The Grouping of Regions are Based on the Unemployment Rate in the Attacking Districts with the K-Means Method

Abstract—Each region experiences the number of idleness faced by the government. Currently, the Government through the Central Bureau of Statistics conducts a National Labor Force Survey to determine the number of unemployed from each region. Banten Province experienced a decline in the Labor Force Participation Rate (LFPR) of 1.60 percent. Currently, the open idleness grouping conducted in Serang District still uses group strata sourced from the office of the Provincial Central Bureau of Statistics. Based on the data already in the group BPS only displays idleness data based on certain criteria and does not show which region is the highest idleness rate. Therefore, it is necessary to classify the region to know which areas have a high and low idleness rate. In this case, the researcher collects data by looking at data already collected by BPS, by interviewing and viewing literature on idleness data in Banten Province. After viewing and collecting idleness data the idleness grouping of open idleness can still be done in another way, namely to see the proximity of the data point distance between one indicator with other indicators, one of them using clustering approach by using k-means method. K-means method is a non-hierarchical clustering method that seeks to partition existing data into one or more forms. By using the method k-means aim in facilitating the grouping of a region by looking at the number of low idleness rates or high level which results is a pie chart that describes the number of areas that have been grouped based on calculations by the k-means method. From the results of the image can be easily seen in the area where the highest and lowest idleness.

Keywords: idleness, clustering, k-means method, data mining

I. INTRODUCTION

Unemployment is one of the main problems faced by the Indonesian government. In August 2015 the number of unemployment based on BPS data was 7.56 million people and increased by 320 thousand compared to the same period in 2014 with 7.24 million people.

Based on data in Banten Province recorded in the National Labor Force Survey (Sakernas 2015), it can be identified as two main groups involved in formal and informal economic activities. Formal activities consist of residents who try to be assisted by permanent workers and those who are workers. While informal group activities are generally outside this status. Workers with the status of workers have the highest number compared to other jobs of 2.83 million. This figure increased by 48 thousand people in the period August 2014-August 2015.

Unemployment is indeed a very complex and very serious problem in Banten Province. This problem is influenced by several economic indicators, among others, the economic growth of a country concerned, the level of inflation, and the amount of wages in force.

If the Banten provincial government grows its economic growth, it is expected to affect the level of decline in disturbances in the Banten provincial government, and this is followed by an increase in the wage level. If the wage level increases it will affect the decrease in the unemployment rate.

The wage rate regulation set by the Banten provincial government will affect the magnitude of the existing unemployment rate. The higher the government sets the wages of workers, the lower the unemployment rate. According to J.R Hicks (in Kaufman and Hotchkiss, 1998), The theory of wage-fixing in a free market is a special case and general value theory. Wages are the price of labor.

Job problems are a problem that often occurs in some areas with an increase in population without the number of jobs that can cause an increase in unemployment, jobs that are unable to accommodate the large number of job seekers based on their level of expertise or education.

At present, the open unemployment grouping conducted in Serang District still uses group strata sourced from the Provincial Statistics Office. The grouping conducted by BPS was only limited to the unemployment rate group based on age indicators, which were divided into nine sub-indicators. The nine sub-indicators are age 15-19,20-24,25-29,30-34,35-39,40-44,45-50,51-54 and more than equal to 50. Even though the grouping of open unemployment rates is still can be done in another way, which is to see the proximity of data points between one indicator with another indicator, one of which uses the clustering approach.

Clustering is an activity of grouping data based on its similarity. This similarity will group one data into one group and other data to another group. In the case of open unemployment rate grouping, the process of determining groups only uses certain data intervals from the nine sub-indicators and then grouped into nine groups with reference to the number of existing indicators.

This research was conducted to classify the level of open unemployment in Serang Regency, Banten Province using the K-means algorithm. K-means is a non-hierarchical data clustering method that attempts to partition existing data into...
one or more clusters. So that data that has the same characteristics are grouped into the same cluster.

