Association Analysis of COVID-19 Outbreak in Thailand Using Data Mining Techniques

Pranomkorn Ampornphan
Faculty of Science and Technology, Valaya Alongkorn Rajabhat University, Thailand
E-mail: pranomkorn@vru.ac.th

Article History
Received: 14 October 2021 **Revised:** 12 December 2021 **Published:** 16 December 2021

Abstract

The objective of this research is to apply data mining techniques to determine the relationships that affect the COVID-19 outbreak among infected people in Thailand. The data mining techniques included cluster analysis using K-means clustering and association rule mining. Cluster analysis was used to classify the infection cases into the appropriate group. Association rule mining was applied to find the relationship between datasets to find patterns associated with each group. The infection causes were obtained from the online data source of the Department of Disease Control, Ministry of Public Health, Thailand. The datasets consisted of 103,639 cases that occurred during the second wave and the third wave (December 2020-May 2021) of the outbreak. The outbreak had the largest number of infected cases and was the most severe since the epidemic occurred in Thailand during the past year. The variables selected for data analysis were age, sex, province of onset, and infection sources. The results from clustering consisted of 7 groups, which were categorized by age group ranging from 0-9 years to 60 years or over. Next, the association rule mining was applied to find the co-occurrences between itemset. The 128 rules were generated, then illustrated the results based on the value of support, and lift by using network visualization. The association network provides the major causes and effects in each group, which can be used as a precaution to prevent the outbreak in the future.

Keywords: Covid-19, Data Mining, Association Rule Mining, Cluster Analysis

Introduction

An outbreak of a novel coronavirus was reported around mid-December, 2019 in Wuhan, China. After the first discovery, the virus spread globally. In February 2020, the disease was named Coronavirus disease 2019 (COVID-19), which is an emerging infectious disease caused by the severe acute repository syndrome coronavirus2 (SARS-CoV-2) (McAleer, 2020). The World Health Organization (WHO) declared the outbreak of "coronavirus disease 2019" or COVID-19 as a Public Emergency of International concern (PHEIC) due to the severity of the outbreak in many countries as well as the increasing cases confirmed (WHO, 2020).

The Ministry of Public Health, Thailand reported the first imported case of COVID-19 from Wuhan, which was the first case outside China, on January 13, 2020. (Emergency Operation Center, 2020 mentioned in Hinjoy et al., 2020). The COVID-19 outbreak in Thailand became part of the global pandemic of coronavirus disease 2019. Later, the MOPH has declared an emerging of COVID-19 as a major public health concern and a national health emergency. The Thailand Department of Disease Control (DDC), the ASEAN Health Cluster, and the WHO took measures to coordinate efforts to stop the outbreak and prevent its further spread (Hinjoy et al., 2020).

Citation Information: Ampornphan, P. (2021). Association Analysis of COVID-19 Outbreak in Thailand Using Data Mining Techniques. *PSAKU International Journal of Interdisciplinary Research, 10*(2), 21-33. https://doi.org/10.14456/psakuijir.2021.3.

Thailand was relatively successful in controlling the epidemic throughout 2020. The outbreak subsided in May 2020 as preventive measures were implemented. The country reported almost no locally transmitted infections. However, the second wave of infections occurred in December 2020-January 2021 due to neglect of the country's migrant workers who often live in squalid, cramped conditions without adequate sanitation and poor access to medical care.

The new outbreak was different from the first wave hit in Thailand last year. The number of infected people was much higher and the virus has spread to many provinces. The primary cluster of migrant workers was in Samut Sakhon province with maximum daily cases in January 2021, then subsiding in February. In April 2021, Thailand has affected by the third wave of infections among local Thais, new clusters have emerged in Bangkok throughout the country, caused by the crowded living condition of people, such as community places, entertainment venues, prisons, and so on. It was the most severe outbreak because of the presence of new variants of coronavirus from the UK, India, Africa, and Brazil. The most worrying infectious disease is both more transmissible and more deadly than the previous dominant strains.

During such outbreaks, a large volume of data can be gathered from various sources, which are available online. These data can be analyzed into useful and effective information to assist the decision-making process (Rotejanaprasert et al., 2020). There are statistical methods that can be analyzed epidemic data using descriptive analysis and preliminary statistical estimation. These methods were generally applied to estimate the disease transmission during the first wave of the outbreak from January-June 2020 to help for prioritizing healthcare and public health resources (Rotejanaprasert et al., 2020).

