173           522
           DKI Jakarta          1412511         15513
Daerah Istimewa Yogyakarta       224307          5928
             Gorontalo            13951           487
                 Jambi            38643           889
            Jawa Barat          1173731         15937
           Jawa Tengah           636409         33489
            Jawa Timur           601545         31764
      Kalimantan Barat            65605          1132
    Kalimantan Selatan            87476          2583
```

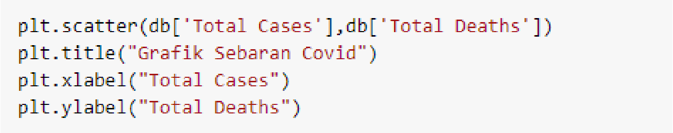
Fig. 4. Initial dataset

Fig. 3. shows the results of Python grouping data by province. There are 34 provinces with varying numbers of new cases and new deaths. To make it clearer, the initial dataset is displayed in excel form in Table 2.

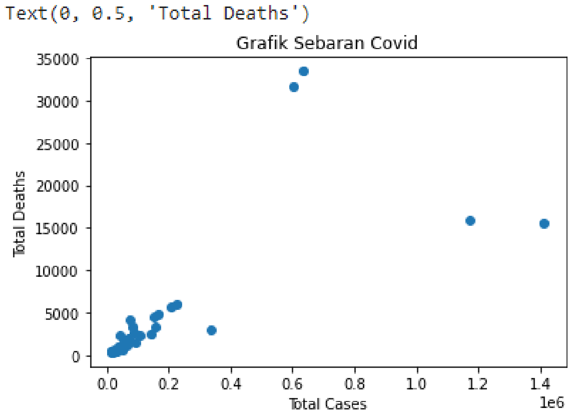
TABLE 2 Total COVID-19 data per province in Indonesia (3 January 2020 – 16 September 2022)

No	Province	Total Cases	Total Deaths
1	Aceh	44038	2223
2	Bali	166831	4731
3	Banten	333875	2945
4	Bengkulu	29173	522
5	DKI Jakarta	1412474	15493
6	Daerah Istimewa Yogyakarta	224307	5928
7	Gorontalo	13951	487
8	Jambi	38643	889
9	Jawa Barat	1173731	15937
10	Jawa Tengah	636409	33480
11	Jawa Timur	601534	31732
12	Kalimantan Barat	65605	1130
13	Kalimantan Selatan	87476	2583
14	Kalimantan Tengah	58217	1563
15	Kalimantan Timur	209017	5726
16	Kalimantan Utara	45417	860
17	Kepulauan Bangka Belitung	66144	1616
18	Kepulauan Riau	70883	1888
19	Lampung	75485	4186
20	Maluku	18736	294
21	Maluku Utara	14595	334
22	Nusa Tenggara Barat	36247	902
23	Nusa Tenggara Timur	94415	1527
24	Papua	49927	579
25	Papua Barat	32170	384
26	Riau	152648	4452
27	Sulawesi Barat	15601	394
28	Sulawesi Selatan	144494	2485
29	Sulawesi Tengah	61099	1733
30	Sulawesi Tenggara	25693	569
31	Sulawesi Utara	52770	1212
32	Sumatera Barat	104640	2371
33	Sumatera Selatan	82198	3376
34	Sumatera Utara	158866	3288

Before the modeling stage is carried out, data segmentation is firstly done in the form of a scatter plot shown in Fig. 5.



```
plt.scatter(db['Total Cases'],db['Total Deaths'])
plt.title("Grafik Sebaran Covid")
plt.xlabel("Total Cases")
plt.ylabel("Total Deaths")
```



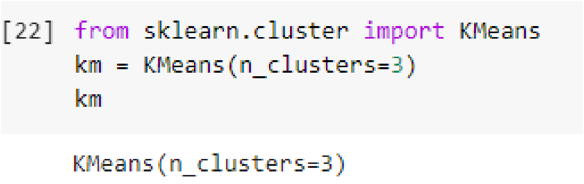
```
Text(0, 0.5, 'Total Deaths')
```

Fig. 5. Segmentation of COVID-19 data

Based on Fig. 5, the data distribution is divided into several segments. There are about 30 provinces that have similar clusters and there are 4 separate data.

D. Modelling

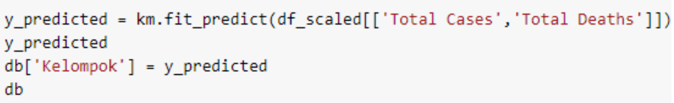
The modeling stage in this study is to group provinces with K-Means to determine provincial clusters. The author uses a value of K = 3 in this study, meaning that the data will be divided into 3 groups. The python command for determining the number of clusters can be seen in Fig. 6.



