ICIC Express Letters
Part B: Applications
Volume **13**, Number **11**, November **2022**

ICIC International ©2022 ISSN 2185-2766
pp. 1133–1141

A COMBINATION OF K-MEANS AND DBSCAN CUSTOMER SEGMENTATION IN B2B BUSINESS: A CASE STUDY IN ELECTRICAL AND MECHANICAL PARTS INDUSTRIES

Voravadhana Asavaphongsavanich, Sotarat Thammaboosadee* and Rojjalak Chuckpaiwong

Information Technology Management Division
Faculty of Engineering
Mahidol University
25/25 Phuttamonthon 4 Road, Salaya, Phuttamonthon, Nakhon Pathom 73170, Thailand
voravadhana.asa@student.mahidol.ac.th; rojjalak.chu@mahidol.ac.th
*Corresponding author: sotarat.tha@mahidol.ac.th

Received March 2022; accepted May 2022

Abstract. *An industry sector is important for the economic growth in Thailand. Among those industries, the electrical and mechanical parts manufacturers are also essential to drive the production process in the factory. Due to the foundation activity supporting, the industrial part manufacturer has become more competitive. The business report in 2019 stated the lost customers and open status of quotations are increasing dramatically. In order to solve and further prevent these problems and gain more competitive advantage, the data mining technique would be necessary to descriptively understand and predict customer behavior which can improve the business strategy to be more effective, which the return-of-investment of the simulated business scenario will prove. The data used in this paper is customer data between 2017 and 2020 in two entities: 1) customer characteristic data, including registered capital, industry code, business type, business size value, and 2) customer transaction data, including purchase history. The combination of descriptive segmentation and predictive modeling toward decision-making strategies that tend to increase the return-of-investment of the industries is challenging, and the main contribution is specified in electrical and mechanical parts manufacturing. The expected results should support the Sales and Marketing team in increasing sales value and new customers and maintaining existing customers by offering highly accurate strategy segmentation.*

Keywords: Data mining, Customer behavior, Strategy selection segmentation, Industrial part manufacturer, k-Means clustering, DBSCAN clustering

1. **Introduction.** To enable Thailand into a high-income country, the industrial sector is one of the factors affecting the country's GDP growth. The Thai government announced the 20-year Thai industrial development strategy (2017-2036), which purposed to upgrade the Thai industry currently in the Industrial 3.0 to become Industrial 4.0 with high technology systems to drive the Thai industry to achieve the needs and changes of innovations in the world. Manufacturers of industrial parts are essential to help the industrial sector develop Thai production technology to be more efficient [1].

For this reason, the electrical and mechanical parts manufacturers become more competitive because having more market shares is an opportunity to increase business growth. Therefore, the business sector focuses on reaching new customers, increasing brand awareness, and creating new customers while maintaining existing customers. Many factors affect the customer's decision, such as the complexity of the purchasing process, the relationship between the buyer and seller, industry type, business size, nationality, registered capital, and the buyer's technology. Referring to the sales report in 2019, the number

DOI: 10.24507/icicelb.13.11.1133

of customers who have bought products without any repeat purchases and the number of quotations that have not been considered are increasing dramatically. Customer data analysis using data mining is vital to understand customer buying behavior in each industry group and predict further purchases of customers in each industry group [2].

Many firms use data analytics to support their decision-making for a more significant outcome. Data analytics also increase the effectiveness of customer relationship management. There are several algorithms used for clustering in data analytics. k-Means is a well-known algorithm for unsupervised machine learning models [3] based on center-based clustering [4] and categorized into partition clustering methods. Another commonly used algorithm for clustering is DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [3]. It is based on density-based development by epsilon (*eps*) and minimum points (*MinPts*). A cluster is a set of data spread over high-density objects, and objects in low density are typically considered noise or outliers [5]. An example of the clustering results of both algorithms is shown in Figure 1. Comparing the performance of the DBSCAN algorithm with a proven segmentation algorithm that utilizes k-Means clustering demonstrated that the DBSCAN algorithm had a higher sensitivity and correctly segmented more swallows. Comparing its performance with a threshold-based algorithm that utilized the quadratic variation of the signal showed that the DBSCAN algorithm offered no direct increase in performance. However, it offered several other benefits, including a faster run time and more consistent performance between patients. All algorithms showed noticeable differentiation from the endpoints [6].

