The data mining analysis to determine the priorities of families who receiving assistance

C Prianto* and N S Harani

Departement of Information Technology, Politeknik Pos Indonesia, Bandung, Indonesia

*Corresponding author’s e-mail: chprianto@gmail.com

Abstract. Government has made efforts to improve the level of social life as an object of national development either by providing direct cash assistance or a program to create jobs. Fund distributed to the society is not small, so the distribution must be done properly. The data mining is a method of extracting data to find the hidden patterns of the data. In this research, the data are processed by using K-Means and C4.5 algorithm. The aim of this research is to classify the society into three clusters based on economic conditions using the K-means algorithm, the three clusters are pre-prosperous families, pre-prosperous families 1 and pre-prosperous families 2. Moreover, the C4.5 algorithm is used to make a decision tree model in determining which priority family gets assistance.

1. Introduction

The most important pillar in the development of a State is the robust social life, including the family economic condition which is the core of society itself. The National Population and Family Planning Agency asserted that there are five types of families based on their economic level: Pre-Prosperous Family, Prosperous Family 1, Prosperous Family 2, Prosperous Family 3, and Prosperous Family 3 plus [1]. Of this type, the pre-prosperous families need receive attention, as they have not been able to meet the minimum basic needs such as food, clothing, shelter, health and worship.

Mapping society into these types of families is very important, the data used must be accurate so that government programs can be implemented on target based on the priority scale. Data is growing so fast, to get accurate data and produce important information, a special technique is needed. A technique that can make it happen is data mining [2]. Data mining is a process of analysing data to find a pattern from the data set. Data mining is able to analyse large data into interesting information in the form of patterns that have meaning for decision supporters [2]. One of the data mining techniques that can be used is the K-means clustering algorithm and the C4.5 algorithm.

The k-means clustering algorithm is used to group data into k clusters [3, 4]. The first aim of this research is to classify the society into three clusters based on economic conditions using the K-means algorithm; the three clusters are pre-prosperous families, pre-prosperous families 1 and pre-prosperous families 2. C4.5 algorithm is one technique in data mining that serves to classify data. In the process of data mining, the C4.5 algorithm helps to make a model that might influence the making of a decision. And in this research the C4.5 algorithm is used to make a decision tree model in determining which priority family gets assistance. In a survey, the C4.5 algorithm is in the first position as the most widely used algorithm in data mining [5].
2. Methods

This study uses the CRISP-DM method that provides standard processes for data mining that can be applied to general problem solving strategies [5]. The CRISP-DM methodology is standardized data mining compiled by three mining market data initiators, namely Daimler Chrysler (Daimler-Benz), SPSS (ISL), NCR. Then developed in various workshops (between 1997-1999). More than 300 organizations contributed to this modelling process and finally CRISP-DM 1.0 was published in 1999 [6].

![CRISP-DM diagram](image)

**Figure 1. CRISP-DM**

The flow in this research, adjusted to the stages in the CRISP-DM method, as seen in Figure 1. based on CRISP-DM:

1. Stage 1 is business understanding, the stage of business understanding is to determine the objectives of the problems to be achieved, in this research the data is collected and then processed according to the aim of research. The aim of this research is to classify the society into three clusters based on economic conditions and make a decision tree model in determining which priority family gets assistance.

2. Stage 2 is data understanding and data preparation. What is done at this stage is collecting and selecting data so that it is in accordance with what is needed. The process of collecting data in this stage is carried out in several ways such as interviews, observation and collection of population data through the local government.

3. Stage 3 is about modelling and evaluation, to achieve the aim of this research we use k-means algorithm and C4.5 algorithm. The evaluation process is carried out so that no mistakes occur at each stage that is carried out

To describe the application of the algorithm used can be seen in the Figure 2.
K-means Clustering is one of algorithms in partition, as K-Means Clustering is based on determining the initial number of groups by defining the initial centroid value [3, 7]. Here are steps in the K-Means Clustering algorithm [3, 7]:

1. Determine k as the number of clusters to be formed
2. Determine k as the initial centroid of the cluster in random
   Determination of initial centroid is made in random from the existing objects as much as k cluster, then to calculate the next i-cluster centroid, the following formula is used:

\[ v = \frac{\sum_{i=1}^{n} x_i}{n} \; ; \; i = 1,2,3,...,n \]  

(1)

Where:
\( v \): centroid on the cluster
\( x_i \): the-i object
\( n \): the number of objects as the cluster members

3. Calculate the distance of each object to each centroid of each cluster. To calculate the distance between objects with the centroid, the Euclidian Distance may use:
\[ d(x, y) = ||x - y|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad ; \quad i = 1, 2, 3, \ldots n \]  

(2)

Where:
- \( x \): the-\( i \) object \( x \)
- \( y \): the-\( i \) power \( y \)
- \( n \): the number of objects

4. Allocate each object into the nearest centroid. To allocate the objects into each cluster during iteration in general, it can be made by means of hard k-means where each object is expressed as a cluster member by measuring the proximity of its nature to the center point of cluster.

