Classification of Generation By Population by Region in Indonesia Using K-Means Algorithm

Ririn Restu Aria¹, Susi Susilowati²

¹,²Universitas Bina Sarana

¹ririn.rra@bsi.ac.id,²susi.sss@bsi.ac.id

Abstract

Population growth caused by the year of birth led to the classification of population groups into several generations. Classification is important because in each generation there is based on population growth has different characteristics and traits in each generation. This research was conducted to try to group generations based on provinces in Indonesia based on the number of residents owned. When researchers analyzed the data obtained from population census data conducted by the central statistics agency (BPS). The method used in generation classification grouping uses the K-Means algorithm method based on 3 clusters. Based on the results of calculations carried out for 3 clusters obtained cluster 1 has 25 provinces, cluster 2 has 3 provinces and cluster 3 has 6 provinces. Based on the 2020 census that has been conducted, the current population is generation Z, generation and Pre Boomer generation is last in line so that from the available data can provide information about mapping in 34 provinces to be able to improve communication patterns between generations and fulfill public facilities that can be used every generation.

Keywords: Generation clustering, K-Means, Clustering

1. Introduction

The 2020 population census conducted by the central statistics agency conducted in February - September 2020 [1], based on the census of the population of Indonesia as many as 270,203,917 people who have a distribution of population that can be classified based on the generation seen based on the year of birth of the population. Based on the results of the 2020 census, Indonesia's population is dominated by generation Z who were born between 1977 and 2012, and then the millennial generation whose population was born from 1981 to 1996. In the process of classifying done for the population group using the literature of William H Frey. In every generation in Indonesia so that from the process can create a good communication process. From this background, generation grouping is needed to make it easier to know the number of generation clustering deployments in provinces in Indonesia. Based on the above assessment several methods can be done to find out the clustering process based on previous research[2][3][4]. Clustering is a method used to analyze data used to solve problems based on data grouping[5][6][7]. For the calculation process, researchers use the K-Means method as an algorithm in the data mining method in the process of grouping data[8]. In this research activity, the data used is divided into post generation Z, Generation Z, Millennial, Generation X, Boomer, Pre Boomer based on the population of 34 provinces, namely the spread of generation with a number that dominates, dominate and less dominates.

2. Research and Methodology

To conduct research is needed by using the overall literature of the recording of total demographic data in Indonesia from the BPS website related to the 2020 census data and also looking for references related to problems from books and related journals to be able
to get problem-solving and using K-Means algorithm in the calculation process based on 3 specified clusters that are very dominating, Dominate and dominate less.

2.1. Data Collection Stages
In the process of collecting data researchers take data from secondary parties based on population surveys conducted from census records conducted from February 2021 to September 2021 conducted online or by BPS officers then the data can be accessed on the BPS website.

2.2. Stages of Data Processing and Analysis
The generation clustering in 34 provinces that have been obtained will be processed first to be able to determine a cluster. The clustering process divides into 3 classes based on the data provided. Then the data is analyzed by calculating the weight of each index by selecting a randomly selected centroid number for the cluster.

2.3. Stages of Application of K-Means Algorithm Method
To be able to complete the K-Means algorithm several stages can be done including:

a) Determining the number of clusters formed from available data is 3 clustering: Very domineering, Dominating, and Less domineering.

b) Determining cluster values randomly, for initial data the specified value comes from West Sumatra Province, Riau Islands Province, and South Kalimantan Province. The results of the cluster value determination can be seen in table 2.

c) From each line that has been calculated, determine the cluster closest to the center of the cluster. This stage can be seen in table 3.

d) Determining the value for the center of the latest cluster to perform recalculation from the initial stage until the overall data from each cluster that we have no change back then the final result can be obtained and we can find out the number of clusters. This can be seen from the processing results with Rapidminer in figures 1, 2 and 3.

