

Thai COVID-19 patient clustering for monitoring and prevention: data mining techniques

Sawitree Pansayta, Wirapong Chansanam

Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen, Thailand

Article Info

Article history:

Received Sep 21, 2022

Revised Mar 13, 2023

Accepted Mar 17, 2023

Keywords:

Clustering

COVID-19

Data mining

Emerging infectious disease

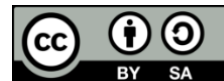
K-means

Thailand

ABSTRACT

This research aims to optimize emerging infectious disease monitoring techniques in Thailand, which will be extremely valuable to the government, doctors, police, and others involved in understanding the seriousness of the spread of novel coronavirus to improve government policies, decisions, medical facilities, treatment. The data mining techniques included cluster analysis using K-means clustering. The infection data were obtained from the open data of the digital government development agency, Thailand. The dataset consisted of 1,893,941 cumulative cases from January 2020 to October 2021 of the outbreak. The results from clustering consisted of 8 groups. Clustering results determined the three largest, three medium-sized, and the two most minor numbers of infected people, respectively. These clusters represent their activities, namely touching an infected person and checking themselves. The components of emerging diseases in Thailand are closely related to waves, gender, age, nationality, career, behavioral risk, and region. The province of onset was mainly in Bangkok and its vicinity or central Thailand, as well as industrial areas. Adult workers aged 19 to 27 years and 43 to 54 years or over were seeds of new infection sources.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Wirapong Chansanam

Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University

Khon Kaen, Thailand

Email: wirach@kku.ac.th

1. INTRODUCTION

In February 2020, the coronavirus disease 2019 was officially named for an emerging infectious disease caused by the severe acute respiratory syndrome coronavirus 2 [1]. The World Health Organization (WHO) declared the outbreak of coronavirus disease 2019 or COVID-19 as a Public Health Emergency Of International Concern (PHEIC) due to the severity of the epidemic in many countries as well as the increasing number of confirmed cases [2]. The COVID-19 outbreak in Thailand became part of the global pandemic of coronavirus disease 2019. Later, the Ministry of Public Health (MoPH) declared the emergence of COVID-19 as a major public health concern and a national health emergency. The Thailand department of disease control (DDC), ASEAN health cluster, and WHO took measures to coordinate efforts to stop the outbreak and prevent its further spread [3], [4]. Some methods were generally applied to estimate the disease transmission during the first wave of the outbreak (January to June 2020) to help prioritize healthcare and public health resources.

Monitoring and preventing the global spread of COVID-19 requires rapid and accurate data analysis, which can be achieved with the help of big data, data mining, machine learning, and other technologies [5]. Currently, big data is being prioritized by scientists, engineers, healthcare administrators, and policymakers. Several researchers have extensively used data mining to discover previously undiscovered patterns in massive datasets [6], [7].

This paper used a data mining technique to monitor novel coronavirus infections in Thailand through cluster analysis. This study aimed to explore the outbreak in Thailand by utilizing data mining methods on the dataset to provide better insights into the COVID-19 outbreak. Since all the datasets were open government data, ethical approval was not required. This study aims to find the characteristics of COVID-19 patients and build a data model to analyze the characteristics of COVID-19 patients in Thailand. The objectives of this study were to optimize emerging infectious disease monitoring techniques in Thailand, which will be highly useful to the government, medical personnel, government officials, and others involved in understanding the seriousness of the spread of novel coronavirus to improve government policies, decisions, medical facilities, treatment. Beyond these was the expectation of a reduction in infections and deaths.

2. LITERATURE REVIEW

2.1. COVID-19 in Thailand

The first suspected case in the first round of COVID-19 outbreak in Thailand was found on January 21, 2020, by a 74 year old Chinese female tourist who arrived in Bangkok on a flight from Wuhan. As of August 31, 2022, there were 4,650,919 cumulative confirmed cases in Thailand, 32,303 deaths, 2,240 new cases, 4,602,862 recovered cases, and a total of 142,712,391 doses of vaccine recipients, as shown in Figure 1 [8]. With the cumulative number of infected people, Thailand was ranked 115th in the world [9]. The tourism sector is Thailand's key economic engine, accounting for 11% of the country's gross domestic product (GDP). The COVID-19 pandemic has left a more serious scar on Thailand's tourism sector than any previous incidents. It is essential for stakeholders to understand the evolution of visitors' demand and Thailand's competitiveness [10].

