TIERS Information Technology Journal
Vol.4, No.1, June 2023, pp. 08~15
ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: 10.38043/tiers.v4i1.4297

Application of The K-Means Clustering Method To Search For Potential Tourists of Bendesa Hotel

I Gede Karang Komala Putra[1], I Gede Wahyu Surya Dharma[2]
Email: igdkarang@@iikmpbali.ac.id[1], suryadharma@iikmpbali.ac.id[2]
[1,2] Faculty of Business, Social, Technology and Humanities, Informatics Study Program,
Universitas Bali Internasional, Bali

ABSTRACT

Hotels play a significant role in the growth of global tourism. With intense competition in the hotel industry, hotels are shifting their focus from solely providing superior services to identifying potential tourists. In a previous study, the J48 algorithm was employed to extract hotel transaction patterns, achieving an accuracy level of 71.6418% by considering gender and age characteristics[1]. In a separate study, foreign guest ratings by province were classified into three clusters. The study concluded that nearly 90% of provinces in Indonesia exhibit low levels of tourism, supported by the analysis of the number of tourists staying, as reported by the statistical center[2]. To identify potential tourists who can bring benefits to the hotel, hotel managers can utilize the k-means algorithm. In this study, a data mining process was conducted using data collected from tourists who stayed at the Bendesa Hotel. The process began with tourist segmentation using the K-means algorithm divided into clusters. Subsequently, the accuracy of the obtained data was calculated. This research employed room class as a reference value to discover tourist characteristics at the Bendesa Hotel. The results of applying the K-means model with 4 clusters indicated that the accuracy level for identifying potential tourists reached 84.4%.

Keywords: Data Mining; Characteristics; Hotels; Clusters; K-Means

Article Info

Accepted	:	10-06-2023
Revised	:	30-04-2023
Published Online	:	25-06-2023

Correspondence Author:

I Gede Karang Komala Putra
Faculty of Business, Social, Technology and Humanities, Informatics Study Program
Universitas Bali Internasional
Jl. Seroja Gang Jeruk No 9A, Denpasar
Email: igdkarang@@iikmpbali.ac.id

1. INTRODUCTION

The range of human activities is highly diverse, and as these activities continue to expand, there is an increasing demand for various forms of media to support them. One particular human activity that requires attention is the field of travel and tourism, which has led to a growing need for hospitality services. To cater to these requirements, numerous hotels have been constructed as accommodations for tourists, resulting in the growth of the hotel industry and the proliferation of tourism destinations in different regions. Additionally, alongside hotel services, other businesses such as restaurants, nightclubs, catering services, bars, pubs, and discotheques have also experienced significant growth.

The crucial aspect that now requires careful consideration is the management of these service industries to foster their advancement as thriving trades. The selection of tourists towards a hotel causes competition in the hospitality business. Not all hotels are successful in competing and dominating the market according to the set targets, especially in the current economic situation. For this reason, a policy and an accurate marketing strategy are needed in the face of increasingly fierce competition. This is where the important task and role of the marketing division of a hotel is to try to fill the rooms at low times, in addition to increasing sales volume from time to time.

Marketing activities in general are different from sales, transactions, or trading. The definition of marketing contains a broader understanding than just sales and advertising. Marketing is a key concept to the success of a business where marketing pays attention to the desires and needs of fulfilling customers to achieve satisfaction and has a positive impact on companies in this era of sophisticated business competition. Marketing is a very important functional area in a business organization as the main support for the operational survival of a business world.

The search for potential hotel tourists staying overnight is one of the strategies that can be used to support the development and sustainability of hotels. Hotel management must be able to recognize potential tourists so that they can maintain their loyalty to the hotel. The marketing strategy is the company's strategy for marketing its products well to achieve the desired profit level. The goal is to be able to compete in every situation, and if the marketing strategy implemented by the company can run smoothly, it will be able to increase sales of the company's services, especially in selling rooms to bring benefits to the hotel. An example would be the development of targeted room sales strategies using the characteristic results obtained through data mining.

