Jurnal Teknik Informatika C.I.T Medicom 14 (1) (2022) 01-09

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www. medikom.iocspublisher.org

Optimization of Determination Against K-Means Cluster Algorithm using Elbow Creation

Melda Pita Uli Sitompul[1*], Opim Salim Sitompul[2], Zakarias Situmorang[3]

[1,2,3]Master of Informatics Engineering, University of North Sumatra, Jl. Dr. T. Mansur No.9, Medan, Sumatera Utara 20222 Indonesia

Email: meldasitompul19@gmail.com[1], opim@usu.ac.id[2], zakarias65@ust.ac.id[3]
** corresponding author*

ARTICLE INFO

***Article history*:**

Received: Sept 20, 2021;
Revised: Oct 06, 2021;
Accepted: Dec 03, 2021;
Available online: Marc 30, 2022

***Keywords*:**

Clustering, K-Means;
RMSSTD;
R squared;
Elbow.

ABSTRACT

Clustering is a data mining method for grouping data that have similar or different characters in each section. One of the methods is using K-Means by measuring the distance between clusters using the shortest distance or Euclidean Distance. K-means entails weakness, which is the determination of clusters in k-means clustering, resulting in the different data grouping and affecting the results of the data cluster distribution. To overcome this issue, the elbow creation method is employed to determine the similarity level in the cluster by observing the comparison between Root Means Square and R Square to measure the homogeneity and heterogeneity of the cluster where this method is applied by considering the changes in the comparison between the RMSSTD (Root Means Square Standard Deviation) and RS (R Squared) values which have the intersection of the RMSSTD and RSquared values. The difference between RMSSTD cluster 1 and RMSSTD cluster 2 was 0.066 and RS cluster 1 and RS cluster 2 was -0.304. Based on those figures, the highest difference was found in cluster 2. All considered, tourist destinations in East Asia frequently visited or interested to visitors are grouped into cluster 2, comprising criteria 6, 7, 8, and 10, or in other words, resort destination, picnic area, beaches, and religious institutions

1. Introduction

In k-means, the cluster determination process is conducted through the discovery number of clusters, data allocation to the existing clusters, and mean calculation of each cluster that is combined, in which the data are observed from the closest distance between clusters and the process will be repeated until there is no change occurring [1] [2] [3]. An adequate cluster is selected by increasing the cluster value so that when the cluster value is formed, the elbow rule does not result in a very different model from the data [4]. Elbow rules looked after to determine the number of clusters from the dataset is based on the minimized total within-cluster variation or total within-cluster variation of square [5] [6] [8].

To choose a tourist recommendation system with EM and SOM groupings, which are related to the dataset variance factor, the elbow rule chosen is in cluster 6 (six) with an eigenvalue of 0.718623 [7][8] [9]. One of the weaknesses possessed by K-means is the determination of clusters in k-means clustering so that the data grouping is different and will affect the results of the data cluster distribution [10] [11]. To overcome this problem, the elbow was performed to determine the homogeneity level in the cluster by observing the comparison between Root Means Square and R Square to measure the homogeneity and heterogeneity of the cluster, in which inversely proportional situation occurs between the decreasing RMSSTD value in cluster addition and the increasing RS value in cluster addition [12].

Obtaining the best cluster can be carried out by comparing the value of RMSSTD and Root Square which shows overlapping values [13] [14].

2. Research Method

The research method is procedures the researcher will perform for the RMSSTD and Rsquare methods. The research methods are as follows:

