Data Mining for Clustering Revenue Plan Expense Area (APBD) by using K-Means Algorithm

Wahyudin[1], I Putu Ari Wijaya[2], and Ida Bagus Alit Swamardika[3]
[1][2] Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
[3] Department of Electrical and Computer Engineering, Udayana University
E-Mail: Wahyukely2014@gmail.com

Abstract: APBD is a systematic detailed list of receipts, expenditures and local spending within a certain period (1 year) arranged in Permendagri No. 16 of 2006, so that the data APBD can be used as guidelines for governments and local expenditures in carrying out activities to raise revenue to maintain economic stability and to avoid inflation and deflation. Government financial institutions in areas such as DPKA kota Bima, experienced difficulties in identifying the relevance of each archive data on a APBD that so much, that results in a data warehouse, in addition to the administration, APBD in the government of Kota Bima have not been effective. To minimize the difficulty in identifying heap data archive APBD, then the data warehouse can be used to produce a knowledge that by using the techniques of Data Mining (DM), the method used is clustering and forecasting, clusterisasi performed using the K-Means Algorithm while for forecasting with multiple linear regression. With this method intended to classify and identify the data in the budget that have certain characteristics in common, and can predict the value of APBD in the future.

Keywords : Clustering, K – Means.

I. INTRODUCTION

A. Background

Growth of Information Technology which fast So very is influencing of human being life in so many area do not aside from in management of data, at this moment so much data which there are in a organization, causing difficulty in the case of group data. But with growth of Technology Information (TI) there are is assorted of solution to overcome the difficulty, one of them is by using technique Data Mining (DM). “DM is represent process seeking of relationships and pattern which hidden in a number of big data as a mean to conduct classification, estimation, prediksi, association of rule, clustering, visualisation and deskripsi" (Han et al, 2001, in Baskoro, 2010).
Data of Revenue Plan Expense Area (APBD) managed by monetary body of area in Town of Bima basically have been grouped pursuant to earnings, indirect and direct expense. However because data managed by monetary organizer body of the area have data which so much, hence it is important to know how relevant between earnings data, indirect and direct expense of institute. One of the used method that is clustering. With clustering meant to identify data of Revenue Plan Expense Area owning equality in certain characteristic.
There are various algorithm which is used in technique DM with method of clustering one of them is algorithm of K-Means. "Algorithm of K-Means is one of the algorithm of unsupervised simplest learning which recognized can finish problems of clustering better" (Mac Queen, 1967). With applied of algorithm of K-Means in course of Revenue Plan Expense Area (APBD) clusterisasi hence expected can group and determine the amount of most precise accurate also calculate assess indirect expense and also assess direct expense to come to data of Revenue Plan Expense Area (APBD) in Town of Bima. Pursuant to the problems hence performed a research entitling "DATA MINING FOR THE CLUSTERING OF REVENUE PLAN EXPENSE AREA (APBD) USE ALGORITHM of K-MEANS".

B. Formula of Problem

As for Formula Internal issue research are :

1. Budget tend to be specified is higher
2. Revenue plan tend to to be specified is lower
3. Lack of integrity, consistency and synchronization of planning between SKPD
4. Lack of integrity, consistency and synchronization of planning between SKPD

C. Target of research

1. Analysing its result to determine definition parameters pursuant to characteristic at each cluster.
2. Can estimate indirect expense value and direct expense in the future.

D. Benefits Research

1. Develop protipe clusterisasi data of APBD Town of Bima
2. With Data Mining and algorithm of K-Means can be used to group and determine the amount of most precise cluster or accurate to data of APBD Town of Bima

II. LITERATURE REVIEW

- **Bima city**

Kota Bima geographically located in the eastern part of the Island of Sumbawa in the position 11841'00"-11848'00" Longitude East and 820'00"-830'00" Bars South. The level of average rainfall 132,58 mm with rainy days: The average 10.08 day/month. While the sun shines heat throughout the season with an average of intensity of the highest flashes on October, with temperature 19.5C to 30,8C. The city of XYZ has a land area: paddy fields area of 1.923 hectares (94,90 percen is paddy field irrigation), the forest area of 13.154 ha, gardens area of 3.632 ha, fields and abundant area of 1.225 ha and coastal areas throughout 26 km. the town of XYZ itself adjacent to Ambalawi Sub-district on the North,West Side of the Gulf of XYZ, The east border with Wawo Subdistrict, while South of the border with Palibelo sub-district,kabupaten Bima [1].

