Comparison of K-Means and Mean-Shift Algorithm in Analyzing the Impact of COVID-19 on Indonesian Tourism

Faustine Ilone Hadinata
Department of Information Systems, Multimedia Nusantara University, Tangerang, Indonesia
faustine.ilone@student.umn.ac.id

Accepted on December dd, 2022
Approved on December dd, 2022

***Abstract* — The COVID-19 pandemic has caused a slump in numerous industries all around the world, such as tourism, automotive, property, and food provision sectors. The purpose of this study is to group tourists based on their behaviours using two different clustering algorithms and compare them. The clusters found are expected to help in gaining the distribution of patterns and correlation between COVID-19 and Indonesian tourism. This study uses a combination of qualitative and quantitative methods. The descriptive data in this paper is based on literature studies in the form of journals, articles, and news. Data from reliable sources are processed as a form of the quantitative method where comparison and correlational techniques are required. The results of the analysis will be discussed in the paper.**

Index Clustering; COVID-19; Indonesia; K-means; Mean-Shift; pandemic; tourism

I. INTRODUCTION

COVID-19 is an infectious disease with the symptom of fever, dry cough, and difficulty in breathing caused by SARS-CoV-2 [1]. The infection spreads from one person to another through droplets that are often produced when coughing or sneezing. Numerous regulations were made by the government in hopes of protecting the people and reducing the chance of spreading the virus. One of them is reducing mobility through the Implementation of Restrictions on Social Activities or Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM). Therefore, this limitation also affects the tourism sector in Indonesia.

Tourism is one of the leading sources of national income and foreign exchange earnings [2]. Along with that, the decline in foreign exchange from this sector is a threat that must be anticipated and now becoming a problem that needs to be solved. According to Sandiaga Uno (2021), the Minister of Tourism and Creative Economy, the number of foreign tourists who visited Indonesia in 2019 was 16,106,954 while in 2020 there were just 4,052,923 of them. This shows a decrease of 74.8%. In addition, the tourism sector's foreign exchange in 2019 was US$16.9 billion, but in 2020 it was just US$3.2 billion, a decrease of 81% from the previous year [3]. Tourism is also a labor-intensive sector that has a workforce of more than 13 million workers. The multiplier effect that follows, including the derivative industries formed underneath, also undergoes the impact of this pandemic [4].

II. LITERATURE REVIEW

A. Machine Learning

Machine learning is a branch of Artificial Intelligence (AI). It aims to create a mathematical model that reflects data patterns so that it is possible for a computer or a program to do automatic learning without specific directions from the user [5]. In order to do that, machine learning needs data that will be grouped based on their usage: training and testing. The purpose of data training is to train the algorithm in finding a suitable model, while data testing is used to test the performance of the model trained at the testing stage [6].

B. Unsupervised Learning

Unsupervised learning is a part of machine learning that aims to obtain pattern or structure recognition of data in unlabeled datasets. In accordance with its name, it is also a machine learning technique that does not require the user to supervise the model as it will learn and discover data patterns on its own through the information it gained [7]. In other words, unsupervised learning facilitates users to do more complex processing compared to supervised learning.

C. Clustering

Clustering is one of the unsupervised learning methods as it does not require an output target from each data [8]. This method aims to cluster data based

on similarities of patterns or distances between them. Thus, it can also find an unrecognized group of data that does not belong in a certain dataset. In other words, clustering is an adaptive procedure in which objects are grouped together based on the principle of maximizing intra-class similarity and minimizing inter-class similarity [9].

D. K-Means

K-Means is one of the clustering algorithms that is popular and has a wide user base as it is a simple and efficient grouping among other algorithms. The Institute of Electrical and Electronics Engineers (IEEE) also recognized it as one of the top 10 data mining algorithms [10]. K-Means groups 'n' objects into 'k' classes or clusters based on their distance from the centroids or cluster centers [11].

E. Mean-Shift

Mean-Shift is a clustering method that groups spatial data through an iterative process of shifting data points to modes which will then assign them to suitable groups [12]. Bandwidth is a single parameter used in Mean-Shift to show the effect on the estimated density generated which will also affect the number of clusters [13].