In the previous study, the J48 algorithm was utilized to search for hotel transactions, achieving an accuracy of 74% by considering attributes such as age and gender[1]. In a separate study, foreign guest ratings by province were classified into three clusters. The study concluded that nearly 90% of provinces in Indonesia exhibit low levels of tourism, supported by the analysis of the number of tourists staying, as reported by the statistical center[2]. However, this led the researchers to explore alternative algorithms and additional attributes to identify potential tourists suitable for specific room types. This information would be beneficial for the hotel in formulating future strategies. The data used for this analysis were collected from staying tourists. The variables used to search for potential tourists are age, gender, profession, continent, and room of the tourists staying overnight. This study uses one of the techniques known in data mining, namely the clustering technique. Clustering is a procedure that involves categorizing and distributing data patterns into multiple data sets. Patterns that exhibit similarities are grouped together in clusters, while distinct patterns separate themselves by belonging to different clusters[3].

2. RESEARCH METHOD

The stages of research that will be carried out by the author in this research process are as follows.

2.1 Research Design

For this research, the authors employed experimental research methods, utilizing a specific experimental design known as the Cross-Industry Standard Process for Data Mining (CRISP-DM). The CRISP-DM process model consists of six distinct stages, which are as follows [4]:

1. Understanding of Business (Bussiness Understanding)

In this phase, there are three stages carried out, viz.

a. Understanding of business goals

Must understand the hotel’s business objectives which will affect the search for the characteristics of tourists who stay. Some of the hotel’s business objectives related to the search for characteristics are.

1. Knowing the characteristics that will be used as attributes in the k-means algorithm.
2. Get a lot of profit from the number of tourists who stay.
3. Create economical and promotional packages to increase the percentage of new tourist visits or regular customers.

b. Situation assessment

Hotel Bendesa is a 3-star hotel in the Kuta area which is located at Jalan Legian Kuta. This hotel was established in 1992 on a land area of 100 square acres and has 49 employees. Hotel Bendesa has a total of 50 rooms with details of 10 deluxe rooms, 25 superior rooms, 10 standard rooms, and 5 economy rooms.

c. Translate business goals into data mining goals

.

At this stage, an understanding of business objectives is required and translated into data mining objectives. One of the goals of data mining is to find potential tourists staying overnight. The characteristic results can be used by management to identify customers, retain customers and increase sales value.

2. Data Understanding

The Data Understanding stage is the second phase of the research, during which the researchers gather the data required for the k-means algorithm. Literature sources, such as textbooks, journals, scientific papers, and relevant websites, were utilized to acquire the necessary information. In this study, the researchers collected data from 4352 tourists who had stayed overnight from January 2019 to December 2019 (prior to the COVID-19 pandemic). Out of the complete dataset, 1000 data were selected and deemed eligible for use in the analysis. In this study, the researchers employed a simple random sampling technique, which involves selecting sample members from a population in a random manner, without considering any specific divisions or strata within that population [5].

3. Data Preparation

The third stage is Data Preparation. The researcher prepares the existing data and then looks for the attributes used to analyze the problem. The data in question are tourists who have stayed at the hotel as a reference in finding potential tourists who will stay overnight. In data preparation, several data preparation techniques are used to process the initial data, namely:

a. Data selection (Data Selection)

In this study, the researcher opted to use a specific dataset comprised of tourist data, including information such as Gender, Age, Continent, Profession, and Room. The data source utilized for this study was the check-in book, which was initially in a manual format. To facilitate analysis and processing, the manual data was converted into a document with a .csv extension, making it easier to work within a digital format.

b. Data Processing (Data Preprocessing)

In order to ensure the high quality of the data, the researchers conducted data cleansing procedures to address concerns such as noisy data and missing values. A sample of 1000 data points was randomly selected using the simple random sampling technique from the total collection of 4352 data points for the purpose of the data mining process.

c. Data transformation (Data Transformation)

Nominal data must be initialized in the form of numbers so that it can be used. Researchers took data such as Gender, Profession, and Continent. All data have the same status in the sense that no data has more or fewer levels compared to other data [6].

d. Modeling

In this study, researchers used the K-means model or algorithm. In using the K-means algorithm, it is necessary to use the best number of clusters [7]. In this study, 4 clusters were used according to the room class used as a reference.

e. Evaluation

The fifth stage is Evaluation. At this stage evaluation of the model is carried out used, whether the model has achieved the objectives to be achieved. In this study, the accuracy of 1000 data was carried out by using room classes in the search for potential tourists who will stay overnight.

f. Deployment

The sixth stage or the last stage viz Spread. At this stage, the use of the model that has been made is carried out build system. In this study, it was only carried out up to the fifth stage namely evaluation.

