Analysis Of Tourist Visit Places Using Principal Component Analysis In The K-Means Method

Sema Oktaviani Tinting[1], Ari Purno Wahyu Wibowo[2],

[1,2]Universitas Widyatama, Bandung 40125, Indonesia
ari.purno@widyatama.ac.id

Abstract

Tourism is a place where domestic and foreign tourists visit with many facilities which aim to increase the number of visits abroad with E-tourism which can further develop with quality internet data services. The opportunity for tourism operators to introduce lesser-known tourism destinations will be greater, thereby increasing the number of tourists visiting. This research will create a system to analyze the level of accuracy of grouping data on tourist visits using the K-Means method and apply the PCA/Principal Component Analysis method in grouping tourist visits using data taken from the Central Statistics Agency from 2020-2022 on the BPS North Kalimantan website, website Ministry of Park and Creative Economy and National Library website. Therefore, tourist visiting places in North Kalimantan will be grouped based on the number of tourist visits into several clusters, namely tourist visiting places that are good, quite good and not so good using the k-means method and the PCA/Principal Component Analysis method in grouping tourist visits. the result obtained is that. There are 3 clusters with the results obtained based on a comparison of tourism clustering of places visited by tourists in North Kalimantan using the k-means method with a silhouette value of 0.945952526 which begins with reducing the dimensions of the dataset using PCA.

Keyword: Tourism, Visiting Places, Tourists, PCA/Principal Component Analysis, K- Means

1. INTRODUCTION

1.1 Background of the problem

Visa-free visits since 2015 for tourists from 169 countries continue to attract foreign visitors to Indonesia, facilitating easier access. Each foreign tourist spends about US $1,200, contributing approximately 11.5% to the Gross Domestic Product (GDP). The demand for 3G and 4G technology in MICE, along with creativity in tourism products, can thrive with support from the creative economy, enhancing welfare and reducing regional disparities. While cultural trade-offs should ideally be independent of market forces, market demand must still be considered in cultural production, as culture extends beyond just art. This relationship between tourism and the creative economy is expected To foster synergy among stakeholders.[1] In December 2022, there were 4,258 foreign tourists visiting, a decrease of 2,965 visits compared to the previous month which reached 1,293 visits. Cumulatively, the number of foreign tourists (tourists) visiting North

V. Mardiansyah and B. A. Prasetyo (eds.), *Proceedings of the Widyatama International Conference on Engineering 2024 (WICOENG 2024)*, Advances in Engineering Research 252,
https://doi.org/10.2991/978-94-6463-618-5_4

Kalimantan in the January – December 2022 period reached 40,047 visits. The large number of tourists can act as a contributor to Gross Domestic Product, and also contribute to foreign exchange and maintain exchange rate stability. [2]

Tourism development involves the business environment, governance, natural and artificial tourism potential, and supporting infrastructure. The tourism index calculation helps local governments recognize competitiveness as a key driver of local and national economic growth. Developing tourist destinations can enhance openness and encourage more visitors, thereby increasing the occupancy rates of star hotel rooms.[3] Additionally, well-managed community-based tourism, coupled with cooperation between central and regional governments, has the potential to boost the regional economy.

This research will create a system to analyze the level of accuracy of grouping data on tourist visits using the K-Means method and apply the PCA/Principal Component Analysis method in grouping tourist visits using data taken from the Central Statistics Agency from 2020-2022, North Kalimantan BPS website, website Ministry of Park and Creative Economy and National Library website. [4], [5] dan [6].

2. THEORETICAL REVIEW

2.1 Tourism

Law Number 9 of 1990 concerning tourism explains that everything related to tourism, business objects and tourist attractions or uniqueness related to that place or region. Various travel activities carried out by a person or group of people from place to place which are usually far from where they live are tourism which has the aim of working as well as vacationing. This visit is temporary or non-permanent or permanent and in time they will return to their original place of residence [7]. There are two very important factors, namely how they travel and where they temporarily live. There were many activities carried out on the trip. A person's journey to work, even though it may be quite far in terms of distance, is certainly not included in the tourist category. In other words, tourism activities are activities that make a person or group happy by having activities that are different from having a place to have fun [8]. The concept of tourism is multidimensional with several definitions of tourism used by practitioners with different perspectives according to the goals to be achieved. The definition of tourism cannot be exactly the same among experts. The following are several definitions of tourism [9] and [10]

2.2 Traveler

Tourists are a person or group of people who go on a tourist trip to a place, whose stay is at least 24 hours in the area or country visited. However, if they live in the area or country they visit for less than 24 hours, they are called travelers[11]. President of the Republic of Indonesia No. 9/1969 written in chapter 1 article 1, says that tourists are everyone who travels from their place of residence to visit one place or another and enjoy the journey of their visit. [1]. According to Spillane, tourists are temporary visitors who stay less or more than 24 hours on an island or country they visit and their trips can be grouped. [10] :

a. Cruises are for recreation, vacation, health, study, religious and sports purposes.
b. Trade relations, relatives, friends, conferences and missions. Tourists in general are a group of people who visit an area to go on a tourist trip, but do not live in the destination area or work to earn wages.