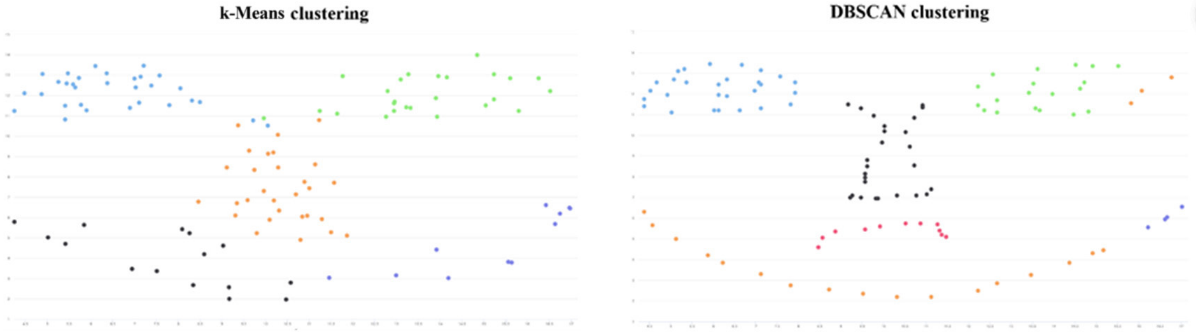


FIGURE 1. Difference between k-Means and DBSCAN clustering

Therefore, the different viewpoints of those two clustering methods are challenging in terms of how to obtain the benefit from the combination of those methods. This paper proposes a heuristic approach to merge the clustering results based on business interpretation of each cluster from each method. The results obtained from k-Means and DBSCAN are reunited into a new cluster that combines them leading to better understand customer behavior in the industrial segment. The outcome of this paper comes out with business action for lost customer, new customer, and loyalty customer to improve marketing strategies for a case study company.

The rest of the paper is organized as follows: the literature review to summarize the research that relates to the clustering and clustering combination in Section 2, the methodology to describe step-by-step of the research in Section 3, the results to show the clustering output and combination outcome in Section 4, and conclusions and further works in Section 5.

2. Literature Review.

Chen et al. [7] mentioned that most previous studies on customer churn prediction have focused on the banking, retail, and telecommunication industries focused on B2C customers. The researchers applied four data mining techniques, including logistic regression (LGR), decision tree analysis (C4.5), artificial neural network (multilayer perceptron, MLP), and support vector machine (SVM). The dataset contained 69,170 business customers in the largest logistic company in Taiwan with 18 variables in three general categories: customer profile, transaction behavior, and quality of delivery between March 2010 and August 2012. The result shows that decision tree analysis (C4.5) is the best algorithm for this research, with a 93.1% accuracy and 93.3% F1 score. The result also identified that the top three variables influenced were recency, length, and monetary, respectively.

Alizadeh and Minaei-Bidgoli [8] proposed evaluating customer loyalty in the banking industry by comparing performance from different algorithms. The predictive model was applied based on demographic variables of customers using various classification methods such as decision tree, artificial neural networks, Naïve Bayes, k-nearest neighbors, and support vector machine. K-medoids, X-means, and k-Means were used for clustering and evaluation by the Davis-Bouldin index. The result showed that artificial neural networks have the most accuracy in predicting loyal customer in the banking industry.

Shirazi et al. [9] developed a customer churn prediction model for the financial sector in Canada due to the threats and disruptions from competitors by using customer data with assets. The model returned four types of customers: churned, potential clients, second potential client segment, and retained clients. There are several results from this study for developing a new strategy that benefits the financial sector to maintain and increase the number of clients. The results showed that clients with long-term relationships or high incomes are less likely to churn.

Monalisa and Kurnia [10] proposed clustering customer behavior from the retail business in Indonesia using the combination of k-Means and DBSCAN applied to the Recency-Frequency-Monetary (RFM) model. The result can be categorized into four customer characteristics: 1) Lost customer, a customer with high recency value and low monetary and frequency value; 2) New customer, a customer with low recency, frequency, and monetary value; 3) Prospect customers with low recency and monetary and high-frequency values; 4) Loyal customer, a customer with a low recency value and high monetary value.