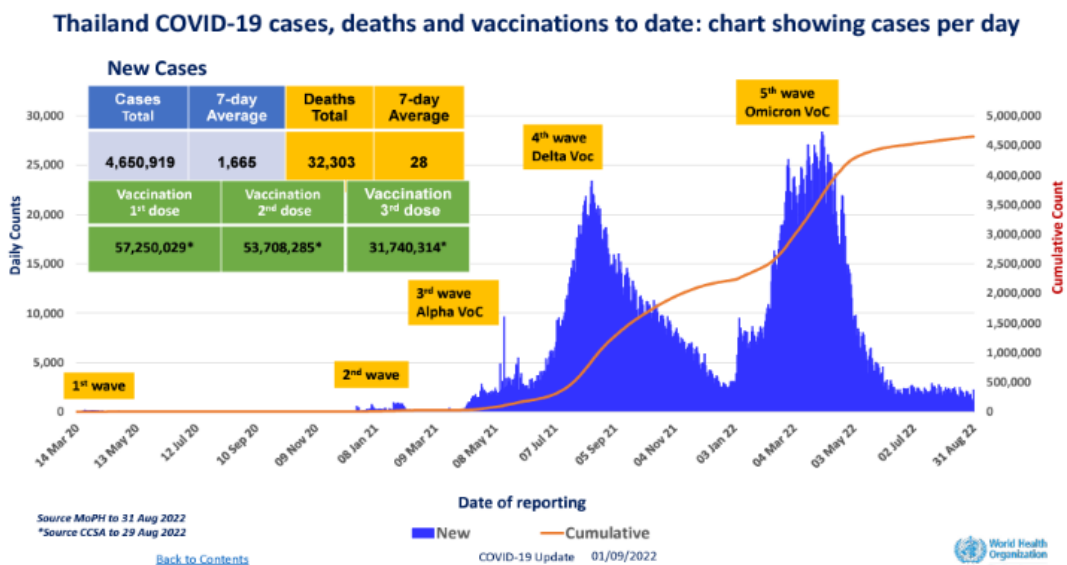


Figure 1. Thailand COVID-19 cases, deaths and vaccinations to date (COVID-19-WHO Thailand situation reports)

2.2. Data mining in COVID-19 research

According to the publication by Abd-Alrazaq *et al.* [11], the White House reached out to the worldwide artificial intelligence (AI) community via technology and research firms to work on various data mining approaches in support of a study of COVID-19 to identify a cure for the pandemic. They investigated the problem of rampant rumors and disinformation during the COVID-19 pandemic, which contributed to the continuation of unsubstantiated practices that aided in spreading the virus and unhealthy mask-wearing behavior. For the detection and removal of non-scientific based online information, Tasnim *et al.* [12] have advised the use of advanced application of data mining approaches such as natural language processing.

Data mining methods can be utilized to analyze and forecast the global expansion and trends of the COVID-19 disease outbreak, as emphasized by Ayyoubzadeh *et al.* [13]. The author analyzed data taken from Google Trends using the long short-term memory (LSTM) data mining model. In order to effectively

address the problems caused by the COVID-19 pandemic, Franch-Pardo *et al.* [14] have advocated an interdisciplinary perspective and methods such as data mining, web-based mapping, and spatiotemporal analysis. The data mining techniques of linear regression and content analysis were used by Li *et al.* [15] to categorize news and user-generated subjects that shed light on the COVID-19 pandemic.

To estimate the number of new cases and confirmed cases of COVID-19, Qin *et al.* [16] used data from social media search indexes (SMSI) for dry cough, fever, chest pain, coronavirus, and pneumonia for 40 days. The authors used the lagging series of SMSI to make predictions about future COVID-19 suspects. Based on medical records, Kumar [17] described how AI can be used with machine learning (ML), natural language processing (NLP), and other modern technologies to combat the COVID-19 pandemic on multiple fronts. Using a technique called publication mining, Ren *et al.* [18] were able to determine when diabetes and COVID-19 were being studied in similar physiological circumstances.

Data mining activities were performed on 485 patients suspected of having COVID-19 and collected from Sina Weibo by Huang *et al.* [19]. Researchers wanted to look at the number of infected people using Sina Weibo to get medical advice. Regarding treatment for COVID-19, Huang *et al.* [19] recommended employing a classification model. Positive-tested patients were identified by Sarker *et al.* [20] using a semi-automated filtering process to analyze tweets retrieved from Twitter on terms linked to COVID-19.

In this article, we examined the data mining methods applied to the research on the COVID-19 epidemic. Data mining is often used to assist companies in making better business decisions, improving customer service, and gaining competitive advantages. This research tries to identify some correlations between different variables, such as the decision-making of individuals, and the strategies used by different groups. Data scientists were tasked with using data mining techniques to investigate the myriad of correlations between the spread of the novel virus and the behaviors of individuals throughout the globe to devise strategies for fighting the pandemic.

2.3. The framework of data mining

Data mining, also known as knowledge discovery in databases, is defined by Usama M. At the international conference, it is considered the extraction of unknown, potentially useful and extraordinary knowledge contained in the database [21]–[27]. In a word, data mining is the process of processing huge amounts of data, including data collection, cleaning, transformation, and integration, as well as data analysis and visual presentation of results through association rules and category division technology of data mining so that people recognize the association relationship hidden in a huge amount of data. This cleansed data is then transformed into appropriate formats that can be understood by other data mining tools, and filtration and aggregation techniques are applied to the data in order to extract summarized data [28].

The main concept of data mining is a method to extract useful latent data and science from large batches of data. Usually, data mining methods are in the form of regression, classification, clustering, and association rules. Due to the characteristic of the Thailand COVID-19 dataset, a clustering approach is considered to be applied to solve the problem [29]–[31].

