Basic and Applied Science Conference (BASC) 2021
Volume 2021
http://dx.doi.org/10.11594/nstp.2021.1112

NST PROCEEDINGS

OPEN ACCESS

Conference Paper

Data Mining Using K-Means Clustering Algorithm for Grouping Countries of Origin of Foreign Tourist

Herliyani Hasanah*, Nugroho Arif Sudibyo, Rhezka Mahendra Galih

Department Informatics Engineering, Duta Bangsa University

**Corresponding author:*
E-mail:
herliyani_hasanah@udb.ac.id

ABSTRACT

Indonesia has enormous potential to develop the tourism sector. The role of the tourism sector in Indonesia's economic development is increasingly important. The contribution has been made by the tourism sector through foreign exchange earnings, regional income, regional development, investment, and employment increment as well as business development across various areas in Indonesia. One of the government's targets in the tourism sector is to increase foreign tourist visits. Grouping or clustering the countries of origin of the tourists need to be done to help the government in determining strategies. This study uses the K-means clustering algorithm to classify the data on the country of origin of tourists and evaluate the clusters using silhouette score for determining the appropriate number of clusters. The result of the silhouette score shows that K = 2 has a value of 0.8, which is the best cluster that can be used to classify data on the country of origin of tourists. Based on the test results of the clusters, both of the clusters were then identified as cluster 1 for the category of low visitors with 206 members and cluster 2 for the category of high visitors with 6 members, namely Malaysia, Singapore, China, Other Asia, Timor Leste, and Australia. The results of the clustering process are expected to be input data for further performance, namely mapping the right marketing strategy for the countries visiting Indonesia so as to increase foreign tourist visits to Indonesia.

Keywords: Tourism, foreign countries, K-means clustering, silhouette score

Introduction

Indonesia is a touristic developing country (Wiratama et al., 2014). The role of the national tourism sector is increasingly important in line with the development and contribution made by the tourism sector, foreign exchange earnings, regional development, employment, and business development. The contribution of the tourism sector to the national Gross Domestic Product (GDP) in 2014 has reached 9% or Rp. 946.09 trillion, while foreign exchange from the tourism sector in 2014 has reached Rp. 120 trillion and the contribution to job opportunities is 11 million people (Anggraini, 2017).

Tourism has become one of the largest and the most dynamic industry sectors in the world (Han & Hyun, 2015; Katrakilidis et al., 2017; Novokreshchenova et al., 2016). In this growing industry, the number of people involved globally is ever-increasing. In Indonesia currently, the tourism sector is significantly increasing job opportunities for residents. According to the Minister of Tourism (Yahya, 2015), Indonesian tourism is considered to have advantages in terms of destinations and prices. The number of foreign tourists visiting Indonesia is increasing, which can be seen in the graph in Figure 1:

How to cite:
Hasanah, H., Sudibyo, N. A., & Galih, R. M. (2021). Data mining using k-means clustering algorithm for grouping countries of origin of foreign tourist. *Basic and Applied Science Conference (BASC) 2021*. NST Proceedings. pages 88-94. doi: 10.11594/ nstp.2021.1112

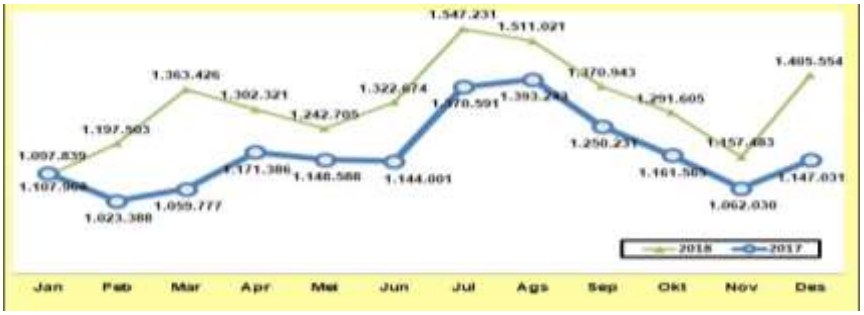

Figure 1. Development of foreign tourists to Indonesia

Figure 1 shows that the development of foreign tourists to Indonesia has a positive growth rate. From February 2017, it showed 1,023,388 people and the same month in 2018 it showed 1,197,503 people until the end of the year, in December 2017, showing 1,147,031 people and in December 2018 with 1,405,554 people, it can be concluded that monthly visits of foreign tourists increased.