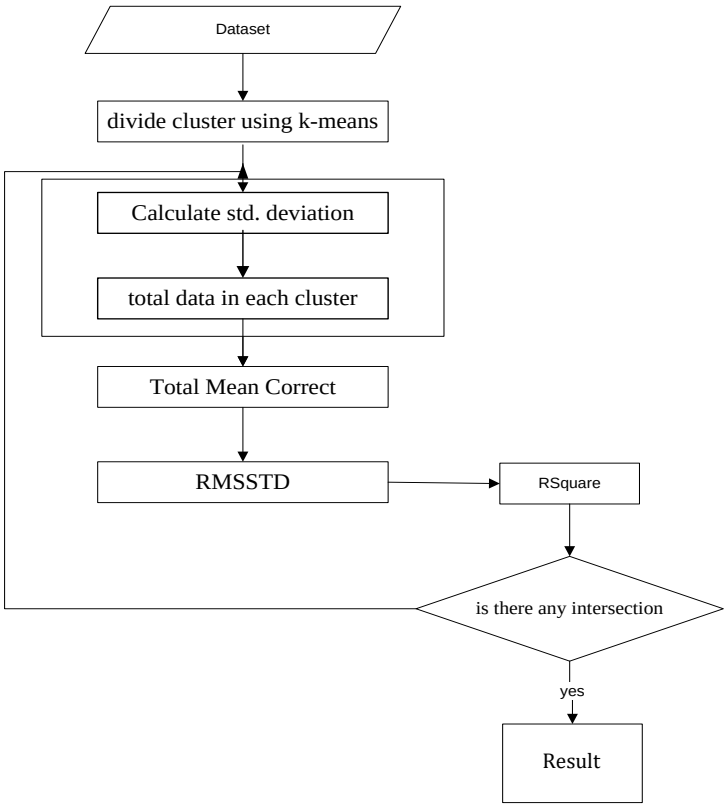

Fig 1. Research Method

2.1 Dataset

For the research, the dataset used is extracted from the University of California Irvine (UCI) Machine Learning on Trip Advisor in East Asia with a total of nine hundred and eighty data and has eleven attributes, comprising attribute 1 user id, attribute 2 feedback from art gallery user, attribute 3 nightclub user, attribute 4 juice bar user, attribute 5 restaurant user, attribute 6 museum, attribute 7 lodging user, attribute 8 picnic area user, attribute 9 beach users, attribute 10 theater user, and attribute 11 religious institution user. The number of trip Advisor data totaling 980 data is presented in Table 1.

Table 1.
Trip Advisor dataset

User ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
User 1	0,93	1,8	2,29	0,62	0,8	2,42	3,19	2,79	1,82	2,42
User 2	1,02	2,2	2,66	0,64	1,42	3,18	3,21	2,63	1,86	2,32
User 3	1,22	0,8	0,54	0,53	0,24	1,54	3,18	2,8	1,31	2,5
User 4	0,45	1,8	0,29	0,57	0,46	1,52	3,18	2,96	1,57	2,86
...	...	...	...	...	...	...	...	...	...	...
User 979	0,93	0,2	0,13	0,43	0,3	0,4	3,18	2,98	1,12	2,46
User 980	0,93	0,56	1,13	0,51	1,34	2,36	3,18	2,87	1,34	2,4

2.2 Cluster Division

In this section, before the determining number of clusters, the first calculation step is to find the distance using Euclidean distance and Complete Linkage [13] [14].
Formula:

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

$$d(1,2) = 1 - 2 \| = \sqrt{(0{,}93 - 1{,}02)^2 + (1{,}80 - 2{,}20)^2 \ldots + (2{,}40 - 2{,}30)^2}$$
$$= 1{,}142$$

the calculation continued until d (980.980) which will be displayed in Table 2.

Table 2.
Euclidean Distance Matrix

	Euclidean Distance									
User	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	1,14	1,13	1,06	1,00	0,99	0,78	0,19	0,19	0,10	0,10
3	2,34	2,32	2,10	1,16	1,16	1,02	0,51	0,51	0,51	0,08
4	2,33	2,28	2,28	1,10	1,10	1,04	0,53	0,53	0,50	0,44
	...	...	...	...	...	...	...	...	...	...
980	1,85	1,85	1,37	0,73	0,72	0,49	0,48	0,48	0,48	0,02

Based on the distance matrix that had been performed, the cluster results can be seen in the dendrogram graphic in Figure 2.

```
# Dendrogram
plot(result)
```