2.4. K-means clustering

MacQueen's K-means clustering algorithm [32] is one of the most popular unsupervised machine learning algorithms for grouping a dataset into a set of k different sounds (clusters). The k represents the total number of groups that the data scientist often supplies. Items are organized into groups based on their similarity in many different aspects. The first thing the K-means algorithm does is figure out how many groups need to be created. Next, k objects are chosen at random to serve as cluster focuses. If there are any residual objects, we place them all at the centroid closest to them according to some distance metric.

Clustering problems are defined as problems when encountering a homogeneity of a group of data points in the submitted dataset. Cluster is the name of each of these groups and can be interpreted as a place where dense local objects are higher than in other areas. Partitional clustering, for example, is the easiest form of clustering, which has the aim of being able to partition the submitted dataset into unincorporated subsets (clusters). The criterion generally applied is the incorrect clustering criteria where every point calculates the squared distance starting from the center of the suitable cluster.

The K-means algorithm is a well-known clustering method that can minimize grouping errors. The meaning is that its performance is highly dependent on its initial state or is called a local search procedure [33]. K-means clustering is a type of partitioning method. The meaning of the K-function is to partition the data into k clusters exclusive to each other and recover the specified cluster index for all observations. The accuracy of large data is generally more accurate using K-means grouping [34], [35]. K-means clustering is a way of measuring the cohesiveness of objects in a dataset. K-means treats each observation in the data as an object located in space. The distance between a point and its cluster center, known as intra-cluster, is a good indicator to measure how tightly knit a cluster is [36]–[38].

3. RESEARCH METHODOLOGY

The research framework is divided into three stages: data acquisition and pre-processing, knowledge discovery using K-means, and discovered insights. At the end of the research, the data interpretation and further implementation are provided. The steps involved in this study are described in five main steps.

3.1. Data acquisition and pre-processing

The daily new cases of the COVID-19 outbreak were obtained from the official website of open government data, available at the digital government development agency official website [39]. The sample size was 1,893,941 cases, gathered from January 2020 to October 2021. A total of 20,100 usable examples were discovered following data cleansing and extraction using the data mining program.

3.2. Knowledge discovery using K-means

The data that is processed in order to give the results of k clusters on a number of n objects is called K-means clustering. Thus, in (3), the objective function of squared error can be minimized. Clustering is a data mining method for dividing datasets into several groups. Cluster analysis uses mathematical models to discover groups of similar patterns from datasets. The datasets that have a high degree of similarity to each other belong to the same group and have a high degree of dissimilarity to the ones another belong to different groups [40].

This study focuses on K-means clustering, in which the clustering procedure follows a simple way to classify a given dataset through a certain number of clusters [40]. K-means clustering method is designed to investigate the grouping or partition of COVID-19 datasets according to a known number of clusters by which asking the end-user to input the number of clusters in advance, then applying performance evaluation or cluster validity to identify the appropriate number of clusters. The closer centroid distance implies the higher similarity, and vice versa. The k centroids or measurement of similarity can be calculated as (1).

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^n (x_j - c_j)^2} \quad (1)$$

Where:

$d(x_j, c_j)$ = data distance x_j to cluster center c_j

x_j = data to j on data attribute to n

c_j = center point to j on data attribute to n

The K-means clustering works in five steps:

- a) Determine the number of k points in the data domain as the initial groups to be clustered;
- b) Choose k random points from data as a centroid;
- c) Assign all data points to the groups closest to the cluster centroid;
- d) When all data points have been assigned, recalculate the position of new centroids;
- e) Repeat step b) to d) until the data points are in their original clusters.

The K-means clustering method has problems determining the appropriate number of clusters. Data analysts must randomly choose the appropriate number of clusters based on their experience to know an ideal value of k. This study proposed the elbow method, one of the commonly used existing approaches, to identify the best number of clusters, the so-called cluster validation process.

The elbow method to determine the appropriate number of k works in five steps [41]:

- Initialize the number of k;
- Increase the value of k;
- Calculate the average within centroid distance from each value of k;
- Analyze the average within centroid distance from k values which are decreased rapidly;
- Locate and plot to find the elbow point from k values.

The elbow method uses an average within centroid distance (awcd) to choose an ideal value of k based on the distance between data points and their assigned clusters. The average within centroid distance can be calculated as (2),

$$awcd = \frac{d_{1c} + d_{2c} + d_{3c} + \dots + d_{nc}}{n} \quad (2)$$

Where: d = distance between the data and the cluster centroid.

The k values and the average within centroid distance will be plotted to see an inflection point that looks like an elbow, i.e., the method's name. The greater the number of k, the smaller number of the average within centroid distance value. The classification of observations into groups requires some methods for computing the distance or the (dis)similarity between each pair of observations. The choice of distance

measures is a critical step in clustering. If there are two elements (x, y) , it will influence the shape of the clusters. The classical method for measuring distance is Euclidean distance, which is defined by (3).

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (3)$$

The choice of distance metrics is important since it significantly impacts the clustering results. The default distance measurement for the majority of clustering tools is the Euclidean distance. Other dissimilarity measures, however, may be selected based on the type of data and research goals, so you should be aware of your possibilities.