4. Clustering Process Using K-Means Algorithm

Clustering is a procedure that involves categorizing and distributing data patterns into multiple data sets. Patterns that exhibit similarities are grouped together in clusters, while distinct patterns separate themselves by belonging to different clusters [2]. The K-Means algorithm is a method of categorizing data by attempting to divide it into multiple groups in such a way that the data within each group share similarities. The algorithm aims to group data points together based on their similarities and differentiate them from data points in other groups [8]. The objective of the K-Means algorithm is to minimize a function by reducing the variation within clusters and maximizing the variation between clusters [9].

The subsequent equation presents the method for calculating distance, the iterative formula, and SSE [6]:

The formula for measuring distance

$$d(x,y) = \sqrt{(xi - yi)^2 + (xi - yi)^2}$$

Iteration formula

$$(i) = \frac{x1 + x2 + x \ldots + xn}{\Sigma \mathrm{x}}$$

SSE Formula

$SSE = (Ci)^2 + (Ci)^2 + (C \dots)^2 + (C \dots)^2$

The K-Means algorithm process is as follows [10]:

1. Randomly select k objects, these objects will be represented as the mean in the Cluster.
2. Each object is included in a cluster with a high level of object similarity to that cluster. The degree of similarity is determined by the object's distance to the mean or centroid of the cluster.
3. Calculate the new centroid value for each cluster.
4. The process is repeated until the members in the Cluster set do not change.

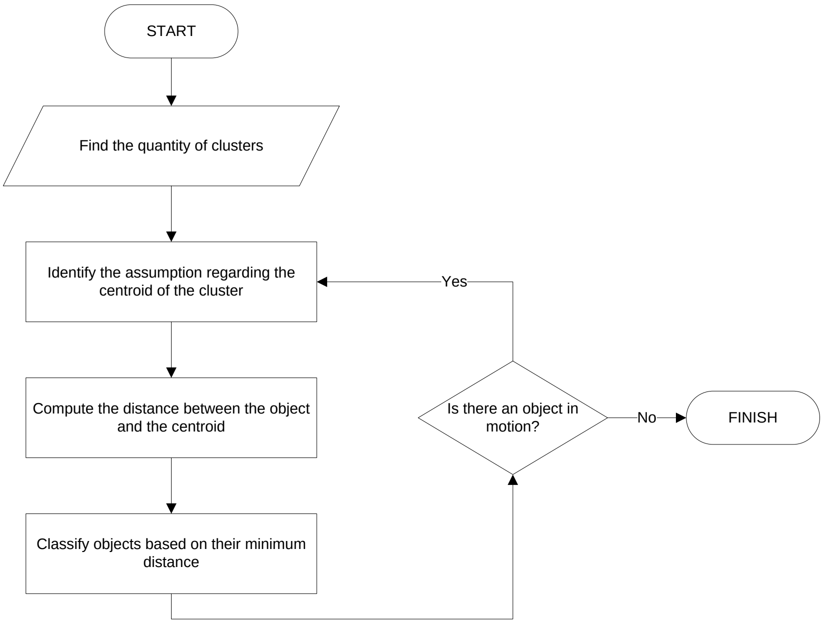

START
Find the quantity of clusters
Identify the assumption regarding the centroid of the cluster
Yes
Compute the distance between the object and the centroid
Is there an object in motion?
No
FINISH
Classify objects based on their minimum distance

Figure 1. K-Means flowchart
[Source : Microsoft VISIO]

3. RESULTS AND DISCUSSION

3.1 Data Description

Data initialization is done by sorting the data from the largest to the smallest [11]. In Gender, the numbers are sorted from the largest based on the amount of data, the largest amount of data is initialed with the number 1, and so on until the smallest amount of data. In addition to gender, the Profession and Continent attributes also include nominal data types so that they are initialized to numeric form. Room data is not changed because it is used as a reference class for clusters.

