Mapping Patterns Achievement Based on CRISP-DM and Self Organizing Maps (SOM) Methods

Santi Ika Murpratiwi[1], A.A Ngurah Narendra[2], and Made Sudarma[3]

[1][2] Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
[3] Department of Electrical and Computer Engineering, Udayana University
Email: santiikamurpratiwi@student.unud.ac.id

Abstract— **Successful of regional development is a reflection of the good government policy. Many programs created by the local government every period but many programs stop without results. Therefore, we need to evaluation its government programs. Evaluation of government programs must do to prevent the failure of the programs. One way to evaluate the government programs is to collect the support data to see the data model and analyzed the data. In this research will conduct an evaluation on the achievement of RPJMD performance of Bali Province by using data mining. The process of data mining used the CRISP-DM method and Self Organizing Maps (SOM) for mapping the achievement patterns. The used data is RPJMD data in the 2013-2018 period. but the data used for data mining process only 3 years data that is 2014-2016 as a middle evaluation. That data clustered into five clusters. The final result of this research are 78% of assessment indicators in the RPJMD program are inconsistent position and 22% are in an inconsistent position. Moreover, there are 84 assessment indicators that have no reached the target. From that results of data, assessment mapping can use as the guidance of Bali Province to catch up the achievement of RPJMD programs and prepare the next strategies to support the success of the RPJMD program during the RPJMD period.**

Index Terms—**Data Mining, CRIPS-DM, SOM, evaluation, Pattern Mapping**

I. INTRODUCTION

The successful of the regional development is seen from the ability the government to develop and build the infrastructures in the region. Each period of government must have work program to develop the region. As well as the Regional Development Planning Board (Bappeda) of Bali province also has many program and activities to build Bali. These activities include creating programs for each organization of Regional Device (OPD), where the programs measured in its implementation every 5 years. Bappeda makes a document on the work program of the Bali Province in the RPJMD. RPJMD is a program that initiated by the governor.

In the implementation of RPJMD program must be evaluated at the beginning, middle, and at the end so the implementation of RPJMD runs optimally and fit with the target achievement. Comparing the value of achievement with the target value achieved is a way to evaluate the RPJMD program. Bappeda Bali province has completed several RPJMD programs in the last periods which it becomes the standard for composing the next RPJMD.

The RPJMD program has 50 programs indicators targeted at regional development. Each indicator must evaluate so as not failed. The total sub-indicators in the RPJMD are 237 data each year with 5 years execution every period.

During this time, Bali province evaluated manually the RPJMD program so that the mapping patterns of achievement results are less visible. The pattern of achievement that can see early can help the government to set the new strategies for finish the government programs. The process of pattern mapping can do during the implementation of RPJMD.

In this research, we will make a pattern mapping by clustering using Self Organizing Maps (SOM) methods and Cross-Industry Standard Process for Data Mining (CRISP-DM) in analyzing the achievement results of RPJMD. The data used is the RPJMD data for the 2013-2018 period and the main data as the material of analysis is data of RPJMD from 2014 until 2016. By making a pattern achievement from RPJMD is expected to help the provincial government of Bali is prepares the new strategy for continuity the RPJMD so that end of the RPJMD all development indicator can be finished with good and satisfactorily.

II. STUDI LITERATURE

A. Literature Review

Much research on data mining that discusses the Self Organizing method and CRISP-DM has been done. Self Organizing Maps (SOM) can apply in data mapping and image. In a research by Ajeissh M about using SOM for clustering documents in a network with ring topology [1]. SOM method is also used to build a recommendation system at the Library and Archive Board of East Java Province [2]. In the recommendation system in the library, SOM method is used to help the library visitor for selecting books to be borrowed. In another case, SOM methods are used for clustering the Indonesia Online News [3] and used for mapping distribution of exam questions [4]. SOM is a suitable method for multilevel management data so it can use for customer segmentation in a supermarket [5] and it used for grouping the data of new student at the Bogor Agriculture Institute [6].