3.3. Discovered insights

In this section, the evaluation of K-means will be determined by using the sum of square error (SSE) value. The generated classifiers from each cluster of COVID-19 datasets will indicate the distance between the intragroup and extra group. The cluster groups are illustrated using a graph-based visualization technique since it assists in the interpretation of clustering analysis and subject matter from the topic. Each visualization will be discussed in the next section.

4. RESULT AND DISCUSSION

In this section, the descriptive knowledge characteristics from the experiment results are discussed. The data mining models were built based on a real dataset. It considered only the demographic data, consisting of COVID-19 waves, gender, age, nationality, occupation, risk group, patient type, and region of life. The 20,110 infectious cases (after the cleaning step) were used as input datasets for the investigation of the frequent patterns of each cluster. In this study, we classified the datasets using K-means clustering by the Waikato Environment for Knowledge Analysis (WEKA). WEKA is a data processing software developed at the University of Waikato in New Zealand. WEKA can perform various common data mining tasks, such as preprocessing, classification, clustering, and regression. It uses machine learning methodologies to solve real-world data mining problems [42].

4.1. Number of clusters determined

The K-means clustering algorithm calculates the optimal number of clusters in a dataset in various ways. Maximum gap statistics (i.e., the largest gap statistics) will indicate the best estimate for the cluster size [43]. To determine the size of the number of good clusters, k is taken into account. The elbow method is one approach for determining k . The elbow method produces information by identifying the best number of clusters and comparing the percentage of clusters that will form an elbow at a given point. This method generates ideas by picking cluster values and then combining them to be utilized as a data model [44]. We have used the elbow method algorithm to calculate the k value in K-means. The graph can be used as the source of information to display different percentage outcomes for each cluster value. The optimum cluster value is derived from the SSE value, which exhibits a notable and elbow-shaped drop [45].

The elbow method is a heuristic method used to determine the optimal number of clusters in a dataset when performing cluster analysis. It works by fitting a model to the data with a range of different numbers of clusters, and then evaluating the model's performance for each number of clusters. The idea is to choose the number of clusters that gives the best trade-off between model complexity and performance. The elbow in the elbow method refers to the point on the curve where the improvement in model performance begins to level off. This is often interpreted as the optimal number of clusters, as adding more clusters beyond this point may not significantly improve the model's performance. However, it is not always easy to identify this elbow point unambiguously, as it can be influenced by various factors such as the shape of the data and the distance metric used. In some cases, there may not be a clear elbow point at all. One way to help identify the elbow point is to plot the model's performance metric (such as the within-cluster sum of squares) as a function of the number of clusters. The number of clusters at which the rate of improvement in the performance metric begins to slow down is often considered to be the elbow point. However, this is still a subjective interpretation and may not always be a reliable indicator of the optimal number of clusters.

In the elbow method, we examine the proportion of variation that can be attributed to each cluster separately. This elbow cannot always be unambiguously identified. The first clusters will add much information, but the marginal gain will drop at some point, giving an angle in the graph. In this research, we utilized the factextra R package to identify the number of potential clusters in the dataset [46]. The result of this analysis for the elbow methods is shown on the left side of Figure 2.

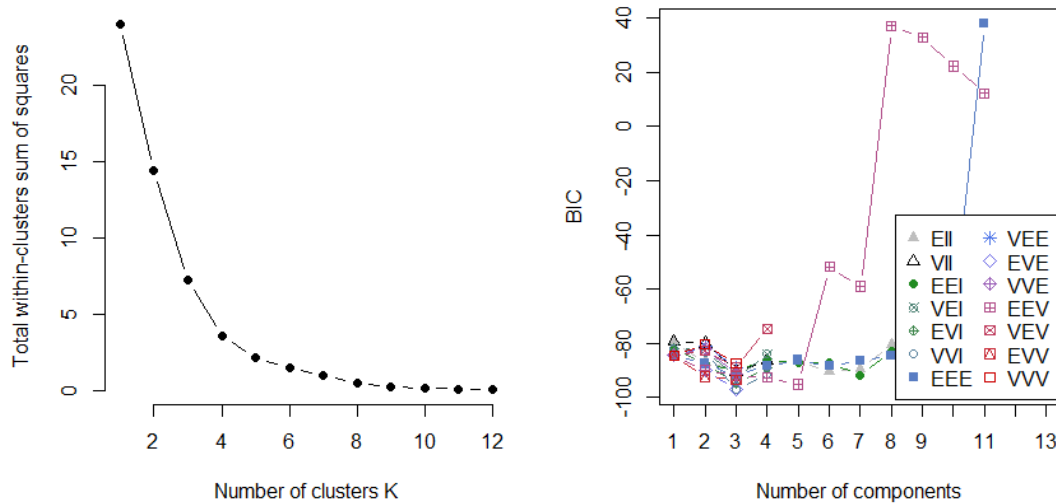


Figure 2. Elbow plot of the best number of clusters and a Bayesian inference criterion (BIC)