In a world dominated by the concept of big data, data mining methods can be used to explain the phenomenon and to predict future steps using the dataset (Buscema et al., 2020). This study aimed to explore the outbreak from the second wave to the third wave (December 2020-May 2021) by utilizing data mining methods on the dataset to provide better insights into the COVID-19 outbreak. The dataset was retrieved from the official website offered by the Digital Government Development Agency (Public Organization), which was reported by the Thai Ministry of Public Health from the daily confirmed cases in 77 provinces in Thailand. Data mining methods are used to classify groups and relationships between potential factors within each group; K-means clustering and Association rule mining. Since all the dataset was open government data, ethical approval was not required.

The remainder of this paper is organized as follows. The second section presents the related work of the data mining approach for epidemic transmission. The third section explains the research methodology, including conceptual framework, data collection and pre-processing, cluster analysis, and association rule mining. The fourth section analyzes and discusses the results derived from the data mining technique. The last section discusses the conclusion and suggestions.

Literature Review

Generally, all public health organizations across the world stored the data in electronic format. The data mainly contains all the information aspects related to the patient as well as the parties involved in the health system (Ahmad et al., 2015). The various types of data are continuously increasing and have become a large collection of big data. The knowledge gained from big data analysis enables the public health organization to make decisions more efficiently and effectively. In Thailand, the novel coronavirus (COVID-19) outbreak was officially becoming part of a global pandemic disease in March 2020 (Triusoke, S. et. al., 2021). The medical records of confirmed COVID-19 patients have been collected daily since the outbreak began. The datasets can be retrieved from the Department of Disease Control and available as Thailand 's open access data.

Researchers in the fields of statistics, mathematics, and other disciplines have used mathematical modeling to identify the transmission of the COVID-19 situation. Also, data mining techniques can be utilized to extract the insight patterns and to understand the situation based on available information.

Data mining techniques are applied to epidemic transmission from particular methods depending on information characteristics. The researches related to this study were summarized as follows:

Ahmad et al. (2015) reviewed various data mining techniques for descriptive tasks (supervised learning method) as well as predictive tasks (unsupervised learning method) in the health domain. Data mining results determine meaningful patterns and can be very useful to improve treatment, fraud detection, and improve customer relationship management. This provides the application, challenges, and future of data mining in healthcare.

Buscema et al., (2020) applied data mining methods called “Topological Weighted Centroid” (TWC) to obtain relevant information from limited and poor datasets, such as the latitude and longitude of cities, in the COVID-19 epidemic in Italy. TWC method shows some very interesting aspects, such as an outbreak zone, estimated epicenter, and the most intense of infectious areas.

Husein et al. (2020) engaged in COVID-19 analysis to general diagnosis using AI technology that consists of various data mining methods to analyze COVID-19 datasets from the unconfirmed and the confirmed group to improve and pave the way for timely clinical interventions.

Bagheri et al. (2020) applied data mining models to forecast the outbreak of Brucellosis (Malta fever) using human Brucellosis cases and climatic parameters as studied data to analyze and compare the performance of various prediction methods. The prediction results identified the influential parameters for the control and prevention of disease.

In summarize, the data mining techniques can be classified into two categories: (1) descriptive tasks, and (2) predictive tasks (Ayyoubzadeh et al., 2020).

The descriptive task focuses on the discovery of interesting patterns or associations relating to the data. It examines the insights from data to answer about "What has happened?", "Where exactly is the problem?", and "What is the frequency of the problem?". Clustering, summarization, and association are the techniques categorized under descriptive tasks. Whereas the predictive tasks involved the prediction and classification of the behavior of the models or training data founded on the current and past data. The models will be used as supervised learning functions to predict future results. It examines the insight from data to answer about "What will happen next?", "What is the outcome of the problem?", and "What actions are required to be taken?". The techniques under the predictive tasks are regression, decision tree, KNN, random forest, naïve bayes, time-series, and so on.

The characteristics of COVID-19 datasets retrieved from open government data consist of age, sex, nationality, date of confirmed COVID-19 positive, province of onset, and infection sources. Therefore, the descriptive techniques will be appropriately used to mine for rules, detect patterns, summarize and group the input data to determine meaningful insights to the users.

Research Methodology

In this study, the proposed methodology consisted of 4 phases; data collection and pre-processing, K-means clustering, Association rules mining, and discovered insights. A conceptual framework is shown in Figure 1.