In this research used CRISP-DM method to help in the analyzing data mapping requirement. CRISP-DM is widely used for data management with data mining. The CRISP-DM method was implemented in data mining for the executive information system in the Department of Marine and Fisheries of Central Java Province [7]. In another research, CRISP-DM used in the forecasting system for predict graduation students in pursuing courses [8].

Therefore, the process of mapping pattern data achievement in the RPJMD Bali Province used to perform performance evaluation of RPJMD by combining the SOM method and CRISP-DM method. The combine of the two methods is the right choice because the CRISP-DM method is used for understanding business process and analysis the Bappeda Bali Province need for evaluating the achievements and SOM method is a method for data management and data mapping analysis with clustering data.

B. Data Mining

Data mining is a process that uses statistical techniques, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from various databases in large capacity [9]. Data mining also called Knowledge Discovery in Database (KDD). KDD is an activity that includes the collection, use of historical data to find a regular pattern or a relationship in large datasets [10].

Outline the training methods used in data mining techniques are differentiated into two approaches [10] ie.

1. Unsupervised learning, this method is applied without any training and without any teacher. The teacher hired is the label of the data. The methods that included in this approach are A Priori, FP-Growth, K-means, K-Medoids, FCM, and SOM.
2. Supervised learning, this method must through the process of learning with the exercise or training and teacher or coach. In this approach to find the decision function, separator function, or regression function is used some sample data that has output or label during the training process. Examples of methods that using supervised learning approach are Native Bayes, C4.5, C5.0, KKN, and linear regression.

Data mining is classified by the usual tasks of description, estimation, classification, clustering, and association [11]. In Figure 1 is the step of processing data mining.
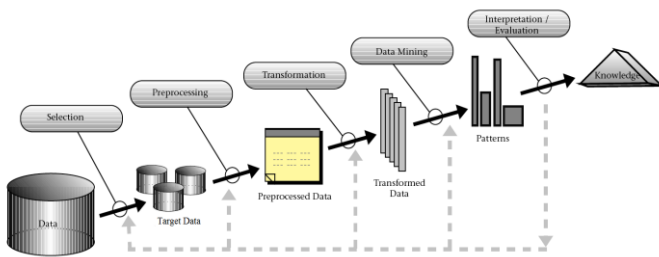

Fig. 1 Stages in Data Mining

Stages in data mining are as follows.
1. Data cleaning is a process of eliminating noise and inconsistent data or irrelevant data.
2. Data integration is a combination of data from various databases into a new database.
3. Data selection, not all data in the database used for analysis. Therefore only the appropriate data to be retrieved from the database.
4. Data transformation, data will alter or merged into the appropriate format for data mining process.
5. The mining process is a major process when methods are applied to find the valuable and hidden knowledge of the data using a particular method.
6. Pattern evaluation is used to identify interesting patterns into the found knowledge base.
7. Knowledge Presentation is the visualization and presentation of knowledge about the methods that used to acquire the knowledge gained by users.

C. CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) developed in 1996 and it used to analysis process of an industry as problem-solving strategies of business or research units [12]. For processed data using the CRIPS-DM method, there is no specific characteristic because the data will be processed again at the phase in the CRISP-DM. There are six phases in CRISP-DM [12] as in Figure 2 below.
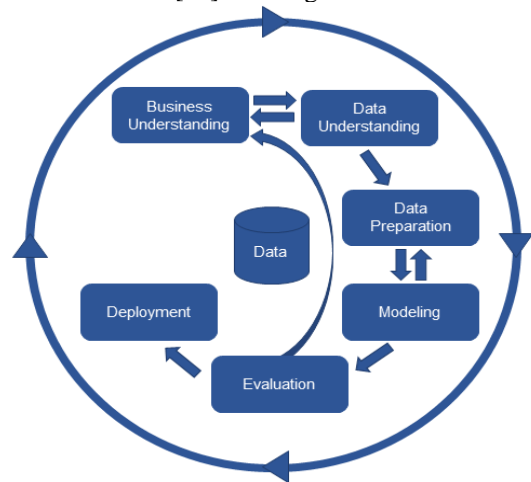

Fig 2. CRISP-DM Process
1. Business Understanding Phase are to determine the project goals, determine the detail requirements and create a formula for data processing / preparing initial data processing strategy.
2. Data Understanding Phase are collected data and select data that will use to develop analysis and evaluate the quality of the data.
3. Data Preparation Phase are selecting cases/variables to be analyzed and repair the data fields if needed.
4. Modeling Phase is applying a predefined modeling method and can back to data preparation if needed for repairing data.