We can infer the optimal number of clusters, k , from the elbow plot, and for this data, $k=4, 5, 6,$ and 7 are all viable options. However, $k=8$ also looks like a strong possibility. This motivates us to consider the alternative method, which is accompanied by a Bayesian inference criterion (BIC) for K-means (right side of Figure 2). It is possible to generate a probability for the Gaussian mix using the K-means model, which is nearly a Gaussian mixture model. The predicting enhanced vegetation index (EVI) model was chosen because it was the most accurate (Equal volume but variable shape and using identity matrix for the eigenvalues). Accordingly, considering this approach, 8 clusters seemed to make the most sense. In addition, the clustering process was executed with the K-means cluster method, and the clustering outcomes were shown based on the number of clusters 4 to 8 executed with WEKA in Figure 3.

numClusters	<input type="text" value="8"/>
numExecutionSlots	<input type="text" value="1"/>
preserveInstancesOrder	<input type="text" value="False"/>
reduceNumberOfDistanceCalcsViaCanopies	<input type="text" value="False"/>
seed	<input type="text" value="12"/>

Figure 3. Weka clusters simple K-means parameter box

4.2. K-means clustering by WEKA

By adopting the K-means clustering method to achieve the smallest possible SSE, we create a K-means cluster that is more adept at locating the best possible cluster center. K-means clustering with the best possible focal point was implemented. SSE is one of the statistical methods used to measure the total difference between the actual value of the value achieved [45]. To calculate SSE using (4).

$$SSE = \sum_{i=1}^n (d)^2 \tag{4}$$

Where: d =the distance between the data and the cluster center

The SEE is a formula for gauging how far actual results deviate from a given model’s predictions. SEE is frequently used as a point of reference in identifying the most effective clusters.

Number of iterations: 5

Within cluster sum of squared errors: 34740.92300781254

The output of K-means by WEKA visualization is presented in Figure 4.

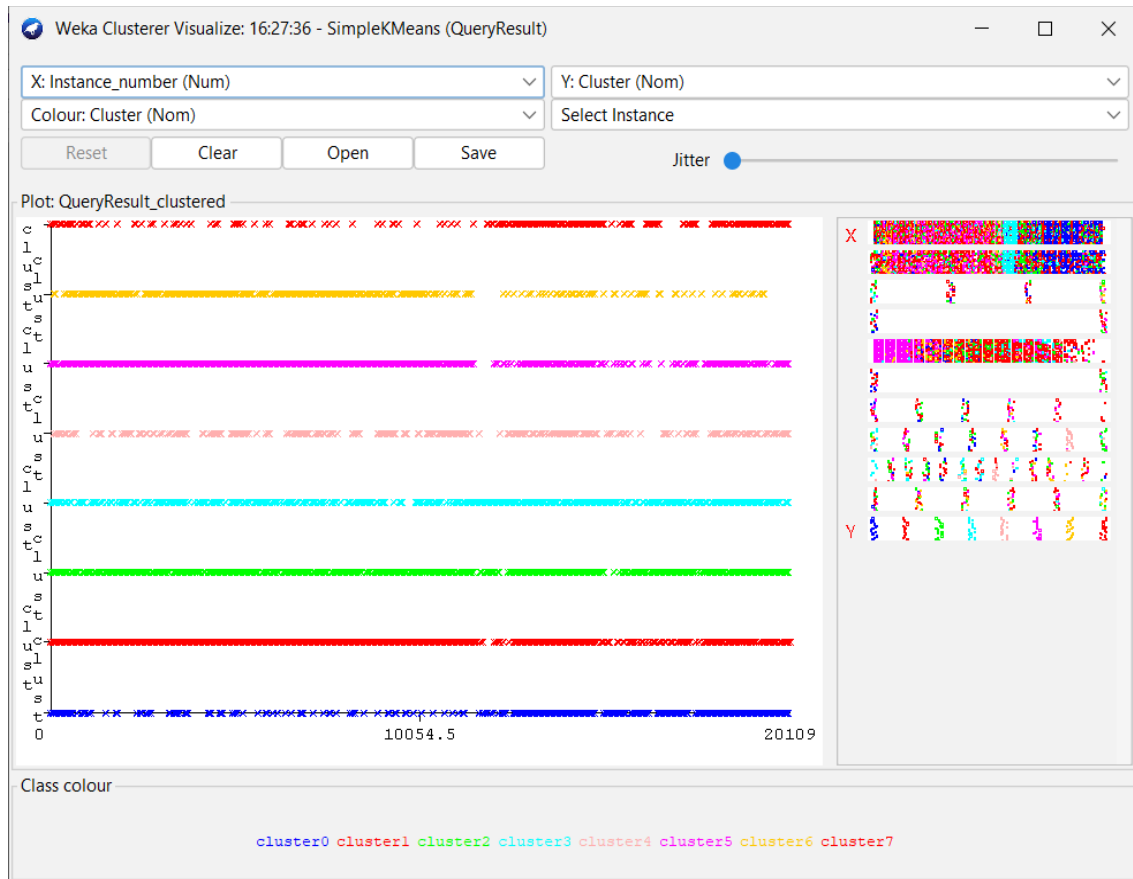


Figure 4. WEKA K-means cluster visualization output

By utilizing each cluster’s centroid, we were able to achieve the result that the characteristics of each type of cluster could be shown in Table 1. Based on the characteristics of Thai patients exposed to COVID-19 in the third wave, mostly are females with an average age of 34.72 who work in the industrial sector, have close contact with a previously confirmed patient and touch an infected person, live in central of Thailand, most recovered and died in a small proportion. However, factors or group characteristics associated with mortality cannot be concluded. Therefore, clustering with cluster=8 has not been able to explain the comorbidity and mortality due to COVID-19.