- **Data Mining**

Data of Mining is with refer toing process to dig added value in the form of information which during the time unknown manually from basisdata. yielded to be information to be obtained by extraction and recognize important pattern or withdraw from data which there are in basisdata. Data Mining is especially used to look for knowledge which there are in big data bases so that often referred as by Knowledge Discovery in Databases (KDD). Process seeking of this knowledge use various techniques study of computer (machine learning) to analyse and extraction it. Process seeking have the character of and iteratif of interactif to find type or pattern which is valid, new, useful, and understood. In its applying of data mining need various software analyse data to find data relationship and pattern to be able be used to make prediksi accurately.

- **Stages of data mining**

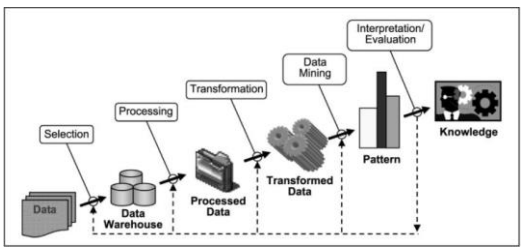

Fig 1. Stages of data mining

a. Data Selection
Selection - selecting or segmenting the data according to some criteria e.g. all those people who own a car, in this way subsets of the data can be determined.

b. Pre-Processing dan Cleaning Data
Preprocessing - this is the data cleansing stage where certain information is removed which is deemed unnecessary and may slow down queries for example unnecessary to note the sex of a patient when studying pregnancy.

c. Transformation
Transformation - the data is not merely transferred across but transformed in that overlays may added such as the demographic overlays commonly used in market research. The data is made useable and navigable.

d. Data Mining
Data mining - this stage is concerned with the extraction of patterns from the data. A pattern can be defined as given a set of facts(data) F, a language L, and some measure of certainty C a pattern is a statement S in L that describes relationships among a subset Fs of F with a certainty c such that S is simpler in some sense than the enumeration of all the facts in Fs.

e. Interpretation / Evaluasi
Interpretation and evaluation - the patterns identified by the system are interpreted into knowledge which can then be used to support human decision-making e.g. prediction and classification tasks, summarizing the contents of a database or explaining observed phenomena.

- **Algoritma K-Means**

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

- **Stages of K-Means Algorithm**

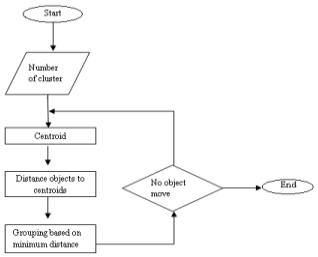

Fig 2. Stages of K-Means Algorithm

Explanation of figura 2

a. Select the number of clusters or segments. There is no magic formula to calculate this before hand and it has to be an iterative process. We will discuss ways and means to evaluate it with an example.
b. Randomly select the number of observations or points, which are equal to the number of clusters. So if we have decided on having 5 segments, randomly select 5 observations from your dataset and assign them as Centroids.
c. Group data by a minimum distance.
d. Repeat Step 3 and Step 4 until the Centroids don't move.

if not, go back to the 2nd way. Until a stable iteration result is found:

- **Clustering**

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.
A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".
A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.
In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering.
Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

- **The Goals of Clustering**

So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection).

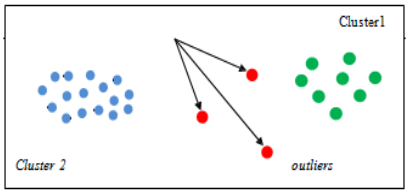

Fig 3. Clustering Example

III. Research Methodology

The research was conducted in the Local Government of Kota Bima which focuses on the main objectives of Clustering of APBD and the Clustering of APBD will be used as reference for the development and utilization of APBD in Local Government of Kota Bima.

A. Place of Research

The city government of Bima, which is based in Sukarno Hatta, focuses on the main objectives of Clustering of APBD and the Clustering of APBD will be used as reference for the development and utilization of APBD in the Regional Government of Kota Bima.

B. Data source

D In this study only the data obtained directly from the stake holders who are related to the data of the Municipal Government of Bima.