The use of the K-Means method will be tested using the clustering algorithm K-Means. The data used consists of seven features or data attributes of each, these features relate to the characteristics of unemployment, namely the sub-district, the number of households, population, population by sex, by age, and the number of unemployed. Then the data is grouped using the K-Means algorithm with the number of clusters specified in this study as many as 2 clusters.

II. ALGORITHM K-MEANS CLUSTERING

K-Means algorithm is a relatively simple algorithm for classifying or grouping a large number of objects with certain attributes into groups as many as K. K-Means is a non-hierarchical data clustering method that attempts to partition existing data into the form of one or more clusters or group. The K-Means algorithm was first introduced by MacQueen JB in 1976. In the K-Means algorithm, the number of K clusters was predetermined.

In the procedure for forming the K-Means Cluster, some steps can be taken, including:

- Determine the number of clusters (k) to be formed
- Awaken k centroidal (average of each cluster).
- Calculate the distance between each object and each centroid and input the object to the appropriate cluster based on the closest distance.
- Determine the centroid of the new cluster.
- Repeat steps 3 and 4 until there is no more object transfer between clusters.

According to Permatadevi, et al. (2013), if the clustering process for each k is completed, then to determine the most optimal number of clusters can be assessed using the Davies-Bouldin Index (DBI). Clustering with an optimal number of clusters is clustering that has a minimum DBI value.

\[
\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} R_{ij} \\
\text{dengan} \\
R_{ij} = \max_{i=1,...,k} R_{ij}, \quad R_{ij} = \frac{n_i + n_j}{d_{ij}} \\
S_i = \left( \frac{1}{n_i} \sum_{x \in S_i} d^2(x, v_i) \right)^{1/2}
\]

Information:
- k = Number of clusters
- Rij = Size of similarity between ni and nj
- Si = Size of i-cluster dispersion, i = 1, 2, ..., k
- dij = distance between centroid cluster i and j cluster centroid
- ni = Number of i-cluster members, i = 1, 2, ..., k
- vi = Centroid cluster of ni

The formula used in this calculation uses the equileidient formula as follows:

\[
d(x_i, y_i) = \sqrt{(x_i - y_i)^2 + (x_i - y_i)^2}
\]

Recalculate the cluster center with the current cluster membership. Cluster center is the average of all objects in a particular cluster. Calculation formula through cluster center determination. The shortest distance between clusters and data will determine the cluster position of a data.

\[
c_1 = \frac{x_1 + x_2 + x_3 + ...}{n}
\]

The center of the cluster will stop if the center of the last cluster with the center of the previous cluster does not change or does not change again.

After the clustering process is complete, the SSE value of each cluster will be calculated. The SSE value depends on the number of clusters and how the data is grouped into these clusters. The goal is to obtain a fixed partition or number of clusters that minimize the total squared error. The smaller the SSE value, the better the SSE value, the better the clustering value. The following formula is used.

\[
SSE = (C_1)^2 + (C_2)^2 + (C_3)^2 + (C_4)^2
\]

III. RESULTS AND DISCUSSION

The mechanism of test would be done was:

To determining the cluster based on data that has been available, it needed a flowchart to easier indeterminate the counting plot as a plot to find the result of cluster implementation to processing the data. There were some steps. They were:

![Flowchart Algoritma K-Means Clustering](image)

The data used consists of seven features or data attributes from each, these features relate to the characteristics of unemployment, namely the sub-district, the number of households, population, population based on sex, based on age, and the number of unemployed. Then the data is grouped using the K-Means algorithm ith the number of clusters specified in this study as many as 2 clusters. The following is the original open unemployment rate data in Serang Banten Regency by District.