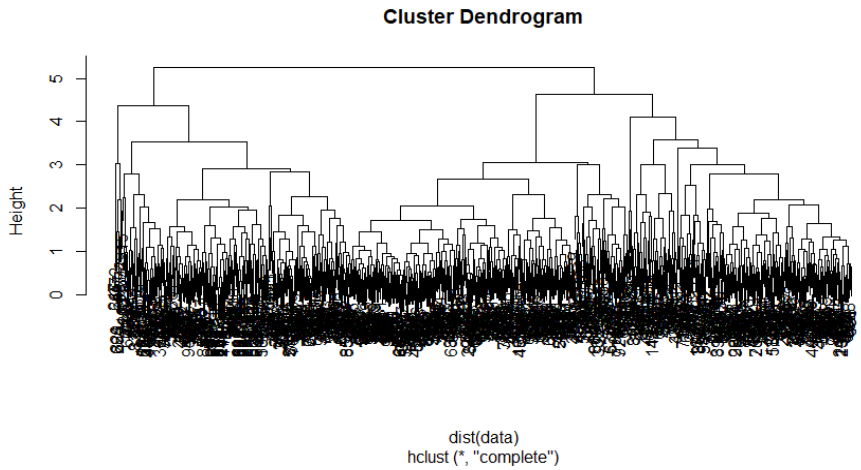

Fig 2. Dendrogram cluster

From the results of the dendogram in Fig 2, it was found cluster division, including: 304 cluster 1, 368 cluster 2, 292 cluster 3, 6 cluster 4, 10 cluster 5. Thus, the results of cluster distribution are described in Table 3.

Table 3.
Results of Cluster Distribution

User ID	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	cluster
1	0,93	1,8	2,29	0,62	0,8	2,42	3,19	2,79	1,82	2,42	1
2	1,02	2,2	2,66	0,64	1,42	3,18	3,21	2,63	1,86	2,32	2
3	1,22	0,8	0,54	0,53	0,24	1,54	3,18	2,8	1,31	2,5	2
4	0,45	1,8	0,29	0,57	0,46	1,52	3,18	2,96	1,57	2,86	1
..	..	..	..	..	..	..	..	..	..	..	..

User ID	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	cluster
979	0,93	0,2	0,13	0,43	0,3	0,4	3,18	2,98	1,12	2,46	1
980	0,93	0,56	1,13	0,51	1,34	2,36	3,18	2,87	1,34	2,4	1
min	**0,34**	**0,00**	**0,13**	**0,15**	**0,06**	**0,14**	**3,16**	**2,42**	**0,74**	**2,14**	
max	**3,22**	**3,64**	**3,62**	**3,44**	**3,3**	**3,76**	**3,21**	**3,39**	**3,17**	**3,66**	
avg	**0,8932**	**1,3526**	**1,0133**	**0,5325**	**0,9397**	**1,8429**	**3,1809**	**2,8351**	**1,5694**	**2,7992**	
median	**0,83**	**1,28**	**0,82**	**0,5**	**0,9**	**1,8**	**3,18**	**2,82**	**1,54**	**2,78**	

2.3 Standard Deviation

The standard deviation was used to find the deviation of each data point subtracted by mean To calculate the standard deviation, the first step is to calculate the variance of each variable. In this step, the subtraction of the mean in each sample should be achieved [15].

Sample variance = (x⟧ _(i-) x_j)⟧ ^2

1st data for x_1

= (0.93 - 0.872)2

= 0.0032, The calculation continued from x_2 to x_{10} which can be seen in Table 4.

Table 4.
Result of sample variance for cluster 1

data	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	0,003	0,236	1,5871	0,007	0,010	0,389	7,829	0,0031	0,0704	0,1486
2	0,1786	0,236	0,547	0,001	0,194	0,075	1,325	0,0130	0,0002	0,0029
3	0,029	0,002	0,656	0,074	0,358	0,0652	0,0001	0,0006	0,46972	0,0988
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
304	0,003	0,568	0,009	0,0004	0,192	0,318	1,3255	0,0005	0,0460	0,1645
total	**33,252**	**69,553**	**206,295**	**28,864**	**48,68**	**88,43912**	**0,018**	**5,186**	**35,615**	**31,182**

Variance = $\left(\frac{\sum_{j=1..d}^{i=1..nc}\sum_{k=1}^{nij}(xk-xj)^2}{(n-1)*2}\right)$

$$= \frac{33,252 + 69,553 + 206,295 + 28,864 + 48,689 + 88,439 + 0,0188 + 5,1864 + 35,615 + 31,182}{(304-1)*2}$$

= 0,1772

St. deviasi (RMSSTD)= $\sqrt{\frac{\sum_{j=1..d}^{i=1..nc}\sum_{k=1}^{nij}(xk-xj)^2}{(n-1)*2}}$

$= \sqrt{0,1772}$

= 0,421

2.4 Elbow Creation

Elbow creation governs the number of clusters based on a change in the ratio between the values of RMSSTD (Root Means Square Standard Deviation) and RS (R Squared)[16].

2.5 Total Mean Correct

The mean correct number is applied to measure the validation of the total standard deviation value.

Σ mean correct =

Total variance cluster 1 + Total variance cluster 2 + Total variance cluster 3 + Total variance cluster 4 + Total variance cluster 5 = 547,089 + 629,428 + 521,010 + 12,071 + 18,581

= 1728,181.