3. Results and Discussion
To conduct the process of grouping generation classification in the territory of Indonesia is done first with the selection of centroid data conducted randomly from 33 provinces from data obtained from BPS.

| Provincial Name        | Post Generation Z | Generation Z | Milenial Generation X | Boomer | Pre Boomer |
|------------------------|-------------------|--------------|------------------------|--------|------------|
| Aceh                    | 1.531.897         | 1.377.887    | 991.294                | 512.865| 78.737     |
| Sumatera Utara          | 2.198.567         | 4.241.259    | 3.791.537              | 2.814.656| 1.569.163  |
| Sumatera Barat          | 728.658           | 1.558.106    | 1.390.340              | 1.074.413| 677.138    |
| Riau                    | 975.045           | 1.831.988    | 1.704.452              | 1.262.954| 566.314    |
| Jambi                   | 501.619           | 975.166      | 940.102                | 728.651 | 361.400    |
| Sumatera Selatan        | 1.243.243         | 2.286.741    | 2.202.735              | 1.700.263| 915.080    |
| Bengkulu                | 280.112           | 553.664      | 532.287                | 407.474 | 211.996    |
| Lampung                 | 1.247.288         | 2.375.721    | 2.335.896              | 1.856.163| 1.033.585  |
| Kepulauan Bangka Belitung | 189.584         | 400.381      | 393.664                | 300.597 | 154.139    |
| Kepulauan Riau           | 306.559           | 562.655      | 578.183                | 430.132 | 168.344    |
| DKI Jakarta              | 1.291.532         | 2.678.252    | 2.816.278              | 2.404.005| 1.227.534  |
| Jawa Barat              | 6.212.835         | 12.965.399   | 12.653.335             | 10.169.066| 5.600.895  |
| Jawa Tengah             | 4.312.777         | 9.023.730    | 9.125.046              | 8.012.090| 5.241.102  |
| DI Yogyakarta           | 391.116           | 835.000      | 859.386                | 823.953 | 619.663    |
| Jawa Timur              | 4.565.674         | 9.643.116    | 10.028.010             | 9.263.150| 6.154.554  |
From the initial data owned then, researchers determined centroid data that became 3 randomly selected clusters, such as the data in table 2.

| Provincial Name | Post Generation Z | Generation Z | Milenial | Generation X | Boomer Z | Pre-Boomer |
|-----------------|--------------------|--------------|----------|--------------|----------|------------|
| Banten          | 1.675.105          | 3.264.335    | 3.258.101| 2.454.483    | 1.138.713| 163.525    |
| Bali            | 475.536            | 1.053.952    | 1.057.293| 982.678      | 630.340  | 117.607    |
| Nusa Tenggara Barat | 821.297          | 1.448.701    | 1.387.755| 1.050.838    | 537.338  | 74.163     |
| Nusa Tenggara Timur | 879.410          | 1.569.178    | 1.316.510| 914.174      | 551.055  | 95.239     |
| Kalimantan Barat | 750.200           | 1.521.612    | 1.452.788| 1.071.008    | 546.225  | 72.557     |
| Kalimantan Tengah | 352.020           | 755.008      | 734.453  | 544.257      | 252.143  | 32.088     |
| Kalimantan Selatan | 606.227          | 1.092.878    | 1.051.899| 848.903      | 425.910  | 47.767     |
| Kalimantan Timur | 502.134           | 1.055.423    | 1.023.266| 778.362      | 362.914  | 43.940     |
| Kalimantan Utara | 92.259            | 205.124      | 194.197  | 139.566      | 62.778   | 7.890      |
| Sulawesi Utara  | 322.731           | 661.469      | 645.872  | 562.155      | 365.293  | 64.403     |
| Sulawesi Tengah | 438.554           | 843.569      | 775.178  | 583.232      | 306.709  | 38.492     |
| Sulawesi Selatan | 1.144.702         | 2.567.400    | 2.312.797| 1.796.402    | 1.060.974| 191.234    |
| Sulawesi Tenggara | 402.455           | 786.855      | 682.934  | 476.912      | 242.303  | 33.416     |
| Gorontalo       | 151.217           | 335.659      | 306.123  | 233.762      | 127.093  | 17.827     |
| Sulawesi Barat  | 218.471           | 432.546      | 364.272  | 254.184      | 130.276  | 19.480     |
| Maluku          | 248.588           | 566.464      | 490.013  | 330.631      | 184.047  | 29.180     |
| Maluku Utara    | 190.938           | 387.963      | 344.657  | 232.997      | 113.825  | 12.557     |
| Papua Barat     | 155.749           | 341.528      | 328.307  | 210.930      | 90.113   | 7.441      |
| Papua           | 665.696           | 1.156.343    | 1.244.419| 852.966      | 351.144  | 33.139     |

