Data modeling COVID-19 patients in Thailand: 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 Feb 13, 2023
Revised Aug 7, 2023
Accepted Aug 30, 2023

Keywords:

Clustering
COVID-19
Data mining
Data modeling
Thailand

ABSTRACT

This study aimed to investigate the characteristics of COVID-19 patients in Thailand and develop a data model for analyzing these characteristics. A total of 1,888,941 cases from the Thailand Department of Disease Control website from January 12, 2020, to October 29, 2021, were analyzed, and 20,110 cases were selected for further analysis. The two-step cluster analysis method was used to cluster the data according to four variables: nationality, occupation, patient type, and risk groups. The results showed the presence of three groups of COVID-19 patients. Group 1 consisted of 5,883 workers in trade and service occupations who had contact with the public and were either Thai nationals or from abroad. Group 2 was the largest cluster, consisting of 7,420 migrant workers classified as foreigners and working in industrial settings. Group 3 consisted of 6,870 cases of indirect transmission, with individuals in this group infected through close contact with family members or individuals in the first two groups. This clustering approach offers valuable insights for pandemic management, aiding in identifying high-risk groups and developing tailored interventions. In future outbreaks with similar characteristics, such as airborne transmission, contact infection, or super spreader events, our model can serve as a valuable tool for devising effective management plans and countermeasures. In conclusion, this study emphasizes the significance of cluster analysis in understanding the dynamics of COVID-19 transmission and highlights its potential for informing effective pandemic management strategies. It also outlines promising avenues for further research to enhance our knowledge of COVID-19's impact on specific populations and inform future public health efforts.

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

COVID-19 is emerging disease cause by coronavirus 2(SARS-COV-2), it’s high infectivity and virus spread can transmission from human to human. That causes of respiratory tract infection, the infection process starts from [1] virus infect by airborne transmission using coughing, sneezing, or other secretions There is an incubation period of 2-14 days [2]. After virus infection the patient was symptomatic as fever, dry cough, fatigue, and dyspnea, however this symptomatic had high efficacy in elderly (60-year-old) who was congenital disease such as lunge and respiratory system. The first confirmed case of COVID-19 was detected in Wuhan, China, and pandemic in worldwide till The World Health Organization (WHO) has declared it a state of emergency. After about two years (December 10, 2021), more than 267 million infections cases and more than 5 million deaths were found worldwide [3]. In Thailand, more than 1.89 million infections cases were confirmed and more than 19,070 deaths [4]. For Thailand is considered a very high number of deaths. In addition to public health problems, COVID-19 emerging still effects to economic, daily life, occupation, and unemployment [5], [6]. The COVID-19 impact to many problems caused by the missing of information about the standard operating procedures for dangerous communicable diseases, similarly if knowing the information about the population risk groups it can be to management and preventing outbreak in the future. This research is divided into 4 parts as follows: Literature and reviews, research methods, results, and conclusions and recommendations.

COVID-19 is a latest emerging disease cause by coronavirus 2 (SARS-COV-2) that can infect from human to humans it’s etiologies of acute respiratory syndrome [1]. Virus is RNA virus type (positive-seneRNA) a biggest virus group and higher diversity of host such as bat cat dog and human all the mammalian [7]. First reported was in Wuhan, Hubei province, China in December 2019 that super spreader in Wuhan markets. Fifty day of super spreader infection has 1,800 deaths caused was confirmed at the moment China government sent information to World Health Organization (WHO) for reporting a new virus emerging to Emergency response on March 21, 2020 and called the respiratory disease caused by new strain coronavirus is coronavirus disease 2019 or COVID-19 [8]. CO instead of corona, VI for Virus, D for disease and 19 for 2019. The timeline of COVID-19's emergence in Wuhan, China and spread to other countries and continued to spread globally, leading to the declaration of a pandemic by the World Health Organization in March 2020 [9].

COVID-19 is effective for respiratory syndrome and easy transmission it’s infecting by airborne transmission such as airborne droplets, cough, or direct contact with patients. The analysis of viral genetic shown relationships with virus in bat causes of pneumonia like the severe acute respiratory syndrome (SARS) disease was emerged in 2003. Initially, it was not clear about the transmission of virus from animal to human, but the virus has zoonosis characterized [10]. Coronaviruses were varying sensitivity it has affected asymptomatic to acute respiratory syndrome and has an effect to organs system such as ageusia, anosmia and dyspnea that severity of disease depends on the congenital disease of patients. Airborne transmission such as droplets from coughing, sneezing or secretions can transmission virus and infection to replicate. After infection there is an incubation period of 2-14 days, the symptoms will be shown in 5 days after infection [2]. The research of Chen Wang and other studies COVID-19 symptoms was reported by The Lancent 41 COVID-19 confirmed cases compared with SARS and middle east respiratory syndrome (MERS) disease all caused from coronavirus that showed similar symptoms including dry cough, dyspnea and fever.

First patient of COVID-19 in Thailand was reported on January 12, 2020 it’s Chinese tourists and first Thai people reported in January 31, 2020 they were Taxi driver, no traveling abroad. After that, Number of patients were increase until the study date (October 29, 2021) more than 246 million cases worldwide and 4.99 deaths [3] were reported for Thailand was reported 1.89 million confirmed cases and 19,070 deaths case [4]. After outbreak situation the government was established centre for administration of situation due of outbreak the communicable disease coronavirus 2019 to manage the COVID-19 outbreak situation.

Data modeling is the process of creating simple diagrams of existing software systems and data elements. It uses text and symbols to represent information and data flow diagram. Data model layouts are made for designing new databases or reconfiguring legacy applications that helps organizations use available data efficiently and recompense need of organize business [11]. The process of creating and developing a data model can be accomplished in many of ways such as Factor Analysis, Cheating Model and Data mining. In this research will use the cross industry standard process for data mining (CRISP-DM) for data mining including 6 steps [12], [13] i) business understanding, ii) data understanding, iii) data preparation, iv) data modeling, v) data evaluation and vi) data development.