To do clustering, first determine the number of clusters that will be used, the research I conducted has three clusters, with 29 sub-districts in Serang Banten and nine attributes from the sub-indicators.
After determining the number of clusters, the amount of data, and the number of attributes, it will form a data table that is ready to be used for clustering using the k-means method. The table that is ready to use the next step is the iteration process. The following iteration calculation process with two clusters that have been determined:

1st iteration
- Determination of the initial center of the cluster. For initial determination: First data is taken as the center of the First Cluster: (2077,1704,974,271,722,387,972,583,1075). The 14th data is taken as the center of the 2nd Cluster: (1972,964,769,398,334,2427,753,605,1654)
- Calculation of cluster center distance
- To measure the distance between the data and the center of the cluster, Euclidian distance is used, then the distance matrix, C1, and C2 will be obtained as follows:

| Kecamatan District | Berdasarkan umur |
|-------------------|-----------------|
|                   | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 | 50-54 | ≥55 |
| 1. Cinangka       | 2077   | 1704   | 974   | 271   | 722   | 387   | 972   | 583   | 1075 |
| 2. Padasarden     | 2762   | 2191   | 1950  | 1263  | 646   | 234   | 254   | 227   | 175  |
| 3. Cimangis       | 974    | 2065   | 877   | 908   | 172   | 292   | 564   | 543   | 154  |
| 4. Pabuaran       | 1200   | 1071   | 911   | 657   | 538   | 557   | 309   | 453   | 104  |
| 5. Gunung Sari    | 768    | 1784   | 197   | 787   | 130   | 231   | 119   | 117   | 64   |
| 6. Baros          | 778    | 3763   | 1773  | 261   | 213   | 452   | 435   | 752   | 375  |
| 7. Petir          | 785    | 2779   | 2391  | 376   | 547   | 646   | 545   | 867   | 360  |
| 8. Tunjung Teja   | 546    | 3925   | 702   | 532   | 134   | 766   | 342   | 152   | 49   |
| 9. Cikusas        | 792    | 2109   | 1197  | 110   | 663   | 853   | 765   | 1972  | 237  |
| 10. Pamarayan     | 654    | 2759   | 4329  | 328   | 324   | 974   | 166   | 129   | 104  |
| 11. Bandung       | 922    | 1789   | 937   | 123   | 518   | 453   | 103   | 487   | 102  |
| 12. Jawian        | 732    | 1765   | 432   | 724   | 325   | 1735  | 1308  | 1975  | 165  |
| 13. Koto          | 2858   | 1027   | 229   | 541   | 977   | 768   | 973   | 574   | 350  |
| 14. Cikande       | 1972   | 964    | 769   | 398   | 334   | 2427  | 753   | 605   | 1654 |

**TABLE 1. DATA FOR CLUSTERING**

Data Survey Pengangguran Terpuka Kecamatan Serang Tahun 2014

| Kecamatan District | Berdasarkan umur |
|-------------------|-----------------|
|                   | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 | 50-54 | ≥55 |
| 1. Cinangka       | 2077   | 1704   | 974   | 271   | 722   | 387   | 972   | 583   | 1075 |
| 2. Padasarden     | 2762   | 2191   | 1950  | 1263  | 646   | 234   | 254   | 227   | 175  |
| 3. Cimangis       | 974    | 2065   | 877   | 908   | 172   | 292   | 564   | 543   | 154  |
| 4. Pabuaran       | 1200   | 1071   | 911   | 657   | 538   | 557   | 309   | 453   | 104  |
| 5. Gunung Sari    | 768    | 1784   | 197   | 787   | 130   | 231   | 119   | 117   | 64   |
| 6. Baros          | 778    | 3763   | 1773  | 261   | 213   | 452   | 435   | 752   | 375  |
| 7. Petir          | 785    | 2779   | 2391  | 376   | 547   | 646   | 545   | 867   | 360  |
| 8. Tunjung Teja   | 546    | 3925   | 702   | 532   | 134   | 766   | 342   | 152   | 49   |
| 9. Cikusas        | 792    | 2109   | 1197  | 110   | 663   | 853   | 765   | 1972  | 237  |
| 10. Pamarayan     | 654    | 2759   | 4329  | 328   | 324   | 974   | 166   | 129   | 104  |
| 11. Bandung       | 922    | 1789   | 937   | 123   | 518   | 453   | 103   | 487   | 102  |
| 12. Jawian        | 732    | 1765   | 432   | 724   | 325   | 1735  | 1308  | 1975  | 165  |
| 13. Koto          | 2858   | 1027   | 229   | 541   | 977   | 768   | 973   | 574   | 350  |
| 14. Cikande       | 1972   | 964    | 769   | 398   | 334   | 2427  | 753   | 605   | 1654 |