Source: BPS Data

| Table 2. Determination of Centroid's initial data |
|-----------------|--------------------|--------------|----------|--------------|----------|------------|
| Post Generation Z | Generation Z | Milenial | Generation X | Boomer Z | Pre-Boomer |
| 728.658          | 1.558.106       | 1.390.340   | 1.074.413 | 677.138     | 105.817  |
| 306.559          | 562.655         | 578.183     | 430.132   | 168.334     | 18.701   |
| 606.227          | 1.092.878       | 1.051.899   | 848.903   | 425.910     | 47.767   |

After determining the centroid center then calculated based on the available data so that 3 clusters were obtained and determined the closest distance from the centroid center and the value of the cluster for each provincial data. The results of the calculation can be seen in table 3.

| Table 3. Iteration Process and Clusterization Results |
|-----------------|--------------------|--------------|----------|--------------|------------|
| C1              | C2                 | C3           | Closest Distance | Cluster |
| 195794,4945     | 1497461,54         | 598969,6967  | 195794,4945     | 1          |
| 4353847,083     | 5925632,89         | 5014308,176  | 4353847,083     | 1          |
| 0               | 1584379,363        | 680667,924   | 0            | 1          |
| 533824,3018     | 2044441,651        | 1139876,411  | 533824,3018    | 1          |
| 904149,0545     | 682713,283         | 236561,9432  | 236561,9432    | 3          |
| 1379920,513     | 294488,962         | 203948,023   | 1379920,513    | 1          |
| 1616795,03      | 73129,3522         | 953011,1436  | 73129,3522     | 2          |
| 1603927,755     | 317232,995         | 2586661,435  | 1603927,758    | 1          |
| 1872011,148     | 301737,456         | 1209063,084  | 301737,456     | 2          |
| 1584379,363     | 0                  | 915370,7266  | 0            | 2          |
| 2382666,446     | 3935593,104        | 3027603,829  | 2382666,446    | 1          |
| 19857869,04     | 2143186,91         | 2051873,83   | 19857869,04    | 1          |
| 14065985,21     | 15637133,62        | 14726769,77  | 14065985,21    | 1          |
| 992926,9785     | 730620,6052        | 443205,6762  | 443205,6762    | 3          |
| 15892659,47     | 17456693,22        | 16549908,2   | 15892659,47    | 1          |
To perform the calculation process with the Rapidminer application, the data we have is carried out the import process into the application by adjusting the data type and determination of the id, as seen in figure 1.

![Figure 1. Transformation Data Process](image)

After doing the process of reading the data, the next step is to determine the results of clustering, with $K = 3$ in the RapidMiner application, thus producing the cluster data output in figure 2.

![Figure 2. Clusterization results with Rapidminer](image)