In recent years, there has been a growing interest in utilizing data mining techniques in the medical field, particularly in the context of the COVID-19 pandemic [14]–[16]. Alimadadi *et al.* [17] noted that the global AI community, in collaboration with technology and research companies, was approached by the White House to develop various data mining techniques to support COVID-19-based studies aimed at finding a solution to the pandemic. Tasnim and colleagues [18] suggested adopting advanced data mining techniques, specifically natural language processing, to identify and eliminate non-scientific online content. Ayyoubzadeh *et al.* [19] stressed the potential of data mining algorithms for studying and predicting the spread and trends of the COVID-19 outbreak worldwide, and used the Long Short-Term Memory (LSTM) data mining model to analyze data from the Google Trends website.

Abd-Alrazaq *et al.* [20] applied text-mining techniques based on data mining to analyze tweets collected between February 2, 2020, and March 15, 2020, utilizing sentiment analysis and topic modeling to gain insights into public opinion regarding the pandemic. Franch-Pardo and colleagues [21] emphasized the importance of adopting interdisciplinary perspectives and utilizing diverse approaches like data mining, web-based mapping, and spatiotemporal analysis to tackle the complex challenges presented by the pandemic. Similarly, Li and colleagues [22] employed linear regression and content analysis methodologies of data mining to analyze information from the Chinese microblogging platform Weibo. Their study offered valuable insights into the COVID-19 outbreak, including classifying news and user-generated topics, contributing to a better understanding of the situation. Qin *et al.* [23] suggested the importance of estimating the number of new and confirmed cases of COVID-19, and worked on data collected from social media search indexes for keywords such as dry cough, fever, and coronavirus to achieve this goal.

Kumar [24] discussed the use of artificial intelligence, including machine learning and natural language processing, in fighting the COVID-19 crisis based on medical data. Han *et al.* [25] used the Latent Dirichlet allocation (LDA) model and the random forest algorithm to analyze text from Sina-Weibo to investigate public opinion, examining the text's space, time, and substance. Finally, Mehrotra and Agarwal [26] reviewed and summarized the data mining techniques used to study the COVID-19 pandemic, noting the potential for data mining in disease prediction and cure. In conclusion, data mining techniques have become an increasingly important area of research in the medical field, particularly with the COVID-19 pandemic.

This study embarked on an ambitious journey to explore the intricacies of COVID-19 infections in Thailand, with its centerpiece being constructing a robust data model. This model, reinforced by meticulous data curation and advanced analytical techniques, provided a comprehensive understanding of the virus's spread within the Thai population. Its insights serve as a beacon of guidance for the country's pandemic response and offer valuable lessons for the global community in the ongoing fight against infectious diseases.

2. RESEARCH METHOD

Unsupervised Learning is separate all data set into groups by using similar data into same groups and different data into another group. Data clustering can be divided for 3 types are i) Hierarchical clustering Analysis, ii) Nonhierarchical clustering analysis for examples K-mean clustering and iii) Two step cluster analysis [27]. Hierarchical clustering analysis is a famous technique for create data group (Case) by using a small of data leases than 200 data and less of variable, that can use for all of data including; i) Select the variables that can are expected to be related and results in the clustering. ii) Select the method for measure distance for similar data (Case). iii) Select the guideline for grouping (Cluster).

Nonhierarchical clustering is an analysis technique that is suitable for data with a big data (Case), more than or equal 200 data. This technique is quantitative information in ratio scale and internal scales; however, this technique needs to know number of group or find K values before that call this technique is K-means clustering and including; i) Identify number of K values, ii) Identify K centroid for each cluster, iii) Determine distance of objects to centroid, iv) Grouping objects based on minimum distance, v) Centroid changes, If the data is not migrated in step 4, the resulting segmentation is assumed to be appropriate and readable. But if the data is migrated Steps 2-3 must be followed until no more groups of data have been moved.

Two step cluster analysis is suitable for clustering analysis of large number of cases. And it is the information in the ordinal scale and the nominal scale. The system will calculate the appropriate number of clusters. The number of groups is not required. The algorithm is as follows; i) Separate data to small groups (Pre cluster the case or records), ii) Combine small group into according to number of groups required. Similar group will together, and different data will create to another group. Following this algorithm [28].

$$d_{(R)(S)} = \xi_R + \xi_S - \xi_{(R,S)} \tag{1}$$

Where $\xi_v = -N_v \cdot \left[\left[\sum_{k=1}^{K^A} \frac{1}{2} \cdot log(\hat{\sigma}_k^2 + \hat{\sigma}_{v\cdot k}^2)\right] + \left[\sum_{k=1}^{K^B} \hat{E}_{v\cdot k}\right]\right]$

and where $\hat{E}_{v\cdot k} = -\sum_{l=1}^{L_k} \left[\frac{N_{v\cdot k\cdot 1}}{N_v} \cdot log\left[\frac{N_{v\cdot k\cdot 1}}{N_v}\right]\right]$

In the analysis, K^A represents the total count of the continuous variable, while K^B signifies the total count of the categorical variable. R_k refers to the interval or range of the kth continuous variable, N represents the total number of observations in the database, and N_k denotes the number of objects in the kth cluster. Furthermore, σˆ_k^2 corresponds to the estimated variance of the kth continuous variable for all data, and σˆ_Rk^2 is the estimated variance of the kth continuous variable within the R cluster. N_Rkl signifies the number of objects in the R cluster. The distance between the R and S clusters is denoted as d_((R)(S)), where R and S represent indices that indicate the clusters formed by joining the clusters R and S. Determining the number of clusters: i) The Bayesian information criterion (BIC), ii) Akaike information criterion (AIC). Based on the Pseudo-F value, the higher the value, the more appropriate [14], [16].

$$BIC_R = -2 \cdot \sum_{i=1}^{R} \xi_R + m_R \cdot \log(N) \quad (2)$$

$$AIC_R = -2 \cdot \sum_{i=1}^{R} \xi_R + 2 \cdot m_R \quad (3)$$

Where $m_R = R \cdot \{2 \cdot K^A + \sum_{k=1}^{K}(L_k - 1)\}$

and where L_k *is number of groups in* k

In this study will use two step cluster analysis because of the larger data (cases) and this data is in the ordinal scale and nominal scale.