Table 1. Characteristics of Thai COVID-19 patients using 8 clusters

Cluster	Characteristic	Percentage of member
0	Ripple_2, female, 27, foreigners, 'industrial career', workplace, 'check by themself', central	17
1	Ripple_3, female, 54, thai, 'Commerce and service careers', 'Close contact with a previous confirmed patient', 'touch an infected person', central	26
2	Ripple_2, male, 43, thai, 'industrial career', 'Close contact with a previous confirmed patient', 'touch an infected person', central	11
3	Ripple_1, male, 24, thai, 'not working', StateQuarantine, 'Thai people come from abroad', central	10
4	Ripple_4, female, 22, thai, 'Commerce and service careers', 'medical and public health personnel', 'medical personnel', north	4
5	Ripple_3, male, 22, thai, 'not working', 'Close contact with a previous confirmed patient', 'touch an infected person', central	17
6	Ripple_3, male, 24, thai, 'industrial career', 'Close contact with a previous confirmed patient', 'risk group survey', east	6
7	Ripple_2, male, 19, foreigners, 'industrial career', workplace, 'risk group survey', central	10

Based on the clustering in Table 1, which contains 6 clusters, it appears that they live in central Thailand. Thus, it is necessary to be more careful and more intensive treatment in patients with these

characteristics. This results in 8 clusters with similar characteristic in the age group range from 19 to 27 years and 43 to 54 years or over, as shown in Table 1.

The clustering results determined the three largest, three medium, and the two smallest numbers of infected people. Cluster 1, 0, and 5 have the largest number of infected people, respectively. These clusters represent patients' activities in touching an infected person and checking themselves. On the other hand, cluster 4 and 6 has the smallest number of infected people, consisting of close contact with a previously confirmed patient. People who are less likely to be infected than those who have had close contact with a previously confirmed patient perhaps their behaviors are to keep a social distancing policy.

5. CONCLUSION

This research has succeeded in analyzing COVID-19 using a data mining approach. Based on the K-means cluster clustering results in the k=8 cluster, it became apparent that the components of an emerging disease in Thailand are closely related to waves, gender, age, nationality, career, behavioral risk, and region. Additionally, males and females were equally infected with COVID-19 disease. The province of onset was mainly in Bangkok and its vicinity or central Thailand, as well as industrial areas. Adult workers in the age group of 19 to 27 years and 43 to 54 years or over were the seeds of new infection sources. Data mining method based on descriptive tasks, i.e., cluster analysis, was possible to analyze the demographic data. Cluster analysis was used to categorize datasets into groups based on similarity using K-means clustering. The age group of infection cases was used as a variable to identify whether the province of onset and infection sources are linked to the transmission of the disease. In future works, data analytics is expected to reveal more influencing factors. Young children, teenagers, pre-aged people, and elderly people tend to be at higher risk of infection by being close to the patients. If more information about the infected person is collected, data analytics could make survival prediction possible. It is certainly possible that data analytics could be used to identify and analyze additional influencing factors related to the COVID-19 pandemic in the future. As more data is collected and analyzed, researchers may be able to identify new patterns and trends that could provide insight into the factors that increase or decrease an individual's risk of infection or severity of illness. For example, in addition to age, other factors that may be important to consider when predicting the likelihood of survival in patients with COVID-19 could include preexisting health conditions, lifestyle factors (such as diet and exercise habits), and the availability of medical resources in the patient's local area. By analyzing data on these and other factors, it may be possible to develop more accurate models for predicting the likelihood of survival in patients with COVID-19, which could be useful for guiding treatment decisions and allocating resources. It is also important to note that while data analytics can be a powerful tool for identifying patterns and trends in large datasets, it is only one part of the equation. Ultimately, the effectiveness of any prediction model will depend on the quality and reliability of the data that is used to develop it, as well as the accuracy and validity of the statistical and machine-learning techniques that are applied.

ACKNOWLEDGEMENT

The authors would like to thank the participants who took part in this study. This work was supported by the Faculty of Medicine, Khon Kaen University, Thailand [grant number 2266].