Data Collection and Pre-processing
Cluster Analysis
Association Rules Mining
Discovered Insights

Figure 1 Conceptual Framework

Step 1. Data collection and pre-processing

The COVID-19 dataset that consists of age, sex, province of onset, and infection source was used as input data. The daily new cases of the dataset were obtained from the official website of open government data, available at https://data.go.th/dataset/covid-19-daily. The sample size was 103,639 cases, gathered from December 2020 to May 2021, during the second wave and third wave of the COVID-19 outbreak. Data pre-processing was performed to prepare raw data in an understandable format and applicable in the data mining process, includes cleaning, instance selection, normalization, transformation, etc. The information characteristics are shown in Table 1.

Table1 Descriptive Information of observation

Variable	Category	Frequency	Percent
Sex	Male	52,333	50.50%
	Female	51,306	49.50%
Age_group	0-9 years	3,872	3.70%
	10-19 years	6,080	5.90%
	20-29 years	31,683	30.60%
	30-39 years	28,057	27.10%
	40-49 years	16,786	16.20%
	50-59 years	10,089	9.70%
	>=60 years	7,072	6.80%
Province_of_onset	Bangkok	31,638	30.50%
	Samut Sakhon	13,923	13.40%
	Nonthaburi	8,360	8.10%
	Samut Prakran	5,965	5.80%
	Chonburi	5,116	4.90%
	Others	70,275	37.30%
Infection_source	Close_to_patient	37,277	36%
	Cluster_SKN	12,870	12.40%
	Cluster_prison	11,256	10.90%
	Disease_investigate	11,107	10.70%
	Proactive_search	6,557	6.30%
	Others	24,572	23.70%

Step 2. Cluster Analysis

Clustering is a data mining method for dividing the datasets into several groups. Cluster Analysis uses mathematical models to discover groups of similar patterns from datasets. The datasets have a high degree of similarity to each other belongs to the same group, and have a high degree of dissimilarity to the ones another belongs to the different groups. Each group is a cluster and usage patterns may be extracted by analyzing each cluster (Kodinariya & Makwana, 2013). Two clustering techniques are widely used: hierarchical clustering and k-

means clustering (Pitchayadejanant & Nakpathom, 2017). This study focuses on k-means clustering in which the clustering procedure follows procedure a simple way to classify a given data set through a certain number of clusters (assume k clusters) (Kodinariya & Makwana, 2013). The K-Means algorithm is the best algorithm in partitional Clustering algorithm and is most often used among other Clustering algorithms, because of its simplicity and efficiency.
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 applies performance evaluation or cluster validity to identify the appropriate number of clusters. The K-means clustering applies statistical analysis to identify k number of centroids (cluster center point), then allocate every data point to the nearest cluster. The closer centroid distance implies the higher similarity, vice versa. The k centroids or measurement of similarity can be calculated as,

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

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 as follows,
(1) Determine the number of k points in data domain, as the initial groups to be clustered.
(2) Choose k random points from data as centroid.
(3) Assign all data points to the groups that are closet to the cluster centroid.
(4) When all data points have been assigned, recalculate the position of new centroids.
(5) Repeat step (2)-(4) until the data points are in their original clusters.
The K-means clustering methods has the problems in determining the appropriate number of clusters, in which the 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 among existing approaches, to identify the best number of clusters, so called "cluster validation process".
The elbow method to determine appropriate number of k works as follows (Syakur et al., 2018),
(1) Initialize the number of k,
(2) Increase the value of k,
(3) Calculate the average within centroid distance from each value of k,
(4) Analyze the average within centroid distance from k values which are decreased rapidly,
(5) Locate and plot to find the elbow point from k values.

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

$$Average\ within\ centroid\ distance = \frac{d_{1c} + d_{2c} + d_{3c} + \cdots d_{nc}}{n}$$

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 name of the method. The greater the number of k, the smaller number of averages within centroid distance value.

In grouping COVID-19 datasets, the researcher used the existing variables -- sex, age group, province of onset and infection source -- to calculate the clusters' similarities and the elbow point can be used to roughly determine "k", as shown in figure 2.