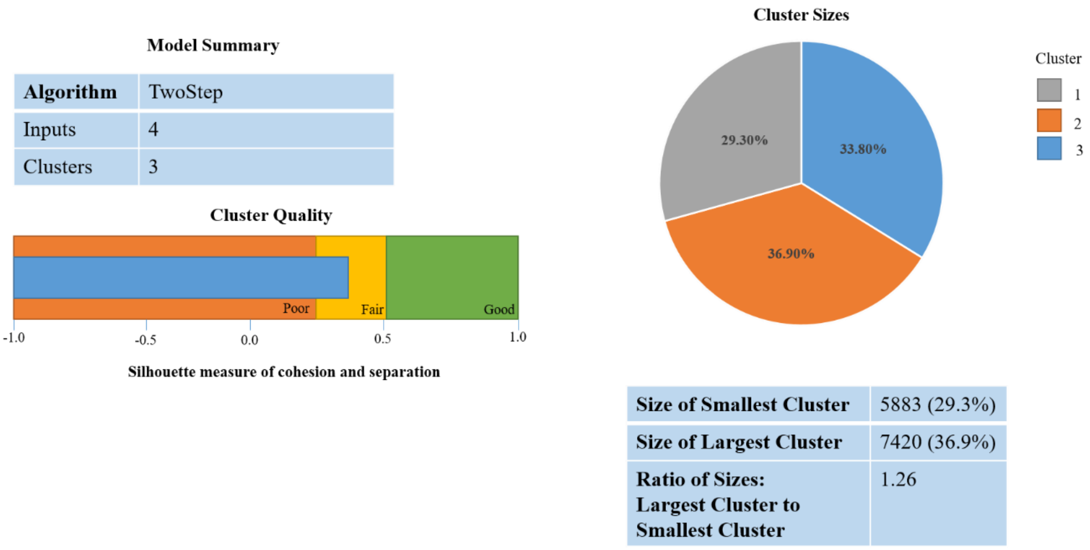
2.1. Data

The data were used in this study from Thai Department of Disease Control, since January 12, 2020, to October 29, 2021, 1,888,941 cases was confirmed in Thailand. After cleaning data showed 20,100 cases that will use to in this study by IBM SPSS statistics.

2.2. Two step clustering analysis

Once cleaned, the data will be analyzed using IBM SPSS Statistics Version 28.0.10 (142). Excel files were imported into SPSS, and the two step analysis method was applied with selected variables categorized as categorical variables. Schwarz's Bayesian criterion (BIC) or Akaike information criterion (AIC) will be used to determine the suitable grouping, choosing the lowest value of BIC and AIC. However, selecting the lowest BIC or AIC may lead to a higher number of groups, complicating the analysis. The Ratio of Distance Measures will also be analyzed, indicating significant distances between groups. This information is derived from reference [29]. Ultimately, the system will automatically determine the appropriate number of groups based on the applied criteria and analysis methods.

The clustering analysis will employ pertinent and thoughtfully selected variables to group the data effectively. Following the analysis, four variables remained significant: nationality, occupation, risk group, and patient type. To present the results visually, we transitioned from the model summary diagram to the Clusters table, as shown in Figure 1. This table offers comprehensive details about each cluster's clustering outcomes and data. Subsequently, we thoroughly examined the data within each cluster to summarize the findings and assign appropriate names to each group. This meticulous process established a well-defined model for categorizing COVID-19 patients in Thailand based on their nationality, occupation, risk group, and patient type.



Model Summary

Algorithm	TwoStep
Inputs	4
Clusters	3

Size of Smallest Cluster	5883 (29.3%)
Size of Largest Cluster	7420 (36.9%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.26

Figure 1. Model summary and cluster sizes diagram viewer

2.3. Data model quality evaluation

Select an assessor by an expert from faculty of Medicine, Khon Kaen University. The selected assessor holds advanced degrees in relevant fields and has a strong track record of research and publications in areas related to infectious diseases, including COVID-19. Their affiliation with Khon Kaen University, a renowned institution in the medical and research communities, further validates their expertise and adds to their credibility as an assessor. Having experience evaluating research methodologies and data analysis techniques, the assessor possesses the skills to examine the data model thoroughly. They are familiar with the complexities and challenges associated with analyzing large-scale datasets, particularly in the context of infectious diseases, and can identify potential biases or limitations in the model's design.

Create the quality assessment of data models referenced by handbook on data quality assessment methods and tools [30]. This assessment process ensures that the data models produce valid and valuable insights, enabling decision-makers and researchers to make informed and evidence-based conclusions. The quality assessment begins with thoroughly reviewing the data model's construction methodology. This includes examining the data sources, data collection processes, and data preprocessing techniques applied. The assessment ensures that the data used in the model are of high quality, free from errors, duplications, or inconsistencies that could adversely affect the model's outcomes.