REFERENCES




- [1] M. McAleer, "Prevention is better than the cure: risk management of COVID-19," *Journal of Risk and Financial Management*, vol. 13, no. 3, p. 46, Mar. 2020, doi: 10.3390/jrfm13030046.
- [2] D. Cucinotta and M. Vanelli, "WHO declares COVID-19 a pandemic," *Acta Biomedica*, vol. 91, no. 1, pp. 157–160, 2020, doi: 10.23750/abm.v91i1.9397.
- [3] S. Hinjoy *et al.*, "Self-assessment of the Thai Department of Disease Control's communication for international response to COVID-19 in the early phase," *International Journal of Infectious Diseases*, vol. 96, pp. 205–210, Jul. 2020, doi: 10.1016/j.ijid.2020.04.042.
- [4] A. Stuckelberger and M. Urbina, "WHO international health regulations (IHR) vs COVID-19 uncertainty," *Acta Bio Medica: Atenei Parmensis*, vol. 91, no. 2, p. 113, 2020, doi: 10.23750/abm.v91i2.9626.
- [5] Y. Shi, "Big data: history, current status, and challenges going forward," *National Academy of Engineering of the National Academies*, vol. 44, no. 4, pp. 6–11, 2014.
- [6] D. L. Olson and Y. Shi, *Introduction to business data mining*, 1st ed. New York, New York, USA: McGraw-Hill/Irwin Professional Publishing, 2007.
- [7] Y. Shi, Y. Tian, G. Kou, Y. Peng, and J. Li, *Optimization based data mining: theory and applications*. Berlin, Heidelberg: Springer London, 2011.
- [8] Y.-H. Kuai and H.-L. Ser, "COVID-19 situation in Thailand," *Progress In Microbes & Molecular Biology*, vol. 4, no. 1, Dec. 2021, doi: 10.36877/pmmb.a0000260.