2.3 Data Mining

Data mining can be interpreted as a way of obtaining information from the results of mining information which can later be used in a larger database. One thing that needs to be paid attention to is the rules for finding high frequency patterns between sets of item sets which are usually called Association Rules [12], [13] and [14].

2.4 Clustering

Clustering is one of the techniques or methods in data mining. The clustering algorithm is an algorithm for grouping a number of data into certain groups, often called clusterers. Clustering is a type of unsupervised learning, which means it does not require a learning process or phase and does not use any teaching in the groups.. [6], [15] and [16].

2.5 K-Means

A method of analyzing data or a method in data mining that carries out an unsupervised modeling process and a method of grouping data with a partition system. [17] and [18].

The aim of this algorithm is to find the distances between objects and the closest centroid, namely by minimizing the objective function J which is formulated as a function of U and V as follows:

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} \mu_{ik} d^2(x_k, v_i)$$

2.6 PCA (Principal Component Analysis)

Principal Component Analysis is used to reduce the dimensions of the observed data into smaller dimensions without losing significant information in describing the entire data. PCA will be used in the last phase of the eigenface, to find the most important part of the eigenvector. Before that, we will explain the basic concepts used in PCA [19] and [20].

$$\begin{aligned}
PC_1 &= a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p + \varepsilon_i \\
PC_2 &= a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p + \varepsilon_i \\
&\ \ \vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
PC_k &= a_{12}X_1 + a_{22}X_2 + \cdots + a_{pk}X_p + \varepsilon_i
\end{aligned}$$

3. METHODOOLOGY

3.1 Curent System Analysis

The system running in this research aims to identify current problems, to produce a system plan that can solve problems that arise in accordance with the objectives. The results obtained in this research are one of the researchers' objectives is to develop the use of data which is expected to achieve more precise accuracy, making it easier to analyze to increase the accuracy of grouping tourist visit data using the K-Means method and applying the PCA/Principal Component Analysis method in grouping. tourist visits.

3.2 K-means

The calculation stage using the K-means method is a data processing process that is intended to find and obtain information based on the calculation.

The initial centroid of the iteration

After the data goes through the stages of data pre-processing, the data will then go through the calculation stage.

3.3 Silhouette Coefficient

3.4 This Silhouette Coefficient evaluation combines two methods, namely the Cohesion method and the Separation method. The Cohesion method here is useful for measuring the distance between objects in a cluster, and the separation method is used to measure the distance from the first cluster to the next cluster. [13]. There are 3 stages in calculating the Silhouette Coefficient method, namely [21] :

1. Calculate the average for each object I with all objects within one cluster scope. So you will get the so-called average value a_i .
2. For each object I, the minimum value of the average distance is calculated from one point to another point in a different cluster. So you will get a minimum average value which will be called b_i .

After the values from the 2 stages are obtained, the Silhouette Coefficient value can be calculated using the equation formula [22]:

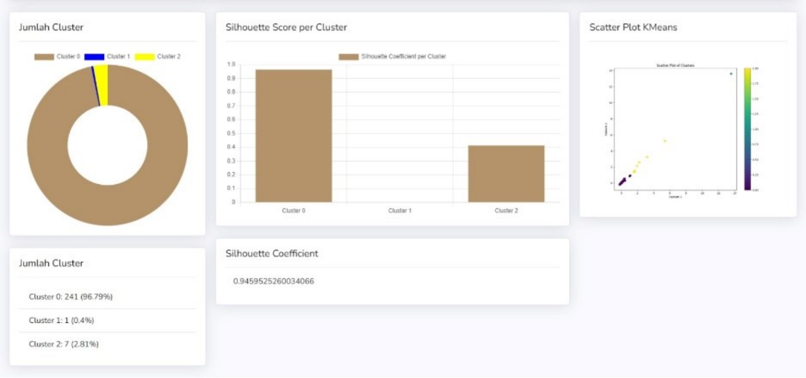
$$Si = \frac{bi - ai}{\max\,(ai - bi)}$$