IV. SOLUTION AND RESULT

Having taken steps collecting data through the source of exist in Government Town of Bima hence steps of process at algorithm of K-Means to getting clustering it will in process

Tahun	Pendapatan	Belanja Tidak langsung	Belanja Langsung
2011	301.553.514.383	99.572.060.106	130.828.573.884
2012	359.246.501.847	151.397.703.745	222.566.731.839
2013	423.651.578.493	170.833.244.772	230.186.897.446
2014	417.827.261.000	167.340.367.941	200.611.893.059
2015	426.719.642.227	190.183.064.040	180.491.575.430
2016	460.831.808.019	222.548.822.824	191.151.334.250

Fig 4. Data APBD of Bima City

- Early Centroid Calculation with 3 Centroid Values. Here's The Formula To Count Initial ClusterCentroid:

$$C(i)=\frac{X1+X2+X3\ldots Xn}{\sum x}$$

Cluster Revenue Value

$$C0 = \frac{301.553 + 359.246}{2} = 330.399$$

$$C(1) = \frac{423.651 + 417.827}{2} = 420.739$$

$$C(2) = \frac{426.\ 719 + 460.831}{2} = 443.775$$

Cluster value indirect expenditure

$$C0 = \frac{99.573 + 151.397}{2} = 125.485$$

$$C(1) = \frac{170.833 + 167.430}{2} = 169.131$$

$$C(2) = \frac{19.183 + 222.548}{2} = 120.861$$

Cluster expenditure value

$$C0 = \frac{130.828 + 222.566}{2} = 176.697$$

$$C(1) = \frac{239.186 + 200.161}{2} = 219.673$$

$$C(2) = \frac{180.491 + 191.151}{2} = 185.821$$

- Early Centroid Calculation with 2 Centroid Values.

Cluster Revenue Value

$$C0 = \frac{130.553 + 359.246 + 423.651}{3} = 304.483$$

$$C(1) = \frac{417.825 + 426.719 + 460.831}{3} = 435.125$$

Cluster value indirect expenditure

$$C(0) = \frac{99.572 + 151.397 + 170.833}{3} = 140.600$$

$$C(1) = \frac{167.340 + 190.183 + 222.548}{3} = 193.357$$

Cluster expenditure value

$$C(0) = \frac{129.828 + 207.566 + 219.186}{3} = 185.526$$

$$C(1) = \frac{199.611 + 152.491 + 178.151}{3} = 176.751$$

From Centroid calculations that have been done on each cluster that has been determined, the following results Calculations

Table 1. Clustering Result

Cluster	3 Centroid			2 Centroid	
	C_0	C_1	C_2	C_0	C_1
Revenue	330.399	420.739	443.775	304.483	435.125
value indirect expenditure	125.485	169.131	120.861	140.600	193.357
expenditure value	176.697	219.673	185.821	185.526	176.751

- Distance Calculation Process.
  In this step is done distance calculation to know each result of data distance at number K in each Centroid

  Distance Calculation Formulas As follows:

$$d(x,v) = \sqrt{(xi - vi}$$

a. Distance Calculation with 3 Centroid values

- The distance between the first data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(301.553+330.399)^2+(99.572+125.485)^2+(130.828+176.697)^2} = 48.561,58$

- The distance between the first data with the fsecond centroid (C_1)
  $d_{1.1=}\sqrt{(301.553+420739)^2+(99.572+169.131)^2+(130.828+1219.673)2} = 152.972,44$

- The distance between the first data with the third centroid (C_2)
  $d_{1.2=}\sqrt{(301.553+443.775)^2+(99.572+120.861)^2+(130.828+185.821)^2} = 174.836,21$

- The distance between the Second data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(359.246+330.399)^2+(151.397+125.485)^2+(130.828+222.566)^2} = 48.561,58$

- The distance between the second data with the second centroid (C_1)
  $d_{1.1=}\sqrt{(359.246+420.739)^2+(151.397+169.131)^2+(222.566+1219.673)2} = 76.630,79$

- The distance between the second data with the third centroid (C_2)

$d_{1.2=}\sqrt{(359.246+443.775)^2+(151.397+120.861)^2+(222.566+185.821)^2} = 128.765,41$

- The distance between the third data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(423.651+330.399)^2+(170.833+125.485)^2+(230.186+222.566)^2} = 114.560$

- The distance between the thirt data with the second centroid (C_1)
  $d_{1.1=}\sqrt{(423.651+420.739)^2+(170.833+169.131)^2+(125.485+1219.673)2} = 11.974,19$

- The distance between the thirt data with the third centroid (C_2)
  $d_{1.2=}\sqrt{(423.651+443.775)^2+(179.833+120.861)^2+(230.186+185.821)^2} = 74.923,53$

- The distance between the fourth data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(415.827+330.399)^2+(167.340+125.485)^2+(200.611+222.566)^2} = 108.254$

- The distance between the fourth data with the secondt centroid (C_1)
  $d_{1.1=}\sqrt{(417.827+420.739)^2+(190.183+169.131)^2+(200.611+1219.673)2} = 11.974,19$

- The distance between the fourth data with the thirtt centroid (C_2)
  $d_{1.2=}\sqrt{(417.827+443.775)^2+(167.340+120.861)^2+(200.611)+185.821)^2} = 55.777,81$