| C1         | C2         | C3         | Closest Distance | Cluster |
|------------|------------|------------|------------------|---------|
| 3068097,015 | 4626298,133 | 3716906,221 | 3068097,015 | 1       |
| 663254,8838 | 1013907,877 | 288424,9161 | 288424,9161 | 3       |
| 204107,6111 | 1493303,542 | 5827315,293 | 204107,6111 | 1       |
| 264546,7401 | 1507648,389 | 627443,0556 | 264546,7401 | 1       |
| 154763,2111 | 1561337,214 | 655495,1809 | 154763,2111 | 1       |
| 129774,283  | 289334,3531 | 634664,7031 | 289334,3531 | 2       |
| 680667,924  | 915370,7266 | 0           | 0               | 3       |
| 793068,9267 | 799331,7683 | 148382,1182 | 148382,1182 | 3       |
| 2219749,733 | 645666,8656 | 1557092,481 | 645666,8656 | 2       |
| 1372711,887 | 270035,4309 | 719386,7004 | 270035,4309 | 2       |
| 1164541,462 | 220441,2154 | 501716,7612 | 422044,2154 | 2       |
| 1648814,168 | 328566,16   | 2318550,763 | 1648814,168 | 1       |
| 132405,505  | 279840,6107 | 666035,3305 | 279840,6107 | 2       |
| 2005075,342 | 435818,4547 | 1343570,992 | 435818,4547 | 2       |
| 1886605,249 | 320706,2214 | 1225199,683 | 320706,2214 | 2       |
| 1681259,467 | 146308,7524 | 1023807,738 | 146308,7524 | 2       |
| 1945721,004 | 37455,5237  | 1283564,704 | 37455,5237  | 2       |
| 2008999,732 | 434203,3346 | 1346268,173 | 434203,3346 | 2       |
| 589290,7715 | 1066645,104 | 224680,0134 | 224680,0134 | 3       |
| 121930015,1 | 123505138,3 | 12591681,8  | 121930015,1 | 1       |
Figure 3 and figure 4 shows provincial data that is a group of calculations with Rapidminer.

![Figure 3](image-url)

**Figure 3.** Provincial data based on clusterization

![Figure 4](image-url)

**Figure 4.** Clustering calculations in diagrams

4. Conclusion

Based on the results of research that has been done can be drawn conclusions:

a. K-Means algorithm used is able to map generation clustering into 3 clusters, namely the dominant cluster has 25 provinces, the dominant cluster has 6 provinces and the non-dominant cluster has 3 provinces obtained from 34 provinces in Indonesia.

b. From the results of the research that has been done, researchers suggest that further research be conducted to provide public facilities owned by a province that can be accessed by every generation.

References

[1] B. P. Statistik, “Sensus Penduduk 2020,” 2021. https://sensus.bps.go.id/topik/tabular/sp2020/85/175748/0.

[2] M. Y. Rizki, S. Maysaroh, and A. P. Windarto, “Implementasi K-Means Clustering dalam Mengelompokkan Minat Membaca Penduduk Menurut Wilayah,” *Just IT J. Sist. Informasi, Teknol. Inf. dan Komput.*, vol. Vol. 11, N, pp. 41–49, 2021.

[3] P. Marpaung and R. F. Siahaan, “Penerapan Algoritma K-Means Clustering Untuk Pemetaan Kepadatan Penduduk Berdasarkan Jumlah Penduduk Kota Meda,” *J. Sains Komput. Inform.*, vol. Volume 5 N, pp. 503–521, 2021.
[4] I. Ali, A. R. Dikananda, F. A. Ma’ruf, and M. Abdurohman, “Pengelompokan Jumlah Penduduk Berdasarkan Kategori Usia 0-18 Tahun dengan menggunakan Algoritma K-Means Untuk Menentukan Pengembangan Potensi desa Wisata di Desa Wisata Di Kabupaten Cirebon,” *J. Manaj. Inform.*, vol. Vol. 8 No., pp. 25–31, 2021.

[5] L. Rahmawati, S. W. Sihwi, and E. Suryani, “Analisa Clustering Menggunakan Metode K-Means dan Hierarchical Clustering (Studi Kasus : Dokumen Skripsi Jurusan Kimia, FMIPA. Universitas Sebelas Maret),” *ITSMART J. Teknol. dan Inf.*, vol. Vol.3 No.2, 2014.

[6] E. Prasetyo, *Data Mining : Konsep dan Aplikasi Menggunakan MATLAB*. 2012.

[7] M. Fauzi and Yudi, “Penerapan Algoritma K-Means Clustering untuk Mendeteksi Penyebaran penyakit TBC (Studi Kasus : Di Kabupaten Deli Serdang),” *J. Tek. Inform. Kaputama (JTIK)*, vol. Vol I No 2, pp. 1–7, 2017.

[8] F. L. Sibue and A. Sapta, “Pemetaan Siswa Berprestasi Menggunakan metode K-Means Clustering,” *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. Vol. IV No, pp. 85–92, 2017.