IMPLEMENTATION OF CLUSTERING METHODS FOR GROUPING TOURISM POTENTIAL IN KABUPATEN SUMEDANG

M Egypt Pratama[1], Alif Finandhita[2]

[1.2]Information Engineering - University Computer Indonesia
Jl. Dipatiukur No. 112-116, Coblong, Bandung 40132
E-mail: Egypt97jip@gmail.com[1], alif.finandhita@email.unikom.ac.id[2]

ABSTRACT

Sumedang has a vision as a cultural center in West Java has the potential of traditional elements, either culture, typical food, traditional crafts, in addition to farming, forestry and agriculture. It became one of the pull factors of tourists to visit Sumedang. The increase in tourist arrivals will give good influence to the surrounding community. On the other hand, in spite of the development efforts in increasing tourist arrivals, there are problems such as lameness number of tourist attractions in the center and edge of town, so it takes even distribution of tourist traffic. Therefore, necessary efforts equalization tourists visiting one of which can be done by classifying tourist attraction that has the potential to be developed. Clustering methods in data mining is the right solution to overcome the problems Attraction grouping. By applying the k-means algorithm Attraction group formation begins by determining the number of clusters that will be in the form, specify the value of the centroid of each cluster, calculate the distance between data and taking into account the minimum distance classifying objects. Through data mining applications with clustering method Attraction grouping process will be easier and produce some recommendation attractions that can be developed. and grouping taking into account the minimum distance of the object. Through data mining applications with clustering method Attraction grouping process will be easier and produce some recommendation attractions that can be developed. and grouping taking into account the minimum distance of the object. Through data mining applications with clustering method Attraction grouping process will be easier and produce some recommendation attractions that can be developed.

Keywords : *Clustering*, Data Mining, Data Potential, K-Means, visit Tourism.

1. INTRODUCTION

Sumedang is one of the districts in West Java and is directly adjacent to the provincial capital, Bandung. Sumedang District has an area of 153 124 ha and a population of nearly 1 million people, holds enough natural resources potential [1], Especially the potential in the tourism sector, special food, traditional crafts, in addition to farming, forestry and agriculture. Sumedang District has a lot of potential to become a tourist area. As seen from the tourist traffic of more than 1 million people [1].

Dinas pariwisata kebudayaan kepemudaan dan olahraga (DISPARBUDPORA) Sumedang is the agency in charge of managing tourism in Sumedang. According to the Chairman of the Section of Objects and attractions in tourism development in Sumedang is still limited in the budget. This is because in the delivery of services divided by 4 fields. With this budget constraint the Department of Tourism made attractions priority groups that will be developed.

From year to year, the number of tourists visiting Sumedang increased compared to the previous year [1]. In the meeting conducted annually attractions DISPARBUDPORA to determine the priority group to develop the agency created manually by choosing the attractions just by guessing if it is developed tourist spot will increase the number of tourist visits.

But there are problems encountered in the development of tourist attractions in Sumedang to absorb more of the number of tourist visits. One of the problems faced by imbalances between the number of tourists visiting attractions at the center and the suburbs, this is indicated by the buildup of tourism activities in the area and surrounding Buahdua. This buildup can be seen by comparing the number of tourists visiting Kampung Toga located in the city center during the year 2018 was 11 421 people, while the number of visits Cigumentong located adjacent to the southern city amounted to 2,535 people. Therefore, necessary efforts equalization tourists visiting one of which can be done by classifying tourist attraction that has the potential to be developed by the method of clustering.