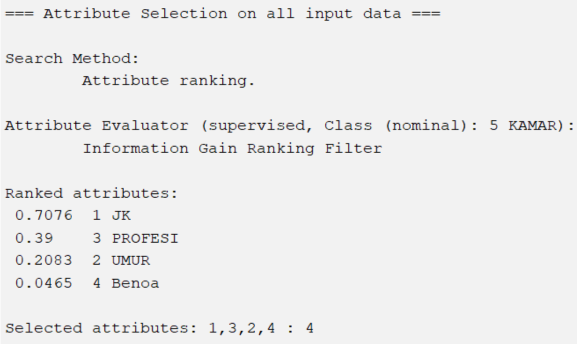
Table 1: Attribute Initialization
[Source: Microsoft EXCEL]

No	Attribute		Amount	Initials
1	Gender	Man	508	1
		Woman	492	2
2	Profession	Private	573	1
		Student	334	2
		civil servant	58	3
		Not Working	35	4
3	Continent	Europe	758	1
		Asia	135	2
		America	62	3
		Australia	45	4

.

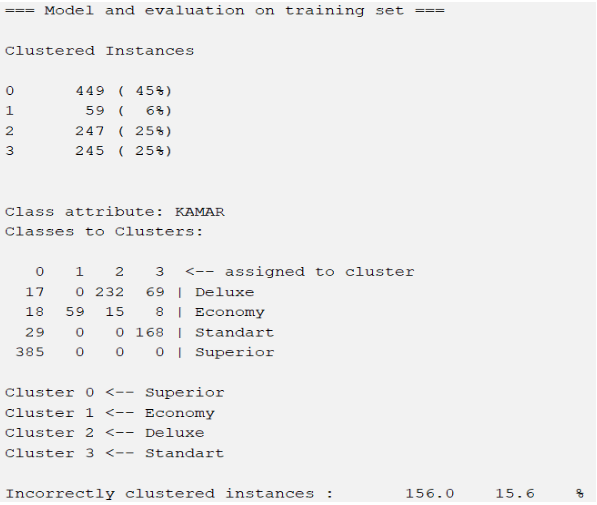
3.2 Discussion

The K-Means clustering process starts with a sample data of 1000 records using Weka 3.8.6. Clustering was tested with 4 clusters according to the room class and the maximum iteration used was 500. The following are some of the results obtained from the test:



```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 5 KAMAR):
        Information Gain Ranking Filter

Ranked attributes:
 0.7076  1 JK
 0.39    3 PROFESI
 0.2083  2 UMUR
 0.0465  4 Benoa

Selected attributes: 1,3,2,4 : 4
```

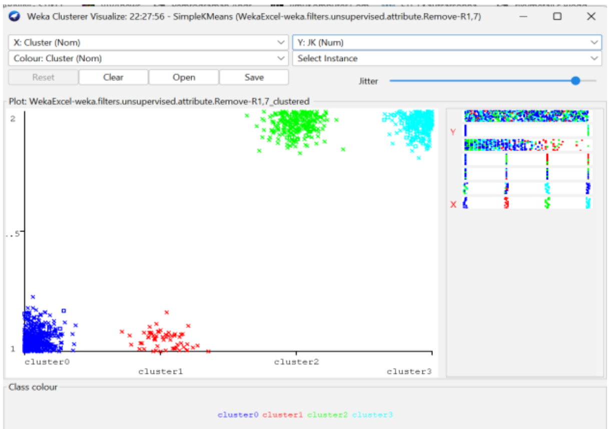
Figure 2. Attributes that affect the Room class
[Source: Weka]



```
=== Model and evaluation on training set ===

Clustered Instances

0       449 ( 45%)
1        59 (  6%)
2       247 ( 25%)
3       245 ( 25%)


Class attribute: KAMAR
Classes to Clusters:

   0   1   2   3  <-- assigned to cluster
  17   0 232  69 | Deluxe
  18  59  15   8 | Economy
  29   0   0 168 | Standart
 385   0   0   0 | Superior

Cluster 0 <-- Superior
Cluster 1 <-- Economy
Cluster 2 <-- Deluxe
Cluster 3 <-- Standart

Incorrectly clustered instances :       156.0     15.6     %
```