Information:
Si : Silhouette Coefficient
b : The average distance of the medoid to objects outside the cluster
a : Average distance between media and objects in the cluster
As for the Silhouette Coefficient value, it will be between -1 and 1. And the following is the Silhouette Coefficient according to Kaufman and Rousseeuw:

0.7 < SC <=1 belonging to the Strong Structure Cluster
0.5 < SC <=0.7 belongs to the Medium Structure Cluster
0.25 < SC <=0.5 classified into the Weak Structure Cluster
SC <= 0.25 belonging to the No Structure Cluster

4. RESULTS

a. K-Means



Jumlah Cluster

Cluster 0: 241 (96.79%)

Cluster 1: 1 (0.4%)

Cluster 2: 7 (2.81%)

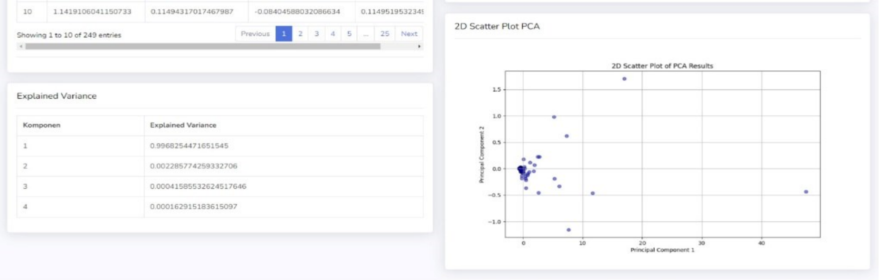
Silhouette Coefficient

0.9459525260034066

fig 1 Result K-means and Silhouette

The analysis reveals three distinct clusters of tourist visits: Cluster 0, characterized by a high percentage of visits, indicates popular tourist hotspots; Cluster 1 comprises locations with low visit numbers, suggesting potential areas for development; while Cluster 2 includes destinations with moderate visitation levels. Notably, the evaluation using the Silhouette Coefficient method yielded an impressive value of 0.945952526, signifying that the clusters are not only well-defined but also exhibit high internal cohesion and separation from one another. This strong clustering result underscores the efficacy of the K-Means algorithm in segmenting tourist visit data, providing valuable insights for stakeholders in tourism management and strategy development. The clarity of these clusters can facilitate targeted marketing efforts and resource allocation, ultimately enhancing the tourism experience in the region.

b. Principal Component Analysis/PCA



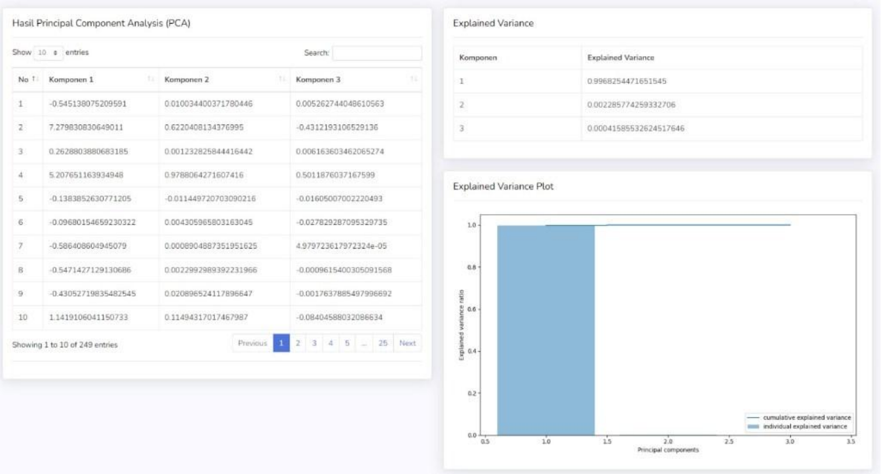
10	1.1419106041150733	0.11494317017467987	-0.08404588032086634	0.114951953234

Showing 1 to 10 of 249 entries

Explained Variance

Komponen	Explained Variance
1	0.9968254471651545
2	0.002285774259332706
3	0.00041585532624517646
4	0.000162915183615097

2D Scatter Plot PCA



Hasil Principal Component Analysis (PCA)