Using the data model of the COVID-19 patient group obtained for experts to evaluate according to the quality assessment of the data model (Model) to consider its validity and suitability. Using the data model of the COVID-19 patient groups obtained from the study, experts in the field were engaged to perform a rigorous evaluation based on the quality assessment of the data model. This evaluation aimed to assess the validity and suitability of the data model for accurately representing the characteristics and patterns observed in the COVID-19 patient population in Thailand. During the evaluation process, experts carefully scrutinized the methodology used in the data model construction, ensuring that it adhered to rigorous scientific standards and statistical techniques. They reviewed the choice of variables, the clustering algorithm employed (Two-step Cluster Analysis), and the overall approach to ensure it was appropriate for the research objectives. The experts also assessed the integrity and reliability of the dataset used to construct the data model. They verified the data sources, data collection methods, and accuracy of the reported COVID-19 cases to confirm the trustworthiness of the information used in the analysis.

3. RESULT AND DISCUSSION

3.1. Two step clustering analysis

The research dataset encompassed a substantial cohort of 20,110 individuals who tested positive for COVID-19. In their quest to gain deeper insights into the characteristics of these patients, the researchers employed the Two-Step Clustering Analysis method, which took into account four essential variables: nationality, occupation, risk group, and patient type. This analytical approach allowed them to discern underlying patterns and relationships within the data, identifying three distinct and meaningful clusters. The first cluster, aptly labeled Group 1, comprised 5,883 confirmed COVID-19 cases, representing approximately 29.3% of the total dataset. This group primarily consisted of individuals with specific attributes related to their nationality, occupation, risk factors, and patient type. These defining characteristics played a significant role in shaping the transmission dynamics and the nature of the cases within this cluster. Group 2 emerged as the largest cluster, encompassing 7,420 confirmed cases, accounting for 36.9% of the dataset. The individuals in this cluster shared common characteristics related to their nationality, occupation, risk group, and patient type. This specific combination of attributes distinguished Group 2 from the other clusters and highlighted their unique role in the overall COVID-19 epidemiology. Group 3, comprising 6,870 confirmed cases, constituted approximately 33.8% of the dataset. These individuals exhibited distinct features regarding nationality, occupation, risk group, and patient type. The defining characteristics of Group 3 set it apart from the other clusters and offered valuable insights into the patterns of infection and transmission within this particular cohort.

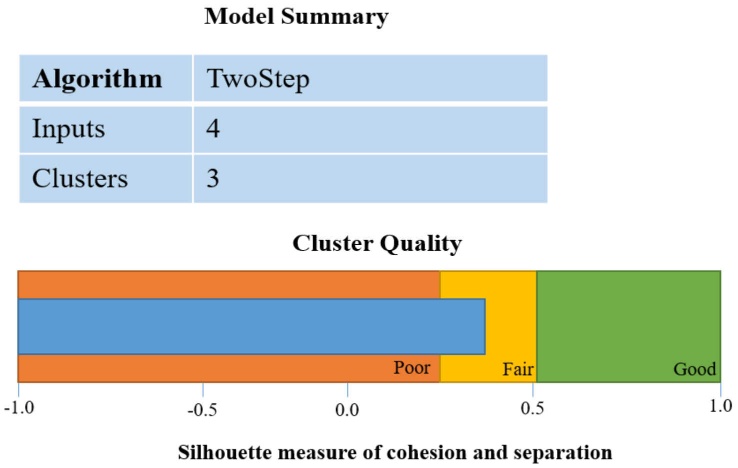
For a more comprehensive understanding of the attributes and dynamics of each group, additional detailed information can be found in Table 1 and Figure 2. These visual representations and tabulated data provide a comprehensive snapshot of the distinctive characteristics and distribution of COVID-19 cases within each cluster, enriching our understanding of the pandemic's impact on different population segments.

Table 1. The results of two step clustering analysis

Cluster	N	% of Combined	% of Total
1	5883	29.3%	29.3%
2	7420	36.9%	36.9%
3	6870	33.8%	33.8%
Combined	20110	100.0%	100.0%
Total	20110		100.0%

3.2. Groups of clusters

Figure 3 and Table 2 presented the results of the data analysis, demonstrating that Group 2 had a higher number of cases than Group 3 and Group 1. The data indicated a distinct flow of cases from Group 2 to the other two groups, suggesting that Group 2 played a significant role in the transmission dynamics of COVID-19 among the identified clusters. Which demonstrated that Group 2 exhibited significantly higher numbers of COVID-19 cases compared to both Group 1 and Group 3. Figure 3 visually presents the distribution of cases across the three identified patient groups. The graph displayed a clear prominence of Group 2, illustrating a substantial proportion of COVID-19 infections being attributed to this cluster. On the other hand, Group 1 and Group 3 appeared to have comparatively lower-case numbers, highlighting their relatively smaller contributions to the overall spread of the virus.