F. Silhouette Coefficient

This is a method used to evaluate clustering algorithms that combine cohessian and separation methods. Cohessian counts all the objects contained in a cluster and separation calculates the average distance of each object in a cluster to the closest cluster [14].

III. METHODOLOGY

A. Object of Research

Tourists is the main object of this research as the goal is to group them based on their behaviours using two different clustering methods and compare them. It also aims to gives an understanding of the correlation and relationship between the number of COVID-19 cases around the world and the number of visits to Indonesia from 2020-2022.

B. Data Collecting Method

The required data for this research is obtained from several reliable open sources on the internet, such as Foreign Tourist Visits per Month based on Nationality from Badan Pusat Statistik (BPS), as well as Daily Cases of COVID-19 from World Health Organization (WHO). Besides the two main data above, there is also supporting data: Foreign Tourist Visits per Month based on Entrance Gate from Badan Pusat Statistik (BPS). All data used in this research is within the time range of 2020-2022.

C. Method of Research

To provide an overview of the research methodologies, the following figure depicts the flowchart of the overall process (Fig. 1).

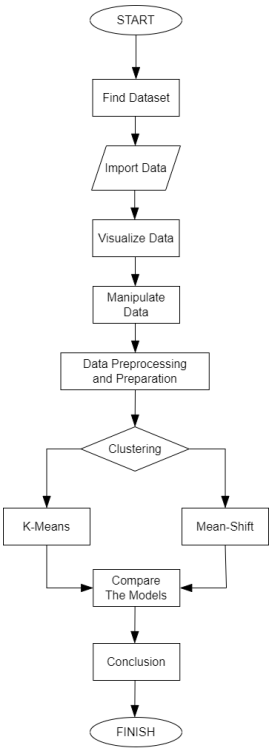

Fig. 1. The Process Flow

The research starts with finding suitable datasets and importing them to Jupyter Notebook. First, we do a visualization using Tableau and Jupyter Notebook to get a clear view and simple observations of the data. We then continue manipulating them to meet our needs. For example, since we use three datasets in total, we have to ensure there is a column with the corresponding values between the three datasets in order to be able to merge them. The next step is defining research boundaries where we filter and just use the countries in which their people visited Indonesia in 2020, 2021, or 2022. The data manipulation process includes renaming values, merging datasets, dropping irrelevant columns, converting data types, and handling outliers. Normalization of the data then follows and encoding is done after. We then compare two clustering algorithms: K-Means and Mean-Shift. We conclude the result at last in hopes of answering the thesis that has been declared before.

This research also uses a combination of qualitative and quantitative methods to ensure a more comprehensive, reliable, and valid result. First, we collected data and information about our topic based on literature studies of journals, articles, and news. We then processed the required data obtained and interpret the results through the statistical method.

IV. Results and Analysis

A. *Exploratory Data Analysis (EDA)*

	Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
0	1/3/2020	AF	Afghanistan	EMRO	0	0	0	0
1	1/4/2020	AF	Afghanistan	EMRO	0	0	0	0
2	1/5/2020	AF	Afghanistan	EMRO	0	0	0	0
3	1/6/2020	AF	Afghanistan	EMRO	0	0	0	0
4	1/7/2020	AF	Afghanistan	EMRO	0	0	0	0