Data mining is a method for finding new information that is useful from a collection of large amounts of data and can help in making decisions [2]. In data mining, there are several methods of estimation, prediction, classification, clustering and

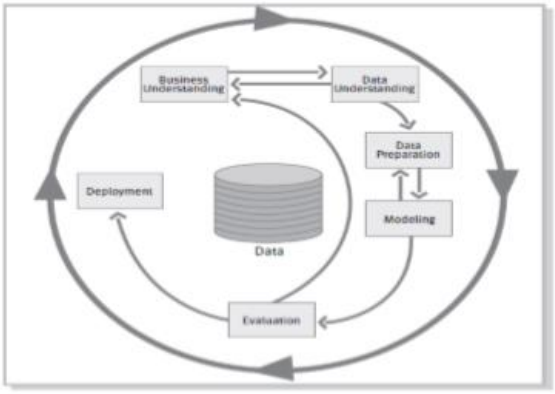
association. Clustering is a technique of classifying data by separating the data into groups according to certain desirable characteristics which the label of any data not yet known and the grouping is expected to know the groups of data for later labeled according problems faced

Based on the description of the background of the problem, then DISPARBUDPORA requires a system that uses clustering method to obtain a knowledge that is used as a reference or recommendation in determining the development planning of tourist attractions that occur equalization tourist visits. With the aim to help the DISPARBUDPORA especially section objects and tourist attraction in the Attraction grouping that will be developed.

2. RESEARCH CONTENT

2.1 Data Mining

The method to be used for the completion of data mining in this study is a framework of CRISP-DM [2], The stages of the CRISP-DM framework can be seen in Figure 1 with includes several processes:



Picture 1, Stages of CRISP-DM

The explanation phase of CRISP-DM framework, among others [3]:

1. Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then transform this knowledge into a data mining problem definition and the initial plan was designed to achieve these objectives. Phase understanding of business is divided into several steps, among others:

a. Determining Business Objectives
At this stage it will produce output in the form of background and objectives of the business.

b. situation assessment
At this stage it will produce output in the form of inventory resources, constraints and risks, as well as the costs and benefits.

c. Determination of Interest Data Mining
At this stage it will generate output target Data Mining.

d. Creating a Project Plan
At this stage it will produce output in the form of a project plan and a preliminary assessment for the use of tools and techniques.

2. Data Understanding

On the data understanding phase starts with the initial data collection and understand the data that will be used in research [3], Phase understanding of data is divided into several steps, among others:

a. Initial Data Collection
This step will be collected at the data to be used in data mining.

b. explanation of Data
This step will be explained on the data collected in the early stages. In the explanation of the data will be described the format of the data, the quantity of data, the number of records and fields in each table, and so on.

c. exploration Data
At this stage of data analysis will be done already described in the previous stage. One way to explore the data is using descriptive statistical analysis [4]. Descriptive statistical analysis relating to the search for values below:

1) The average value using equation (2.1) the following:

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{n}$$

2) Standard deviation value using equation (2.2) the following:

$$\sigma = \sqrt{\frac{\sum(xi - \bar{x})^2}{n}}$$

3) Value quartile 1, 2, and 3 by using equation (2.3) the following:

$$Letak\ Q_i = \frac{i(n+1)}{4}$$

4) Value interquartile range (IQR) using equation (2.4) the following:

$$IQR = Q_3 - Q_1$$

d. Verification of Data Quality
Stage to determine the quality of the data.

e. Data conversion
Stage of normalization with the min-max method.

3. Data Preparation

Data preparation phase covers all activities necessary to build the data to be used in modeling. The preparation phase data is divided into several steps, among others:

a. Data Selection
Stage to select attributes of the data [4],
b. Data cleaning
Data cleansing phase aims to eliminate data.
c. Preparation of Preliminary Data
At this stage will be set up data that has been selected and cleared for use in the modeling stage.
d. Data integration
At this stage the data is aggregated according to the need for modeling.
e. Data format
At this stage formatted data structure will be adjusted to the data required for modeling.

4. modeling

At this stage it will be election and the application of the appropriate model based on objectives to be achieved [5], Modeling stage is divided into several steps, among others:

a. Selection Modeling Techniques
This stage is used to select the appropriate modeling techniques to the problems and objectives to be achieved.
b. Modeling
At this stage will be described modeling techniques have been at the stage of selecting modeling techniques.
c. testing Model
At this stage the model that has been chosen will be applied to the test case.