5. Evaluation Phase is an evaluation of the model. Whether the model is in accordance with the requirement and goals that have been determined in the beginning of CRISP-DM.
6. Deployment Phase is using a model that has been producing to presented so that it becomes a new knowledge.

D. Clustering Analysis

Clustering is a grouping of records, observations or cases into classes that have similar objects. Cluster are collections of similar records that not similar to another cluster. Clustering is different with classification because clustering not has target variables to make a cluster.

Clustering does not classify, forecasting, or predicting the value from the target variable. The clustering algorithm is used to define the whole segment of the data set into a subgroup with relatively similar records. With the similar record in the cluster being maximized and the record similar in the out of cluster is minimized [11].

The main method of clustering can be classified into the following categories[13].

1. Partition Method, suppose there is a database containing n objects. Partitioning methods build k partitions on the database with each partition representing cluster and $k \leq n$. The partition formed must meet the requirement that each cluster must contain at least one object and each object must include exactly one cluster.
2. The hierarchical method is making a hierarchical decomposition of the data set (or object) using some criteria.
3. A method based on concentration is an approach based on connectivity and density function.
4. A grid-based method is an approach based on the structure of multiple-level granularity.
5. The model-based method is a model hypothesized for each cluster and the basic idea is to find a suitable model for each cluster.

E. Self Organizing Maps (SOM)

Kohonen Network was introduced by Teuvo Kohonen who was a Finnish scientist in 1982. Kohonen Network gives a type for SOM that is a special class of artificial neural network approach [13].

Self Organizing Maps (SOM)is one model of artificial neural network using the method of learning without supervision (unsupervised learning). The advantage of SOM algorithm is able to map high-dimensional data into the form of a low dimensional map. The mapping process occurs when a pattern of free dimension is projected from the input space to the position of one or two dimensional array. Here is a calculation formula Self Organizing Maps (SOM).

$$W_{ij}(t+1) = W_{ij}(t) + \beta NS(X(t) - W_{ij}(t)) \quad (1)$$

Where W_{ij} is a neighbor function whose value changes according to the distance of the neuron with the winning neuron

and β is the learning rate. The value of the activation function determines the magnitude of the weighted value of the neuron of the winner and its neighbor neuron.

III. RESEARCH METHODOLOGY

A. Data Source

In this research, the data used is data achievement performance of Bappeda Bali Province in accordance with an indicator of assessment which has been determined. In Table 1 below is a table of indicator assessment list.

TABLE I
INDICATOR ASSESSMENT OF BAPPEDA BALI

Code	Aspect Indicator
A.	Community welfare aspects
A.1	Focus on Welfare and Economic Equity
A.2	Focus on People's Welfare
A.3	Focus on Cultural and Sports Art
B.	Aspects of General Services
B.1	Focus on Mandatory Public Service
B.2	Focus on Public Service
C.	Aspect of Regional Competitions
C.1	Focus on Economic Capability
C.2	Focus on Facility of infrastructure
C.3	Focus on Climate of Investing

The data that will use is data result of performance achievement of Bali Province on 2014-2016 where every year there are around 237 assessment data. Then the data will be clustered based on the variable of years and each assessment indicator.