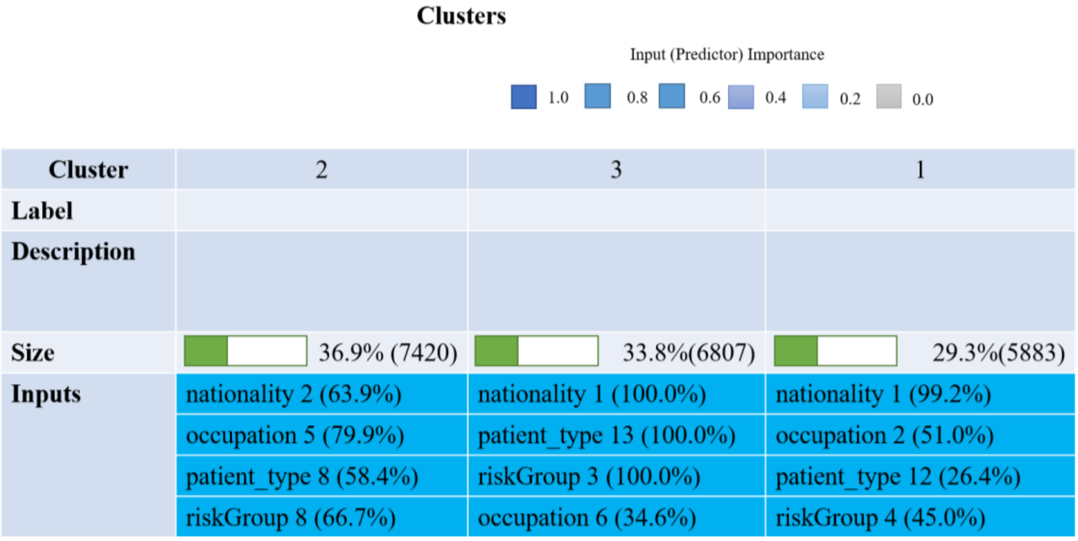


Model Summary

Algorithm	TwoStep
Inputs	4
Clusters	3

Cluster Quality

Figure 2. The two step clustering analysis result



Clusters

Input (Predictor) Importance

■ 1.0 ■ 0.8 ■ 0.6 ■ 0.4 ■ 0.2 ■ 0.0

Cluster	2	3	1
Label			
Description			
Size	36.9% (7420)	33.8%(6807)	29.3%(5883)
Inputs	nationality 2 (63.9%)	nationality 1 (100.0%)	nationality 1 (99.2%)
	occupation 5 (79.9%)	patient_type 13 (100.0%)	occupation 2 (51.0%)
	patient_type 8 (58.4%)	riskGroup 3 (100.0%)	patient_type 12 (26.4%)
	riskGroup 8 (66.7%)	occupation 6 (34.6%)	riskGroup 4 (45.0%)

Figure 3. Cluster table

Table 2. Grouping results

Factor	cluster 1 5,883 cases	cluster 2 7,420 cases	cluster 3 6,807 cases
nationality	Thai (99.2%)	foreigners (63.9%)	Thai (100%)
occupation	Commerce and service careers (51.0%)	industrial career (79.9%)	not working (34.6%)
patient_type	Thai people come from abroad (26.4%)	risk group survey	
(58.4 %)	touch an infected person (100%)		

3.3. The details of data analysis used two step clustering analysis

Group 1 was Thai nationality, occupation: trading and services, risk groups they influenced from abroad and communities spread such as restaurants, boxing stadiums and entertainment places that called this group is dealer and service. The data analysis revealed that Group 1 significantly impacted the spread of the virus through multiple pathways. Firstly, due to their occupation in trading and services, they frequently interacted with people in public settings such as markets, restaurants, and entertainment venues. This close contact with customers and visitors in high-traffic places facilitated the potential transmission of the virus, leading to infections both among themselves and within the broader community.

Group 2 migrant workers was foreigner, occupation: industrial workers, immigrants worked, Risk groups was infected from industries or markets. This group is biggest cluster. These individuals were employed as industrial workers, primarily in various industries within the country. The data analysis revealed that Group 2 had a notable influence on COVID-19 transmission, particularly in the context of infections arising from industrial settings and markets, which were identified as significant risk factors. As foreign migrant workers, this group faced unique challenges and vulnerabilities that contributed to their increased risk of infection. Often residing in close-knit communities or dormitory-style accommodations, these workers shared living spaces and work environments, facilitating the virus's rapid spread.

Group 3 indirect infection transmission this group was infected by family members or persons in the cluster of two groups before. There were infected by contagiousness or close contact with patients. Unlike Groups 1 and 2, which acquired infections through direct interactions in occupational or community settings, Group 3's infections were primarily the result of close contact with family members or individuals from the two previously identified clusters. The study findings highlighted that individual in Group 3 contracted the virus from family members who had already been infected or from persons who belonged to either Group 1 (Thai nationals in trading and services) or Group 2 (foreign migrant workers in industrial settings). This indirect transmission occurred through contagiousness, indicating that the virus was transmitted within households or close-knit social circles. The close contact and prolonged interactions within family settings made Group 3 particularly vulnerable to infection from infected family members. As COVID-19 is highly contagious, individuals within households were at increased risk of contracting the virus from an infected family member, especially if proper preventive measures were not followed. Individuals in Group 3 were also at risk of infection from people in the other two clusters. This may have occurred through interactions at social gatherings, community events, or shared public spaces where members of different clusters came into contact.

The results of the analysis showed that there are three main groups of COVID-19 patients in Thailand: i) workers in the service industry who are either Thai or from abroad, ii) migrant workers who are foreigners and work in the industrial sector, and iii) individuals who were indirectly infected through close contact with patients from the first two groups.

3.4. Data model quality evolution

The experts evaluated the quality of the data model as follows: statistical methodology (sound methodology) proper, appropriate statistical procedures data models to classify COVID-19 patients in medicine and public health that can be used to plan defenses treatment and patient care (relevance). The results of the analysis and processing will show the model summary (model summary) obtained and the cluster quality graph used to indicate the suitability of the grouping analysis is at a fair level, can be accepted (accuracy). Time and punctuality (timeliness/punctuality) are appropriate, considering the resulting data model and methods for selecting variables, which are patient data sets already in the database. The model can indicate coherence and comparability.

> *"The result of this study can be clearly explained easily accessible can be utilized, however the experts* have *suggested that the source of the information should be explained. and the scope of the data to be analyzed from which organize that will make this study more complete (Accessibility/Clarity) Summary Two step cluster analysis can be used to analyze data"*.