5. Evaluation

At this stage of evaluation of the models that have been made in the modeling stage, to ensure that the model really achieve business goals. In the evaluation phase, there are several steps, among others:

a. evaluation Results
b. The review process
c. Determining the Next Steps

6. deployment

Bedasarkan model creation analysis. In the deployment phase, there are several stages:

a. planning deployment
At this stage it will produce output in the form of planning will be done at this stage of deployment.
b. Monitoring and Maintenance Plan
At this stage the plan will be monitored and maintained.
c. Creating a Final Report
At this stage it will produce output in the form of a final report and final presentation.
d. review Project
At this stage will result in a review of the project documentation has been done.

2.2 Data Preprocessing

In data mining, data quality that will be used to note. Stage in the attention to the quality of data that is data Preporcessing. [2], Some of the activities that will be done preprocessing of data in this study are as follows[5]:

a. Handling Missing value
b. handling Outliers

2.3 Clustering

Clustering or clustering is the process of classifying the data set into several groups or clusters so that objects within a cluster have high kimiripan, but very different from the other cluster [5], Clustering is widely used in the field with a variety of vital applications, such as market research, which is used to market and customer segmentation.

2.4 K-Means

Method of k-Means clustering algorithm is the oldest and most widely used in a variety of small to medium applications [6],

This method of classifying bedasarkan same data characteristics.

Grouping the data with K-means:

1. Determine the number of clusters.
2. Determining the value of the centroid in each cluster. To determine the centroid of the initial value of the iteration is done randomly.
3. Calculate the distance of each data at each centroid using Euclidean formula, the following formula:

$$d(x_j, c_j) \sqrt{\sum_{j=1}^{n} (x_j - c_j)^2}$$

d = distance j = the number of data c = centroid x = Data

4. Grouping objects to determine the cluster members to calculate the distance to the cluster center.
5. Go back to step 2, if there are data to move the group, until there is no more data to move the group.
6. Show Results as for the selection flowchart of K-Means algorithm can be seen in Figure 2:

Picture 2, Flowchart Algorithm K-Means

2.5 Understanding Business

The business goals of DISPARBUDPORA is to improve and develop the attraction.

Target setting data mining consists of goals and success criteria for data mining. Here is the target of data mining:

1. The purpose of data mining
   The application of data mining can help section objects and attractions within attractions clustered Attraction so that development can be precisely targeted and meet the target.
2. Criteria for data mining success
   Criteria for success against this research that the group that was formed could be recommendations to set priorities in the development of tourist attractions.

2.6 Understanding Data

This stage is the second stage in the method of the CRISP-DM, several processes including:

1. Data collection
   Data used in the study of the traffic data objects and data tourist attraction potential in 2018. Data object visit a tourist attraction that will be used in the form of Excel files (* .xls format).
2. explanation of Data
   Visit data objects used tourist attraction has 7 attributes, the number of records 36 while data on the potential of having five attributes with the record number of tourist visits 36. Data Attributes: An object name, location, Quarter 1 Quarter 2 Quarter 3 Quarter 4 Total , Average. Tourism potential data attributes: Name ODTW, Object Type, Facilities / infrastructure, attractiveness, Distance from City Center Sumedang.
3. exploration Data
   This phase helps achieve the goal of data mining, analysis descriptive statistics were used in this study [7] :

Table 1 exploration Data

No.	Attribute name	value min	Max value	Average value	standard deviation	IQR
1	total	2535	19213	9315	3477	3688
2	Infrastructure	1	5	3.2	1.2	2
3	Attractiveness	1	5	2.6	0.7	1
4	Distance	2	5	4.3	0.7	1

4. Verification of Data Quality
   This phase will examine the missing values and outliers. This stage uses interquartile range. The results of this phase:

Table 2 Verification of Data Quality

No.	Attribute name	Upper limit	Lower limit
1	total	21 728	-400
2	Infrastructure	10	-2
3	Attractiveness	6	0
4	Distance	8	2

With the verification of data quality can be concluded:

1. In the attributes of the number of visits is no outlier.
2. In the attributes of facilities / infrastructure no outlier.
3. In the appeal attribute no outlier.
4. In the distance attribute was no outlier.
5. Attraction of traffic data and the data does not have the potential for missing Value

2.7 Preparation of Data

This phase covers all activities to produce the final dataset [8], Some of these activities:

1. Data Selection
   Attributes used in the final dataset that will be used on the model number, facility / infrastructure, distance, allure
2. Data cleaning
   This stage is based on the stage of verification of data.
3. Data development
   This phase will be the construction of a data selection stage bedasarkan.
4. Merging Data
   Data that is designed to be implemented in this hold.
5. Data format
   The final stage before the start modeling.

2.8 Modeling

Modeling of the fourth stage in the method of CRISP-DM [8], This stage consists of modeling techniques, the analysis of the modeling and model testing

1. Modeling techniques
   The model used is k-Means. Modeling
2. testing Model

Model K-Means has stages and parameters.

3. testing Model

Clustering method aims for the establishment of a data group (cluster) of the data set is not known factions.

Step-by-step process of analysis using k-means algorithm is as follows:

1. Determine the number of clusters is 3 clusters (k = 3) will be made.
2. Determine the centroid of each cluster.
3. Random initialization centroid point. Following this initial centroid point to be used

Table 4 Centroid early

	JK	J	S / P	DT
C1	.280	0,000	0,000	0,000
C2	0.403	0.547	0.547	0.399
C3	1	1	1	1

4. Calculate the distance data by centroid. The formula for calculating the distance is as follows: An example of the distance calculation using data from the centroid Euclidean equation below in the picture

Jarak data dengan cluster 1 adalah :

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^{n}(1-0{,}280)^2 + (1-0)^2 + (1-0)^2 + (0{,}5-0)^2} = 1{,}664$$

Jarak data dengan cluster 2 adalah :

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^{n}(1-0{,}403)^2 + (1-0{,}547)^2 + (1-0{,}547)^2 + (0{,}5-0{,}399)^2} = 0{,}881$$

Jarak data dengan cluster 3 adalah :

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^{n}(1-1)^2 + (1-1)^2 + (1-1)^2 + (0{,}5-1)^2} = 0{,}500$$

Picture 3 Calculate the distance Centroid

5. Calculate the shortest distance to the centroid point, then classify the data according to the smallest distance to the centroid point.

Table 5 Iteration 0

Name Attractions	JK	J	S / P	DT	**c1**	**c2**	**c3**	class
Cipanas Sekarwangi	1	1,000	1,000	.500	**1.664**	**.881**	**.500**	3
Cipanas Cileungsing	0.767	1,000	1,000	.500	**1,577**	**0.744**	**0.552**	3
CURUG SINDULANG	0.403	.500	.500	.250	**.760**	**0.163**	**1,191**	2
MOUNTAIN KEY	.536	.500	.500	.250	**0.792**	**.210**	**1,131**	2
TOMB Dayeuh LUHUR	0.398	.250	.250	.500	**0,624**	**0.433**	**1,318**	2
TOMB OF THE MOUNTAIN LINGGA	.371	.250	.250	.250	**.443**	**0,447**	**1,443**	1
TOMB MARONGGE	0,308	.250	.250	.250	**.434**	**.456**	**1,472**	1
Museum Prabu Gues Ulun	37	0	0	0	**20**	**0**	**0**	
Prince rock	0.279	.250	.250	0,000	**0.354**	**0.592**	**1,626**	1
VILLAGE TOGA	0.533	.750	.750	.500	**1,200**	**0.331**	**.770**	2

6. Process back to step No. 2. For the next step, a new centroid is calculated by calculating the average value in each cluster. If the new centroid is different from the centroid at the beginning, then the process is continued to the next step. However, if the newly calculated centroid equal to the centroid at the beginning, then the clustering process is completed. Examples of new centroid calculation using the equation for Total Visits

Examples of new centroid calculation using the equation for Distance.

$$C1 = \frac{0{,}371 + 0{,}308 + 0{,}279 + 0{,}369 + 0{,}294 + 0}{6} = 0{,}270$$

$$C2 = \frac{9{,}876}{27} = 0{,}365$$

$$C3 = \frac{1 + 0{,}767 + 0{,}609 + 0{,}620}{4} = 0{,}749$$

Picture 4 Counting the new centroid

Examples of new centroid calculation using the equation for Distance.

