

The Grouping Of Types Of Tax Revenue In Binjai City Uses The K-Means Clustering Method Algorithm

Anita Renni Arianti¹, Novriyenni², Indah Ambarita³

¹²STMIK Kaputama

anitarenni017@gmail.com^{1*}; novriyenni.sikumbang@gmail.com²; yesnovada@gmail.com³

Abstract

Taxes are a major source of state revenue. Without taxes, most activities of the country are difficult to implement. The use of tax money includes ranging from employee shopping to financing various development projects. Public utilities such as roads, Bridges, schools, hospitals/center facilities, police stations are financed with tax returns. Tax money is also used for financing to provide a sense of security for all levels of society. Every citizen from birth to death enjoys facilities or government services all funded with tax money.

Keywords: *Datamining, Grouping Types of Tax Revenue, K – Means*

1. Introduction

Data of tax payment by tax payers will be recorded into the bpkpad database. Tax payments will continue to grow each year, this pile of data will be very useful if data retrieval or data management can provide new sources of information, making it invaluable for the development of tax revenue in the city of binjai.

Based on those data, the author is interested in conducting research to manage tax type acquisition data using data mining techniques. Data mining is a process of enormous amounts of information and data gathering, or big data, to manage tax revenues data so it can produce useful information. Thus, a method of getting through the process of clustering is needed. Clustering methods constitute a function that leads to the identification process of groups, then products or items that have special characteristics. It is usually used in knowing specific groups in their spread. It is hoped that this method will be able to data collection of tax revenues in the city of binjai. The result of this grouping will result in a type of tax collection, whereby obedient tax payments are paid according to the type of tax effort, and bpkpad can also identify the most extensive tax collection and few tax deposits. Data of tax payment by tax payers will be recorded into the bpkpad database. Tax payments will continue to grow each year, this pile of data will be very useful if data retrieval or data management can provide new sources of information, making it invaluable for the development of tax revenue in the city of binjai. The results of this grouping will produce information on the grouping of types of tax revenue, namely those who pay taxes based on the type of business of the taxpayer, besides that BPKPAD can find out the grouping of types of tax revenue that have the most and pay little taxes [3].

2. Research Methods

2.1. Data Mining

The process of scientific discovery on the database is often called mining data or data mining. Data mining is an analytical step toward a generally large cluster of data to connect those data to and summarize them in easily understood and used forms. The hedges and husks generated are typically models (models) or patterns (patterns) [5].

2.2. Clustering

Clustering is a data analysis method, which is often included as one of the data mining methods, the purpose of which is to grouping data with the same characteristics [1].

According to hermawati [6] analysis of the cluster (clustering) is finding objects in one group of equals (getting in touch) and being (getting in touch point) with objects in another group. The primary purpose of clustering is the grouping of a number of data/objects into clusters so that each cluster will contain as closely data as possible. Clustering methods seek to place similar objects (close distances) in front of groups and draw group distances as far as possible. It means that objects in one group are very similar to each other and differ from objects in other groups [6].

2.3. K-Means Algorithm

The k-means algorithm is an infinitely simple algorithm for classifying or grouping large amounts of objects with certain attributes into groups (clusters) as much as k. In k-means, the number of k clusters has been determined in the past [2].

According to Prasetyo (2012, h.189) it stated that k-means is one of the methods of non hierarchical grouping (sephos) that attempts to partition the data into clusters so that the data that has the same characteristics will be included in one cluster and that data that has different characteristics are grouped into another group [4].

The K-Means algorithm in cluster formation is as follows:

1. Suppose given a matrix $X = \{X_{ij}\}$ data of size $n \times p$ with

$i = 1, 2, \dots, n, j = 1, 2, \dots, p$ and assume the number of initial clusters is K.

2. Determine the centroids.

3. Calculate the distance of each object to each centroid using the deEuclidean distance or can be written as follows:

$$J(x_i c_i) = \sqrt{(x_i - c_i)^2} \quad (1)$$

4. Each object is arranged to the nearest centroid and the collection of objects will form a cluster.

5. Determine the new centroid of the cluster that will be formed, in which object the new centroid is obtained from the average of each that lies in the same cluster.