B. Research Stages

The research stage in this research is shown in Figure 3. This research will combine CRISP-DM method and SOM method where the research started with study literature by looking for research and supporting methods. The next stage is the business process understanding and information that needed for analysis in Bappeda Bali. Then do the necessary data retrieval and understanding the data to determine the variables that will be analyzed. After that, the next stage is preprocessing data. Processing data followed by eliminating unnecessary data and adding information if needed. And then is the data modeling process. In the data modeling process will implement SOM method for clustering the data. The final stage is the evaluation phase of the results followed by deployment results to Bappeda Bali for help Bappeda Bali to create strategies to reach the RPJMD targets if needed.
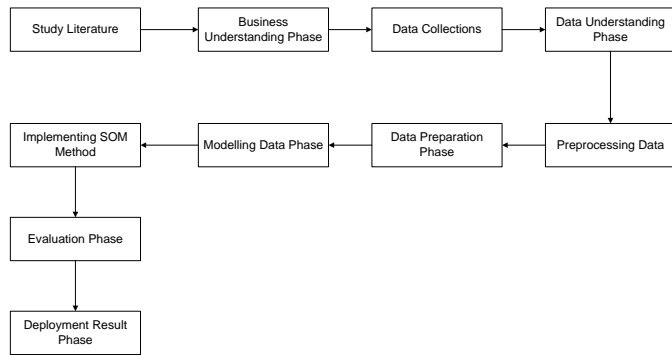
Fig 3. Research Stages

C. Self Organizing Maps (SOM) Algorithm

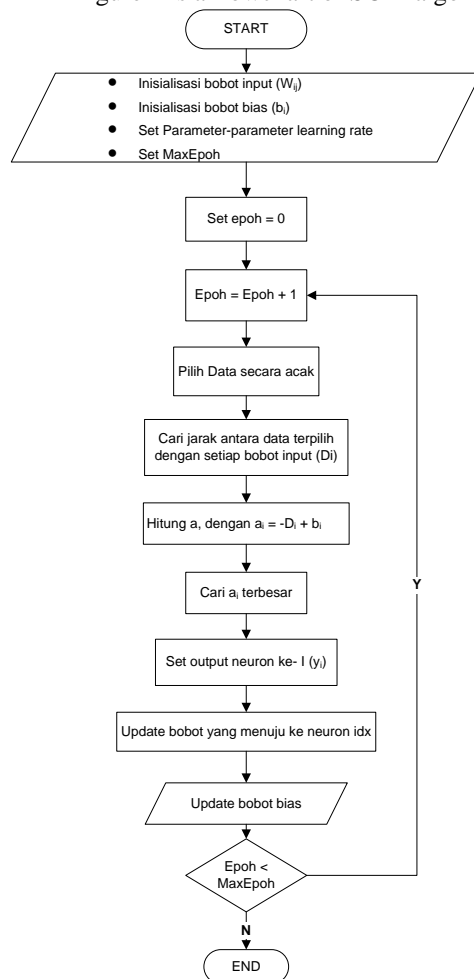In Figure 4 is a flowchart of SOM algorithm.

Fig. 4 Flowchart of SOM Algorithm

The following are the steps of Self Organizing Maps Algorithm.

1. Initialization, the output layer vector weight (W_{ij}) with a random value, earning rate, and neighbor function used.
2. Calculate the level of similarity or weight of the neuron in the output layer with the data entry and selected the

winning neuron (BMU).

3. Update the learning rate and calculate the neighbor's function.
4. Perform steps 2 through 3 until the value of interaction (epoch) is reached.

IV. RESULTS AND DISCUSSION

A. Data Preparing

In the data preparation phase, the data has been cleaned and transformed according to the analysis requirements such as providing the identity of the data or the removal of empty and useless fields. The data used is performance RPJMD period 2014-2016. This process of analysis is conducted in the middle of the implementation period of RPJMD 2013-2018 with the purpose is to see the progress of the performance achievement of Bappeda Bali Province in three years of implementation RPJMD. The amount of data used is 237 data per year. The data will be clustered based on variable level of achievement that is "consistent and inconsistent". In addition, there is also an analysis per year to see achievement every year. The following table 2 shows the data of RPJMD achievement of Bali Province for the indicators aspect of community welfare.

TABLE II
INDICATOR ACHIEVEMENT OF COMMUNITY WELFARE ASPECTS

B. SOM Variables Initialization Process

The clustering process using SOM is assisted with excel applications from Faculty of Engineering of the University of Buenos Aires [14]. In the first step that must initialize the SOM variable shown in Figure 5.