4. DISCUSSION

Using data from the Thailand Department of Disease Control website between January 12, 2020, and October 29, 2021, a total of 1,888,941 COVID-19 cases were analyzed using data selection, data cleaning, and data transformation techniques. The data was then clustered using the two-step cluster analysis method based on four variants: nationality, occupation, patient type, and risk groups. The results identified three distinct groups of COVID-19 patients. (i) The first group, comprising 5,883 cases (29.3% of total cases), included workers who had to deal with other people's services and were either Thai nationals or from abroad with a trade and service occupation. These cases were mainly contracted at community places. Likewise, studies by Setyowati *et al.* [31] found that the impact of COVID-19 on the informal business sector, including online motorcycle taxi drivers, affects efforts to control the spread of the disease. Kshirsagar *et al.* [32] highlight the pandemic's impact on workplace learning and businesses, while Furuse *et al.* [33] identified 22 probable primary case-patients for the clusters, most of whom were 20-39 years of age and presymptomatic or asymptomatic at virus transmission.

(ii) Group 2: comprises migrant workers who are qualified as foreigners and work in the industrial sector. This group was significantly impacted by the COVID-19 pandemic, with 7,420 reported infections accounting for 36.9% of infections in the worker area and direct contact with patients. According to Kerwin and Warren [34], foreign-born workers play a critical role in industries such as healthcare, agriculture, meatpacking and poultry processing, construction, child care, and critical retail, which have been deemed essential during the pandemic. The study highlights the high rates of foreign-born workers employed in these industries in the United States. The COVID-19 pandemic has also had a significant impact on the tourism industry in Malaysia, particularly in the airline and hotel businesses. Foo *et al.* [35] discuss the stimulus packages offered by the Malaysian government to ensure the sustainability of the tourism industry in the country.

In Bolivia, the first cases of COVID-19 were imported from Italy and Spain, according to Escalera-Antezana *et al.* [36], the patients showed symptoms similar to those previously reported, such as fever, cough, and shortness of breath. One patient required intensive care due to severe pneumonia. The study emphasizes the importance of implementing measures to prevent the importation of COVID-19 cases and detect and control clusters of cases. In Italy, the COVID-19 outbreak began in February 2020 and quickly spread across the country, leading to a nationwide lockdown in March 2020. The first known instance of local transmission occurred on February 20, 2020, in the Lombardy region. Livingston and Bucher [37] provide an overview of the outbreak in Italy, including the number of cases, deaths, and recoveries, as well as the impact on the healthcare system and economy.

(iii) Group 3: indirect infection transmission, this group was infected by family members or people in the first two groups. They were infected through contagiousness or close contact with patients. The results showed that 6,870 confirmed cases (33.8%) were Thai nationals who had contact with confirmed cases. If there is an outbreak with similar characteristics to COVID-19, such as airborne transmission, contact infection, and super-spreaders, our model can be used to develop a management plan, make decisions, and take countermeasures against emerging disease outbreaks. A study by Cai *et al.* [38] in Wenzhou, China, indicated that indirect transmission of the causative virus occurred. This may have been due to virus contamination of common objects, virus aerosolization in a confined space, or spread from asymptomatic infected persons. Similarly, a study by Danis *et al.* [39] found that a cluster of cases was linked to a British businessman who attended a chalet in the French ski resort of Contamines-Montjoie. Several other individuals who were staying in the same chalet were also infected.

The investigation identified a total of 13 confirmed cases, including four children, one of whom was asymptomatic. The virus that causes COVID-19 spreads easily among people, primarily through respiratory droplets released when someone with the virus coughs or sneezes. Close contact with an infected person or touching a surface contaminated with the virus and then touching one's mouth, nose, or eyes can also lead to transmission. A study by Jang *et al.* [40] found that in February 2020, 112 people were infected with SARS-CoV-2 associated with fitness dance classes at 12 sports facilities in Cheonan, South Korea. Vigorous exercise in densely populated sports facilities is suggested to increase the risk of infection. The transmission was suspected in the presymptomatic phase in 7 cases, and the longest period before symptom onset was five days. The study highlights the importance of implementing preventive measures in sports facilities to prevent the spread of COVID-19.

In conclusion, the COVID-19 outbreak in Cheonan was part of a larger outbreak in South Korea, with 10,765 confirmed cases by April 30, 2020. Most of the cases were from Daegu and North Gyeongsang provinces. The studies cited above suggest that indirect transmission of the COVID-19 virus can occur through a variety of means, including virus contamination of common objects, virus aerosolization in a confined space, and spread from asymptomatic infected persons. Close contact with an infected person or touching a surface contaminated with the virus and then touching one's mouth, nose, or eyes can also lead to transmission. The findings of these studies highlight the importance of implementing preventive measures to reduce the risk of COVID-19 transmission, such as social distancing, hand washing, and wearing a mask.

There are four possible directions for future research based on this study: i) Further analysis of the characteristics of each group: It would be interesting to delve deeper into the specific characteristics of each group, such as the specific occupations and risk factors that contribute to their higher likelihood of infection. This information could help to identify specific strategies for mitigating the spread of the virus within each group. ii) Comparison with other countries: It would be useful to compare the findings of this research with similar studies conducted in other countries to see if the same patterns hold true or if there are differences that could shed light on the unique characteristics of the COVID-19 pandemic in Thailand. iii) Developing interventions and prevention strategies: Based on the identified clusters and risk factors, it would be important to develop targeted interventions and prevention strategies to help reduce the spread of the virus within each group.

This could include education and awareness campaigns, testing and quarantine measures, and other public health interventions. iv) Continued monitoring and evaluation: As the COVID-19 pandemic continues to evolve, it will be important to continue monitoring and evaluating the effectiveness of different interventions and prevention strategies in reducing the spread of the virus. This could involve regularly updating the data and re-running the analysis to see if there are any changes in the clusters or risk factors over time.