Zhang et al. [11] proposed another combination approach of k-Means and DBSCAN. They provide the application of electric power consumer behavior segmentation in China by using k-Means and then applying DBSCAN to identifying outliers in each cluster. This approach is developed since the k-Means algorithm has a problem where it groups the instances into the optimum clusters in existing clusters, even if the correlation is not great. The result of this complimentary method can analyze consumer behavior in more details compared with the traditional approach.

After reviewing the previous study on customer segmentation, the researchers found that most studies are focused on B2C business, and customer segmentation can result in lost customers, new customers, prospect customers, and loyal customers. Moreover, the combination of center-based and density-based clustering can improve customer segmentation results. Finally, the result from the clustering combination can be used to develop and improve marketing strategies for a case study company.

3. Methodology.

This paper proposed to analyze customer data to understand customer behavior by customer transaction data combined with customer characteristics based on five processes as shown in Figure 2 consisting of 1) define business goal, 2) data collection, 3) data transformation, 4) data modeling, and 5) deployment.

3.1. Business goal defining.

This process proposed defining the business objective and problem statement to align with company goals by setting up a meeting with the Sales and Marketing team. Referring to the sales report for 2019-2020, the number of customers who have bought products without any repeat purchases and the number of quotations that have not been considered are increasing dramatically. Data mining could be another

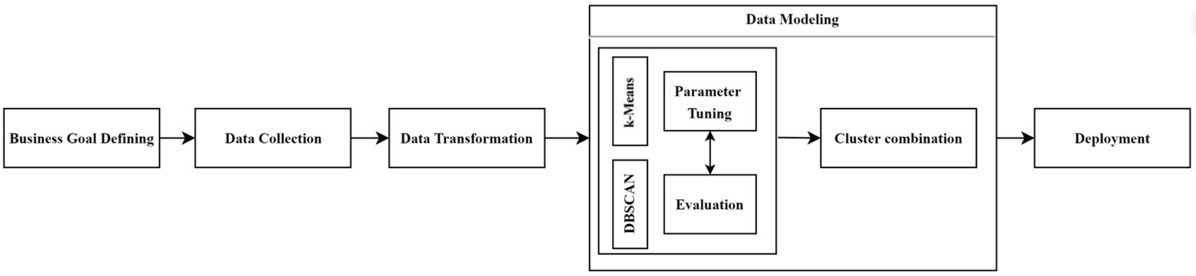


FIGURE 2. Research methodology

tool for the Sales and Marketing team to maintain customer retention before noticing when customers are likely to churn and increase sales in the future.

3.2. **Data collection.** After understanding the business objective, the following process is to create a data list and define the structure of the data collected from different sources. This process helps us understand which data we have and which data we could collect more to achieve the purpose of this research. This paper collected two types of data: customer transactions from a case study company specializing in movement cable and plastic bearing and customer characteristics from the government agency, which is the Department of Business Development, Ministry of Commerce, between 2017-2020. Data collected in this paper does not include private person customer and personal data.

3.3. **Data transformation.** This step is to convert the data collected into data that can be analyzed in the next step. Converting this data may need to do data cleaning because all data may not be of good quality [12], such as removing ordinary customers, converting item code to the product name, and generating new attributes such as Recency, Frequency and Monetary. Recency, (R) can be described as the length of time from the last purchased date, and Frequency (F) describes the average number of transactions in the period. Monetary (M) describes the amount of purchased value [13].

3.4. **Data modeling.** This step will be the data analysis process using the suggested data analysis techniques to analyze customer segment and customer behavior. This paper used a clustering model including k-Means and DBSCAN to find the segment of customers in the case study company. This paper increases the performance rate for each model by doing parameter tuning and measuring the performance of each model.

k-Means process can be described into five steps: 1) Define the number of clusters by user, 2) Select the initial centroid randomly according to the number of clusters, 3) Calculate the distance of data to the nearest centroid, 4) Calculate the new centroid again, and 5) Repeat until there are no changes to the medoid so that clusters and cluster members are obtained. In contrast, DBSCAN in this paper is setting epsilon (eps) or a number of radiuses from neighborhood to core point and minimum points (MinPts) or minimum data point to define the core point of the cluster [14].