Fig. 2. Output of Importing the COVID-19 Dataset

The figure above is the output of importing the COVID-19 dataset into the Jupyter Notebook. The data is in the form of CSV and then kept inside a data frame named 'covid'. The head() function is used to show the first five rows of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 246480 entries, 0 to 246479
Data columns (total 8 columns):
 #   Column             Non-Null Count   Dtype 
---  ------             --------------   ----- 
 0   Date_reported      246480 non-null  object
 1   Country_code       245440 non-null  object
 2   Country            246480 non-null  object
 3   WHO_region         246480 non-null  object
 4   New_cases          246480 non-null  int64 
 5   Cumulative_cases   246480 non-null  int64 
 6   New_deaths         246480 non-null  int64 
 7   Cumulative_deaths  246480 non-null  int64 
dtypes: int64(4), object(4)
memory usage: 15.0+ MB
```

Fig. 3. COVID-19 Dataset Information

The figure above is the output of the function info() on the 'covid' data frame. This function shows the data frame's shape, its column names, and the data type of each column.

	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
count	2.464800e+05	2.464800e+05	246480.000000	2.464800e+05
mean	2.553436e+03	9.535013e+05	26.688758	1.455616e+04
std	2.879783e+04	4.601037e+06	251.297505	6.454205e+04
min	-8.261000e+03	0.000000e+00	-60.000000	0.000000e+00
25%	0.000000e+00	7.070000e+02	0.000000	7.000000e+00
50%	1.800000e+01	1.946350e+04	0.000000	2.420000e+02
75%	4.390000e+02	2.593308e+05	5.000000	4.021000e+03
max	5.528680e+06	9.620643e+07	23283.000000	1.060430e+06

Fig. 4. COVID-19 Dataset Description

The figure above is the output of the function describe() on the 'covid' data frame. It is used for calculating numerical data and generating its count, mean, std, min, Q1, Q2, Q3, and max.

This process will be repeated for all the datasets that must be imported, such as the data of Foreign Tourist Visits per Month based on Nationality and Foreign Tourist Visits per Month based on Entrance Gate from 2020-2022.

B. *Data Visualization*

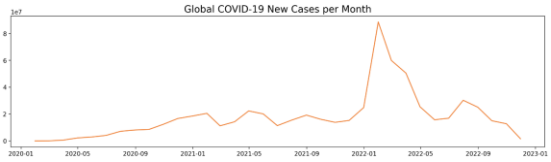

Fig. 5. Line Plot of the Global COVID-19 New Cases per Month

The figure above is a line plot of the global COVID-19 cases all around the world. From the graph, it can be seen that the highest number of new cases is in January 2022 with more than 88 million cases which details can be seen in Table 1.

TABLE I. Top 5 Months with the Highest Global COVID-19 New Cases

No	Month Reported	Number of COVID-19 New Cases
1	January 2022	88,651,375
2	February 2022	59,968,561
3	March 2022	50,546,467
4	July 2022	30,314,162
5	April 2022	25,455,096

All five months are in 2022 with the range of COVID-19 new cases from 25 million to 88 million. This is also in accordance with the line plot above (Fig. 5).

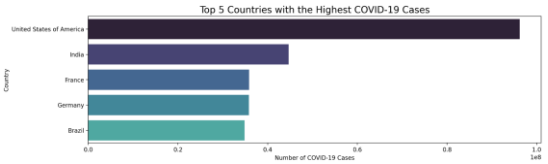

Fig. 6. Bar Plot of Top 5 Countries with the Highest COVID-19 Cases

The figure above is a bar plot of the top five countries with the overall highest COVID-19 cases of all time. The United States of America has the 1st place with a total of more than 96 million cases, followed by India, France, Germany, and Brazil respectively. While for Indonesia, it is in the 20th place with a total of around 6.5 million cases.

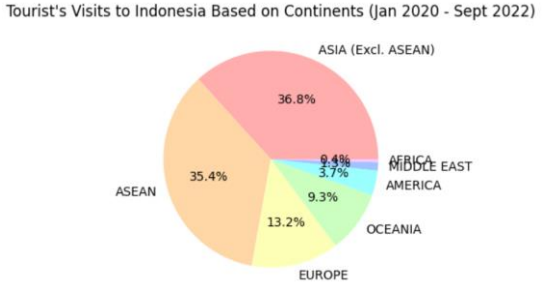

Fig. 7. Pie Chart of Tourist's Visits to Indonesia based on Continents (Jan 2020 – Sep 2022)

From the pie chart above, it can be seen that most tourists who visit Indonesia live in the continents of Asia (36.8%) and ASEAN (35.4%). We can also see that Africa (0.4%) is the continent with the least number of tourists to Indonesia.

	Continent	Number of Visits 2020	Number of Visits 2021	Number of Visits 2022	Total Number of Visits
0	TOTAL ASIA (Excl.ASEAN)	3184074	1809682	896182	5889938
1	ASEAN	3042894	1056452	1565218	5664564
2	TOTAL EUROPE	882358	111812	1113774	2107944
3	TOTAL OCEANIA	595572	71008	816062	1482642
4	TOTAL AMERICA	268144	50146	280446	598736
5	TOTAL MIDDLE EAST	99562	11278	90862	201702
6	TOTAL AFRICA	33242	4682	31818	69742

Fig. 8. Tourist's Visits to Indonesia based on Continents (2020 – 2022)

The figure above is a table that shows the distribution of visits to Indonesia from 2020, 2021, 2022, and the total visits based on continents. In line with the previous visualization (Fig. 7), Asia is the continent with the highest contributions with a total of 5,889,938 visits, while there are just 69,742 visits from the African continent.

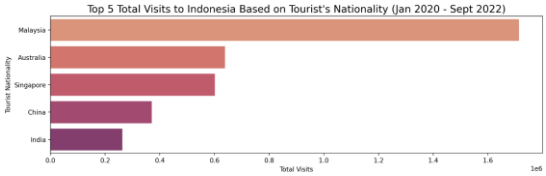

Fig. 9. Bar Plot of Top 5 Total Visits to Indonesia based on Tourist's Nationality (Jan 2020 – Sept 2022)

The figure above shows the highest number of visits to Indonesia comes from Malaysians, followed by Australians, Singaporeans, Chinese, and Indians.