From the Indonesian Central Bureau of Statistics, data can be obtained through the website bps.go.id which includes data on the number of foreign tourists visits according to nationality, the data obtained may increase over time. This study used data residing in tourist arrivals in the region of Southeast Asia which includes Brunei Darussalam, Malaysia, Philippines, Singapore, Thailand, Vietnam, Laos, Cambodia, and Myanmar in 2018.

The main elements that must be considered to support tourism development in tourist destinations include three elements, namely infrastructure, accommodation, promotion. According to Novak, promotion is an activity to introduce products, convince and remind the product to the target buyer, probably they will be moved and want to buy the product (Novak, 2011). The purpose of promotion is to inform, influence, and persuade and remind target customers about the company and its marketing mix (Helaey, 2013).

Tourist attractions need promotion to be known throughout the world. The more often the promotion is carried out, the greater the scope of information about tourist attractions is known to people around the world which makes tourists interested in visiting these tourist attractions. Promotion can be done by advertising, sales promotion, public relations, direct marketing, and personal selling which can be used to attract tourists (Karunanithy & Sivesan, 2013). The implementation of tourism planning requires a tourism development strategy to accommodate the roles and tasks of tourism elements by empowering tourism potential related to marketing strategies.

One of the marketing strategies to do an effective promotion is by finding a potential target market. Finding the appropriate target market can be done by grouping the customer (i.e. tourist) by their attribute. By the grouping technique, it is expected that the promotion strategy can be further improved especially for tourists from certain countries.

Data mining is not a new topic in the research, usually used to increase the accuracy of the previous techniques which are expected to solve various problems that are often encountered (Maulana & Fajrin, 2018) Cluster analysis is a multivariate analysis method that functions to group an object (Windarto, 2017). It can be applied to various topics or research themes for example similarity of watersheds (Aytac, 2020), DOS attacks (Iswardani & Riadi, 2016), course grouping (Rustam & Annur, 2019), poverty rates (Sudibyo et al., 2020), and many more. One of the most famous among several clustering methods is K-means clustering. K-means is a cluster analysis that is currently receiving much attention (Ramadani et al., 2019).

Research on clustering has often been done by researchers. As an example similarity of watersheds (Aytac, 2020), DOS attacks (Iswardani & Riadi, 2016), course grouping (Rustam & Annur,

2019), poverty rates [16], and many more. However, research on grouping in tourism is quite often done. For example, research was conducted to classify tourist visits in the province of DJKI Jakarta (Maulida, 2018). Furthermore, the research was conducted on the recommendation of Umbul tourism with K-means clustering (Pamungkas et al., 2020). Next, Research was conducted to determine superior destinations (Seimahuira, 2021).

Material and Methods

This study implements the K-Means algorithm for data clustering of foreign tourist arrivals to Indonesia in publications. The stages of the research are shown in Figure 2.

