P-ISSN: 2528-0260 E-ISSN: 2579-5392
p.723-728

CLUSTERING FOR SEARCHING TYPE OF HOUSE SUITABLE FOR NEW CONSUMER CANDIDATES USING K-MEANS CLUSTERING METHOD (CASE STUDY OF PT. MAXIMA JAYA PERKASA)

[1]WIWIET HERULAMBANG, [2]EKO PRASETYO, [3]AZZIYATI NUR

[1]Informatics Engineering Study Program, Faculty of Engineering

Bhayangkara University – Surabaya

e-mail: herulambang@ubhara.ac.id, eko@ubhara.ac.id, azziyatinur7709@gmail.com

ABSTRACT

For some Indonesian people, housing is one of the secondary needs, so that in choosing the right housing must be in accordance with the wishes of consumers. With the existence of PT. Maxima Jaya Perkasa, which was pioneered since 2012, in which the data on housing sales in the company has increased rapidly each year. Then data mining analysis can be done using the K-means Clustering method. K-means Clustering is a method of clustering non-hierarchical data which seeks to partition existing data into two or more groups. This method partitioned the data into groups so that the data with the same characteristics were entered into the same group and the data with different characteristics were grouped into other groups. This study uses data such as salary income, age, status, house prices and mortgage payments. The results of this study were conducted twice using 12 training data training data and 100 training data plus 1 as test data and obtained an accuracy value of 83% and error of 17%.

Keywords: *K-means Clustering, Houses, Euclidean, Salary, House Prices, KPR Payment*

1. INTRODUCTION

The house is a necessity secondary, so in the selection of housing that is right to fit the desires of consumers. In choosing a type of home there are several criteria used [1] such as building area, house price and mortgage installments. Existing home buyer consumer data on PT. Maxima Jaya Mighty background and different income because of that in this study how to process data that has been obtained from consumers who buying a house beforehand like salary income, age and status to find out consumer patterns based on the type of house from the data. By utilizing the K-Means method Clustering uses 100 existing consumer data at PT. Maima Jaya Perkasa as training data, the company is expected to be able to determine grouping house types based on data patterns previous consumers with potential new customers.

2. THEORETICAL BASIS

K-means is a method of clustering non-hierarchical data which seeks to partition existing data into two or more groups. This method partitioned data into groups so that data with the same characteristics were put in the same group and data with different characteristics were grouped into other groups [2].

General data grouping algorithm for the K-Means method:

a. Determine the number of groups

b. Randomly allocate data into groups

c. Calculate the centre of the group (centroid / average) of the data in each group

Formula:

$$C_i = \frac{1}{M}\sum_{j=1}^{M} xj$$

Information :

C_i = centroid

M = amount of data

d. Allocate each data to the nearest centroid.

Distance of data to centroid using Euclidean parameter formula:

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

Information :

D = distance between data x_2 and x_1

|. | = absolute value.

Data is explicitly re-allocated to groups that have centroids the closest distance from the data. Allocating data using formulas [3] :

$$a_{ij} = \begin{cases} 1 & d = min\{x_1 C_1\} \\ 0 & etc \end{cases}$$

Information :

a_{ij} = membership value of point xi to center group C1

d = shortest distance from data xi to K group after compared

Ci = centroid (group center) i

e. Return to step 3, if there is still data who moved the group, or if there is one changes in centroid value above the threshold value specified, or if the value changes on the objective function used still above the specified threshold value. The formula for calculating the objective function values:

$$J = \sum_{i=1}^{N}\sum_{l=1}^{K} a_{ic} D(x_1, C_1)^2$$

Information :

N = number of data K = group

ail = membership value of data point xi to the centre of group C1

C1 = centre of the 1st group

D (xi, C1) = the distance of point xi to group C1

a has a value of 0 or 1. If a data is a member of a group, ail value = 1. If not, the value of ail = 0. Iteration is stopped if the objective function value = 0.

3. SYSTEM DESIGN

Flowcharts are graphical depictions from the steps and the sequence of procedures from a program. Flowcharts usually make it easier solving a problem, especially a problem which needs to be studied further.

P-ISSN: 2528-0260 E-ISSN: 2579-5392
p.723-728

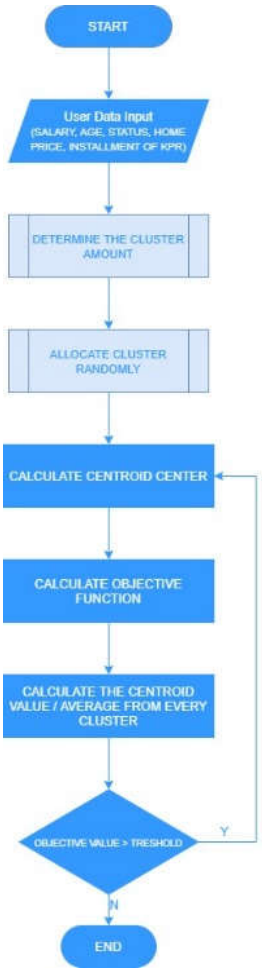

Figure 1. *System Flowchart*

The flowchart in Figure 1 above illustrates the process of grouping new customers using the k-means method. The first stage is the input of new consumer data in which there is income / salary, age, living status, house prices, home payments. The next stage will be processed using the k-means clustering method.

3.1. Data Flow Diagrams (DFD)

3.1.1. DFD Level 0

Admin as a provider of input and management of data that is in the system or admin's position is behind the system (backend). Home consumers as users to obtain information on the system or the position of prospective home buyers in front of the system (frontend).