Figure 3. K-Means Results
[Source: Weka]

In Figure 2, the attributes that most influence the selection of rooms at the Bendesa hotel are Gender with a value of 0.7076, Profession with a value of 0.39, Age with a value of 0.2083, and Continent with a value of 0.0465. Meanwhile, from Figure 3 above, it can be seen that the results obtained by calculating K-Means using the room class attribute get incorrectly clustered instances of 15.6%, which means the accuracy rate reaches 84.4%. After knowing the level of accuracy in the k-means method, the visualize cluster assignment process is carried out to see the results of the clusters formed from each attribute.



Weka Clusterer Visualize: 22:27:56 - SimpleKMeans (WekaExcel-weka.filters.unsupervised.attribute.Remove-R1,7)

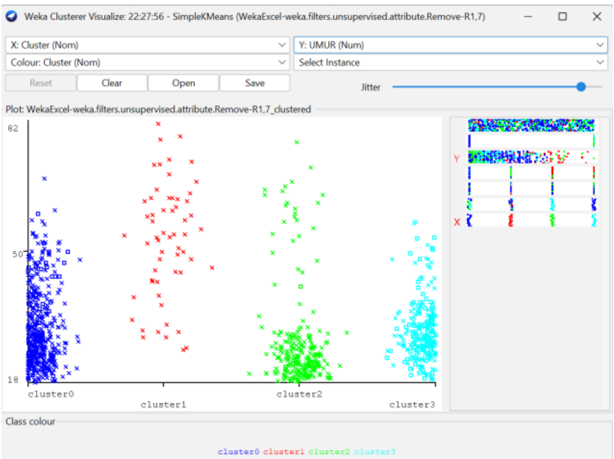
X: Cluster (Nom)
Y: JK (Num)
Colour: Cluster (Nom)
Select Instance

Plot: WekaExcel-weka.filters.unsupervised.attribute.Remove-R1,7_clustered

Figure 4. Visualize the Gender Attribute Cluster
[Source: Weka]

Some of the information obtained is as follows:

1. In Cluster 0 (449 data) and Cluster 1 (59 data) the dominant sex is male.
2. In Cluster 2 (247 data) and Cluster 3 (245 data) the dominant sex is female.



Weka Clusterer Visualize: 22:27:56 - SimpleKMeans (WekaExcel-weka.filters.unsupervised.attribute.Remove-R1,7)

X: Cluster (Nom)

Y: UMUR (Num)

Colour: Cluster (Nom)

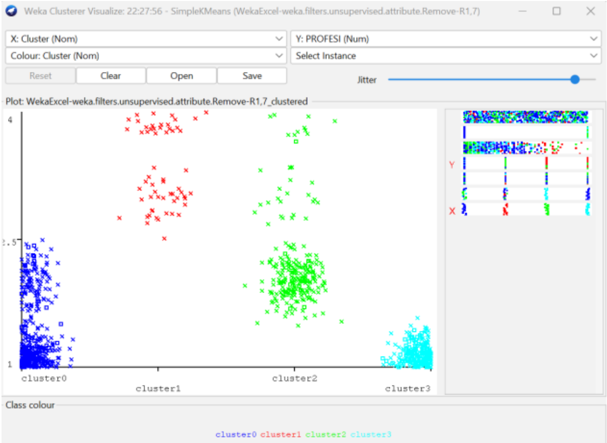
Select Instance

Plot: WekaExcel-weka.filters.unsupervised.attribute.Remove-R1,7_clustered

Figure 5. Visualize the Age Attribute Cluster
[Source: Weka]

Some of the information obtained is as follows:

1. Cluster 0 has an age range between 18-59 years (449 data).
2. Cluster 1 has an age range between 22-79 years (59 data).
3. Cluster 2 has an age range between 18-82 years (247 data).
4. Cluster 3 has an age range between 20-57 years (245 data).