No	Komponen 1	Komponen 2	Komponen 3
1	-0.545138075209591	0.010034400371780446	0.005262744048610563
2	7.279830830649011	0.6220408134376995	-0.4312193106529136
3	0.2628803880683185	0.001232825844416442	0.006163603462065274
4	5.207651163934948	0.9788064271607416	0.5011876037167599
5	-0.1383852630771205	-0.011449720703090216	-0.01605007002220493
6	-0.09680154659230322	0.004305965803163045	-0.027829287095329735
7	-0.586408604945079	0.0008904887351951625	4.979723617972324e-05
8	-0.5471427129130686	0.0022992989392231966	-0.0009615400305091568
9	-0.43052719835482545	0.020896524117896647	-0.0017637885497996692
10	1.1419106041150733	0.11494317017467987	-0.08404588032086634

Showing 1 to 10 of 249 entries

Explained Variance

Komponen	Explained Variance
1	0.9968254471651545
2	0.002285774259332706
3	0.00041585532624517646

Explained Variance Plot

Fig 2 result Principal Component Analysis

The analysis reveals that Principal Component Analysis (PCA) effectively reduced the dimensionality of the dataset on tourist visits in North Kalimantan, generating three principal components with eigenvalues of -0.5451380, 0.0100344, and 0.0052627. The first principal component captures a significant portion of the variance, while the second and third contribute minimally. Importantly, over 90% of the total variance was retained, and the classification error accuracy remained below 10%, indicating high precision in the results. These findings underscore the efficacy of PCA in simplifying complex datasets and enhancing understanding of tourism dynamics in the region, thereby informing strategic planning for tourism management.

5. CONCLUSION

Based on the results of the research conducted, the study aims to categorize tourist destinations in North Kalimantan into several clusters: good, quite good, and less good. This categorization was achieved using the K-Means clustering method in conjunction with Principal Component Analysis (PCA) to effectively manage the dimensionality of the dataset. The analysis revealed three distinct clusters of tourist destinations. Taman Nasional Borneo emerged as a key destination, notable for its rich biodiversity, including orangutans and various endangered species. This indicates a strong attraction for ecotourism. Pulau Derawan was identified as another significant cluster, recognized as a prime location for diving and beach activities, featuring stunning coral reefs and an abundance of marine life, making it highly appealing for water sports enthusiasts. Additionally, Sungai Mahakam provides a unique river tourism experience, where visitors can explore traditional villages while enjoying the natural scenery, emphasizing cultural tourism.The clustering process yielded a silhouette value of 0.945952526, indicating a high level of confidence in the distinctiveness of the clusters formed. The initial step of reducing the dataset's dimensions using PCA facilitated this analysis by preserving over 90% of the variance while minimizing classification error to below 10%. These findings highlight the effectiveness of combining K-Means and PCA in analyzing tourism data, offering valuable insights for stakeholders in the tourism sector to enhance destination marketing and management strategies.

6. REFERENCES

[1] F. N. A. Sahara, M. Iqbal, and B. Sanawiri, “ANALISIS MOTIVASI BERKUNJUNG WISATAWAN DAN TINGKAT PENGETAHUAN WISATAWAN TENTANG PRODUK INDUSTRI KREATIF SEKTOR KERAJINAN (Studi pada Wisatawan Domestik di Kota Batu, Jawa Timur),” J. Adm. Bisnis, vol. 35, no. 2, pp. 146–154, 2016.

[2] D. N. Saksono, A. Y. Sari, and K. R. Dwi, “Rekomendasi Lokasi Wisata Kuliner Menggunakan Metode K-Means Clustering Dan Simple Additive Weighting,” J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 10, pp. 3835–3842, 2018.

[3] Z. M. Muktaf and E. R. Zulfiana, “Persepsi Wisatawan Asing Terhadap Wisata Indonesia,” J. Cakrawala ISSN, vol. 1693, pp. 83–106, 2018.

[4] Badan Pusat Statistik, “Jumlah Kunjungan Wisatawan Mancanegara per Bulan ke Indonesia Menurut Pintu Masuk, 2017-2020,” Badan Pusat Statistik. 2020.

[5] I. K. J. Arta, G. Indrawan, and G. Rasben Dantes, “Data Mining Rekomemdasi Calon Mahasiswa Berprestasi di STMIK Denpasar Menggunakan Metode

Technique For Other Reference By Similarity to Ideal Solution," J. Ilmu Komput. Indones., vol. 4, no. 1, pp. 11–21, 2019.

[6] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," J. Nas. Teknol. dan Sist. Inf., vol. 5, no. 1, pp. 17–24, 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

[7] R. W. Sari and D. Hartama, "Data Mining : Algoritma K-Means Pada Pengelompokkan Wisata Asing ke Indonesia Menurut Provinsi," Semin. Nas. Sains Teknol. Inf., pp. 322–326, 2018.