	Kebangsaan	Number of Visits 2020	Number of Visits 2021	Number of Visits 2022	Total Number of Visits
1	Malaysia	980118	480723	252728	1713569
3	Australia	256291	3196	379382	638869
2	Singapore	280492	18704	302563	601759
4	China	239768	54713	76254	370735
6	India	111724	6670	144534	262928

Fig. 10. Tourist's Visits to Indonesia based on Nationality (2020 – 2022)

Even though Malaysia's number of visits decreased from year to year, its total number of visits to Indonesia remains the highest compared to other nations. Complimenting the bar plot above (Fig. 9), the 1,713,569 visits from Malaysia are the highest number of visits among other countries.

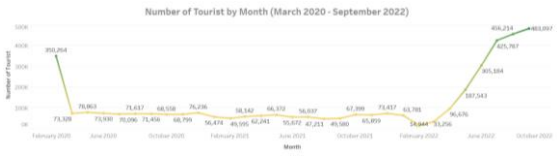

Fig. 11. Line Plot of the Number of Tourist to Indonesia by Month (March 2020 – September 2022)

From the line plot above, we can see that there is a massive decrease in the number of visits to Indonesia from February 2020 to March 2020 as it was the time when COVID-19 first appeared. The number has been at a constant low since then. We can also see that there is a further decline in tourist visits from January 2022 until February 2022. This is the same period of time when global COVID-19 new cases experienced a significant increase (Fig. 5). However, the good news is that the number has been gradually increasing little by little since March 2022. In April 2022, it even reached its highest point since March 2020 with a total of 96,676 visits. The number of visits to Indonesia continues to increase while the global COVID-19 new cases also undergo a reduction. Therefore, the two variables do have a correlation with each other.

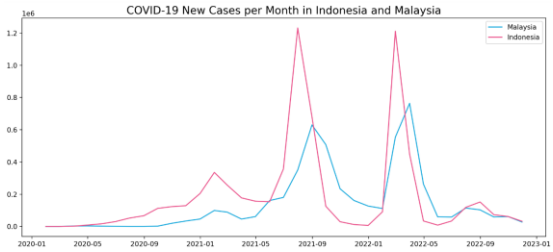

Fig. 12. Line Plot of the COVID-19 New Cases per Month in Indonesia and Malaysia

We are using Malaysia as an example as it has the highest number of visits to Indonesia (Fig. 9). Both of the nations have a similar period of time when experiencing an increase in COVID-19 new cases. One of the peaks is also in accordance with the global one, which is at the beginning of 2022 (Fig. 5). The cause of the significant increase is the Omicron variant that first entered Indonesia on 16th December 2021 [15]. The number of infected people has been increasing since then until it reached its highest point in February 2022. The same thing happened to Malaysia in the same month with the highest record of 24,599 new cases that surpassed its highest point on 26th August 2021 [16].

However, the first peak (July 2021 – August 2021) does not cause tourist visits to decrease as it does on the second peak (January 2022 – February 2022). This shows that there is not much correlation between the number of COVID-19 cases in the visited nation and the origin nation with the number of tourist visits.

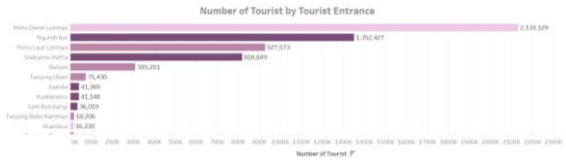

Fig. 13. Bar Plot of the Number of Tourists by Tourist Entrance

The figure above is a bar plot that shows the distribution of entrance gates the tourist tends to use. 'Pintu Darat' or land has the biggest number of visitors which means it is the most common entrance gate for tourists to enter Indonesia.

C. *Data Manipulation*

	Date_reported	Country	New_cases	jumlah_wisatawan
0	2022-09-30	Albania	331027	200
1	2022-09-30	Andorra	46227	45
2	2022-09-30	Angola	103131	111
3	2022-09-30	Argentina	9708420	2838
4	2022-09-30	Armenia	442875	305
...	...	...	...	...
112	2022-09-30	United Arab Emirates	23679346	3149
114	2022-09-30	Uruguay	985422	747
115	2022-09-30	Uzbekistan	244218	809
116	2022-09-30	Vanuatu	11974	50
117	2022-09-30	Zambia	333503	139

Fig. 14. Fixed Data Frame

After a long process of renaming columns and values, merging tables, and dropping irrelevant data, the fixed data frame is made and named ‘fixed_df’. It consists of the reported date, country, the number of COVID-19 cases, and the number of tourists.

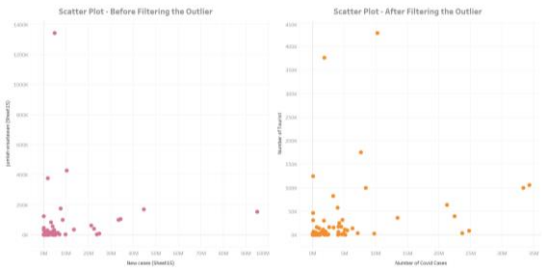

Fig. 15. Scatter Plot of Before and After Filtering the Outliers

Before going further with the research, outliers need to be handled for a better and more accurate result. The figure above is the comparison of a before and after visualization of the outliers using a scatter plot. The scatter plot on the left shows the distribution of data before outliers were handled and the one on the right shows the data distribution after outliers handling.

	New_cases	jumlah_wisatawan
New_cases	1.00000	0.30014
jumlah_wisatawan	0.30014	1.00000

Fig. 16. Correlation between COVID-19 Cases and the Number of Tourist

Using corr() function can help see the correlation between numeric variables. From the figure above, it can be seen that the correlation of COVID-19 cases with the number of tourists is a weak positive as it falls below 0.5. This is because the closer the correlation number is to 1, the stronger the correlation will be.

D. *Data Preprocessing and Preparation*