**TABLE 2. DETERMINATION OF THE INITIAL CENTER OF THE CLUSTER**

| Center 1 | 2077 | 1704 | 974 | 271 | 722 | 387 | 972 | 583 | 1075 |
|----------|------|------|-----|-----|-----|-----|-----|-----|------|
| Center 2 | 1972 | 964  | 769 | 398 | 334 | 2427| 753 | 605 | 1654 |
TABLE 3. FIRST ITERATION

| Kecamatan District | Berdasarkan umur | C1 | C2 | Jarak Tepanek |
|-------------------|-----------------|----|----|--------------|
|                   | 15 - 19 | 20 - 24 | 25 - 29 | 30 - 34 | 35 - 39 | 40 - 44 | 45 - 49 | 50 - 54 | 55 + |
| 1                 |          |        |        |        |        |        |        |        |       |
| C1                |          |        |        |        |        |        |        |        |       |
| C2                |          |        |        |        |        |        |        |        |       |
| Jarak Tepanek     |          |        |        |        |        |        |        |        |       |

A. Determination of the new cluster center

After knowing the members of each cluster, then the new cluster center is calculated based on the data of each cluster member according to the cluster member center formula. So we get the following calculation:

TABLE 4. DETERMINATION OF NEW CLUSTERS FROM ITERATION 1

|               | C1         | C2         |
|---------------|------------|------------|
| C1 :          | 1452.833   | 626.7916667 |
| C2 : 664.8    | 310.6      | 406.2      |
| 2030.6        | 1484       | 1111.8     |
| 2396.8        |            |            |

TABLE 5. SECOND ITERATION

| Kecamatan District | Berdasarkan umur | C1 | C2 | Jarak Tepanek |
|-------------------|-----------------|----|----|--------------|
|                   | 15 - 19 | 20 - 24 | 25 - 29 | 30 - 34 | 35 - 39 | 40 - 44 | 45 - 49 | 50 - 54 | 55 + |
| 1                 |          |        |        |        |        |        |        |        |       |
| C1                |          |        |        |        |        |        |        |        |       |
| C2                |          |        |        |        |        |        |        |        |       |
| Jarak Tepanek     |          |        |        |        |        |        |        |        |       |

Advances in Social Science, Education and Humanities Research, volume 410

180
Because the 3rd iteration = the 4th iteration has the same member, there is no need to do iteration again.

**B. Data grouping**

Distance calculation results will be compared and selected the closest distance between the data and the center of the cluster, this distance indicates that the data is in one group with the closest cluster center.
To simplify the process of calculating k-means, a display program is made using the PHP language using the MySQL database. This data form display provides information about unemployment data sourced from BPS Banten Province.

In addition to displaying unemployment data that has been inputted, if there are weaknesses and advantages in inputting unemployment data, then in this main menu we can do input again and delete the data. If we want to delete our data, simply point the cursor to the data and a red cross will appear to be ready to be deleted, and if we want to add data then we click the add data button and the display will appear as follows.
C. Iteration Process

On this page store data that has been inputted according to data from BPS Banten Province. How to use the iteration form by selecting the first cluster center and the second cluster center. After selecting the center of the cluster, unemployment will be shown based on age criteria. The iteration process will stop if it has met the maximum iteration. The data that I will do iteration in Cinagka and Cikande sub-districts will produce four iterations as shown in the following figure.

If the selection of sub-districts to be included in the first cluster and the second cluster has been determined, the next step is to click the data process so that it can run the iteration process which will be shown in the figure below.
D. Iteration Process Results

On the results page will be displayed in the form of tables that have been colored according to the sub-districts included in cluster one and the second cluster. In addition to displaying in the form of a table, it will also display in the form of pie charts according to the clusters' presentations that are colored for clusters of one orange and two clusters of light blue.