For doing clustering using Self Organizing Maps (SOM) we must input the amount of data that will observe and training, besides that must input the number of variables. Then input the number of clusters to built in n, and input the number of training reply. Learning parameter also entered forgive limit start from 1% - 99% because of the used data in the form of a percentage. The last step is built in a cluster and get mapping pattern of the data.

Fig 5. Initialization SOM Variables

C. Clustering Process

The clustering process is done by looping according to SOM initialization input. The data will be inserted into the clusters according to the information contained. The number of the cluster formed according to input in the application. In Figure 6 is an example of clustering process detail in the SOM method.

Fig 6. Process Clustering Used SOM method

D. Pattern Mapping Results

Furthermore, from the distance obtained between the next clusters will be implemented in a graph model as in Figure 7 below.
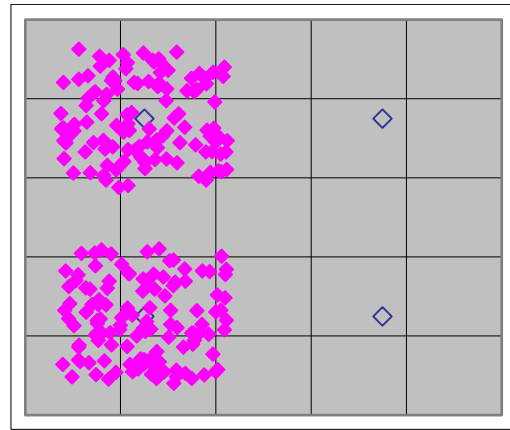

Fig 7. Pattern Mapping

Clusters are represented in a matrix graph according to the number of clusters formed. In Figure 7 clusters that shown in the 5x5 matrix format. From the pattern, the mapping can see that the dispersion of more patterns in clusters 3, 4, 5. It can be concluded that dissemination of achievement of RPJMD performance of Bali Province is good because of the achievement more than 50%.

E. Results Analysis

From the results of pattern mapping then can be view in some graph for help the user read the information from the pattern mapping. The graphic can use as the guidance to evaluate the achievement of the performance of Bali province in accordance with data RPJMD period 2013-3018. In figure 8 is shown the graphic achievement of the performance Bali province every year.
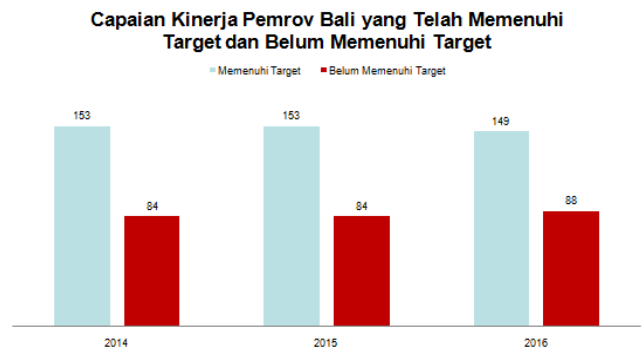
Fig 8. Graph of Achievement of RPJMD Bali Province

From the graph in Figure 8 can see that the achievement of indicators that have reached the target and not reach the target from 2014-2016 is relatively stable. It can be under the spotlight of the government as its annual performance is not yet clearly, so it needs hard effort from the government to reach the target indicator because the remaining period of the RPJMD 2013-2018 only two years. In Figure 9 is a consistent and not consistent achievement of performance RPJMD Bali Province.

In Figure 9 it can be seen that the achievement of the RPJMD of Bali Province from 2014-2016 is more consistent

than inconsistent. Consistent here means that the assessment indicator always increases every year or at least does not decrease. 78% of the RPJMD Bali Province assessment score is at a consistent level and 22% of indicators assessment are not consistent.
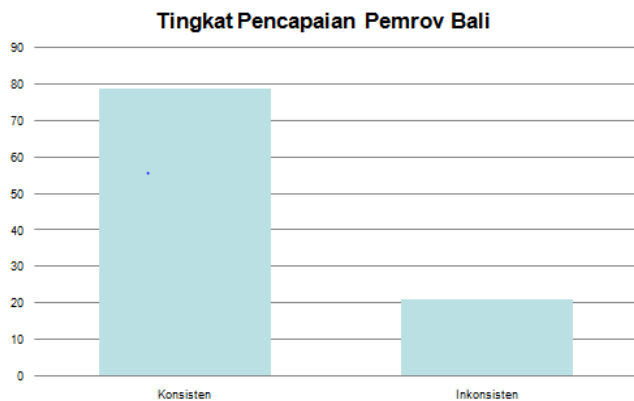