```
In [107]: target = fixed_df.iloc[:, [2,3]].values
          target = np.nan_to_num(target)
          target

Out[107]: array([[ 331027,      200],
                 [  46227,       45],
                 [ 103131,      111],
                 [9708420,     2838],
                 [ 442875,      305],
                 [10266007,   428590],
                 [5114943,    11305],
                 [ 821311,      489],
                 [ 679912,      658],
                 [2025197,     7218],
                 [ 994037,     2988],
                 [4548712,    17659],
                 [  62075,      103],
                 [ 398668,      300],
                 [ 229665,     2486],
                 [1258670,     1879],
                 [  21631,       44],
                 [ 137870,     2963],
                 [4234020,    24707],
```

Fig. 17. Creating a Model

Before performing clustering, a model has to be made. The figure above shows the process of creating a new data frame named ‘target’ that gets just the third and fourth columns of ‘fixed_df’. Both of the columns have numerical values.

	Date_reported	Country	New_cases	jumlah_wisatawan	Tourist Binned
5	2022-09-30	Australia	10266007	428590	High
93	2022-09-30	Singapore	1843685	376509	High

Fig. 18. Binning on the Number of Tourists

Binning is a technique that can be used to group data into smaller amounts of categories. In this case. the number of tourists is categorized into three groups: ‘Low’, ‘Medium’, and ‘High’. The figure above is the output of filtering the ‘Tourist Binned’ value to ‘High’. There are just two countries in the ‘fixed_df’ data frame with a high number of tourists.