This paper used the elbow method to determine the best number of k-Means clusters. The elbow method focused on the percentage of the comparison between the number of clusters that will form an elbow at a point. The best number of clusters can be found when the value of the first and second clusters gives the angle on the graph or the value decreases dramatically [15].

This paper purposed to increase the performance of DBSCAN by the knee method. This method determines k-nearest neighbor distances. The process is to calculate the average distance from every point. The value of k will correspond to minimum points. After that k-distances will be plotted to determine the "knee" corresponding to epsilon.

3.5. **Deployment.** Data modeling in this research was applied in the case study company to analyze purchasing behavior so that the Sales and Marketing team can prepare strategies to retain customers and tackle customers who have a chance of being lost.

4. **Results.** This paper proposed to use registered capital, business size value, industry code, business type, payment, Recency, Frequency, and Monetary as an attribute to cluster customer data by using k-Means and DBSCAN. Customer data in this paper detected an outlier using local outlier factor (LOF), an algorithm to calculate outlier based on density between each data point and its neighbor points. The lower the density of the point, the more likely it is to be identified as the outlier [16]. Customer data will be removed if the outlier value is over 2.5. Customer data also used the normalization function to convert different ranges to the same range [17]. In the final step, this paper proposes the combination of k-Means and DBSCAN clusters to summarize customer behavior, as shown in Figure 3.

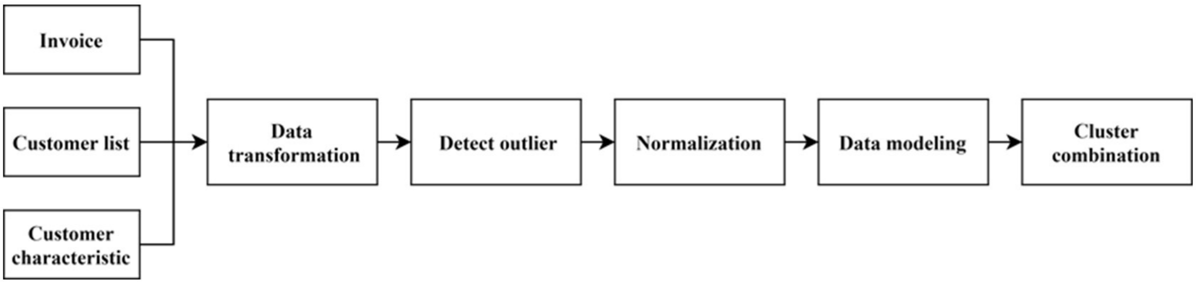


Figure 3. Clustering process

4.1. **k-Means clustering results.** The k-Means clustering results in a centroid value were indicated in numerical. So, this paper proposed to convert a numerical value into text form, as shown in Table 1. Table 1 shows the final result of k-Means clustering after running the model several times and still finding some outliers cluster, which contained only two members. Finally, this paper has 7 clusters left for the k-Means method. Cluster 0 represents a prominent manufacturer company in the rubber and plastic, metalworking, and electronic equipment industry that has been considered a lost customer. Cluster 1 can be represented as a microservice company in metalworking, electronic equipment, and retail industry considered a lost customer. Cluster 2 can be represented as a microservice company in machine maker, retailing, and other industries considered new customers.

Table 1. k-Means clustering meaning

Cluster	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Registered capital	Large	Micro	Micro	Small	Micro	Large	Medium
Business size value	Large	Micro	Micro	Small	Micro	Large	Medium
Industry	RP/MW/EE	MW/EE/RT	MM/RT/OT	RT/PF/OT	MW/MM/MRO	FB/RP/AT	RP/GC/MM
Business type	Manufacturer	Service	Service	Service	Manufacturer	Manufacturer	Manufacturer
Payment	Cash	Cash	Cash	Credit term	Cash	Credit term	Credit term
Recency	High	High	Low	Low	Low	Low	Low
Frequency	Low	Low	Low	Low	Low	Low	High
Monetary	Low	Low	Low	Low	Low	Low	High
Customer type	Lost customer	Lost customer	New customer	New customer	New customer	New customer	Loyalty customer