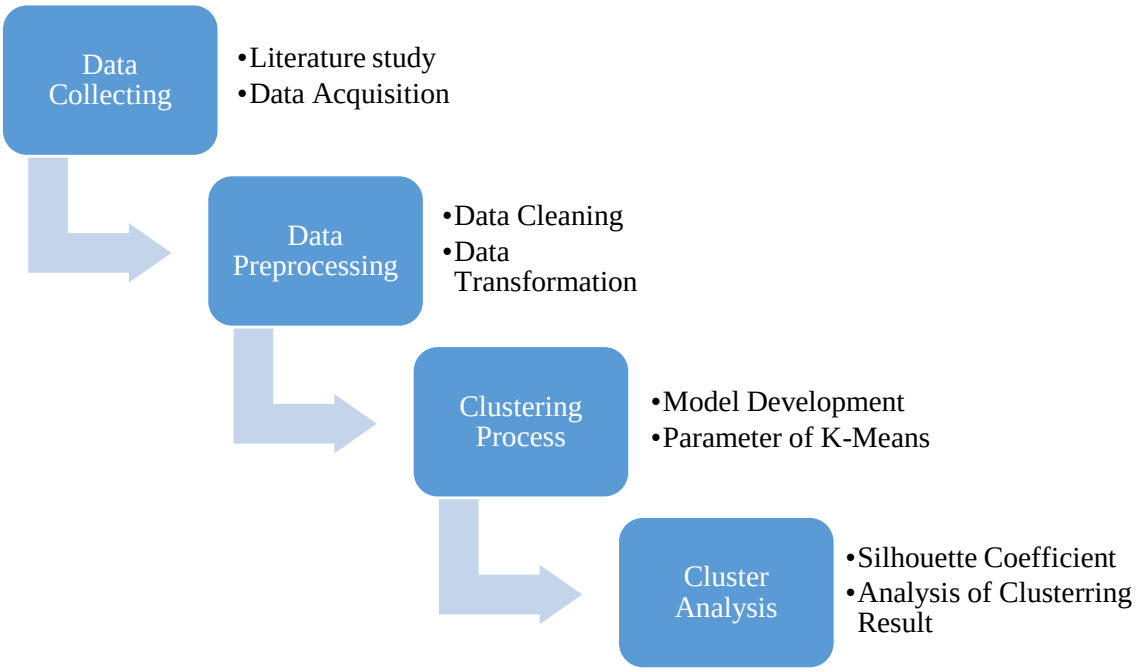

Figure 2. The stages of research

Data collecting

Data is provided by the Indonesian Government Central Bureau of Statistics, which can be accessed through the website https://bps.go.id. The data used records of foreign tourists visiting Indonesia from January to December 2018. That consists of attributes namely the Origin country, and the number of visitors for each month of a year. The total amount of data records is 245.

Data preprocessing

Preprocessing data consist of data cleaning. During data cleaning, records with missing data and attributes are simply removed. Furthermore, some of the data is not valid, such as within one year there are no foreign tourists to Indonesia in a country. After data cleaning, data checking is then carried out. The result of the data from the given dataset, the attributes in the dataset do not have null values.

Clusterring process

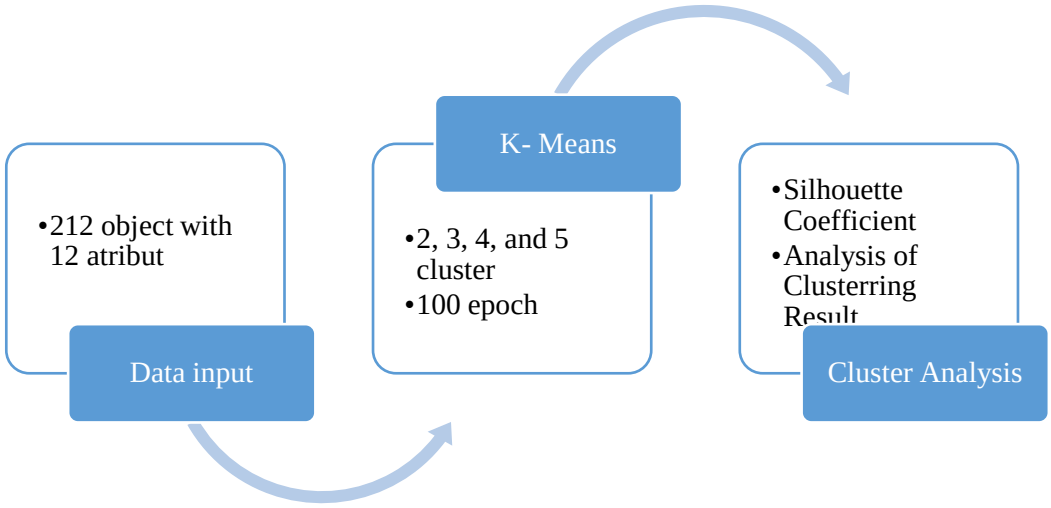

Figure 3. Clustering process

The K-means clustering parameters used are the number of clusters (K), where K={2, 3, 4, 5}, and the number of epochs is 100. Each number of K would be tested and analyzed to find the best clusters K-means implements an unsupervised learning algorithm. This is a simple way of classifying observational data into clusters (Awat & Ballera, 2019). The K-means steps applied in this study are as follows:

1) The first step is to determine the value of the centroid C using the following formula

$$d(i, C) = \frac{1}{|A|} \sum_{j \in C} d(i, j) \quad (1)$$

Where, $d(i, C)$ is the average distance of object i to all objects in other clusters C where A ≠ C.

2) The second step is to calculate the distance between the data points to all the centroids using the Euclidean Distance, with the following formula (Faisal & Zamzani, 2020).

$$d(c, x) = \sqrt{\sum_{i=1}^{n} (c_i - x_i)^2}$$

Where c is the centroid of a cluster, x is a data point inside a cluster, and n is the total number of attributes of the dataset.

3) Moving a data point to the nearest centroid.
4) Calculate the new cluster centroids using the means of each attribute value of all data points in the corresponding clusters
5) Back to step 2.

Cluster analysis

In testing a model, the aim is to obtain information on how close the relationship between one object and other objects in a cluster is and how far between one cluster and another cluster. The test method used in this case is to look for the Silhouette Coefficient where this method is a combination of the other two methods, namely the Cohesion method which is useful in measuring how close the relationship between one object is with other objects in a cluster and the Separation method which is useful for calculating how far a cluster is separated from other clusters or the extent to which a cluster is separated from other clusters. In the given case study, the result of the clustering is intended to be used for deciding the right marketing strategy for the corresponding countries to improve the quality of services and escalate the number of visitors.

Results and Discussion

The input of the K-means is 212 data points with 12 attributes of numeric variables. The data attributes for clustering are January, February, March, April, May, June, July, August, September, October, November, and December. The clustering process uses Rapidminer as shown in Figure 5. The results of the clustering are distributed into 4 cluster division scenarios, namely 2, 3, 4, and 5 clusters. The distribution of the clustering process with scenarios 2,3,4 and 5 clusters is shown in Table 1.

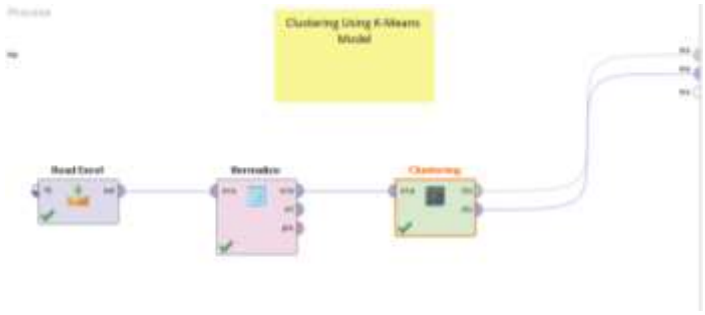

Figure 4. Clustering Using K-Means model

Table 1. Cluster distribution

Number of Clusters	Cluster Distribution
2	Cluster 1 = 206 , Cluster 2 = 6
3	Cluster 1 = 198 , Cluster 2 = 6, Cluster 3 = 8
4	Cluster 1 = 198 , Cluster 2 = 2 , Cluster 3 = 4, Cluster 4 = 8
5	Cluster 1 = 198 , Cluster 2 = 3 , Cluster 3 = 1, Cluster 4 = 2 , Cluster 5 = 8

The results of the silhouette score can be seen in Figure 6 which shows K = 2 has a value of 0.8, in this case, the use of 2 clusters is the best cluster that can be used.