2.6 Root Means Square Standard Deviation (RMSSTD).

RMSSTD (Root Means Square Standard Deviation) is a tool to measure the homogeneity level of the data contained within clusters found. The lower the RMSSTD value, the more homogenous the data in the cluster are found [17] [18] [19].

$$RMSSTD = \sqrt{\frac{\sum_{j=1..d}^{i=1..nc}\sum_{k=1}^{nij}(xk-\bar{x}j)^2}{\sum_{j=1..d}^{i=1..nc} nij-1)}} \quad (2)$$

Based on the results of the st. deviation calculation above, then the amount of RMSSTD is presented in Table 5.

Table 5.
Results of RMSSTD Value

Cluster	RMSSSTD
1	0,421
2	0,347
3	0,328
4	0,313
5	0,302

2.7 Root Square (RS).

The next step is to calculate the RS used to measure the homogeneity and heterogeneity between clusters.

The next step is to calculate the Root Square to measure the homogeneity and heterogeneity between clusters.

RS contains a value between 0 and 1. The value of 0 is for the same cluster and 1 for a completely different cluster. RS was calculated using the following formula [20].

$$S = \frac{\left\{\sum_{j=1..v}^{d}\left[\sum_{k=1}^{n_j}(x_k-\overline{x_j})^2\right]\right\}-\left\{\sum_{j=1..v}^{i=1..c}\left[\sum_{k=1}^{n_{ij}}(x_k-\overline{x_j}\right]^2\right\}}{\sum_{j=1..v}\left[\sum_{k=1}^{n_j}(x_k-\overline{x_j})^2\right]} \quad (3)$$

For 1 cluster

RS $= \frac{1728{,}181-1728{,}1181}{1728{,}181}$

$= 0$

For 2 clusters

RS $= \frac{1728{,}181-1176{,}517}{1728{,}181}$

$= 0{,}319$

For 3 clusters

RS $= \frac{1728{,}181-1150{,}438}{1728{,}181}$

$= 0{,}334$

For 4 clusters

RS $= \frac{1728{,}181-533{,}081}{1728{,}181}$

$= 0{,}6915$

For 5 clusters

RS $= \frac{1728{,}181-30{,}653}{1728{,}181}$

$= 0{,}982$, Given the results of the above calculations, the number of R square is shown in Table 6.

Table 6.
Results of R Square Value

cluster	RS
1	0
2	0,319
3	0,334
4	0,6915
5	0,982

2.8 Determination of the best cluster.

Determination of the best cluster can be done if it has an intersection point between the RMSSTD and R square values. Otherwise, the process will repeat the 4th (fourth) stage. If there is an intersection point in the cluster, finding the best cluster can proceed [11] [3].

In K-Means modeling, determining the number of cluster can be performed by comparing the values of Root Mean Square Standard Deviation (RMSSTD) and Root Square (RS). If those two values have inversely proportional values that intersect each other, then the cluster is the best. The comparison of values between RMSSTD and R square can be seen in Table 7.

Table 7.
RMSSTD and R Square Comparison Table

Number of Cluster	RMSSTD	R Square
1	0,421	0
2	0,347	0,319
3	0,328	0,334
4	0,313	0,6915
5	0,302	0,982

3. Results and Discussion

In this study, the number of clusters was obtained using R Studio Programming. Researchers observed the graphs if the two values between RMSSTD and R Square had inversely proportional values that intersected in determining the number of clusters.

3.1 Testing

Tests were conducted to understand how this research works. In this stage, researchers divided the data into two; Training Data consisting of 784 data, Evaluation data comprising 196 data. The number of clusters formed is 5 clusters.

a. Training Data

In the training data, 80% of the selected datasets were tested, totaling 784 data, with ten criteria which can be seen in Table 8.