*RP = Rubber and plastic, MW = Metalworking, EE = Electronic equipment, RT = Retailing, MM = Machine maker, OT = Other, PF = Port facilities, MRO = Maintenance-Repair-Operating supplies, FB = Food and Bev, AT = Automotive, GC = Glass & Ceramic

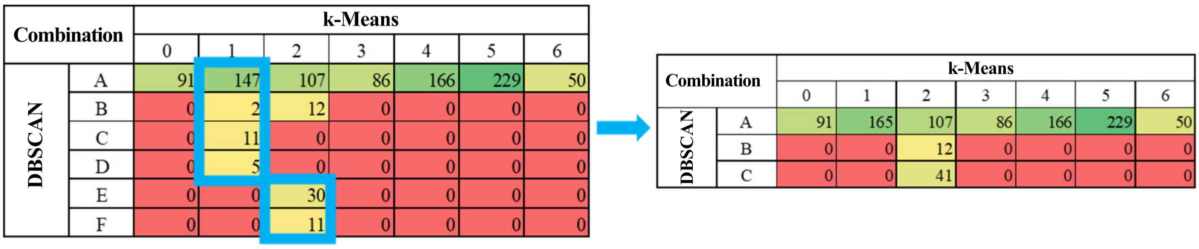
Cluster 3 can be represented as a small service company in retailing, port facilities, and other industries considered new customers. Cluster 4 can be represented as a micro manufacturer company in metalworking, machine maker, and MRO industry that has been considered a new customer. Cluster 5 represents a prominent manufacturer company in the food and beverage, rubber and plastic, and automotive industry that has been considered a new customer. Cluster 6 is a medium manufacturer company in rubber and plastic, glass and ceramic, and machine maker industry that has been considered loyal customers.

4.2. **DBSCAN clustering results.** Table 2 shows the final result of DBSCAN clustering after running the model; some clusters can be merged as the distance is very close, and there is no significant meaning to separate the cluster. Finally, this paper has 6 clusters left for the DBSCAN method. Cluster A could not be identified customer behavior. Cluster B can be represented as a microservice company that has been considered a new customer. Cluster C can be represented as a microservice company considered a lost customer. Clusters D and E can represent a microservice company considered a lost customer. Cluster F represents a small service company considered a lost customer.

Table 2. Summary of customer characteristics by DBSCAN clustering

Cluster	**Cluster A**	**Cluster B**	**Cluster C**	**Cluster D**	**Cluster E**	**Cluster F**
Business size	Others	Micro	Micro	Micro	Micro	Small
Business type	Others	Service	Service	Service	Service	Service
Payment	Others	Cash	Cash	Cash	Cash	Cash
Customer type	Others	New customer	Lost customer	Lost customer	Lost customer	Lost customer

4.3. **Combination of k-Means and DBSCAN.** This paper obtains the results of the k-Means and DBSCAN clustering. The k-Means is an algorithm for clustering based on distance segmentation. The k-Means also contained centroid point and distance to form a new cluster, which is appropriate with a set of data in a circle. The DBSCAN is an algorithm for clustering based on segmentation. Since DBSCAN is working with data density, the DBSCAN comes out with the best result for unshaped data. DBSCAN Cluster A is not able to identify the customer characteristics. Thus, the results obtained from k-Means and DBSCAN are reunited into a new cluster that combines k-Means and DBSCAN. For example, Cluster A0 is a customer group member of Cluster A of DBSCAN and Cluster 0 of k-Means, as shown in Figure 4. The result also found that some k-Means and DBSCAN have similar characteristics. Figure 4 also shows that some customers were categorized in the k-Means and DBSCAN clusters with similar characteristics. For example, Clusters A, B, C and D of DBSCAN and Cluster 1 of k-Means are lost customers who are microservice companies. Clusters E and F of DBSCAN and Cluster 2 of k-Means



Combination		**k-Means**						
		0	1	2	3	4	5	6
DBSCAN	A	91	147	107	86	166	229	50
	B	0	2	12	0	0	0	0
	C	0	11	0	0	0	0	0
	D	0	5	0	0	0	0	0
	E	0	0	30	0	0	0	0
	F	0	0	11	0	0	0	0

Combination		**k-Means**						
		0	1	2	3	4	5	6
DBSCAN	A	91	165	107	86	166	229	50
	B	0	0	12	0	0	0	0
	C	0	0	41	0	0	0	0

Figure 4. Combination of k-Means and DBSCAN clustering

are lost customers who are microservice companies. So, this paper proposed to group a new cluster as shown in Figure 4.