5. Perform iterations, and then determine the new centroid position by using the equation. Repeat step 3 if the new centroid position is not the same.

The construction of classification modelling in this research is by using C4.5 Algorithm. In general, to make a decision tree using C4.5 algorithm is as follows:

a. Select attribute for root
b. Create a branch for each value
c. Divide the case in the branch
d. Repeat the process for each branch until all cases on the branch have the same class.

The selection of attribute for root is based on the highest gain value of the attributes. To calculate the gain, the following equation is used:

\[
Gain(S, A) = Entropy(s) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i)
\]

(3)

Information:
- \( S \): The set of cases
- \( A \): attribute
- \( n \): number of attribute partition \( A \)
- \( |S_i| \): the number of cases on the-\( i \) partition
- \( |S| \): the number of cases in \( s \)

To calculate the entropy value, the following equation may be used:

\[
Entropy(s) = \sum_{i=1}^{n} -p_i \times \log_2 p_i
\]

(4)

Information:
- \( S \): The set of cases
- \( A \): feature
- \( n \): number of partitions \( S \)
- \( p \): the proportion of \( S \) to \( S \)

3. Result and Discussion

The data used in this research is population data from Cigugur Girang West Bandung which consists of 4000 households. If we take the smallest value allowed in the sampling (10%) of the total population, the number of samples to be used is approximately 400 households. and after going through the selection process, the data that will be used is 374 household data.

The attributes data that are used and the number of scoring on the values of each attribute can be seen from the following Table 1.
The process of data mining using the K-means algorithm is done by using Weka 3.80 software, and the results can be seen in Table 2.

Data processing using the K-Means algorithm results that the category of pre-prosperous families is in cluster 1 with 79 instances, prosperous families 1 gather in cluster 2 with 190 instances, and prosperous families 2 gather in cluster 3 with 105 instances.

The data collected in the cluster of pre-prosperous families (cluster 1) is then re-processed by the C45 algorithm; the results can be seen in Figure 3,
From the above decision tree results we will find a rule that to determine which families receive assistance, the first is seen from the education factor, which is not included in the priority is the head of the family who have high school education level and university, while which fall into the category of priority is the head of the family who has a primary education level.

If there is a family head that has a junior high school level then it should be considered based on the next factor that is the area of the house. If the head of household has a house area of 120 m² and 54 m², then entering into the scale is not a priority, and when the head of the family has a house area of 36 m², and 45 m² then enter into the priority scale get assistance.

If there is a family head that has a house area of 60 m² then the next factor should be considered if the number of dependents is less than or equal to 4 then it is not a priority, and if the number of dependents of the head of household is more than 4 then enter into the priority scale of assistance.

4. Conclusion
Based on the results of the analysis and discussion, the following conclusion is obtained:
1. Using K-Means clustering algorithm, the population data may be clustered into Pre-
Prosperous Family, Prosperous Family 1, and Prosperous Family
2. Model to determine priorities of aid recipients may be produced by data derived from the pre-
prosperous families clustered into a rule of decision tree

5. References
[1] M H Permatasari 2009 Konsep Keluarga Sejahtera (Jakarta: Universitas Indonesia)
[2] Hand D J. Principles of data mining. Drug safety. 2007 Jul 1;30(7):621-2.
[3] Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. Data &
Knowledge Engineering. 2007 Nov 1;63(2):503-27.
[4] Riza LS, Awaludin R, Sutarno H, Munir, Wibawa A. A model for auto generating sets of
examination items in educational assessment by using fuzzy c-means. World Transactions on
Engineering and Technology Education. 2017; 15(2): 114-119.
[5] Fadillah A P 2015 Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ) *Jurnal Teknik Informatika dan Sistem Informasi*, 1.

[6] Suyanto S M 2017 *Data Mining Untuk Klasifikasi dan Klatering Data* (Bandung: Informatika)

[7] Kabade RS, Gaikwad MS. Segmentation of brain tumour and its area calculation in brain MR images using K-mean clustering and fuzzy C-mean algorithm. *International Journal of Computer Science & Engineering Technology*. 2013 May;4(05).