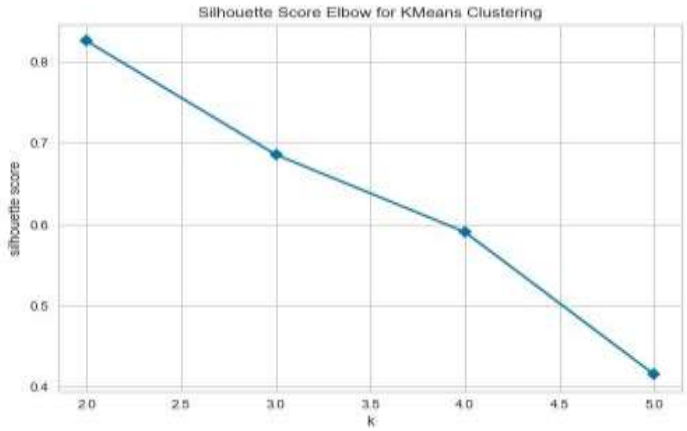

Figure 5. Silhouette score for each K clusters

The distribution of the 2 clusters is shown in table 2 below. Cluster 1 has 206 members and cluster 2 has 6 members. Membership of each cluster in detail as in table 2, while the centroid of each cluster can be seen in table 3. Cluster 2, the average value of cluster centroid is higher than cluster 1.

Table 2. Data points id number of cluster members

No	Cluster	Cluster Members
1	Cluster 1	1,3,5 – 18, 21 – 31, 33 – 142, 144 – 212 **(206)**
2	Cluster 2	2, 4, 19, 20, 32, 143 **(6)**

From the cluster analysis, it can be seen that cluster 1 consists of countries that have relatively few tourists to visits Indonesia. Cluster 2, on the other hand, consists of a list of countries that have relatively high citizens to visits Indonesia, namely Malaysia, Singapore, China, Other Asia, Timor Leste, and Australia.

Tabel 3. Centroid of cluster

Atribut	Cluster 1	Cluster 2
January	2358,602	136224,167
February	2393,660	146186,333
March	2785,408	164016,833
April	2770,728	155981,167
May	2605,718	148304,667
June	2695,155	156922,333
July	3612,286	170755,5
August	3381,752	173588,833
September	2913,5	162209,5
October	2861,738	150975,0
November	2315,034	143319,5
December	2706,505	174218,833

The results of the clustering process are expected to be an input for further marketing analysis to find the right strategy for tourist promotion. For example, increase promotions, discounts, or free visas to countries that have low arrival rates to Indonesia.

Conclusion

The clustering model using the K-Means algorithm was built to group foreign countries that frequently visit Indonesia. The input data consists of 212 records with 12 attributes. The K-Means algorithm for this clustering process produces scenarios of 2, 3, 4, and 5 clusters with 100 epochs. The results of the silhouette score show that k = 2 has a value of 0.8, in this case, the use of 2 clusters is the best cluster that can be used. The results of the clustering are grouped into 2 clusters, namely, cluster 1 consisting of countries that have a relatively low level of tourist visits to Indonesia with 206 members, and cluster 2 consists of countries that have relatively high tourist visits to Indonesia with 6 members, namely Malaysia, Singapore, China, Other Asia, Timor Leste, and Australia. The results of the clustering process are expected to be input data for further marketing strategy for the countries visiting Indonesia to increase the tourism quality.

References

Anggraini, D. (2017). Analisis hubungan komplementer dan kompetisi antar destinasi pariwisata (Studi kasus: 10 Destinasi Pariwisata Prioritas Di Indonesia). *Tesis MPKP FEB UI*.

Awat, K. A. S., & Ballera, M. A. (2018). Applying K-Means clustering on questionnaires item bank to improve students' academic performance. *In 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, 1-6.

Aytaç, E. (2020). Unsupervised learning approach in defining the similarity of catchments: Hydrological response unit based k-means clustering, a demonstration on Western Black Sea Region of Turkey. *International Soil Water Conserv. Res.*, *8*(3), 321–331. doi: 10.1016/j.iswcr.2020.05.002.

Faisal, M., & Zamzami, E. M. (2020). Comparative analysis of inter-centroid K-Means performance using euclidean distance, canberra distance and manhattan distance. *Journal of Physics: Conference Series, 1566* (1), 012112.

Han, H., & Hyun, S. S. (2015). Customer retention in the medical tourism industry: Impact ofquality, satisfaction, trust, and price reasonableness. *Tourism Management, 46* (2015), 20-29.