K	Avg.within centroid distance
2	110.713
3	87.249
4	63.319
5	43.994
6	25.510
7	10.169
8	7.515
9	5.277
10	5.783
11	4.735
12	4.175
13	4.400

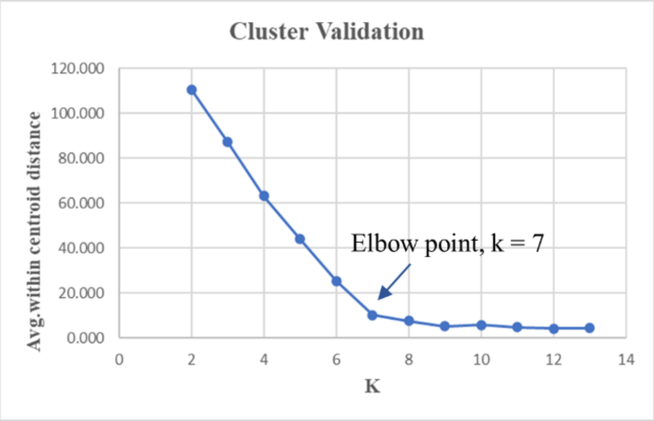


Figure 2 Cluster Validation Process

Step 3. Association rule mining

Association rule mining is a data mining method that aims to observe a frequently occurring pattern, co-occurrences, or associations from datasets. The patterns represent the relationship between datasets. Association rule consists of an antecedent and a consequence that can be represented by the if-then rule. Market basket analysis is a popular application of association rule mining. For example, people who buy diapers are likely to buy baby powder, the if-then rule can be: If (people buy diapers), then (they buy baby powder). The well-known association rule algorithms are the Apriori algorithm and the FP-Growth algorithm. The performance of FP-Growth is better than the Apriori algorithm by which it consuming less memory as well as less time spending (Tan, 2006 mentioned in Pitchayadejanant & Nakpathom, 2017).

In this study, the clusters from the K-means clustering method are used to discover the interesting rules between co-occurrences related to COVID-19 infection cases. The variables within each cluster (such as sex, age, province of onset, etc.) are used to discover the interesting relations. The measures of the effectiveness of the rules are support, confidence, and lift.

Support is measuring the frequent occurrence of the items, can be calculated as:

$$Support(X) = \frac{Frequency\ of\ item\ (X)}{Total\ number\ of\ items\ (N)}$$

The higher value of support, the more frequency the items occur. Rules with high support are preferred since they are likely to apply to a large number of future transactions.

Confidence is measuring the reliability of association rules, can be calculated as:

$$Confidence\ (X \rightarrow Y) = \frac{Support\ (X, Y)}{Support\ (X)}$$

The confidence value ranges from 0 to 1. The higher confidence, the more reliable of association rules. Rules with a high support value are preferred since they define the likelihood of occurrences in the future.

Lift is measuring the importance of association rules or the strength of association between items on the left- and right-hand side of the rule. The lift value is a ratio of confidence of the rule and the expected confidence of the rule. can be calculated as:

$$Lift\ (X \rightarrow Y) = \frac{Confidence\ (X \rightarrow Y)}{Support(Y)}$$

The lift value ranges from 0 to infinity. If a lift value is more than 1, the association between items of $(X \rightarrow Y)$ appears more often than expected. The degree of association between occurrences is dependent on each other and it is useful to predict future consequences.

Step 4. Discovered insights

In this section, the evaluation of association rules will be determined by using the value of confidence, support, and lift. The generated rules from each cluster of COVID-19 datasets will indicate the relationship between sex, age, province of onsets, and infection sources. The association rules are illustrated by using a graph-based visualization technique since it assists the interpretation of association analysis and subject matter from the rules. Each visualization will be discussed in the next section.

Research Results

In this section, the descriptive information characteristics from the experiment results are discussed. The data mining models were built based on simple datasets. It considered only the demographic data, consisting of age, sex, province of onset, and infection source. The 103,639 infectious cases were used as input datasets for the investigation of the frequent patterns of each cluster. In this study, the researcher classified the datasets using K-means clustering. This results in 7 clusters with the similar characteristic of the age group range from 0-9 years, 10-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, and 60 years or over, shown in Table 2.

The clustering results determined the four largest numbers of infected people as well as the three smallest numbers of infected people. Cluster 3, 4, 5, and 6 have the largest number of infected people, respectively. These clusters represent working people, consisting of young workers, adult workers, senior workers, and pre-aging. On the other hand, cluster 1, 2, and 7 has the smallest number of infected people, consisting of young children, teenagers, and aging people. These people are less likely to be infected than working-age people.

Association rules mining was applied to each cluster to find the insights. The results after generating association rules based on minimum support at 0.10 and minimum confidence at 0.50 are shown in Table 2. These values implied the frequency of itemset at least 10% and the probability of co-occurrence between itemset at least 50%, consequently.