Table 6 new Centroid

	JK	J	S / P	DT
C1	0,270	0,125	0,125	.208
C2	0,365	.574	.574	0,407
C3	.749	1,000	1,000	0,625

7. Then repeat steps 4 through centroid unchanged.
8. Clustering process is completed. Here are the results of clustering

Table 7 Results Clustering

Name Attractions	**C1**	**C2**	**C3**
Cipanas Sekarwangi			*
Cipanas Cileungsing			*
CURUG SINDULANG		*	
MOUNTAIN KEY		*	
TOMB Dayeuh LUHUR	*		
TOMB OF THE MOUNTAIN LINGGA	*		
TOMB MARONGGE	*		
Museum Prabu Gues Ulun		*	
Prince rock	*		
VILLAGE TOGA		*	
total	**4**	**4**	**2**

Conclusions from the analysis that has been done is every tourist spot placed on the appropriate cluster cluster parameters

Table 8 Conclusion of clustering

cluster	Id_tempatwisata
cluster 1	5,6,7,9,11,12,16, 17,18,19,21,26,
cluster 2	3,4,8,10,13,14,22,24 , 25,27,28,31,32,33,34,35,36,37
cluster 3	1,2,15,20,23,29,30

Thus it can be interpreted as follows:

Table 9 Conclusions clustering

cluster	Results Cluster
cluster 1	Clusters that have a high priority for development
cluster 2	Clusters that have the potential for an event held to increase the number of visits and popularity
cluster 3	Clusters that have the potential and the high number of visits

2.9 Functional Needs Analysis

1. diagram Context

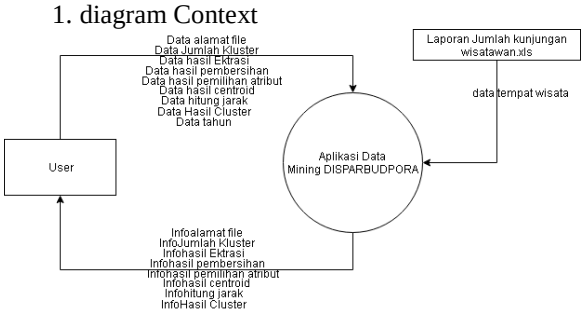


Figure 5 Diagram Context

2. ERD

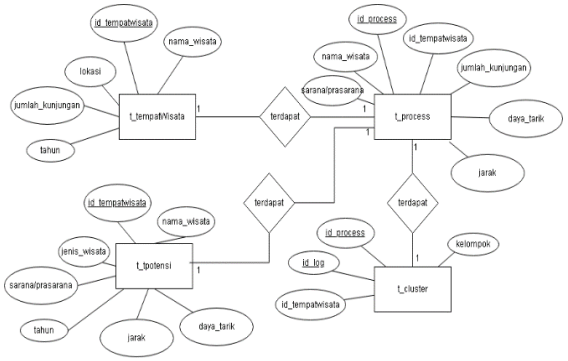


Figure 6 ERD

3. Relationship diagram

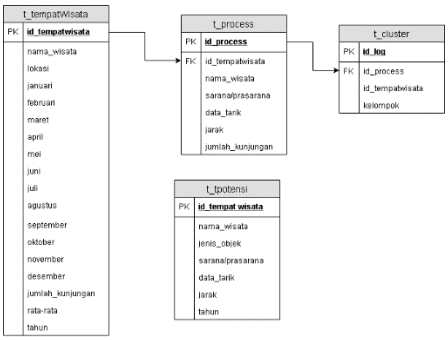


Figure 7 Diagram Context

2:10 System Implementation

Interface implementation is created and encoding form of program files.