From the results of the iteration obtained, it can be concluded that in the cluster of one district which has a high unemployment rate and low workers with a cluster center (1125; 1530,864; 1029,773; 664,2727; 645,7727; 620,3636; 419,727273; 753, 72738; 481,8181818) namely Cinangka, Padarincang, Cionas, Pabuaran, Gunung Sari, Baros, Lightning, Tunjung Teja, Cikeusal, Pamarayan, Bandung, Jawilan, Koppo, Kibin, Bojonegara Districts, Lebak Wangi, Carenang, Binuang, Cangusal, Bandung Tirtayasa, Tanara, Whereas at the center of the two clusters there is a low unemployment rate and a high level of workers with the cluster center point (1845,857; 644,7143; 388,2857; 7814, 4286; 406.8571; 1637,143; 14436,28571; 1170,286; 2346,714286) namely the Districts of Cikande, Keragilan, Waringin Kurung, Mancak, Anyar, Keramat Watu, Ciruas. In this paper, the researchers hope to the Banten Provincial Government that the results of this study can be used as information to improve the condition of human resources in each district so that every community member can work and advance the region so as not to lag behind other districts.
REFERENCES

[1] Sari, Mayang, Rahayu., Prediksi Data Anggaran Pendapatan Belanja Daerah Menggunakan Algoritma K-Means. SATIN - Sains dan Teknologi Informasi, Vol. 1, No. 2, Desember 2015.

[2] Metisen Melpa Benri, Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila. Jurnal Media Infotama Vol. 11 No. 2, September 2015.

[3] Murniati, Sri, Kurnang, Novaria, Yendi, Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Keabsahan Mahasiswa Pada Universitas Bina Darmo Palembang. Seminar Nasional Informatika 2013 (semnasIF 2013) UPN “Veteran” Yogyakarta, 18 Mei 2013, ISSN: 1979-2328.

[4] Johan Oscar Ong, Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University, Jurnal Ilmiah Teknik Industri, Vol. 12, No. 1, Juni 2013, ISSN 1412-6869.

[5] Ramaraj, E. Dr., Napoleon, D., An Efficient Segmentation of Remote Sensing Images For The Classification of Satellite Data Using K-Means Clustering Algorithm, IJIRST –International Journal for Innovative Research in Science & Technology| Volume 1 | Issue 6 | November 2014, ISSN (online): 2349-6010.

[6] Chanu, Jina, Yambem, Manglem, Khumanthem, Dhanachandra, Nameirakpam. Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm, 1877-0509 © 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license.

[7] I.C., Obagbua, O.O Oladipupo., O.J., Oyelade., Application of K-Means Clustering algorithm for prediction of Students' Academic Performance, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010.

[8] Arora, Kumar, Rakesh., Badal, Dharmendra, Dr., Evaluating Student’s Performance Using k-Means Clustering., IJCST Vol. 4, Issue 2, April - June 2013.

[9] Sulianta, Feri., Customer profiling pada supermarket menggunakan algoritma k-means dalam memilih Produk berdasarkan selera konsumen dengan daya beli maksimum., Jurnal Ilmiah Teknologi Informasi Terapan (ISSN:2407-3911) Volume 1 No.1, Desember 2014.

[10] Ratnsisih, Luh, Ni., Astiti, Made, Ni., Perancangan Sistem Pendukung Keputusan Penerimaan Karyawan Dengan Metode K-Means Clustering Berbasis Web, Jurnal Media Aplikom ISSN : 2086 - 972X Juni 2018.

[11] Jasmir., Analisis Profil Akademik Alumni Dengan Menggunakan Metode Klasifikasi K-Means Untuk Menentukan Strategi Promosi STIKOM Dinamika Bunga Jambi., MEDIA SISFO Vol.11, No.1, April 2017.