[8] K. T. Simangunsong, "Analisis Aktivitas Wisatawan Saat Berkunjung Ke Pantai Di Daerah Istimewa Yogyakarta," Kepariwisataan J. Ilm., no. iv, pp. 220–229, 2023, [Online]. Available: http://ejournal.stipram.ac.id/index.php/kepariwisataan/article/view/224.

[9] S. A. Andina and I. Aliyah, "Faktor-Faktor Yang Mempengaruhi Minat Wisatawan Dalam Mengunjungi Wisata Budaya Candi Borobudur," J. Cakra Wisata, vol. 22, no. 3, pp. 27–38, 2021.

[10] I. Junaid, "Jurnal Kepariwisataan Indonesia," J. Penelit. dan Pengemb. Kepariwisataan, vol. 9, no. 2, pp. 119–234, 2016.

[11] D. N. Jannah, H. N. Kristiansen, and M. S. Wibowo, "Pengaruh Kearifan Lokal Terhadap Minat Pengunjung Di Desa Wisata Nongkosawit," J. Pariwisata, vol. 11, no. 1, pp. 28–39, 2024, doi: 10.31294/par.v11i1.21456.

[12] L. Maulida, "Penerapan Datamining Dalam Mengelompokkan Kunjungan Wisatawan Ke Objek Wisata Unggulan Di Prov. Dki Jakarta Dengan K-Means," JISKA (Jurnal Inform. Sunan Kalijaga), vol. 2, no. 3, p. 167, 2018, doi: 10.14421/jiska.2018.23-06.

[13] J. R. S. Penda Sudarto Hasugian, "Penerapan Data Mining Untuk Pengelompokan Siswa Berdasarkan Nilai Akademik dengan Algoritma K-Means," KLIK Kaji. Ilm. Inform. dan Komput., vol. 3, no. 3, pp. 262–268, 2022, [Online]. Available: https://djournals.com/klik.

[14] H. Susanto and S. Sudiyatno, "Data mining untuk memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu," J. Pendidik. Vokasi, vol. 4, no. 2, pp. 222–231, 2014, doi: 10.21831/jpv.v4i2.2547.

[15] B. Melpa Metisen and H. Latipa Sari, "Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila," J. Media Infotama, vol. 11, no. 2, pp. 110–118, 2015.

[16] D. ayu Ilfiana, "Pengklasteran Puskesmas di Kabupaten Kudus Menggunakan Metode K-Means dengan Perbandingan Jarak Euclidean dan Chebyshev," Prism. Pros. Semin. Nas. Mat. 5, vol. 5, pp. 787–798, 2022.

[17] K. Handoko, "Penerapan Data Mining Dalam Meningkatkan Mutu

Pembelajaran Pada Instansi Perguruan Tinggi Menggunakan Metode K-Means Clustering (Studi Kasus Di Program Studi Tkj Akademi Komunitas Solok Selatan)," J. Teknol. dan Sist. Inf., vol. 02, no. 03, pp. 31–40, 2016, [Online]. Available: http://teknosi.fti.unand.id/index.php/teknosi/article/view/70.

[18] W. Mega, "Clustering Menggunakan Metode K-Means Untuk Menentukan Status Gizi Balita," J. Inform., vol. 15, no. 2, pp. 160–174, 2015.

[19] S. Manullang, D. Aryani, and H. Rusydah, "Analisis Principal Component Analysis (PCA) dalam Penentuan Faktor Kepuasan Pengunjung terhadap Layanan Perpustakaan Digilib," Edumatic J. Pendidik. Inform., vol. 7, no. 1, pp. 123–130, 2023, doi: 10.29408/edumatic.v7i1.14839.

[20] M. Wangge, "Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktor-faktor yang Mempengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA," J. Cendekia J. Pendidik. Mat., vol. 5, no. 2, pp. 974–988, 2021, doi: 10.31004/cendekia.v5i2.465.

[21] A. Masruro, K. Kusrini, and E. Luthfi, "Decision Support System Determination of Tourism Sites Using K-Means Clustering and Topsis," Data Manaj. dan Teknol. Inf., vol. 15, no. 4, p. 1, 2014.

[22] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," PENA Tek. J. Ilm. Ilmu-Ilmu Tek., vol. 6, no. 2, p. 48, 2021, doi: 10.51557/pt_jiit.v6i2.659.