1. Interface Home

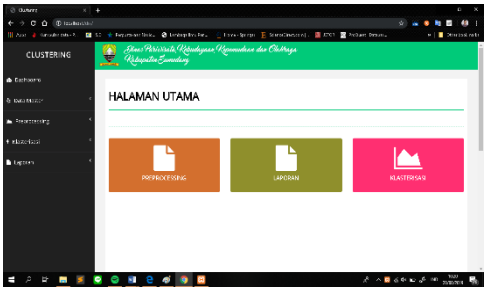


Figure 8 Home

2. Input data travel kunjugan

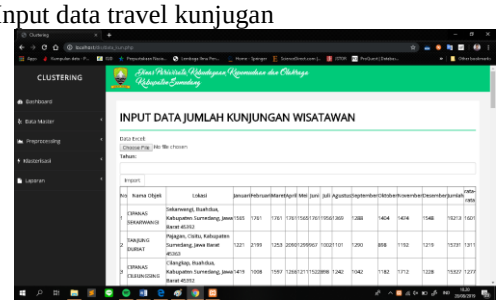


Figure 9 Input data tourist arrivals

3. Input data tourism potential

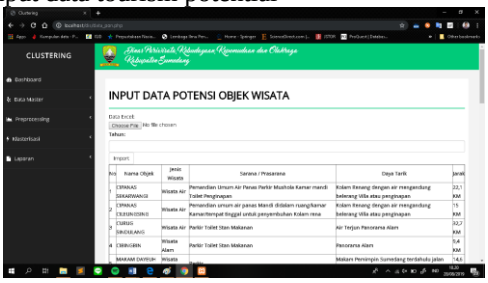


Figure 10 Input data Tourism Potential

4. Praprocessing

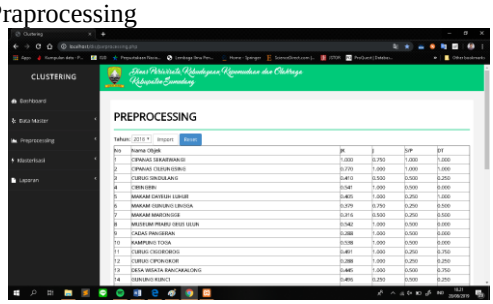


Figure 11 Praprocessing

5. clustering

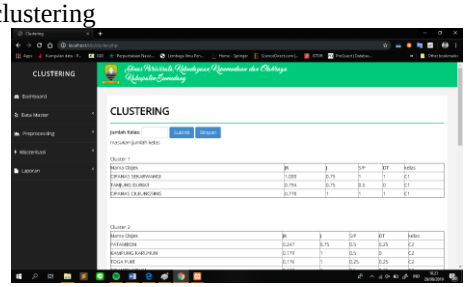


Figure 12 Kalterisasi

6. Report

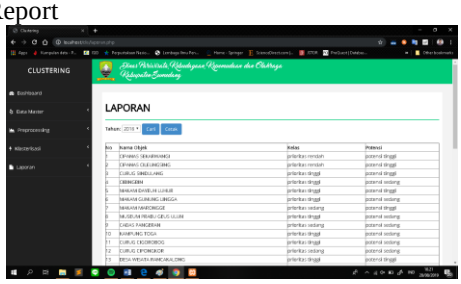


Figure 13 Reports

3. COVER

3.1 Conclusion

The conclusion from this study that the application is built for a section objects and attractions can establish the priority groups attractions will be developed according to traffic data in travel and tourism potential.