Helaey, S. C. (2013). Marketing module series. In Cornel University (Ed.), *Marketing Module 8: Promotion*. Ithaca.

Iswardani, A., & Riadi, I. (2016). Denial of service log analysis using density K-means method. *J. Theor. Appl. Inf. Technol.*, 83(2), 299–302,

Karunanithy, M., & Sivesan, S. (2013). An empirical study on the promotional mix and brand equity: Mobile service providers. *Industrial Engineering Letters*, 3(3), 1-9. https://doi.org/2225-0581

Katrakilidis, C., Konteos, G., Sariannidis, N., & Manolidou, C. (2017). Investigation of convergence in the tourist markets of greece. *European Research Studies Journal, 20*(4A), 707-729.

Maulana, A., & Fajrin, A. A. (2018). Penerapan data mining untuk analisis pola pembelian konsumen dengan algoritma Fp-Growth pada data transaksi penjualan spare part motor. *Klik - Kumpulan Jurnal Ilmu Komputer, 5*(1), 27. Doi: 10.20527/klik.v5i1.100.

Maulida, L. (2018). Penerapan datamining dalam mengelompokkan kunjungan wisatawan ke objek wisata unggulan di Prov. Dki Jakarta Dengan K-Means. *JISKA (Jurnal Inform. Sunan Kalijaga)*, *2*(3), 167. doi: 10.14421/jiska.2018.23-06.

Mowforth, M., & Munt, I. (2015). *Tourism and sustainability: Development, globalisation and new tourism in the third world.* Publisher: Routledge, 476.

Novak, D. (2011). Promotion as instrument of marketing mix. *I International Symposium Engineering Management and Competitiveness*, *2011,* 505-510. Retrieved from http://www.tfzr.uns.ac.rs/emc/emc2011/Files/G 06.pdf

Novokreshchenova, A. O., Novokreshchenova, A. N., Terehin, E. S. (2016). Improving Bank's Customer Service on the Basis of Quality Management Tools. *European Research Studies Journal, 19*(3), 19-38.

Pamungkas, F., Nugroho, D., & Utami, Y. R. W. (2020). Rekomendasi wisata umbul dengan K-Means Clustering. *J. Teknol. Inf. dan Komun., 8*(2), 11–18, 2020, doi: 10.30646/tikomsin.v8i2.478.

Ramadani, S., Ambarita, I., & Pardede, A. M. H. (2019). Metode K-Means untuk pengelompokan masyarakat miskin dengan menggunakan jarak kedekatan Manhattan City Dan Euclidean (Studi kasus kota binjai). *Inf. Syst. Dev.*,*04*(2), 15–29.

Rustam, S., & Annur, H. (2019). Akademik Data Mining (ADM) K-Means Dan K-Means K-Nn untuk mengelompokan kelas mata kuliah kosentrasi mahasiswa semester akhir. *Ilk. J. Ilm.*, *11*(3), 260–268, 2019, doi: 10.33096/ilkom.v11i3.487.260-268.

Seimahuira, S. (2021). Implementasi datamining dalam menentukan destinasi unggulan berdasarkan online reviews tripadvisor menggunakan algoritma K-Means. *Technologia,* 12(1), 53–58.

Sudibyo, N. A., Iswardani, A., Sari, K., & Suprihatiningsih, S. (2020). Penerapan data mining pada jumlah penduduk. *Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat.*, 1(3), 199–207.

Windarto, A. P. (2017). Implementation of data mining on rice imports by major country of origin using algorithm using K-Means clustering method. *International Journal of Artificial Intelligence Research, 1*(2), 26. doi: 10.29099/ijair.v1i2.17.

Wiratama, C., Kurniawaty, N., Febriane, F., Putri, R., & Haekal, H. (2014). The golden line of indonesian tourism. I*nternational Proceedings of Economics Development and Research,* IACSIT Press, 76. Doi: 10.7763/IPEDR. 2014. V76. 2

Yahya, A. (2015). *Tourism became the mainstay of foreign exchange penders country*. Available at: http://www.kemenpar.go.id/asp/detil.asp?c=16&id=2959