Table 8.
Training Data

User	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	0,93	1,8	2,29	0,62	0,8	2,42	3,19	2,79	1,82	2,42
2	1,02	2,2	2,66	0,64	1,42	3,18	3,21	2,63	1,86	2,32
3	1,22	0,8	0,54	0,53	0,24	1,54	3,18	2,8	1,31	2,5
4	0,45	1,8	0,29	0,57	0,46	1,52	3,18	2,96	1,57	2,86
5	0,51	1,2	1,18	0,57	1,54	2,02	3,18	2,78	1,18	2,54
6	0,99	1,28	0,72	0,27	0,74	1,26	3,17	2,89	1,66	3,66
..	...	...	...	...	...	...	...	...	...	...
980	0,93	0,56	1,13	0,51	1,34	2,36	3,18	2,87	1,34	2,4

Table 9.
Results of RMSSTD and RSquare

Cluster	1	2	3	4	5
RMSSTD	0,426	0,35	0,33	0,317	0,305
RS	0	0,326	0,399	0,447	0,488
Subtraction RMSSTD	0,426	0,076	0,02	0,013	0,012
Subtraction RS	0	-0,326	-0,073	-0,048	-0,041

From the table above, it shows that Root Means Square Standard Deviation (RMSSTD) and RSquare values in cluster 2 experienced the biggest downturn among other clusters, namely RMSSTD of 0.076 and RS of -0.326. The results of the graph intersection between the RMSSTD and R Square values are shown in Figure 3.

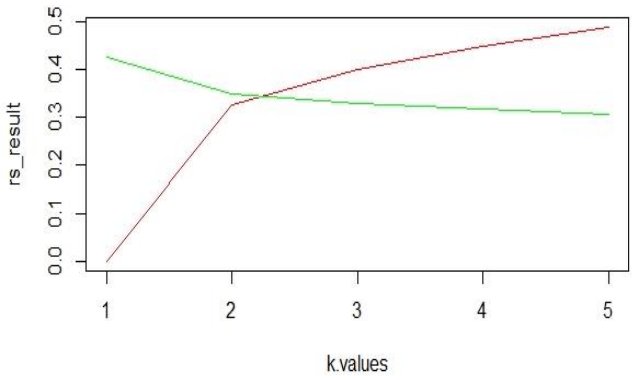

Fig 3. Graph of RMSSTD and R Square values for 784 data

In Figure 3, the graph of the RMSSTD and R Square values shows that the correct cluster can be found in cluster 2 which has an intersection point.

b. Evaluation Data

In the evaluation data, 20% of the dataset were tested, totaling 196 data under ten criteria which is shown in Table 10.

Table 10.
Evaluation Data

User	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
101	0,9	0,96	0,38	0,5	0,78	2,1	3,18	2,9	1,79	2,82
102	0,64	1,12	0,4	0,37	0,64	1,12	3,18	2,94	1,76	3,06
103	0,99	2,12	1,41	0,65	0,96	2,94	3,19	2,82	1,41	2,42
104	0,88	1,2	0,16	0,45	0,5	1	3,18	2,67	1,79	2,86
...	...	...	...	...	...	...	...	...	...	...
445	0,72	0,96	1,57	0,62	1,58	1,86	3,18	2,82	1,25	2,72
446	1,25	0,96	0,43	0,19	0,58	1,06	3,17	2,7	1,18	3,06

3.2 Result

From the results of the table above, it implied in cluster 2 the Root Means Square Standard Deviation (RMSSTD) and RSquare values experienced the most dominant decrease among other clusters, which obtained RMSSTD 0.066 and RS -0.304 as shown in Table 11.

Table 11.
Results of RMSSTD and R Square

Cluster	**1**	**2**	**3**	**4**	**5**
RMSSTD	0,401	0,335	0,312	0,295	0,282
RS	0	0,304	0,398	0,458	0,508
Subtraction RMSSTD	0,401	0,066	0,023	0,017	0,013
Subtraction RS	0	-0,304	-0,094	-0,06	-0,05

3.3 Analysis

From the results of Table 11, it can be determined that cluster 2 experiences a prominent decline compared to the other clusters. Based on the graph, the intersection between the RMSSTD and RSquare values can be observed as presented in Figure 4, in which R-square is red and RMSSTD is green.

```
plot(k.values,rs_result,type="l",col="red")
lines(k.values,rmsstd_result,col="green")
```