- The distance between the fifth data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(426.719+330.399)^2+(190.183+125.485)^2+(180.491+222.566)^2} = 125.433,39$

- The distance between the fifth data with the second centroid (C_1)
  $d_{1.1=}\sqrt{(426.719+420.739)^2+(190.183+169.131)^2+(180.491+1219.673)2} = 64.193,69$

- The distance between the fifth data with the third centroid (C_2)
  $d_{1.2=}\sqrt{(426.719+443.775)^2+(190.183+120.861)^2+(180.491)+185.821)^2} = 24.826,64$

- The distance between the sixth data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(460.831+330.399)^2+(222.548+125.485)^2+(191.151+222.566)^2} = 167.125,6$

- The distance between the sixth data with the second centroid (C_1)
  $d_{1.1=}\sqrt{(460.831+420.739)^2+(222.548+169.131)^2+(191.151+1219.673)2} = 74.497,18$

- The distance between the sixth data with the third centroid (C_2)
  $d_{1.2=}\sqrt{(460831+443.775)^2+(222.548+120.861)^2+(191.151)+185.821)^2} = 24.826,64$

b. Distance Calculation with 2 Centroid Values

- The distance between the first data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(301.553+304.483)^2+(99.572+140.600)^2+(130.828+185.526)^2} = 83.851,95$

- The distance between the first data with the second centroid (C_1)
  $d_{1.1=}\sqrt{(301.553+435.125)^2+(99.572+193.357)^2+(130.828+176.751)2} = 164.339,91$

- The distance between the second data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(359.246+304.483)^2+(151.397+140.600)^2+(222.566+185.526)^2} = 24.530,92$

- The distance between the second data with the second centroid (C_1)
  $d_{1.1=}\sqrt{(359.246+435.125)^2+(151.397+193.357)^2+(222.566+176.751)2} = 111.175,03$

- The distance between the third data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(423.651+304.483)^2+(170.833+140.600)^2+(230.186+185.526)^2} = 76.373,57$

- The distance between the third data with the secondt centroid (C_1)
  $d_{1.1=}\sqrt{(423.651+435.125)^2+(170.833+193.357)^2+(230.186+176.751)2} = 56.939,98$

- The distance between the fourth data with the first centroid (C_0)
  $d_{1,0}=\sqrt{(417.827+304.483)^2+(167.340+140.600)^2+(200.611+185.526)^2} = 71.326,25$

- The distance between the fouth data with the second centroid (C_1)
  $d_{1.1=}\sqrt{(417.827+435.125)^2+(170.833+167.340)^2+(230.186+185.526)2} = 37.185,21$

- The distance between the fifth data with the first centroid C_0)
  $d_{1,0}=\sqrt{(426.719+304.483)^2+(190.183+140.600)^2+(180.491+185.526)^2} = 97.339,52$

- The distance between the fifth data with the second centroid (C_1)
  $d_{1.1=}\sqrt{(462.791+435.125)^2+(190.183+167.340)^2+(180.491+185.526)2} = 25.689,03$

- The distance between the sixth data with the first centroid (C_0)
  $d_{1,0=}\sqrt{(460.831+304.483)^2+(222.548+140.600)^2+(191.151+185.526)^2} = 133.966,38$

- The distance between the sixth data with the second centroid (C_1)
  $d_{1.1=}\sqrt{(460.831+435.125)^2+(222.548+167.340)^2+(191.151+185.526)2} = 355.555,12$

From the calculation of Distance data from Value K with each Centroid of each Cluster we can see in Table below.

Tabel 2. Clustering 2 and 3 Centroid

Data	3 Centroid			2 Centroid	
	C_0	C_1	C_0	C_1	C_2
1	48.561,58	152.972,44	174.836,21	83.851	164.339,91
2	48.561,58	76.630,79	128.765,41	24.530	111.475.03
3	114.560,35	11.974,19	74.923,53	76.373	56.939,98
4	108.254,7	11.974,19	55.777,81	71.326	37.185,21
5	125.433,39	64.193,69	24.826,64	97.339	25.689,03
6	167.125,60	74.497,18	24.826,64	133.996	35.555,12

c. Iteration Value

Based on the workings of the K-Means algorithm after the determined value k then calculate the centroid value and the distance between data on each centroid each. In this stage, the calculation of centroid values on each cluster is called iteration, until the centroid value has not changed from before.
Pada cluster Pendapatan iterasi pertamanya sebagai berikut:

Tablel 3. The first iteration on the revenue cluster

Year	Data	Centroid		
		C_0	C_1	C_2
2011	301.553	48.561,58	152.972,44	174.836,21
2012	359.246	48.561,58	76.630,79	128.765,41
2013	423.651	114.560,35	11.974,19	74.923,53
2014	417.827	108.254,7	11.974,19	55.777,81
2015	426.719	125.433,39	64.193,69	24.826,64
2016	460.831	167.125,60	74.497,18	24.826,64

$$C0 = \frac{301.553 + 359.246}{2} = 330.399$$

$$C(1) = \frac{423.651 + 417.827}{2} = 420.739$$

$$C(2) = \frac{426.719 + 460.831}{2} = 443.775$$

Tablel 4. The first iteration on value indirect expenditure cluster

Year	Data	Centroid		
		C_0	C_1	C_2
2011	99.572	48.561,58	152.972	174.836
2012	151.397	48.561,58	76.630	128.765
2013	170.833	114.560	11.974,19	74.923
2014	167.340	108.254	11.974,19	55.777
2015	190.183	125.433	64.193	24.826,64
2016	222.548	167.125	74.497	24.826,64

Table 5. The first iteration on expenditure Value cluster

Year	Data	Centroid		
		C_0	C_1	C_2
2011	130.828	48.561,58	152.972,44	174.836,21
2012	222.566	48.561,58	76.630,79	128.765,41
2013	230.186	114.560	11.974,19	74.923,53
2014	200.611	108.254	11.974,19	55.777,81
2015	180.491	125.433	64.193,69	24.826,64
2016	191.151	167.125	74.497,18	24.826,64

Because at the first iteration the value of centroid center does not change the same as the previous centroid value, then the iteration process is stopped. Next will be done Iteration Calculation process at the value of 2 Centroid. The process is recalculated from the first centroid, the steps must be in accordance with the conditions set by the K-Means Algorithm. Perhitungan iterasi pertama pada cluster pendapatan dengan 2 nilai centroid.

Tabel 6. The first iteration on the revenue cluster

Year	Data	Centroid	
		C_0	C_1
2011	301.553	83.851	164.339
2012	359.246	29.476	111.475
2013	423.651	76.373	56.939
2014	417.827	71.326	37.185
2015	426.719	97.339	25.689
2016	460.831	133.996	35.555

$$C(0) = \frac{301.553 + 359.246}{2} = 330.399$$

$$C(1) = \frac{423.651 + 417.827+426.719+460.831}{4} = 332.257$$

The process of searching for the next iteration remains done The Centroid recalculation process corresponds to the specified cluster

V. CONCLUSION

- In this study using the number of k as much as 3, the income cluster, indirect expenditure cluster and direct shopping cluster. As well as testing 2 parameters of the centroid value are 3 centroid values and 2 centroid values.
- The experimental results of the clustering method using the K-Means algorithm with 3 better centroid values compared with 2 centroid values

REFERENCES

[1]. www. Bimakota.go.id
[2]. Adiningsih, N. (2011). Penggunaan K-Means Clustering untuk Pelabelan Fonem Sinyal Ucapan. Skripsi Tidak Terpublikasi. Bandung:Institut Teknologi Bandung.
[3]. Agusta, Y. (2007). K-Means-Penerapan,Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika Vol.3 : 47-60
[4]. Johan Oscar (2013) Implementasi algoritma k-means clustering untuk menentukan strategi marketing president university
[5]. Han, Jiawei dan Kamber, Micheline. 2006. Data Mining Concepts and Techniques 2nd Edition. San Fransisco : Morgan Kaufmann publisher.
[6]. Larose, Daniel T. (2005). Doscovering Knowledge in Data : An Introduction to Data Mining. John Willey & Sons, Inc.
[7]. Wahyuni, Santi Febriana. (2013). Penggunaan Cluster-Based Sampling Untuk Penggalian Kaidah Asosiasi Multi Obyektif. Jurnal. Malang: Jurusan Teknik Informatika, Fakultas Teknologi Industri. Institut Teknologi Nasional.
[8]. Johan Oscar (2013). implementasi algoritma k-means clustering untuk menentukan strategi marketing president university
[9]. Tahta Alfina, Budi Santosa, Ali Ridho Barakbah. (2012) Analisa Perbandingan Metode Hierarchical Clustering, K-Means dan Gabungan Keduanya dalam Cluster Data (Studi Kasus: Problem Kerja Praktek Teknik Industri ITS)
[10]. B. Santosa, Data Mining. Teknik Pemanfaatan Data untuk Keperluan Bisnis, First Edition ed. Yogyakarta: Graha Ilmu,
[11]. K. Arai and A. R. Barakbah, "Hierarchical K-means: algorithm for centroids initialization for Kmeans,"