Table 2 The number of rules of each cluster

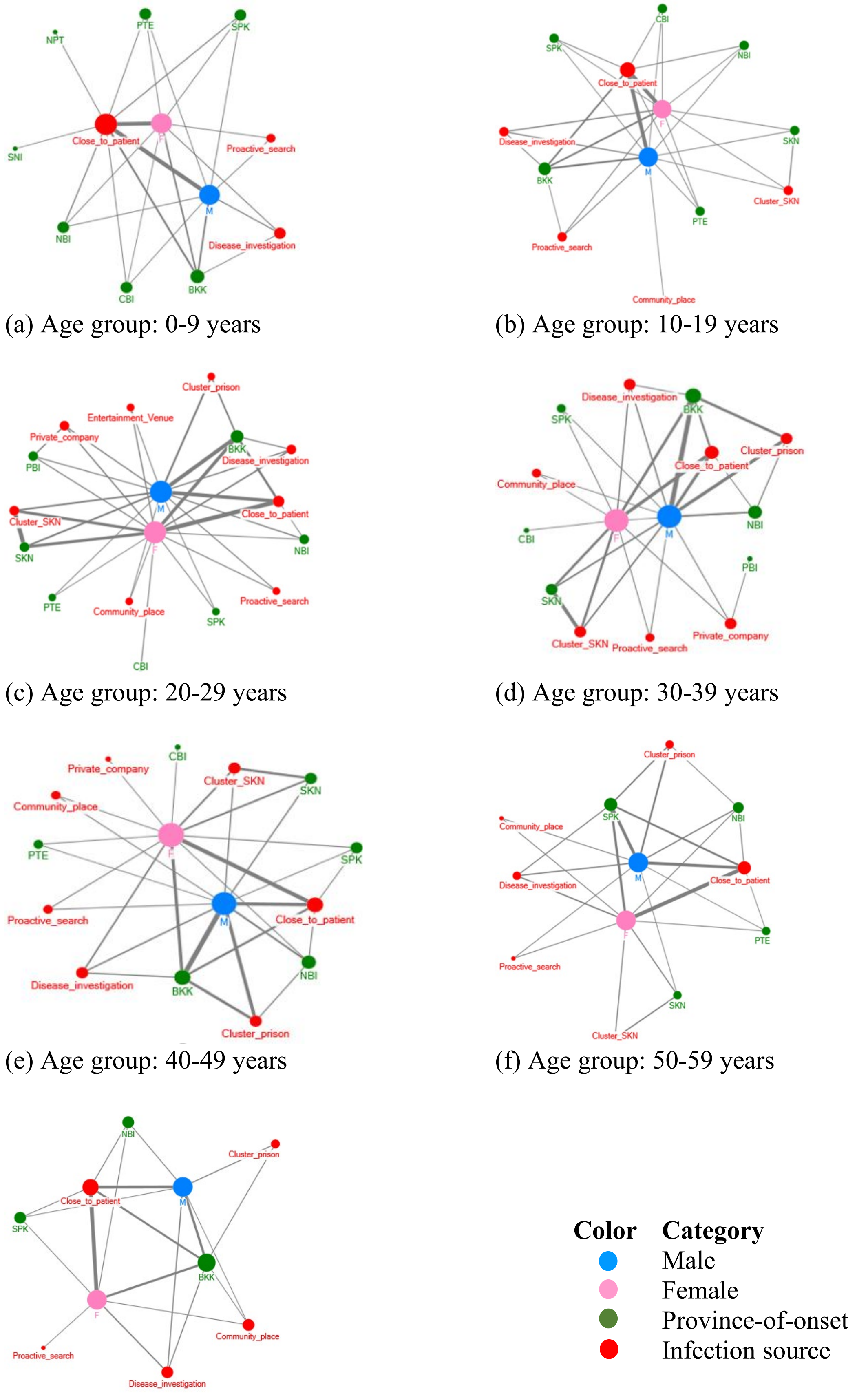
Cluster	Age_group (yrs)	No. of datasets	No. of rules
1	[0-9]	3,872	17
2	[10-19]	6,080	11
3	[20-29]	31,683	24
4	[30-39]	28,057	22
5	[40-49]	16,786	21
6	[50-59]	10,089	20
7	[>= 60]	7,072	13