5. CONCLUSION

This study utilized data from the Thailand Department of Disease Control website from January 12, 2020, to October 29, 2021, involving a thorough analysis of 1,888,941 COVID-19 cases through data selection, data cleaning, and data transformation techniques. Employing the two-step cluster analysis method with the variables of nationality, occupation, patient type, and risk groups, the research categorized the cases into three distinct groups: Group 1: direct service workers and community infections (29.3% of total cases): This cluster encompassed 5,883 cases involving Thai nationals and foreigners engaged in service-oriented jobs with potential exposure to people. Primarily contracting the virus at community places, this group's susceptibility was tied to their occupation and interactions. Similar studies corroborated that the pandemic adversely affected the informal business sector, particularly service providers like online motorcycle taxi drivers. The significance of workplace learning and its influence on businesses was also underscored.

The primary cases within this group were predominantly young adults and virus transmission often occurred while being presymptomatic or asymptomatic. Group 2: migrant industrial workers (36.9% of total cases): This category consisted of 7,420 infections among foreign migrant workers employed in industrial sectors. The pandemic's impact on these workers was highlighted, particularly in essential fields like healthcare, agriculture, and retail. Group 3: indirect infection transmission (33.8% of total cases): comprising 6,870 Thai nationals, this cluster represented individuals infected through close contact with confirmed cases from group 1 or 2. This transmission type was commonly observed, with viral spread occurring through various modes like shared objects, confined spaces, or contact with asymptomatic carriers. This study referenced prior research validating such modes of transmission and highlighted the importance of preventive measures like mask-wearing, hand hygiene, and social distancing. This study emphasized the need for effective management strategies and countermeasures against emerging disease outbreaks similar to COVID-19.

Furthermore, for future research: Detailed Group Analysis: The investigation could investigate each group's specific characteristics and occupational factors to design targeted strategies for curbing viral spread. Cross-Country Comparisons: A comparative study across countries would help validate the findings' universality and uncover potential unique aspects of Thailand's COVID-19 situation. Interventions and Prevention: The identified clusters and risk factors could drive the development of focused interventions, encompassing education, testing, and quarantine measures tailored to each group's vulnerabilities. Ongoing Evaluation: Given the evolving nature of the pandemic, continuous assessment of intervention efficacy and risk factors would be crucial for adapting strategies over time. This study employed sophisticated data analysis techniques to categorize COVID-19 cases into three distinctive groups based on occupation, nationality, and transmission routes. Each group represented a unique population segment with varying risks and challenges. This study emphasized the importance of targeted interventions, global comparisons, and ongoing evaluation in the face of a dynamically changing pandemic landscape. This study underscored the significance of understanding transmission modes and risk factors, offering valuable insights into managing and mitigating the spread of COVID-19 and future emerging diseases.

REFERENCES

[1] G. Spagnuolo, D. De Vito, S. Rengo, and M. Tatullo, “COVID-19 outbreak: an overview on dentistry,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 6, p. 2094, Mar. 2020, doi: 10.3390/ijerph17062094.

[2] S. Banava, S. A. Gansky, and M. S. Reddy, “Coronavirus disease update on epidemiology, virology, and prevention,” *Compendium of Continuing Education In Dentistry (Jamesburg, N.J. : 1995)*, vol. 42, no. 6, pp. 280–289, 2021.

[3] T. Okubo, A. Inoue, and K. Sekijima, “Who got vaccinated for COVID-19? Evidence from Japan,” *Vaccines*, vol. 9, no. 12, p. 1505, 2021. doi: 10.3390/vaccines9121505

[4] Department of Disease Control, “DDC COVID-19 Interactive dashboard,” *Department of Disease Control*, 2021. https://ddc.moph.go.th/covid19-dashboard/ (accessed Oct. 30, 2021).

[5] S. Tadesse and W. Muluye, “The impact of COVID-19 pandemic on education system in developing countries: a review,” *Open Journal of Social Sciences*, vol. 08, no. 10, pp. 159–170, 2020, doi: 10.4236/jss.2020.810011.

[6] N. Chairassamee and O. Hean, “Effects of the COVID-19 pandemic on the labour market in Thailand,” *Journal of Southeast Asian Economies*, vol. 39, no. 3, pp. 330–341, 2022.

[7] S. Manmana, S. Iamsirithaworn, and S. Uttayamakul, “Coronavirus Disease-19 (COVID-19),” *Journal of Bamrasnaradura Infectious Diseases Institute*, vol. 12, no. 2, 2020.

[8] C. Tanking, “How COVID-19 affects cancer and cardiovascular disease,” *The Journal of Chulabhorn Royal Academy*, vol. 2, no. 3, pp. 18–26, 2020.

[9] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, “A novel coronavirus outbreak of global health concern,” *The Lancet*, vol. 395, no. 10223, pp. 470–473, Feb. 2020, doi: 10.1016/S0140-6736(20)30185-9.

[10] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, “COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses,” *Journal of Advanced Research*, vol. 24, no. July, pp. 91–98, Jul. 2020, doi: 10.1016/j.jare.2020.03.005.

[11] C. Stedman, “Definition data modeling,” *Tech Target*, 2021. https://www.techtarget.com/searchdatamanagement/definition/data-modeling (accessed Nov. 13, 2022).

[12] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[13] J. Wirth, R., & Hipp, “CRISP-DM : towards a standard process model for data mining,” in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, no. 24959, pp. 29–39, [Online]. Available: http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

[14] V. Gayathri, M. C. Mona, and S. B. . Chitra, “A survey of data mining techniques on medical diagnosis and research,” *International Journal of Engineering*, vol. 6, no. 6, pp. 301–310, 2014, [Online]. Available: www.iaaet.org/sjsr.