Fig. 9. Level of Achievement of Bali Province

From the result analysis, the government of Bali Province should start to make efforts to reach the target of RPJMD indicator before its time is over. Many things still need to be evaluated because there are still 84 aspect indicators that not yet reach the target.

V. CONCLUSION AND FUTURE WORKS

The result of this research by combining CRISP-DM and SOM method to perform the performance evaluation process of RPJMD Bali Province is found that from 2014-2016 the achievement of the government is still same and no change. This is seen from the number of assessments that reach and not reach the target. In addition, 78% of the indicator assessments of RPJMD get consistent and 22% get the inconsistent assessments.

From the analysis results, the Provincial Government of Bali has a big duty and obligation to improve the development of the area of Bali in order to support the vision and mission, Bali Province.

For future research can implement CRISP-DM and SOM in a desktop application or in website application by using particular programming language for easier in the analysis and equipped with application testing. Another future research, we can build a forecasting application for prediction to look the possibility the achievement of program Bali Province.

REFERENCES

[1] M. Ajeissh and S. Harikumar, "An Adaptive Distributed approach of a Self Organizing Map Model for Document Clustering using Ring Topology," *Conf. Adv. Comput. Commun. Informatics*, pp. 776–781, 2016.

[2] A. Riyandwyana and E. S. Mukhlason, "Pengembangan Sistem Rekomendasi Peminjaman Buku Berbasis Web Menggunakan Metode Self Organizing Map Clustering Pada Badan Perpustakaan Dan Kearsipan (BAPERSIP) Provinsi Jawa Timur," *J. Tek. ITS*, vol. 1, no. 1, 2012.

[3] Ambarwati and E. Winarko, "Pengelompokan Berita Indonesia Berdasarkan Histogram Kata Menggunakan Self-Organizing Map," *IJCCS*, vol. 8, no. 1, 2014.

[4] Pujiadi and Widyaiswara, "Penggunaan Metode Artificial Neural Network Dengan Algoritma Self Organizing Maps Untuk Membantu Guru Dalam Melakukan Pemetaan Soal Ujian Nasional," pp. 1–24, 2010.

[5] A. A. G. B. Ariana, "Customer Segmentation Dengan Metode Self Organizing Map (Studi Kasus : UD . Fenny)," *Lontar Komput. Vol.2*, vol. 2, no. 1, pp. 1–7, 2011.

[6] I. Hermadi and I. S. Sitanggang, "Clustering Menggunakan Self Organizing Maps (Studi Kasus: Data PPMB IPB)," no. Wisnujati, 2006.

[7] I. Purnama, R. Saputra, and A. Wibowo, "Implementasi Data Mining Menggunakan Crisp-Dm Pada Sistem Informasi Eksekutif Dinas Kelautan Dan Perikanan Provinsi Jawa Tengah," 2014.

[8] A. P. Fadillah, "Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ)," *J. Tek. Inform. dan Sist. Inf.*, vol. 1, pp. 260–270, 2015.

[9] E. Turban, J. E. Aronson, and T.-P. Liang, "Decision Support Systems and Intelligent Systems," p. 960, 2005.

[10] B. Santosa, *Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu, 2007.

[11] D. T. Larose, *Discovering Knowledge in Data : An Introduction to Data Mining*. John Willey & Sons, Inc, 2005.

[12] P. (NCR) Chapman *et al.*, *Crisp-dm 1.0 : Step by Step Data Mining Guide*. SPSS Inc., 2000.

[13] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed. USA: Morgan Kaufmann is an imprint of Elsevier, 2012.

[14] F. UBA, "Self Organizing Maps (SOM)." [Online]. Available: materias.fi.uba.ar/7550/NNclust.xls.