Association rules from each cluster are ranked based on the value of support and lift. The selected rules have the lift values greater than 1, meaning that the occurrences between

premises and conclusions are dependent on each other. These rules are potentially useful for predicting future consequences. The examples of some association rules are shown in Table 3. According to the results in Table 3, the existing rules can be described, as follows;
1) The existing rules that were common in clusters 1-2 or groups of young children and teenagers implied the same conclusions. Both male and female young children and teenagers who lived in Bangkok and vicinity areas -Samut Prakhan (SPK), Nonthaburi (NBI), and Samut Sakhon (SKN)-were more likely to infect COVID-19 from being close to the patient.
2) The existing rules that were common in cluster 3-5 or the young working people, for examples:
{sex = F, infection_source = Cluster_SKN} → {province_of_onset = SKN},
{sex = M, infection_source = Cluster_SKN} → {province_of_onset = SKN},
These rules can be implied that among the workers from the age group 20-29 years, 30-39 years, and 40-49 years, the major infected disease in both male and female were
from Cluster_SKN (or foreign workers in Samut Sakhon).
3) The infection source in "Cluster_prison" that occurred in cluster 3, 5, 7 for examples:
{sex = M, infection_source = Cluster_prison} → {province_of_onset = BKK},
{sex = M, province_of_onset = NBI} → {infection_source = Cluster_prison},
And those occurred in cluster 6, for examples:
{infection_source = Cluster_prison} → {sex = M, province_of_onset = BKK},
{infection_source = Cluster_prison, province_of_onset = NBI} → {sex = M}.
These rules can be implied that the male prisoners in the age group from 20 to over 60 years were the infected group that made the prisons in Bangkok (BKK), and Nonthaburi (NBI) were at risk of spreading disease.
4) The additional rules that showed in cluster 7 or the elderly people, for examples:
{sex = F, province_of_onset = NBI} → {infection_source = Close_to_patient},
{province_of_onset = SPK} → {infection_source = Close_to_patient,
{infection_source = Close_to_patient} → {sex = F}.
These rules can be implied that the aging females in vicinity areas, such as Nonthaburi (NBT), and Samut Prakhan (SPK) were likely to infected disease from being close to patient.
Next, association rules from each cluster can be visualized using the network graph. Figure 3 (a)-(g) shows the network graph for association rules of particular age group clusters, ranked according to the support and the lift values.

Table 3 Examples of association rules of each cluster classified by Age-group

Cluster	Age_group	Premises	Conclusions	Support	Confidence	Lift
1	[0-9]	sex = F, province_of_onset = SPK	infection_source = Close_to_patient	0.028	0.859	1.213
		sex = F, province_of_onset = NBI	infection_source = Close_to_patient	0.027	0.858	1.211
		sex = M, province_of_onset = NBI	infection_source = Close_to_patient	0.030	0.817	1.153
		sex = M, province_of_onset = SPK	infection_source = Close_to_patient	0.026	0.800	1.129
2	[10-19]	sex = F, province_of_onset = SKN	infection_source = Cluster_SKN	0.027	0.794	14.545
		sex = M, province_of_onset = SKN	infection_source = Cluster_SKN	0.025	0.788	14.437
		province_of_onset = NBI	infection_source = Close_to_patient	0.041	0.755	1.327
		province_of_onset = SPK	infection_source = Close_to_patient	0.037	0.713	1.254
3	[20-29]	sex = F, infection_source = Cluster_SKN	province_of_onset = SKN	0.104	0.985	5.639
		sex = M, infection_source = Cluster_SKN	province_of_onset = SKN	0.059	0.976	5.591
		sex = M, infection_source = Cluster_prison	province_of_onset = BKK	0.050	0.727	2.650
4	[30-39]	sex = F, infection_source = Cluster_SKN	province_of_onset = SKN	0.098	0.983	5.714
		sex = M, infection_source = Cluster_SKN	province_of_onset = SKN	0.059	0.976	5.671
		sex = M, province_of_onset = NBI	infection_source = Cluster_prison	0.037	0.577	3.519
		sex = M, infection_source = Cluster_prison	province_of_onset = BKK	0.103	0.698	2.138
5	[40-49]	sex = M, infection_source = Cluster_SKN	province_of_onset = SKN	0.038	0.967	8.141
		sex = F, infection_source = Cluster_SKN	province_of_onset = SKN	0.066	0.960	8.082
		sex = M, infection_source = Cluster_prison	province_of_onset = BKK	0.098	0.689	2.006
6	[50-59]	infection_source = Cluster_SKN	sex = F, province_of_onset = SKN	0.044	0.626	12.024
		infection_source = Cluster_prison	sex = M, province_of_onset = BKK	0.067	0.612	3.324
		infection_source = Cluster_prison, province_of_onset = NBI	sex = M	0.030	1.000	2.049
7	[>= 60]	sex = M, infection_source = Cluster_prison	province_of_onset = BKK	0.038	0.704	2.042
		sex = F, province_of_onset = NBI	infection_source = Close_to_patient	0.027	0.689	1.286
		province_of_onset = SPK	infection_source = Close_to_patient	0.039	0.602	1.124
		infection_source = Close_to_patient	sex = F	0.317	0.592	1.100