3.2 advice

There are some suggestions that can be done to develop applications forming study groups, among others:

1. Input format for import into the system not only excel format (.xls), but could be in the form of pdf, csv or .txt format.
2. Applying text mining to convert the data that the data contain free text (free text entries) into numerical data

Similarly, suggestions to the authors give, hopefully these suggestions bias used as fill material that can be useful for writers in particular and generally for the public at large

THANK-YOU NOTE

The author realizes that in this research would not be possible without the support, help and input from various parties. Therefore, in this moment I also wanted to convey a sense of respect and gratitude - as possible to:

1. Allah SWT has given mercy, grace, health, smoothness and ease for writers in completing this thesis research.
2. Both parents, and extended family who continues to provide support in the form of prayers, motivation and enthusiasm and support material that cannot be counted.
3. Dr. Ir. H. Eddy Soeryanto Soegoto, M.Sc., as the Rector of the University Computer Indonesia.
4. Dr. Ir. Herman S. Soegoto, MBA., As Dean of the Faculty of Engineering and Computer Science, University Computer Indonesia.
5. Mrs. Nelly Indriani Widiastuti, S.Si., MT, as Chairman of Information Engineering Program, University Computer Indonesia.
6. Mr. Alif Finandhita, Kom., MT as the lecturers who have provided guidance, direction, feedback and all the sciences to the author until the preparation of this thesis can be resolved properly.
7. Ms. Dian Dharmayanti, ST, M. Kom., As faculty reviewer who has been providing direction in the preparation of this thesis.
8. Mrs. Ken Kyla P, Kom., MT, as the examiner who has been providing direction in the preparation of this thesis.
9. Mr. Taryana Suryana, ST, M. Kom., As the guardian lecturer IF-11 class.
10. Mr. Ajat Sudrajat, S.Ip., as Head Section Pull Objects and travel which has provided support to carry out this research.
11. Friends - Friends IF-11 class of 2015 continues to provide prayer and support to writers in completing this thesis.
12. Friends of Mr. Alif Finandhita student guidance, Kom., MT, which continues to provide support, share your thoughts, support and prayers to the author so you can complete this thesis.

REFERENCES

[1] B. Sumedang Sumedang Regency in Figures 2018, Sumedang: BPS Sumedang, 2018.

[2] APW a. DHMG Sadewo, "Application of Datamining In Meat Broiler population in Indonesia by Province Using K-Means," InfoTekJar (Journal Nas. Inform. And Teknol. Network), vol. 2, pp. 60-67, 2018.

[3] R.Wulan, "Data Mining Algorithm K-Means Clustering On Foreign Travel to Indonesia by Province," Pemantangsiantar.

[4] Dian dd Dharmayanti, "Modeling of Data Warehouse At the Department of Information Engineering Master," UNIKOM Scientific Magazine, vol. 12, no. 2, pp. 151-168, 2015.

[5] Irawan AHAFSA Afrianto, "E-Document Autentification With Digital Signature," pp. 1-8, 2018.

[6] II Alif Finandhita, "Development of E-Diplomas System Model with Digital Signature Authentication," IOP Conf. Series: Materials Science and Engineering, vol. 407, pp. 1-7, 2018.

[7] M. Ginting, "clustering Shopping Cart Sales Transactions Using K-Means Clustering," Medan.

[8] AMB Nadiar Ahmad Syaripul, "Open Data

Interactive Data Visualization city administration," Journal of Information Systems (Journal of Information Systems), vol. 2, no. 12, pp. 82-89, 2016.

[9] MDHS Dr. Riduwan, Statistics for Research Education, Social, Economics, Communication, and Business, Bandung 2009.

[10] AP Windarto, "Application of Data Mining In Fruit Exports by Country of Destination Using K-Means Clustering," Techno.COM, vol. 16, pp. 348-357, 2017.

[11] J. S, Data Mining Algorithm and Implementation with PHP Programming ", Jakarta, in 2019.

[12] KT Suryana., Internet applications using HTML, CSS, and JavaScript, Bandung, 2015.

[13] M. Sidi, "Testing Applications Using Black Box Testing Boundary Value Analysis," 2015.

[14] L. Maulida, "Application of Datamining In Grouping tourist arrivals All Attractions Featured In Prov. DKI Jakarta With the K-Means, "in 2018.