```
[22] from sklearn.cluster import KMeans
     km = KMeans(n_clusters=3)
     km

     KMeans(n_clusters=3)
```

Fig. 6. K-Means Initialization

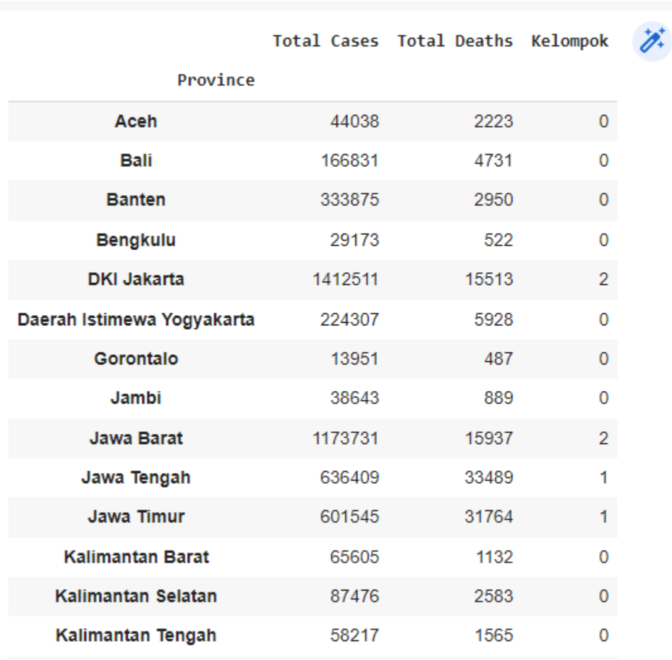
The next step is to do clustering using the python command shown in Fig. 7.



```
y_predicted = km.fit_predict(df_scaled[['Total Cases','Total Deaths']])
y_predicted
db['Kelompok'] = y_predicted
db
```

Fig. 7. Clustering process

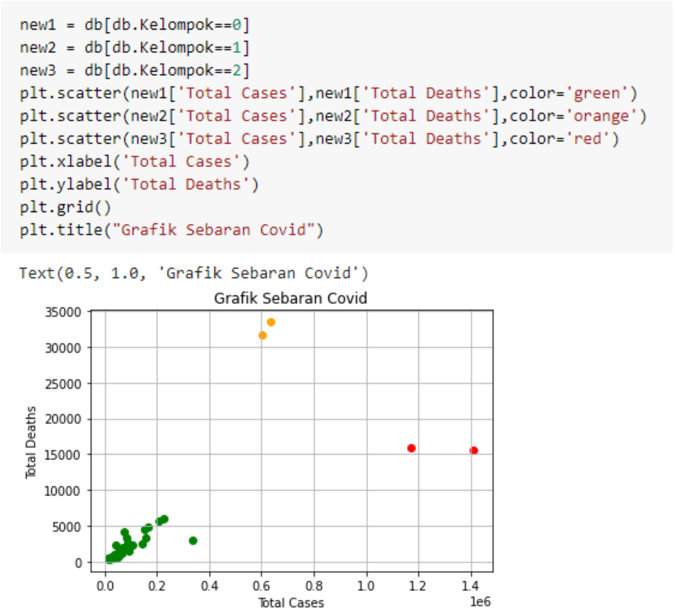
In Fig. 8, a new attribute is added with the name of the group. The cluster attribute is used to record cluster results. The results of the grouping can be seen in Fig. 8.



Province	Total Cases	Total Deaths	Kelompok
Aceh	44038	2223	0
Bali	166831	4731	0
Banten	333875	2950	0
Bengkulu	29173	522	0
DKI Jakarta	1412511	15513	2
Daerah Istimewa Yogyakarta	224307	5928	0
Gorontalo	13951	487	0
Jambi	38643	889	0
Jawa Barat	1173731	15937	2
Jawa Tengah	636409	33489	1
Jawa Timur	601545	31764	1
Kalimantan Barat	65605	1132	0
Kalimantan Selatan	87476	2583	0
Kalimantan Tengah	58217	1565	0

Fig. 8. Clustering results

Based on Fig. 8, the cluster results are obtained, namely cluster 0, cluster 1, and cluster 2. From the results of the cluster, the next step is to visualize it in the form of a scatter plot graph shown in Fig. 9.



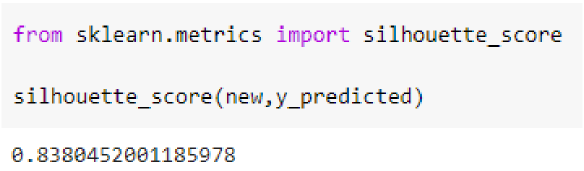
```
new1 = db[db.Kelompok==0]
new2 = db[db.Kelompok==1]
new3 = db[db.Kelompok==2]
plt.scatter(new1['Total Cases'],new1['Total Deaths'],color='green')
plt.scatter(new2['Total Cases'],new2['Total Deaths'],color='orange')
plt.scatter(new3['Total Cases'],new3['Total Deaths'],color='red')
plt.xlabel('Total Cases')
plt.ylabel('Total Deaths')
plt.grid()
plt.title("Grafik Sebaran Covid")