(a) Age group: 0-9 years

(b) Age group: 10-19 years

(c) Age group: 20-29 years

(d) Age group: 30-39 years

(e) Age group: 40-49 years

(f) Age group: 50-59 years

(g) Age group: >= 60 years

Figure 3 (a)-(g) Network graph for association rules of Age group

According to association rules network Figure 3, the size of graph nodes is based on support values and the size of edges is based on lift values. The larger nodes imply higher support while the thicker edges imply the stronger of rules. All of the rules in each cluster can be visually illustrated, as follows;

1) The age group of 0-9 years as well as the age group of 10-19 years, consists of young children and teenagers. Most of the young children and teenagers live in Bangkok (BKK) and surrounding provinces, such as Samut Prakhan (SPK), Samut Sakhon (SKN), Nonthaburi (NBI), Pathumtani (PTE), and Chonburi (CBI). The infection was due to being close to the patient.

2) The age group of 20-29 years, consists of young workers. Most of the infected cases in both male and female were Cluster SKN or foreign workers in Samut Sakhon (SKN). This was due to being close to the patient, followed by people in Bangkok (BKK) that the infection was found from disease investigation. The infection sources include private companies, community places, entertainment venues, and prisons.

3) The age group of 30-39, consists of adult workers. The major infection sources were from Cluster SKN or foreign workers in Samut Sakhon (SKN) caused by being close to the patient. In Bangkok (BKK) and Nonthaburi (NBI), the infection source was found in prisons from disease investigation and proactive search. Most of them were male prisoners.

4) The age group of 40-49 years, consists of senior workers. The infection sources were mainly in Bangkok (BKK) and Nonthaburi (NBI). Males were infected diseases from prisons, while females were infected from being close to the patient.

5) The age group of 50-59 years and 60 years or over were pre-aging and aging people, respectively. The infection cases are caused by being close to the patient. The infection sources were mainly in Bangkok (BKK), Nonthaburi (NBI), Samut Prakhan (SPK), and Samut Sakhon (SKN).

In summary, there was the number of infected cases in all clusters, both male and female have similar numbers. The infection sources of this outbreak occurred in Bangkok and its vicinity, including industrial areas, such as Samut Sakhon (SKN), Samut Prakhan (SPK), Pathumtani (PTE), and Chonburi (CBI), then spread to many provinces in Thailand. The working people, included the age group of 20-29 years, 30-39 years, and 40-49 years, were infected with disease from community places, private companies, entertainment venues, foreign workers in Samut Sakhon, and prisons. The young children, teenagers, and pre-aging, aging people, in the age group of 0-9 years, 10-19 years, 50-59 years, and 60 years or over, were infected disease from being close to the patient, which is expected to be a working group of people in the family.

Conclusion

This research presented the COVID-19 in Thailand from the second wave to the third wave of an outbreak from December 2020-May 2021. The data mining methods based on descriptive tasks, i.e., cluster analysis, and association rule mining were possible to analyze the demographic data. Cluster analysis was used to categorize datasets into groups based on similarity, using K-means clustering. The number of clusters was obtained from the cluster validation process or elbow method. In this case, the age group of infection cases was used as a variable to identify the cluster profiles whether the province of onset and infection sources are linked to the transmission of disease. Next, association rule mining was used to investigate co-occurrence between itemset of each cluster. There were 128 rules generated by association rule mining, which helped to find patterns of COVID-19 infection. The whole set of derived rules can also be visualized in form of a network graph, differentiated by the size of nodes and thickness of edges using the measurement of the rules, i.e., the support and the lift, so that we can illustrate the most of infection source as well as province of onset and the importance of rules that existed in each cluster. The association network can be used to identify by the nontechnician to easily understand a point of view and pave the way for a precaution the outbreak in future.