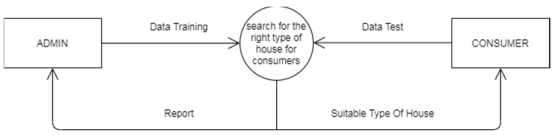

Figure 2. *DFD Level 0*

3.1.2. DFD Level 1

Admin performs authentication by logging in, then entering the training data and randomly assigning groups of data into the cluster and then stored in the database. From the perspective of prospective buyers (consumers) entering the test data and storing it into the database, then proceed with the normalization process and obtain the value of the test data entered into the system.

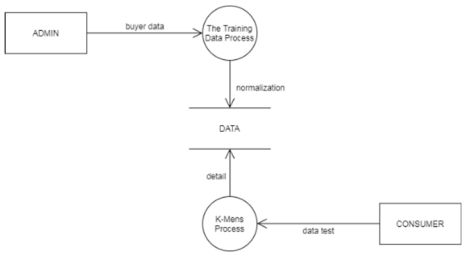

Figure 3. *DFD Level 1*

3.1.3. DFD Level 2

a. The Training Data Process

The first step taken by the admin is to log in, then enter the training data, from the normalized data then saved to the database.

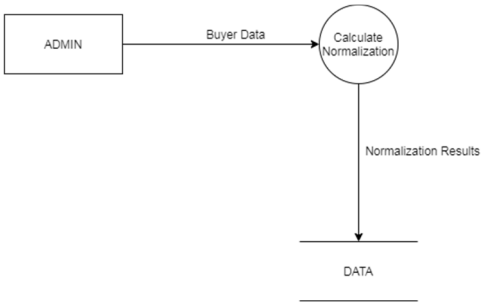

Figure 4. *DFD Level 2 Training Data Process*

b. K-Means Clustering Process

Initialize all the training data to get the initial cluster (group), then each feature of each data is summed and divided by the number of data in one group. Then each data is calculated against each value each centroid uses the Euclidean distance to get an objective function value, function value this objective determines the cessation of iteration calculation if the threshold = 0.

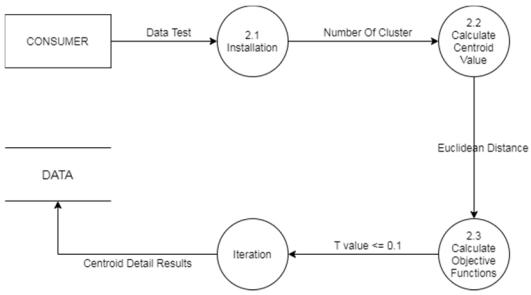

Figure 5. *DFD Level 2 K-Means Clustering*

4. Testing

The first and second testing uses data same test.

Figure 6. *Test Data*

P-ISSN: 2528-0260 E-ISSN: 2579-5392
p.723-728

a. First Testing Result

The final cluster results that can be recommended for potential new customers are based on 12 training data and 1 test data, namely cluster 4. Seen in Figure 7.

Figure 7. *First Testing Results*

b. Second Objective Results

The final cluster results are fast recommended for potential new customers based on 100 training data and 1 test data, namely cluster 3.

Figure 8. *Second Test Results*

5. ANALYSIS OF RESULTS

Based on the results of testing that has been conducted, in the first test there were 12 training data which in which there are 2 data that move clusters and there are 10 data that don't move clusters at the moment 88 training data were added second.

Table 1. *Analysis of Final Results*

DATA	Cluster Results		RESULTS
	Testing 1	Testing 2	
1	4	4	PERMANENT
2	1	1	PERMANENT
3	4	3	MOVE
4	4	4	PERMANENT
5	1	1	PERMANENT
6	1	1	PERMANENT
7	4	4	PERMANENT
8	3	3	PERMANENT
9	1	3	MOVE
10	2	2	PERMANENT
11	2	2	PERMANENT
12	3	3	PERMANENT

So the more training data that has the characteristics included, the better the validity value recommended for potential new customers.

6. CONCLUSION

1. Determination of clustering in the early stages of the K-Means algorithm is very influential on the results of the cluster, as in the results of tests conducted using 12 training data and 100 training data with the amount of data in each different cluster produces different cluster results as well.
2. The K-Means method can classify new potential customers based on variables of salary, age, house price, status and payments with a large amount of data but is not efficient in classifying potential customers appropriately.
   The search system for the type of house that is suitable for potential new customers using the K-Means clustering method has a 83% success rate with 2 tests.

REFERENCES

[1] A. Amborowati, "Sistem Penunjang Keputusan Pemilihan Perumahan dengan metode AHP menggunakan Expert Choice," *J. DASI*, vol. 9, no. 1, 2008.

[2] E. Prasetyo, *Data Mining Konseop dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andy Offset, 2012.

[3] J. B. MacQueen, *Some Methods for classification and Analysis of Multivariate Observations*, 1 ed. California: Berkeley, University of California Press, 1967.

[4] Suprawoto dan Totok, "Klasifikasi data mahasiswa menggunakan metode K-means untuk menunjang pemilihan strategi pemasaran," *J. Inform. dan Komput.*, vol. 1, no. 1, hal. 12–18, 2016.

[5] Ong dan J. Oscar, "Implementation of the K-Means Clustering Algorithm to determine marketing strategies," *J. Ind. Eng. Sci.*, vol. 12, no. 1, hal. 10–20, 2013.