[15] M. Herland, T. M. Khoshgoftaar, and R. Wald, “A review of data mining using big data in health informatics,” *Journal Of Big Data*, vol. 1, no. 1, pp. 1–35, Dec. 2014, doi: 10.1186/2196-1115-1-2.

[16] M. Sultana, A. Haider, and M. S. Uddin, “Analysis of data mining techniques for heart disease prediction,” in *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, Sep. 2017, pp. 1–5, doi: 10.1109/CEEICT.2016.7873142.

[17] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, “Artificial intelligence and machine learning to fight COVID-19,” *Physiological Genomics*, vol. 52, no. 4, pp. 200–202, Apr. 2020, doi: 10.1152/physiolgenomics.00029.2020.

[18] 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.

[19] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R Niakan Kalhori, “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. e18828, Apr. 2020, doi: 10.2196/18828.

[20] 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. e19016, Apr. 2020, doi: 10.2196/19016.

[21] 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, no. October, p. 140033, Oct. 2020, doi: 10.1016/j.scitotenv.2020.140033.

[22] 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. e18700, Apr. 2020, doi: 10.2196/18700.

[23] 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.

[24] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, “Chronic kidney disease analysis using data mining classification techniques,” in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Jan. 2016, pp. 300–305, doi: 10.1109/CONFLUENCE.2016.7508132.

[25] X. Han, J. Wang, M. Zhang, and X. Wang, “Using social media to mine and analyze public opinion related to COVID-19 in China,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 8, p. 2788, Apr. 2020, doi: 10.3390/ijerph17082788.

[26] A. Mehrotra and R. Agarwal, “A review of use of data mining during COVID-19 pandemic,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 6, pp. 4547–4552, 2021.

[27] K. Vanichbuncha, “*SPSS for Windows*,” Bangkok: Department of Statistics Faculty of Commerce and Accountancy Chulalongkorn University, 2010.

[28] T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris, “A robust and scalable clustering algorithm for mixed type attributes in large database environment,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2001, pp. 263–268, doi: 10.1145/502512.502549.

[29] X. Wu, F. Benjamin Zhan, K. Zhang, and Q. Deng, “Application of a two-step cluster analysis and the Apriori algorithm to classify the deformation states of two typical colluvial landslides in the Three Gorges, China,” *Environmental Earth Sciences*, vol. 75, no. 2, p. 146, Jan. 2016, doi: 10.1007/s12665-015-5022-2.

[30] Brawijaya Professional Statistical Analysis Malang, “TwoStep cluster analysis,” *Brawijaya Professional Statistical Analysis Malang*, 2011. https://arifkamarbafadal.files.wordpress.com/2011/09/ebook-038-tutorial-spss-two-step-cluster-analysis.pdf (accessed Oct. 29, 2021).

[31] D. L. Setyowati, S. Paramita, R. H. Ifroh, T. Asrianti, E. Fitriani, and W. Rahman, “Work readiness during COVID-19 among taxibike online drivers in Samarinda, Indonesia,” *International Journal of Public Health Science (IJPHS)*, vol. 10, no. 3, p. 617, Sep. 2021, doi: 10.11591/ijphs.v10i3.20870.

[32] A. Kshirsagar, T. Mansour, L. McNally, and M. Metakis, “Adapting workplace learning in the time of coronavirus,” *McKinsey & Company*, 2020. https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/adapting-workplace-learning-in-the-time-of-coronavirus/ (accessed May 03, 2023).

[33] Y. Furuse *et al.*, “Clusters of coronavirus disease in communities, Japan, January–April 2020,” *Emerging Infectious Diseases*, vol. 26, no. 9, pp. 2176–2179, Sep. 2020, doi: 10.3201/eid2609.202272.

[34] D. Kerwin and R. Warren, “US foreign-born workers in the global pandemic: essential and marginalized,” *Journal on Migration and Human Security*, vol. 8, no. 3, pp. 282–300, 2020, doi: 10.1177/2331502420952752.

[35] L.-P. Foo, M.-Y. Chin, K.-L. Tan, and K.-T. Phuah, “The impact of COVID-19 on tourism industry in Malaysia,” *Current Issues in Tourism*, vol. 24, no. 19, pp. 2735–2739, Oct. 2021, doi: 10.1080/13683500.2020.1777951.

[36] J. P. Escalera-Antezana *et al.*, “Clinical features of the first cases and a cluster of Coronavirus Disease 2019 (COVID-19) in Bolivia imported from Italy and Spain,” *Travel Medicine and Infectious Disease*, vol. 35, p. 101653, May 2020, doi: 10.1016/j.tmaid.2020.101653.

[37] E. Livingston and K. Bucher, “Coronavirus disease 2019 (COVID-19) in Italy,” *JAMA*, vol. 323, no. 14, p. 1335, Apr. 2020, doi: 10.1001/jama.2020.4344.
[38] J. Cai, W. Sun, J. Huang, M. Gamber, J. Wu, and G. He, “Indirect virus transmission in cluster of COVID-19 Cases, Wenzhou, China, 2020,” *Emerging Infectious Diseases*, vol. 26, no. 6, pp. 1343–1345, Jun. 2020, doi: 10.3201/eid2606.200412.
[39] K. Danis *et al.*, “Cluster of coronavirus disease 2019 (COVID-19) in the French Alps, February 2020,” *Clinical Infectious Diseases*, vol. 71, no. 15, pp. 825–832, Jul. 2020, doi: 10.1093/cid/ciaa424.
[40] S. Jang, S. H. Han, and J.-Y. Rhee, “Cluster of coronavirus disease associated with fitness dance classes, South Korea,” *Emerging Infectious Diseases*, vol. 26, no. 8, pp. 1917–1920, Aug. 2020, doi: 10.3201/eid2608.200633.

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, data analytics especially in the digital humanities. He can be contacted at email: wirach@kku.ac.th.