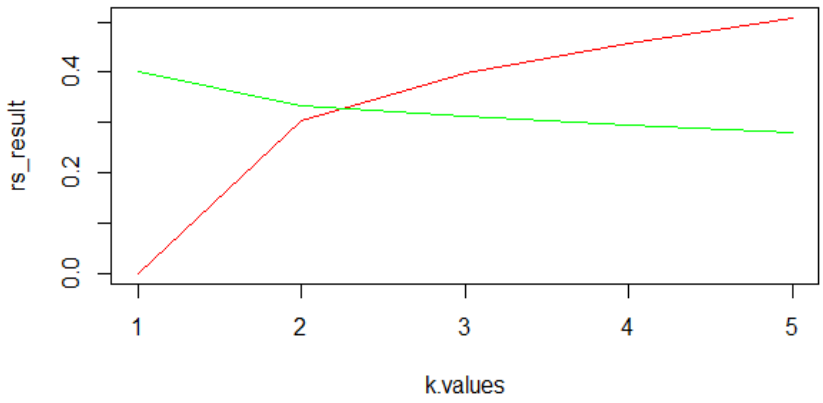

Fig 4. Graph of RMSSTD and RSquare values for 196 data

In Figure 4, the graph of the RMSSTD and RSquare values shows that the correct cluster is in cluster 2 which has an intersection point. Rsquare is red and RMSSTD is green.

4. Conclusion

Based on the research results of the research, the difference between RMSSTD cluster 1 and RMSSTD cluster 2 was 0.066, and RS cluster 1 and RS cluster 2 was -0.304. Hence, these figures imply that the highest difference occurred in cluster 2. Finally, tourist destinations located in East Asia that are frequently visited or interested to visitors are grouped into cluster 2, which are 6, 7, 8, and 10 or resort destination, picnic area, beaches, and religious institutions.

References

[1] A. Agrawal and H. Gupta, "Global K-means (GKM) clustering algorithm: a survey," *Int. J. Comput. Appl.*, vol. 79, no. 2, 2013.

[2] Y. Agusta, "Minimum message length mixture modelling for uncorrelated and correlated continuous data applied to mutual funds classification." Monash University, 2004.

[3] S. B. Sutono, "Selection of representative Kansei adjectives using cluster analysis: a case study on car design," *Int. J. Adv. Eng. Manag. Sci.*, vol. 2, no. 11, p. 239691, 2016.

[4] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and k-means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, 2014.

[5] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, vol. 6. Pearson London, UK:, 2014.

[6] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J.*, vol. 1, no. 6, pp. 90–95, 2013.

[7] B. Everitt, "Cluster Analysis, 5th edn John Wiley & Sons," *Ltd New York.[Google Sch.*, 2011.

[8] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "Reviews-the elements of statistical learning: data mining, inference and prediction," *Math. Intell.*, vol. 27, no. 2, pp. 83–84, 2005.

[9] M. Nilashi, K. Bagherifard, M. Rahmani, and V. Rafe, "A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques," *Comput. Ind. Eng.*, vol. 109, pp. 357–368, 2017.

[10] B. R. Jipkate and V. V Gohokar, "A comparative analysis of fuzzy c-means clustering and k means clustering algorithms," *Int. J. Comput. Eng. Res.*, vol. 2, no. 3, pp. 737–739, 2012.

[11] V. K. Panchal, H. Kundra, and J. Kaur, "Comparative study of particle swarm optimization based unsupervised clustering techniques," *Int. J. Comput. Sci. Netw. Secur.*, vol. 9, no. 10, pp. 132–140, 2009.

[12] A. Singh, A. Yadav, and A. Rana, "K-means with Three different Distance Metrics," *Int. J. Comput. Appl.*, vol. 67, no. 10, pp. 13–17, 2013, doi: 10.5120/11430-6785.

[13] T. S. Madhulatha, "An overview on clustering methods," *arXiv Prepr. arXiv1205.1117*, 2012.

[14] S. Renjith, A. Sreekumar, and M. Jathavedan, "Evaluation of partitioning clustering algorithms for processing social media data in tourism domain," in *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 2018, pp. 127–131.

[15] J. V. De Oliveira and W. Pedrycz, *Advances in fuzzy clustering and its applications*. John Wiley & Sons, 2007.

[16] A. Bhagat, *Mobile intensive care unit relocation modeling using cluster analysis and linear optimization*. State University of New York at Binghamton, 2009.
[17] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2, pp. 107–145, 2001.
[18] W. Niyagas, A. Srivihok, and S. Kitisin, "Clustering e-banking customer using data mining and marketing segmentation," *ECTI Trans. Comput. Inf. Technol.*, vol. 2, no. 1, pp. 63–69, 2006.
[19] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Clustering algorithms and validity measures," in *Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001*, 2001, pp. 3–22.
[20] W. Yotsawat and A. Srivihok, "Rules mining based on clustering of inbound tourists in Thailand," in *Advanced Computer and Communication Engineering Technology*, Springer, 2015, pp. 693–705.