Weka Clusterer Visualize: 22:27:56 - SimpleKMeans (WekaExcel-weka.filters.unsupervised.attribute.Remove-R1,7)

X: Cluster (Nom)

Y: PROFESI (Num)

Colour: Cluster (Nom)

Select Instance

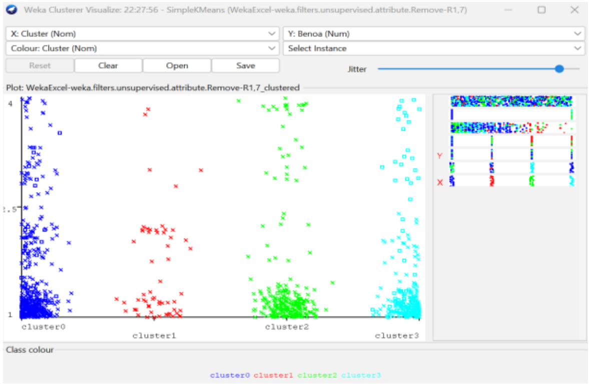
Plot: WekaExcel-weka.filters.unsupervised.attribute.Remove-R1,7_clustered

Figure 6. Visualize Clusters of Profession Attributes
[Source: Weka]

Some of the information obtained is as follows:

1. In Cluster 0, the dominant professions are Profession 1 (Private: 328 data) and Profession 2 (Students: 121 data).
2. In Cluster 1, the dominant professions are Profession 3 (Civil Servant: 35 data) and Profession 4 (Not Working: 24 data).
3. In Cluster 2 the dominant professions are Profession 2 (Students: 213 data), and Profession 3 (Civil Servant: 23 data). And Profession 4 (Not Working: 11 data).
4. In Cluster 3, the dominant profession is Profession 1 (Private: 245 data).

.



Weka Clusterer Visualize: 22:27:56 - SimpleKMeans (WekaExcel-weka.filters.unsupervised.attribute.Remove-R1,7)

X: Cluster (Nom)

Y: Benoa (Num)

Colour: Cluster (Nom)

Select Instance

Plot: WekaExcel-weka.filters.unsupervised.attribute.Remove-R1,7_clustered

Figure 7. Visualize the Continent Attribute Cluster
[Source: Weka]

Some of the information obtained is as follows:

1. Cluster 0 is dominated by Continent 1 (Europe: 322 data), Continent 2 (Asia: 77 data), Continent 3 (America: 34 data), and Continent 4 (Australia: 16 data).
2. Cluster 1 is dominated by Continent 1 (Europe: 35 data), Continent 2 (Asia: 19 data), Continent 3 (America: 3 data), and Continent 4 (Australia 2 data).
3. Cluster 2 is dominated by Continent 1 (Europe: 206 data), Continent 2 (Asia: 12 data), Continent 3 (America: 12 data), and Continent 4 (Australia: 7 data).
4. Cluster 3 is dominated by Continent 1 (195 data), Continent 2 (27 data), Continent 3 (13 data), Continent 4 (10 data).

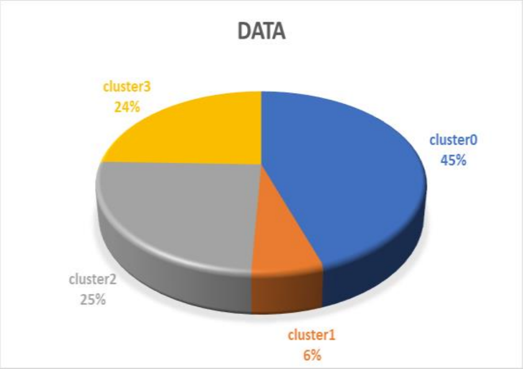


Figure 7. Number of Clusters Diagram
[Source: Microsoft EXCEL]

From the data obtained through clustering, the following results are obtained.

1. Cluster 0 has data with a total of 449 data with the superior room.
2. Cluster 1 has data with a total of 59 data with the economy room.
3. Cluster 2 has data with a total of 247 data with the deluxe room.
4. Cluster 3 has data with a total of 245 data with the standard room.