6. Repeat step 3, if the initial and new centroids are not the same.

There are several appropriate ways to measure the distance of data to the center of the group, including Euclidean, Manhattan/City Block, and Minkowsky, each method has advantages and disadvantages. Measuring distances in the de-Euclidean distance space uses the formula:

$$D(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^p |x_{2j} - x_{1j}|^2} \quad (2)$$

Definition 2.1 :

D is the distance between data x_2 and x_1 , and $|\cdot|$ is an absolute value. Distance measurement on the Manhattan distance using the formula:

$$D(x_2, x_1) = \|x_2 - x_1\|_1 = \sum_{j=1}^p |x_{2h} - x_{1h}| \quad (3)$$

Definition 2.2 :

Measuring distances on the Minkowsky distance using the formula:

$$D(x_2, x_1) = \|x_2 - x_1\|_\lambda = \sqrt[\lambda]{\sum_{j=1}^p |x_{2h} - x_{1h}|^\lambda} \quad (4)$$

Definition 2.3 :

λ is the Minkowsky distance parameter. The most widely used method is Euclidean and Manhattan. Euclidean is the choice if we want to provide the shortest distance between two points (straight distance), as shown in the Euclidean formula. Whereas Manhattan provides the furthest distance in the two data [7].

3. Results And Discussion

3.1. Research methods

In solving a problem in research, of course the researcher must have a way or a method that will be applied in solving the problem so that the research carried out can be completed properly and in accordance with the expected results. The research method is carried out to look for something systematically by using the scientific method and applicable sources. On the basis of the research methodology used in this study, namely the grouping of types of tax revenue in Binjai City using the k-means clustering method.

3.2. Process Design

The system process design flow that will be built with data mining grouping types of tax revenue in Binjai City using the clustering method can be described in the form of a flowchart as shown below.

3.3. Flow chart

The process or workflow of the tax revenue grouping system in Binjai City using the clustering method can be seen in the form of a flowchart as shown in the figure below [8].

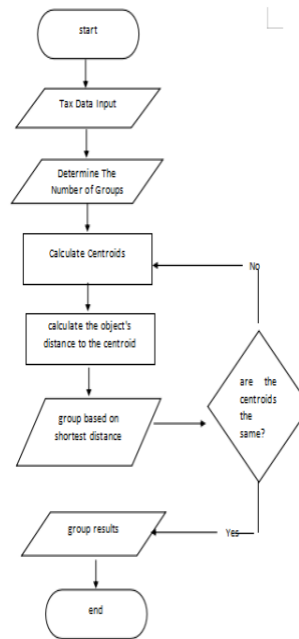


Fig. 1: Flowchart of the Clustering Method

3.4. Research Supporting Data

To analyze data in a study, supporting data is needed so that a research can run as expected. The K-Means algorithm clustering method in obtaining new information about the grouping of types of tax revenue in Binjai City. The data provided are data on tax revenue data in the City of Binjai. Based on research that has been carried out at the Regional Income and Asset Financial Management Agency, data is obtained which will later be used to analyze the classification of types of tax revenue in Binjai City. The data is as in the table below.

Table 1: Tax Revenue Data

No	Object	type of Tax Revenue	Type of business	Total Tax Revenue (Rp)
1	A	Entertainment Tax	Entertainment	174.301.400,00
2	B	Entertainment Tax	Entertainment	6.050.000,00
3	C	Entertainment Tax	Entertainment	1.100.000,00
4	D	Entertainment Tax	Entertainment	45.703.200,00
5	E	Advertisement tax	Retail Stores	3.185.753
6	F	Advertisement tax	resto	11.696.953
7	G	Advertisement tax	resto	1.175.300
8	H	Advertisement tax	resto	11.696.953
9	I	Swallow's Nest Tax	Retail Stores	804.000
10	J	Swallow's Nest Tax	Swallow	800.000
11	K	Swallow's Nest Tax	Swallow	4.500.000
12	L	NON PLN Street Lighting Tax	factory	867.516
13	M	NON PLN Street Lighting Tax	Retail Stores	408.240
14	N	NON PLN Street Lighting Tax	Retail Stores	3.606.120
15	O	Restaurant tax	Retail Stores	793.691.054
16	P	Restaurant tax	resto	793.691.054
17	Q	Restaurant tax	resto	40.290.840
18	R	Groundwater Tax	factory	132.640
19	S	Groundwater Tax	Retail Stores	2.154.556
20	T	Groundwater Tax	factory	2.074.875

a. Application of the Method

In using the clustering method, the initial process carried out to form clusters is to transform data into numeric form with predetermined codes, then determine the number of groups (K), calculate the centroid, calculate the distance of the object to the centroid and then group it based on the closest distance. if no objects are moved or grouped then the iteration is finished.