```
In [108]: from sklearn.preprocessing import MinMaxScaler
          scaler = MinMaxScaler()

          normalized_target = scaler.fit_transform(target)
          normalized_target

Out[108]: array([[9.65031791e-03, 4.06007037e-04],
                 [1.34764006e-03, 4.43341018e-05],
                 [3.00654308e-03, 1.98336771e-04],
                 [2.83026277e-01, 6.56144707e-03],
                 [1.29109847e-02, 6.51011284e-04],
                 [2.99281422e-01, 1.00000000e+00],
                 [1.49114199e-01, 2.63181228e-02],
                 [2.39434012e-02, 1.08035206e-03],
                 [1.98212440e-02, 1.47469223e-03],
                 [5.90398816e-02, 1.67816242e-02],
                 [2.89788237e-02, 6.91145313e-03],
                 [1.32607059e-01, 4.11443798e-02],
                 [1.80965143e-03, 1.79669781e-04],
                 [1.16222331e-02, 6.39344415e-04],
                 [6.69534589e-03, 5.74009950e-03],
                 [3.66935798e-02, 4.32374161e-03],
                 [6.30601210e-04, 4.20007280e-05],
                 [4.01927737e-03, 6.85311879e-03],
                 [1.23432950e-01, 5.75899982e-02],
```

Fig. 19. Data Normalization

The figure above shows the process of data normalization using MinMaxScaler(). The purpose of normalization is to convert numeric fields in a dataset into a common scale and avoid distorting differences in the range of values [17].

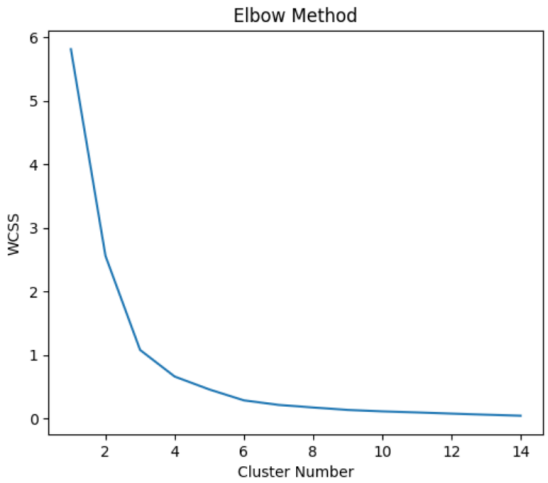

Fig. 20. Elbow Method

The Elbow Method is a technique to find the optimal number of clusters to be made in clustering. This can be done by choosing a point where the graph starts sloping or forming a straight line. In this case, 4 is the number of clusters chosen.

E. Clustering Using K-Means

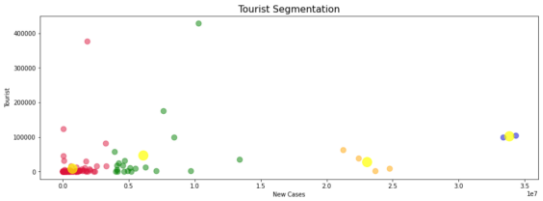

Fig. 21. Clustering Using K-Means

The figure above shows the clustering result using K-Means where we need to define the number of clusters as the 'k'. There are four clusters in total with the characteristics as follows.

- **Cluster 1 (Red)**
  - Low number of COVID-19 cases.
  - Low to high number of tourists.
  - Countries: Albania, Bangladesh, Brunei Darussalam, Cambodia, Croatia, Finland, Egypt, Ireland, Myanmar, New Zealand, and others.
- **Cluster 2 (Orange)**
  - Medium to high number of COVID-19 cases.
  - Low number of tourists.
  - Countries: France, Germany, Italy, Japan, South Africa, and United Arab Emirates.
- **Cluster 3 (Green)**
  - Low to medium number of COVID-19 cases.
  - Low to high number of tourists.
  - Countries: Argentina, Austria, Belgium, Canada, China, South Korea, United Kingdom, Thailand, and others.
- **Cluster 4 (Blue)**
  - High number of COVID-19 cases.
  - Low to medium number of tourists.
  - Countries: Australia, Singapore

```
array([[0.0192919 , 0.00690478],
       [0.77625253, 0.12436766],
       [0.17651487, 0.93923778],
       [0.15676489, 0.07784358]])
```

Fig. 22. Centroids in K-Means Clustering

These are the four centroids of each cluster found in K-Means clustering.