4. CONCLUSION

Based on the results obtained from the research, it can be concluded as follows:

1. The number of clusters used is 4 according to the room class (as a reference).
2. The K-Means algorithm can be implemented in finding the characteristics of tourists who stay based on 4 main attributes
3. Tests for the attributes that have the most influence on room class are Gender with a value of 0.7076, Profession with a value of 0.39, Age with a value of 0.2083, and Benoa with a value of 0.0465
4. From the test results with K-Means, the accuracy rate reached 84.4% of the 1000 data used. The characteristics that are formed from each cluster are generated as follows.
    a. Cluster 0: Tourists who stay overnight have male sex characteristics, aged 18-59 years, private professions and students, European continent, superior rooms with a total of 449 data.
    b. Cluster 1: Tourists who stay overnight have male sex characteristics, aged 22-79 years, civil servant profession and not working, European and Asian continents, an economy class with a total of 59 data.

c. Cluster 2: Tourists who stay overnight have female gender characteristics, aged 18-82 years, slow students and civil servants, European and Asian continents, and deluxe rooms with a total of 247 data.

d. Cluster 3: Tourists who stay overnight have characteristics of the female sex, aged 20-57 years, private profession, the European American continent, standard room with a total of 245 data

ACKNOWLEDGEMENT

Thank you very much to The Hotel Bendesa which has assisted the authors in conducting this research, especially in collecting the sample data of tourists who stay overnight.

REFERENCES

[1] K. F. Apriyana, I. G. Komala Putra, and G. Indrawan, "*Teknik Data Mining untuk Mendapatkan Pola Transaksi Hotel Bendesa dengan Algoritma J48*," *SENAPATI*, pp. 201–205. 2016.

[2] R. Wulan Sari and D. Hartama, "*Data Mining: Algoritma K-Means Pada Pengelompokkan Wisata Asing ke Indonesia Menurut Provinsi". Seminar Nasional Sains & Teknologi Informasi (SENSASI)*. 2018.

[3] A. Wahyuni and S. Anggraini, "*Implementasi Algoritma J48 Data Mining Untuk Inovasi Bisinis Perhotelan Di Masa Pandemi Covid- 19*". *Jurnal Ilmu Komputer dan Bisnis*, vol. 13, no. 1, pp. 182–192. 2022.

[4] T. Hardiani, "*Analisis Clustering Kasus Covid 19 di Indonesia Menggunakan Algoritma K-Means*". *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 2, pp. 156–165. 2022.

[5] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta. 2017.

[6] L. Petra Refialy, H. Maitimu, and M. Soyano Pesulima, "*Perbaikan Kinerja Clustering K-Means pada Data Ekonomi Nelayan dengan Perhitungan Sum of Square Error (SSE) dan Optimasi nilai K cluster*". Techno.COM, Vol. 20, No. 2. 2021.

[7] R. Kesuma Dinata, N. Hasdyna, and N. Azizah, "*Analisis K-Means Clustering pada Data Sepeda Motor*". Informatics Journal Vol. 5 No. 1. 2020.

[8] K. Aprilia and F. Sembiring, "*Analisis Garis Kemiskinan Makanan Menggunakan Metode Algoritma K-Means Clustering*. SISMATIK (Seminar Nasional Sistem Informasi dan Manajemen Informatika). 2021.

[9] A. Rohmah *et al.*, "*Implementasi Algoritma K-Means Clustering Analysis Untuk Menentukan Hambatan Pembelajaran Daring (Studi Kasus: Smk Yaspim Gegerbitung*)". SISMATIK (Seminar Nasional Sistem Informasi dan Manajemen Informatika). 2021.

[10] Z. Nabila, A. Rahman Isnain, and Z. Abidin, "*Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means*". *Jurnal Teknologi dan Sistem Informasi (JTSI)*, vol. 2, no. 2, p. 100. 2021.

[11] R. T. Vulandari, *Data Mining: Teori dan Aplikasi Rapidminer*. Yogyakarta: Gava Media. 2017.

.