To determine the group of an object, the first thing to do is measure the Euclidean distance between two object points (X, Y and Z) which is defined as follows:

Table 2: Types of Tax Revenue

code	Type of tax revenue (X)
1	Groundwater Tax
2	Entertainment Tax
3	NON PLN Street Lighting Tax
4	Advertisement tax
5	Restaurant tax
6	Swallow's Nest Tax

Table 3: Types of Business

code	Type of Business (Y)
1	agency
2	Swallow
3	Entertainment
4	Service
5	Contractor
6	Automotive
7	Factory
8	Banking
9	Company
10	construction
11	resto
12	Retail
13	Retail Stores

Table 4: Total Tax Receipts

code	Total tax revenue (Z)
1	0 - Rp. 200.000.000
2	Rp. 200.000.001 - Rp. 400.000.000
3	Rp. 400.000.001 - Rp. 600.000.000
4	Rp. 600.000.001 - Rp. 800.000.000
5	Rp. 800.000.001 - Rp. 1.000.000.000
6	> Rp. 1.000.000.000

Then transform the criteria data above to be calculated using the clustering method. The transformation data from the data above can be seen in the table below.

Table 5: Data Transformation

No	Object	X	Y	Z
1	A	2	3	2
2	B	2	3	2
3	C	2	3	2
4	D	2	3	2
5	E	4	13	4
6	F	4	11	4
7	G	4	11	4
8	H	4	11	4
9	I	6	13	6
10	J	6	2	6
11	K	6	2	6
12	L	3	7	3
13	M	3	13	3
14	N	3	13	3
15	O	5	13	5
16	P	5	11	5
17	Q	5	11	5
18	R	1	7	1
19	S	1	13	1
20	T	1	7	1

Then form a cluster into 3 groups ($K = 3$) and determine the centroid center point. The clustering calculation process is as follows.

$K=3$ Centroids

$C1 = (5, 11, 1)$ is taken from Q data

$C2 = (1, 7, 1)$ is taken from R data

$C3 = (1, 13, 1)$ is taken from S data

So $K=3$ Centroids

$C1 = (4.63; 11.75; 1.25)$

$C2 = (2.78; 4.11; 1)$

$C3 = (2.33; 13; 1)$

Iteration ke 3

From the calculation above, the results of the iteration 2 calculation are obtained, which are as shown in the table below.

Tabel 1: Hasil Iterasi 3

No	Object	X	Y	Z	C_1	C_2	C_3	Group
1	A	2	3	1	9,14	1,36	10,01	2
2	B	2	3	1	9,14	1,36	10,01	2
3	C	2	3	1	9,14	1,36	10,01	2
4	D	2	3	1	9,14	1,36	10,01	2
5	E	4	13	1	1,42	8,97	1,67	1
6	F	4	11	1	1,01	7	2,60	1
7	G	4	11	1	1,01	7	2,60	1
8	H	4	11	1	1,01	7	2,60	1

9	I	6	13	1	1,88	9,45	3,67	1
10	J	6	2	1	9,85	3,85	11,60	2
11	K	6	2	1	9,85	3,85	11,60	2
12	L	3	7	1	5,03	2,90	6,04	2
13	M	3	13	1	2,07	8,89	0,67	3
14	N	3	13	1	2,07	8,89	0,67	3
15	O	5	13	2	1,51	9,22	2,85	1
16	P	5	11	2	1,13	7,31	3,48	1
17	Q	5	11	1	0,88	7,24	3,33	1
18	R	1	7	1	5,98	3,39	6,15	2
19	S	1	13	1	3,84	9,06	1,33	3
20	T	1	7	1	5,98	3,39	6,15	2

After calculating using the cluster formula in iteration 2, the groups based on the minimum distance to the nearest centroid are:

old group : (2 2 2 2 1 1 1 1 1 2 2 2 3 3 1 1 1 2 3 2)

new group : (2 2 2 2 1 1 1 1 1 2 2 2 3 3 1 1 1 2 3 2)

After calculating using the existing cluster formula, in iteration 2 it is the same as in iteration 1 and there is no data that moves groups again so that the calculation can be stopped. So that a cluster graph can be made for grouping types of tax revenue in Binjai City using the clustering method.