Cluster A0 is a prominent manufacturer in the rubber and plastic, metalworking, electrical and electronic equipment industry that has been considered a lost customer. Cluster A1 is a microservice company in metalworking, electronic equipment, and retail industry considered a lost customer. Cluster A2 is a microservice company in machine maker, retailing, and other industries considered new customers. Cluster A3 can represent a small service company in retailing, port facilities, and other industries considered new customers. Cluster A4 is a micro manufacturer company in the metalworking, machine maker, and MRO industry that has been considered a new customer. Cluster A5 is a prominent manufacturer in the food and beverage, rubber and plastic, and automotive industry that has been considered a new customer. Cluster B2 is a microservice company in machine maker, retailing, and other industries considered new customers. Cluster C2

Table 3. Business action for each cluster

Customer type	**Customer group**	**Action**
Lost customer	Large manufacturer with a cash payment (Cluster A0)	The case study company may contact the customer and collect why they are not continuing to purchase any products because most large businesses require credit term payment, so Sales and Marketing can prepare the proper action for the customer.
	Microservice in machine maker, SI, machinery/retailing with cash (Cluster C2)	The case study company may contact the customer and collect why they are not continuing to purchase any products and approach the customer with a unique selling point and benefit from the product so that Sales and Marketing can prepare the proper action for the customer.
	Microservice in metalworking/electrical & electronic/retailing with cash (Cluster A1)	The case study company may contact the customer and collect why they are not continuing to purchase any products and approach the customer with a unique selling point and benefit from the product so that Sales and Marketing can prepare the proper action for the customer.
New customer	Large manufacturer with credit term (Cluster A5)	Dig deeper and find a new opportunity for cross-selling. Offer new products and innovations related to their industry.
	Small service with credit term (Cluster A3)	Offer a free inspection package to customers, as this cluster contains the port industry.
	Micro manufacturer with a cash payment (Cluster A4)	Dig deeper and find a new opportunity for cross-selling.
	Microservice with cash payment (Cluster A2)	Offer designing support to the customer. Offer new products and innovations related to their industry.
Loyalty customer	Medium manufacturer with credit term (Cluster A6)	The case study company may follow up and keep in contact with the customer to maintain customer relationships. The case study company may set up seminar sessions about innovations to keep them updated.

is a microservice company in machine maker, retailing, and other industries considered lost customers. Cluster A6 is a medium manufacturer in rubber and plastic, glass and ceramic, and machine maker industry that has been considered loyal customers.

4.4. **Business action.** The result from clustering showed three types of customers: loyalty customers, lost customers, and new customers. These three customer types need to act differently from the case study company, as shown in Table 3.

5. **Conclusions.** Since the Thai government announced the 20-year Thai industrial development strategy (2017-2036), it proposed upgrading the Thai industry from Industrial 3.0 to Industrial 4.0 to increase production efficiency in the factory. So, electrical and mechanical parts manufacturers have become more competitive in the market because every brand focuses on creating new customers and maintaining existing customers. Referring to the sales report in 2019, the number of customers who have bought products without any repeat purchases and the number of quotations that have not been considered are increasing dramatically. To solve this problem, the researchers applied k-Means and DBSCAN clustering algorithms to segmenting purchasing behavior in the case study company. This paper collected two types of data: customer transactions and customer characteristics between 2017 and 2020. The result from the combination of k-Means and DBSCAN can be identified by three types of customers, i.e., loyalty customers, lost customers, and new customers. This result also gives a better understanding of the customer character of each segment. In addition, the Sales and Marketing team in the case study company can also develop and improve their strategies to handle their customers.

This paper collected customer transactions from a case study company between 2017-2020 which included the pandemic of the COVID-19 period. So, customer transactions may differ from the typical situation due to their purchase policy and can affect the clustering result. This research can evaluate the satisfaction and possibilities of applying recommended business strategies with the Sales and Marketing team in the case study company. Alternatively, this research can continue to evaluate the return of investment (ROI) after implementing recommended business strategies. However, the results are different since this research focuses on the B2B business. So, this paper proposed combining similar clusters to minimize the number of clusters. In the future, work may not need to do clustering combinations if another research focuses on other businesses, which has the same result.