```
In [111]: from sklearn.metrics import silhouette_score
          score = silhouette_score(normalized_target, pred)
          print(score)

          0.6902215709927271
```

Fig. 23. Silhouette Coefficient of K-Means Clustering

The silhouette coefficient of this K-Means clustering is 0.69.

F. Clustering Using Mean-Shift

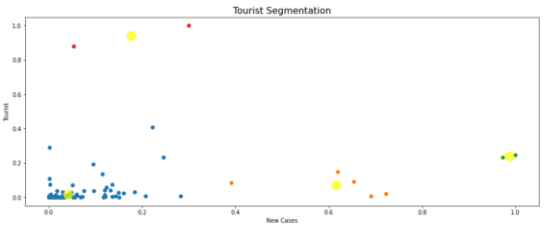

Fig. 24. Clustering Using Mean-Shift

The figure above shows the clustering result using Mean-Shift where bandwidth acts as the parameter then the number of clusters will follow. There are also four clusters in total with different characteristics as follows.

- **Cluster 1 (Blue)**
  - Low to medium number of COVID-19 cases.
  - Low number of tourists.
  - Countries: Albania, Bangladesh, Brunei Darussalam, Cambodia, Croatia, Finland, Egypt, Ireland, Myanmar, New Zealand, and others.
- **Cluster 2 (Orange)**
  - Medium to high number of COVID-19 cases.

- Low number of tourists.
- Countries: France, Germany, Italy, Japan, South Africa, and the United Arab Emirates.

- **Cluster 3 (Green)**
  - High number of COVID-19 cases.
  - Low to medium number of tourists.
  - Countries: Argentina, Austria, Belgium, Canada, China, South Korea, United Kingdom, Thailand, and others.
- **Cluster 4 (Red)**
  - Low to medium number of COVID-19 cases.
  - High number of tourists.
  - Countries: Australia, Singapore

```
array([[0.04067159, 0.01513485],
       [0.61525367, 0.07044269],
       [0.98612501, 0.23925831],
       [0.17651487, 0.93923778]])
```

Fig. 25. Centroids in Mean-Shift Clustering

These are the four centroids of each cluster found in Mean-Shift clustering.

```
In [120]: score = silhouette_score(normalized_target, mean_result)
          print(score)