Text(0.5, 1.0, 'Grafik Sebaran Covid')
```

Fig. 9. Visualization of clustering results

Based on Fig. 7 the cluster consists of 3 groups, where group 0 is orange, group 1 is green and group 2 is red. The green color shows that 30 provinces have several cases below 200.000 and the number of deaths below 10000, in the red color two provinces have the highest number of cases, namely above 600000, but the number of deaths is still below 15000, the provinces are DKI Jakarta and West Java. It shows that even though these 2 provinces have the highest cases, the number of deaths is not too high compared to the 2 provinces with orange color. In the other hand, the orange color shows that there are 2 provinces that have the number of cases below 500000 but a death rate above 30000. These two provinces are Central Java and East Java.

E. Evaluation

The evaluation stage is carried out by calculating the silhouette index score. Silhouette values are in the range of -1 to 1. Silhouette values are closer to 1, indicating better clustering results. In this study, the silhouette value can be seen in Fig. 10.



```
from sklearn.metrics import silhouette_score

silhouette_score(new,y_predicted)

0.8380452001185978
```

Fig. 10. Silhouette score

Based on Fig. 10, the silhouette value is close to the value of 0.84, which means that the clustering process for the COVID-19 data is quite good.

IV. Conclusion

Based on the results of clustering on the kaggle dataset in this study, there were 3 regional groups produced, namely group 1 which showed that there are 30 provinces that have the number of covid cases below 200,000 and the number of deaths below 10,000. In group 2 there were two provinces that had the number of cases. the highest is above 600,000, but the number of deaths is less than group 3, which is 15,000, the provinces are DKI Jakarta and West Java, this shows that the 2 provinces in group 2 even though they have the highest cases but the number of deaths is not too high compared to the 2 provinces in group 3. In group 3 there are 2 provinces that have the number of cases below 500,000 but the death rate above 30,000, these two provinces are Central Java and East Java. So overall, the highest mortality rates are Central Java and East Java, while the highest number of cases are DKI Jakarta and West Java.

V. Acknowledgment

I thank to researchers partner for supporting me to complete this paper.

VI. References

[1] WHO, "Data Covid 19 WHO," *Web WHO Covid*. https://covid19.who.int/ (accessed Jul. 27, 2022).

[2] Noviyanto, "Penerapan Data Mining dalam Mengelompokkan Jumlah Kematian," *J. Inform. dan Komput.*, vol. 22, no. 2, pp. 183–188, 2020.

[3] S. M. Dewi, A. P. Windarto, I. S. Damanik, and H. Satria, "Analisa Metode K-Means pada Pengelompokan Kriminalitas Menurut Wilayah," *Semin. Nas. Sains Teknol. Inf.*, pp. 620–625, 2019.

[4] Hendratno, "Covid-19 Dataset Indonesia." https://www.kaggle.com/ (accessed Jul. 20, 2022).

[5] A. A. Rizal *et al.*, "Jurnal Widya Laksmi (Jurnal Pengabdian Masyarakat) | 13 Peningkatan Efektifitas Programming dengan Pelatihan Python for Data Science bagi Komunitas Programming Pondok Pesantren Nahdlatul Wathan Anjani," *J. Widya Laksmi*, vol. 1, no. 1, pp. 13–19, 2021, [Online]. Available: http://jurnalwidyalaksmi.com.

[6] R. Akbar, F. Arif Deliyus, F. Adeliani, and Z. Olviana, "Implementasi Bussinesee Intelligence Pada Analisis Peningkatan Sarana Perairan Kota Padang Tahun 2013 – 2015 Menggunakan Aplikasi Tableau," *KOPERTIP J. Ilm. Manaj. Inform. dan Komput.*, vol. 1, no. 2, pp. 59–62, 2017, doi: 10.32485/kopertip.v1i02.11.

[7] A. Rohmah, "Analisis Penentuan Hambatan Pembelajaran Daring Dengan Metode Algoritma K-Means Clustering (Studi Kasus : Smk Yaspim Gegerbitung)," *J. Rekayasa Teknol. Nusa Putra*, vol. 4, no. 2, pp. 30–35, 2021, doi: 10.52005/rekayasa.v4i2.122.

[8] M. Sholeh, R. Y. Rachmawati, and E. N. Cahyo, "Penerapan Regresi Linear Ganda Untuk Memprediksi Hasil Nilai Kuesioner Mahasiswa Dengan Menggunakan Python," *J. Din. Inform.*, vol. 11, no. 1, pp. 13–24, 2022.

[9] D. Hartama, "Analisa Visualisasi Data Akademik Menggunakan Tableau Big Data," *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*, vol. 3, no. 3, p. 46, 2018, doi: 10.30645/jurasik.v3i0.65.

[10] I. Effendy, Q. Widayati, and R. Sepriansyah, "Pemanfaatan Software Tableau dalam Pembuatan Dashboard Bencana Karhutla di BPBD Sumatera Selatan," *JPKMBD (Jurnal Pengabdi. Kpd. Masy. Bina Darma)*, vol. 1, no. 2, pp. 132–141, 2021.