REFERENCES

[1] Ministry of Industry, *Thai Industrial Development Strategy (2017-2036) in Thai*, Accessed on 12 August 2021.

[2] F. Safari, N. Safari and G. A. Montazer, Customer lifetime value determination based on RFM model, *Mark. Intell. Plan.*, vol.34, no.4, pp.446-461, 2016.

[3] E. M. Cherrat, R. Alaoui and H. Bouzahir, Improving of fingerprint segmentation images based on k-Means and DBSCAN clustering, *Int. J. Electr. Comput. Eng.*, vol.9, no.4, pp.2425-2432, 2019.

[4] R. Scitovski and K. Sabo, A combination of k-Means and DBSCAN algorithm for solving the multiple generalized circle detection problem, *Adv. Data Anal. Classif.*, vol.15, no.1, pp.83-98, 2021.

[5] X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao and J. Z. Huang, PurTreeClust: A clustering algorithm for customer segmentation from massive customer transaction data, *IEEE Trans. Knowl. Data Eng.*, vol.30, no.3, pp.559-572, 2018.

[6] J. M. Dudik, A. Kurosu, J. L. Coyle and E. Sejdi, A comparative analysis of DBSCAN, k-Means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals, *Comput. Biol. Med.*, vol.59, pp.10-18, 2015.

[7] K. Chen, Y. H. Hu and Y. C. Hsieh, Predicting customer churn from valuable B2B customers in the logistics industry: A case study, *Inf. Syst. E-Bus. Manag.*, vol.13, no.3, pp.475-494, 2015.

[8] H. Alizadeh and B. Minaei-Bidgoli, Introducing a hybrid data mining model to evaluate customer loyalty, *Eng. Technol. Appl. Sci. Res.*, vol.6, no.6, pp.1235-1240, 2016.

[9] F. Shirazi, M. Mohammadi, T. Rogers and I. Technology, A big data analytics model for customer churn prediction in the retiree segment, *Int. J. Inf. Manage.*, vol.48, no.2, pp.238-253, 2019.

[10] S. Monalisa and F. Kurnia, Analysis of DBSCAN and k-Means algorithm for evaluating outlier on RFM model of customer behaviour, *Telkomnika (Telecommunication Comput. Electron. Control)*, vol.17, no.1, pp.110-117, 2019.

[11] L. Zhang, S. Deng and S. Li, Analysis of power consumer behavior based on the complementation of k-Means and DBSCAN, *2017 IEEE Conf. Energy Internet Energy Syst. Integr.*, no.1, pp.2-6, 2017.

[12] R. Rajamohamed and J. Manokaran, Improved credit card churn prediction based on rough clustering and supervised learning techniques, *Cluster Comput.*, vol.21, no.1, pp.65-77, 2018.

[13] I. Karacan, I. Erdogan and U. Cebeci, A comprehensive integration of RFM analysis, cluster analysis, and classification for B2B customer relationship management, *Proc. of Int. Conf. Ind. Eng. Oper. Manag.*, pp.497-508, 2021.

[14] X. Han, C. Armenakis and M. Jadidi, DBSCAN optimization for improving marine trajectory clustering and anomaly detection, *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. – ISPRS Arch.*, vol.43, no.B4, pp.455-461, 2020.

[15] R. Nainggolan, R. Perangin-Angin, E. Simarmata and A. F. Tarigan, Improved the performance of the k-Means cluster using the sum of squared error (SSE) optimized by using the elbow method, *J. Phys. Conf. Ser.*, vol.1361, no.1, 2019.

[16] Z. Cheng, C. Zou and J. Dong, Outlier detection using isolation forest and local outlier, *Proc. of 2019 Res. Adapt. Converg. Syst. (RACS2019)*, pp.161-168, 2019.

[17] R. W. S. Brahmana, F. A. Mohammed and K. Chairuang, Customer segmentation based on RFM model using k-Means, K-Medoids, and DBSCAN methods, *Lontar Komput. J. Ilm. Teknol. Inf.*, vol.11, no.1, p.32, 2020.