- [9] R. Yorsaeng *et al.*, “The impact of COVID-19 and control measures on public health in Thailand,” *PeerJ*, vol. 10, p. 12960, Feb. 2022, doi: 10.7717/peerj.12960.
- [10] R. Klinsrisuk and W. Pechdin, “Evidence from Thailand on easing COVID-19’s international travel restrictions: an impact on economic production, household income, and sustainable tourism development,” *Sustainability*, vol. 14, no. 6, p. 3423, Mar. 2022, doi: 10.3390/su14063423.
- [11] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, “Top concerns of tweeters during the COVID-19 pandemic: infoveillance study,” *Journal of Medical Internet Research*, vol. 22, no. 4, p. 19016, Apr. 2020, doi: 10.2196/19016.
- [12] S. Tasnim, M. M. Hossain, and H. Mazumder, “Impact of rumors and misinformation on COVID-19 in social media,” *Journal of Preventive Medicine and Public Health*, vol. 53, no. 3, pp. 171–174, May 2020, doi: 10.3961/jpmph.20.094.
- [13] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalthori, “Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study,” *JMIR public health and surveillance*, vol. 6, no. 2, p. 18828, Apr. 2020, doi: 10.2196/18828.
- [14] I. Franch-Pardo, B. M. Napoletano, F. Rosete-Verges, and L. Billa, “Spatial analysis and GIS in the study of COVID-19. A review,” *Science of The Total Environment*, vol. 739, p. 140033, Oct. 2020, doi: 10.1016/j.scitotenv.2020.140033.
- [15] J. Li, Q. Xu, R. Cuomo, V. Purushothaman, and T. Mackey, “Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: retrospective observational infoveillance study,” *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. 18700, Apr. 2020, doi: 10.2196/18700.
- [16] L. Qin *et al.*, “Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, p. 2365, Mar. 2020, doi: 10.3390/ijerph17072365.
- [17] S. Kumar, “Monitoring novel corona virus (COVID-19) infections in India by cluster analysis,” *Annals of Data Science*, vol. 7, no. 3, pp. 417–425, May 2020, doi: 10.1007/s40745-020-00289-7.
- [18] X. Ren *et al.*, “Identifying potential treatments of COVID-19 from traditional chinese medicine (TCM) by using a data-driven approach,” *Journal of Ethnopharmacology*, vol. 258, p. 112932, Aug. 2020, doi: 10.1016/j.jep.2020.112932.
- [19] C. Huang *et al.*, “Mining the characteristics of COVID-19 patients in China: analysis of social media posts,” *Journal of Medical Internet Research*, vol. 22, no. 5, p. 19087, May 2020, doi: 10.2196/19087.
- [20] A. Sarker, S. Lakamana, W. Hogg-Bremer, A. Xie, M. A. Al-Garadi, and Y.-C. Yang, “Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource,” *Journal of the American Medical Informatics Association*, vol. 27, no. 8, pp. 1310–1315, Jul. 2020, doi: 10.1093/jamia/ocaa116.
- [21] S. Sumathi and S. N. Sivanandam, *Introduction to data mining and its applications*. Springer Berlin Heidelberg, 2006.
- [22] L. Huancheng, W. Tingting, and Á. Rocha, “An analysis of research trends on data mining in Chinese academic libraries,” *Journal of Grid Computing*, vol. 17, no. 3, pp. 591–601, Aug. 2018, doi: 10.1007/s10723-018-9461-3.
- [23] G. Shmueli, P. C. Bruce, I. Yahav, N. R. Patel, and J. Kenneth C. Lichtendahl, *Data mining for business analytics: concepts, techniques, and applications in R*. New Delhi, India: John Wiley & Sons, 2017.
- [24] K. J. Cios and L. A. Kurgan, “Trends in data mining and knowledge discovery,” in *Advanced Information and Knowledge Processing*, Springer London, 2005, pp. 1–26.
- [25] M. R. K. Mookiah, U. R. Acharya, C. M. Lim, A. Petznick, and J. S. Suri, “Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features,” *Knowledge-Based Systems*, vol. 33, pp. 73–82, Sep. 2012, doi: 10.1016/j.knsys.2012.02.010.
- [26] U. Fayyad, “Knowledge discovery in databases: an overview,” in *Relational Data Mining*, Springer Berlin Heidelberg, 2001, pp. 28–47.
- [27] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan Kaufmann, 2022.
- [28] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 3rd ed. Morgan Kaufmann, 2006.
- [29] J. Han and M. Kamber, *Data mining: concepts and techniques*, 2nd ed. Morgan Kaufmann, 2006.
- [30] M. J. Zaki and J. Wagner Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.
- [31] A. Bao, W. G. Pan, and W. H. Wang, “Advances in data mining and applications in power plants,” *Advanced Materials Research*, vol. 347, pp. 487–493, Oct. 2011, doi: 10.4028/www.scientific.net/amr.347-353.487.
- [32] M. J., “Classification and analysis of multivariate observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.
- [33] A. Likas, N. Vlassis, and J. J. Verbeek, “The global K-means clustering algorithm,” *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, Feb. 2003, doi: 10.1016/s0031-3203(02)00060-2.
- [34] I. G. Costa, F. de A. T. de Carvalho, and M. C. P. de Souto, “Comparative analysis of clustering methods for gene expression time course data,” *Genetics and Molecular Biology*, vol. 27, no. 4, pp. 623–631, 2004, doi: 10.1590/s1415-47572004000400025.
- [35] K. Koonsanit, “Determination of the initialization number of clusters in K-means clustering application using Co-occurrence statistics techniques for multispectral satellite imagery,” *International Journal of Information and Electronics Engineering*, vol. 2, no. 5, pp. 785–789, 2012, doi: 10.7763/ijeee.2012.v2.208.
- [36] S. P. Adhau, R. M. Moharil, and P. G. Adhau, “K-Means clustering technique applied to availability of micro hydro power,” *Sustainable Energy Technologies and Assessments*, vol. 8, pp. 191–201, Dec. 2014, doi: 10.1016/j.seta.2014.09.001.
- [37] G. B. Mufti, P. Bertrand, and L. El Moubarki, “Determining the number of groups from measures of cluster stability,” in *Proceedings of international symposium on applied stochastic models and data analysis*, 2005, pp. 17–20.
- [38] S. Ray and R. Turi, “Determination of number of clusters in K-means clustering and application in colour image segmentation,” in *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, 2000, p. 143.
- [39] “Thailand’s daily COVID-19 information report,” *Digital Government Development Agency*, 2021. <https://data.go.th/dataset/covid-19-daily> (accessed Jul. 20, 2022).
- [40] T. M. Kodinariya and P. R. Makwana, “Review on determining number of cluster in K-means clustering,” *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90–95, 2013.
- [41] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, “Integration K-means clustering method and elbow method for identification of the best customer profile cluster,” *IOP Conference Series: Materials Science and Engineering*, vol. 336, no. 1, p. 12017, Apr. 2018, doi: 10.1088/1757-899x/336/1/012017.
- [42] R. R. Bouckaert, E. F. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, “WEKA manual for version 3-7-8,” *The University of Wakaito*, 2013. https://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf (accessed Jul. 20, 2022).




- [43] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "NbClust: an R package for determining the relevant number of clusters in a data set," *Journal of Statistical Software*, vol. 61, no. 6, pp. 1–36, 2014, doi: 10.18637/jss.v061.i06.
- [44] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, pp. 1–16, Dec. 2021, doi: 10.1186/s13638-021-01910-w.
- [45] T. Thinsungnoen, N. Kaongku, P. Durongdumronchai, K. Kerdprasop, and N. Kerdprasop, "The clustering validity with silhouette and sum of squared errors," 2015, doi: 10.12792/iciae2015.012.
- [46] Al. Kassambara and F. Mundt, "Package 'factoextra': extract and visualize the results of multivariate data analyses," vol. 76, no. 2, 2017.

BIOGRAPHIES OF AUTHORS



Sawitree Pansayta    received an M.S. degree in information science from Khon Kaen University, Khon Kaen, Thailand. Her research interests include (but are not limited to) data analytics and information science. She can be contacted at email: sawipa@kku.ac.th.



Wirapong Chansanam    received his Ph.D. in information studies from Khon Kaen University, Khon Kaen, Thailand. He was with Chaiyaphum Rajabhat University as a lecturer for nine years. In 2019, he joined Khon Kaen University, Khon Kaen, Thailand, where he is currently an associate professor in the Faculty of Humanities and Social Sciences. His research interests include (but are not limited to) information science, especially in the digital humanities. He can be contacted at email: wirach@kku.ac.th.