Clustering Graph

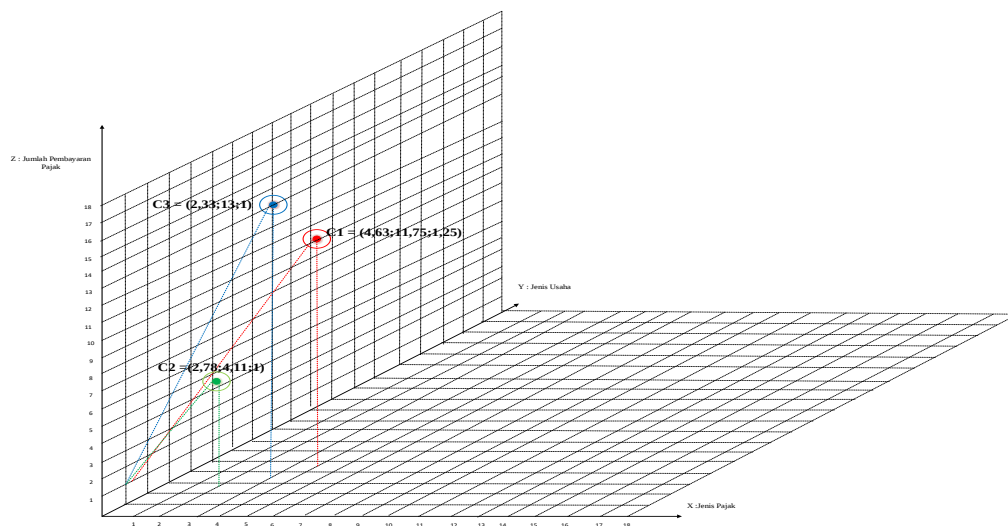


Fig.2: Cluster Graph

Cluster 1 : 4,63; 11,75; 1,25 (5, 12, 1)

Cluster 2 : 2,78; 4,11; 1 (3, 4, 1)

Cluster 3 : 2,33; 13; 1 (2, 13, 1)

From the 20 data grouping types of tax revenues in Binjai City using the clustering method at the Regional Revenue and Assets Financial Management Agency, 3 groups were obtained, cluster 1 contained 5 data, cluster 2 contained 9 data, and cluster 3 contained 6 data.

1.Clusters 1

It can be seen in cluster 1, the cluster results are centered on 5, 12, 1, namely the type of tax revenue, namely restaurants with the type of restaurant business and the amount of tax revenue is Rp. 0 - 200,000,000, - with a total of 8 data.

2.Clusters 2

It can be seen that in cluster 2, the cluster results are centered on 3, 4, 1, namely the type of tax revenue, namely the non-PLN street lighting tax with the type of service business and the amount of tax revenue of Rp. 0 - 200,000,000, - with a total of 9 data.

3. Clusters 3

It can be seen that in cluster 3, the cluster results are centered on 2, 13, 1, namely the type of tax revenue, namely the entertainment tax with the type of retail store business and the amount of tax revenue of Rp. 0 - 200,000,000, - with a total of 3 data.

4. Conclusions

From the cluster data above, it can be concluded that the type of grouping with the highest amount of data is centered on cluster 2 which is centered on 3, 4, 3, namely the type of tax revenue, namely the Non-PLN street lighting tax with the type of service business and the amount of tax revenue Rp. 0 - 200,000,000, - with a total of 9 data.

References

- [1] Sariani, Saripurna, D., & Saragih, R. (2022). Pengelompokan Data Pengguna Narkoba yang Melakukan Program Rehabilitas Rawat Jalan Menggunakan Metode Clustering (Studi Kasus : BNN Kota Binjai), 1-11.
- [2] Siahaan, M. (2022). Data Mining Strategi Pembangunan Infrastruktur Menggunakan Algoritma K-Means, 317-323.
- [3] Masruroh Siti, & Zulaikha. (2013). Pengaruh Kemanfaatan NPWP, Pemahaman Wajib Pajak, Kualitas Pelayanan, Sanksi Perpajakan terhadap Kepatuhan Wajib Pajak (Studi Empiris pada WP OP di Kabupaten Tegal). Diponegoro Journal of Accounting..
- [4] M. W. Talakua., Z. A. Leleury., A. W. Talluta. (2017). Analisis Cluster Dengan Menggunakan Metode K-Means Untuk Pengelompokan Kabupaten/Kota Di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014
- [5] Khakik, N., Ginting, s., & Novriyenni. (2022). Data Mining Pengaduan Hukum Berdasarkan Kasus Dengan menggunakan Metode Clustering
- [6] Hermawati, F. A. (2013). Data Mining. CV. Andi Offset, Yogyakarta.
- [7] Adi Setiawan., (2022). Perbandingan Penggunaan Jarak Manhattan, Jarak Euclid, dan Jarak Minkowski dalam Klasifikasi Menggunakan Metode KNN pada Data Iris. Vol.5, No.1
- [8] Wulan Alinda Wahyuni., Sudin Saepudin. (2021). Penerapan Data Mining Clustering Untuk Mengelompokkan Berbagai Jenis Merk Mesin Cuci.