          0.8410561983639996
```

Fig. 26. Silhouette Coefficient of Mean-Shift Clustering

The silhouette coefficient of this Mean-Shift clustering is 0.84.

V. CONCLUSION

Based on the explanation above, there are a few conclusions that can be made, such as:

- The number of global COVID-19 new cases has a negative correlation with the number of tourists who visit Indonesia which can be seen from the line plot (Fig. 5 & Fig. 11). When global COVID-19 new cases increases, the number of tourists visiting Indonesia decreases. However, there is not much correlation between the number of COVID-19 cases in the visited nation and the origin nation with the number of tourist visits (Fig. 11 & Fig. 12). The correlation test conducted in Jupyter Notebook also indicated the same thing, which is just 0.3 out of 1 (Fig. 16).
- 'Pintu Darat' has the biggest number of visitors as it is the most common entrance gate for tourists to enter Indonesia. The COVID-19 pandemic does not cause a significant change in this part.
- Clustering using Mean-Shift is more accurate in this case as it has a higher silhouette coefficient of 0.84 (Fig. 26) compared to the one of K-Means (Fig. 23). The results obtained are also more reasonable. For example, Australia and Singapore are in Cluster 4 with one of the characteristics of having a high number of tourists (Fig. 24). This is true seeing from the results of binning on the number of tourists (Fig. 18). It is also in accordance with the top five countries that have the highest number of visits to Indonesia (Fig. 9). In contrast, both countries in K-Means clustering are said to have the characteristics of having a low to medium number of tourists.

ACKNOWLEDGMENT

We would like to thank Mr. Samuel Ady Sanjaya, the Information Systems lecturer teaching Data Analysis, who has guided us from the beginning of our research and is willing to give out advice and input from time to time.

REFERENCES

[1] Palukota.go.id. (2021). Pengertian Virus Corona (COVID-19). URL: https://palukota.go.id/pengertian-virus-corona-covid-19/

[2] Dpr.go.id. (2020). Sektor Pariwisata Berikan Devisa Terbesar Untuk Negara. URL: https://www.dpr.go.id/berita/detail/id/30968/t/Sektor+Pariwisata+Berikan+Devisa+Terbesar+Untuk+Negara

[3] Ramadhian, N. (2021). Menparekaf Sandiaga Sebut Devisa Sektor Pariwisata Menurun. Kompas.com. URL: https://travel.kompas.com/read/2021/08/19/153200127/menparekraf-sandiaga-sebut-devisa-sektor-pariwisata-menurun?page=all

[4] Sugihamretha, I. D. G. (2020). Respon Kebijakan: Mitigasi Dampak Wabah Covid-19 Pada Sektor Pariwisata. *The Indonesian Journal of Development Planning, 4*(2), 191-206.

[5] Sitanggang, S. S., Defit, S., & Ramadhan, M. (2021). Analisis Optimasi Fungsi Pelatihan Machine Learning Neural Network dalam Peramalan Kemiskinan. *Jurnal Edukasi dan Penelitian Informatika, 7*(3), 359-369.

[6] Saifudin, A. (2018). Metode Data Mining untuk Seleksi Calon Mahasiswa pada Penerimaan Mahasiswa Baru di Universitas Pamulang. *Jurnal Teknologi, 10*(1), 25-36.

[7] Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *Indonesian Journal on Computer and Information Technology, 5*(1), 75-82.

[8] Afthoni, R., Hamdhani, M., Ardianto, Karimah, A. F., & Patria, H. (2021). Pemanfaatan Algoritma Machine Learning untuk Segmentasi Pelanggan Berbasis Data Konsumsi Listrik di PT PLN XYZ. *Prosiding Seminar Nasional Teknik dan Manajemen Industri dan Call for Paper, 1*(1), 222-231.

[9] Wirayasa, I. K. A. & Santoso, H. (2022). Analisis Employee Satisfaction Menggunakan Teknik Clustering dan Classification Machine Learning. *Jurnal Ilmiah Komputer, 18*(1), 1-10.

[10] Gustientiedina, Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means untuk Clustering Data Obat-Obatan pada RSUD Pekanbaru. *Jurnal Nasional Teknologi dan Sistem Informasi*, *5*(1), 17-24.

[11] Priyatman, H., Sajid, F., & Haldivany, D. (2019). Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa. *Jurnal Edukasi dan Penelitian Informatika*, *5*(1), 62-66.

[12] Sadewo, H., Satria, Y., & Burhan, H. (2021). Application of Mean Shift Clustering to optimize matching problems in ridesharing for maximize the total number of match. *Journal of Physics: Conference Series*, *1821*(1), 1-7.

[13] Awangga, R. M., Pane, S. F., Tunnisa, K., & Suwardi, I. S. (2018). K Means Clustering and Meanshift Analysis for Grouping the Data of Coal Term in Puslitbang tekMIRA. *Telecommunication Computing Electonics and Control*, *16*(3), 1351-1357.

[14] Paembonan, S. & Abduh, H. (2021). Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat. *Jurnal Ilmiah Ilmu-Ilmu Teknik*, *6*(2), 48-54.

[15] Permana, R. H. (2022). Data Lonjakan Corona Harian Sejak Temuan Omicron Hingga Rekor 15 Februari. Detik.com. URL: https://news.detik.com/berita/d-5943789/data-lonjakan-corona-harian-sejak-temuan-omicron-hingga-rekor-15-februari

[16] Azizah, N. (2022). Kasus Baru Covid-19 Malaysia Cetak Rekor Tertinggi. Republika.co.id. URL: https://sindikasi.republika.co.id/berita/r7eo5y463/kasus-baru-covid-19-malaysia-cetak-rekor-tertinggi

[17] Permana, I. & Salisah, F. N. (2022). Pengaruh Normalisasi Data terhadap Performa Hasil Klasifikasi Algoritma Backpropagation. *Indonesian Journal of Informatic Research and Software Engineering*, *2*(1), 67-72.