The results highlighted the information characteristics of each age group by identifying relationships between sex, province of onset, and infection sources. Males and females were equally infected with COVID-19 disease. The province of onset was mainly in Bangkok-vicinity areas as well as industrial areas, such as Nonthaburi, Samut Sakhon, Samut Prakran, Chonburi, Pathumtani, etc. Adult workers in the age group of 20-29 years, 30-39 years, 40-49 years were the seeds of new infection sources. Apart from the common sources of being close to the patient, disease investigation, and proactive search, the new infection sources were from the entertainment venue, the private company, the community place, and the cluster prison. Young children, teenagers, pre-aged people and aging people were more likely to infect disease from being close to the patients. This might be from the family members who are working people, as mentioned previously. If there are more information related to the infected person is collected, then the data analytics is expected to reveal many influencing factors. Apart from demographic data, the information characteristics of common laboratory values, such as serum calcium levels, temperature, age, lymphocyte count, smoking, hemoglobin levels, and oxygen saturation could make the prediction of survival possible.

Acknowledgment

The author thanks to Department of Disease Control, the Ministry of Public Health, and Thailand's open government data center for providing data.

References

Ahmad, P., Qamar, S., & Rizvi, S. (2015). Techniques of Data Mining in Healthcare: A Review. *International Journal of Computer Application, 120*(15), 38-50.

Ayyoubzadeh, S., Ayyoubzadeh, S., Zahedi, H., Ahmadi, M., & Kalhori, S. (2020). Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR Public Health and Surveillance, 6*(2), e18828.

Bagheri, H., Tapak, L., Karami, M., Hosseinkhani, Z., Najari, H., Karimi, S., & Cheraghi, Z. (2020). Forecasting the monthy incidence rate of brucellosis in west of Iran using time series and data mining from 2010 to 2019. *PLoS ONE, 15*(5), e0232910.

Bangkokpost. (2020). *Coronavirus outbreak.* Retrieved from www.bangkokpost.com/topics/1844044/coronavirus-outbreak.

Buscema, P., Torre, F., Breda, M., Massini, G., & Grossi, E. (2020). COVID-19 in Italy and extreme data mining. *Journal of Physica A, 557*, 124991.

Department of Disease Control. (2020). *COVID-19 easing measure.* Retrieved from https://ddc.moph.go.th/viralpneumonia/gui_covid19_phase.php.

Digital Government Development Agency (Public Organization). (2021). *Covid-19-daily.* Retrieved from https://data.go.th/dataset/covid-19-daily.

Hinjoy, S., Tsukayama, R., Chuxnum, T., & et. al. (2020). 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 Disease, 96*(2020), 205-210.

Husein, I., Noerjoedianto, D., Sakti, M., & Jabbar, A. (2020). Modeling of Epidemic Transmission and Predicting the Spread of Infectious Disease. *Journal of Systematic Reviews in Pharmacy, 11*(6), 188-195.

Kodinariya, T., & Makwana, P. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Sciences and Management Studies, 1*(6), 90-95.

McAleer, M. (2020). Prevention is better than the cure: Risk management of COVID-19. *Risk and Financial Management, 13*(3), 46.

Ministry of Public Health. (2020). *Coronavirus disease 2019 (COVID-19) Thailand situation.* Retrieved from https://media.thaigov.go.th/uploads/public_img/source/310763.pdf.

Pitchayadejanant, K., & Nakpathom, P. (2018). Data Mining approach for arranging and clustering the agro-tourism activities in orchard. *Kasetsart Journal of Social Sciences, 39*(2018), 407-413.

Rotejanaprasert, C., Lawpoolsri, S., Pan-ngam, W., & Maude, R. (2020). Preliminary estimation of temporal and saptiotemporal dynamic measures of COVID-19 transmission in Thailand. *PLoS ONE, 15*(9), e0239645.

Syakur, M. (2018). Integrating K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series Material Science and Engineering, 336*(1), 012017.

Triukose, S., Nitinawarat, S., Satian, P., & et al. (2021). Effects of public health interventions on the epidemiological spread during the first wave of the COVID-19 outbreak in Thailand. *PLoS ONE, 16*(2), e0246274.

World Health Organization. (2020). *Novel Coronavirus-Thailand (ex-China)*. Retrieved from https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON234.

Wikipedia. (2020). *COVID-19 pandemic in Thailand.* Retrieved from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Thailand.

Zainol, Z., Wani, S., Nohuddin, P., & et al. (2018). Association Analysis of Cyberbulling on Social Media Using Apriori Algorithm. *International Journal of Engineering & Technology, 7*(4.29), 72-75.