**MATICS**: Jurnal Ilmu Komputer dan Teknologi Informasi ISSN : 1978-161X(p); 2477-2550(e)
(*Journal of Computer Science and Information Technology)* *Volume* 14, No. 2(2022), pp 35-40
**Received**: October 26th 2021; **Accepted**: August 9th 2022; **Avalaible Online**: October 5th 2022

# CRISP-DM Method On Indonesian Micro Industries (UMKM) Using K-Means Clustering Algorithm

Cahya Wulandari, Yusuf Ansori, and Khadijah Fahmi H.H

***Abstract*—UMKM plays an important role in supporting the economy in Indonesia. As one of the steps to reduce poverty, the government should pay more attention to the growth of its UMKM based on existing data. Data of UMKM collected in 2018 in several economic sectors such as the Leather industry, Metal Industry, Woven Industry, Pottery Industry, Fabric Industry, Food and Beverage Industry, and Other Industry can be used as government guidelines in efforts to solve poverty problems by processing them using k-means clustering algorithm. The research was carried out using the CRISP-DM method and k-means clustering algorithm to determine the cluster of provinces so that the policy or decision making can be made more wisely. By using Rapid Miner, data processing can be done quickly. The result of the study shows that DBI values of 0.175 using k=3. Jawa Tengah and Jawa Timur plays important role in the development of UMKM in Indonesia especially in Fabric industry, the wood industry and food and beverage industry.**

***Index Terms*— UMKM, K-Means, CRISP-DM, RapidMiner**

## I. INTRODUCTION

The industrial groups categorized as micro, small, and medium enterprises are the largest economic foundation in Indonesia [1]. According to the UUD 1945 and TAP MPR No.XVI/MPR-RI/1998 on political economy in the context of Economic Democracy, Micro, Small, and Medium Enterprises (Usaha Mikro, Kecil dan Menengah, UMKM) need to be empowered as an integral part of the people's economy which has position, role, and strategic potential to realize the balanced, developed, and fairness of the structure of the national economy. The meaning of UMKM according to UU No.9 Tahun 1999 then changed to UU No.20 Pasal 1 Tahun 2008 on Micro, Small, and Medium Enterprises is:

Micro Enterprise is a productive business owned by an individual and/or an individual business entity that meets the criteria for Micro Enterprise as regulated in this UU (Law).

Small Enterprise is a productive business that is independent, which is carried out by an individual or an individual business entity that is not a subsidiary or a branch of a company that is owned, controlled, or is a part either directly or indirectly of a medium or large business that meets the criteria for Small Business as regulated in this UU.

Medium Enterprise is a productive business that is independent, which is carried out by an individual or business entity that is not a subsidiary or a branch of a company that is owned, controlled, or is a part either directly or indirectly of small or large business with total net assets or annual sales proceeds as regulated in this UU.

Large Enterprise is a productive business carried out by a business entity with a net worth or annual sales proceeds greater than Medium Enterprise. Large enterprises include national state-owned enterprises, joint ventures, and foreign businesses that run their economic activities in Indonesia.

Business World is Micro, Small, Medium, and Large Enterprises that carry out economic activities in Indonesia and are domiciled in Indonesia.

A total of 64,2 million UMKMs in Indonesia (99 percent of total business units) were recorded in 2018 and have a workforce of 116,98 million people (97 percent of the total workforce in the economic sector) 60 percent of Indonesia's Gross Domestic Product comes from UMKM. This suggests that UMKM plays an important role in maintaining the domestic economy [2].

UMKM is one of the efforts to reduce poverty and unemployment in Indonesia. Since February 2005, the President at that time Susilo Bambang Yudhoyono has planned a policy of a program to reduce poverty and unemployment by empowering UMKM. The

Manuscript received October 26, 2021. This work was supported in part by Informatics Engineering Department of Maulana Malik Ibrahim Islamic State University..

Cahya Wulandari, Author is with the Informatic Engineering Departement of Maulana Malik Ibrahim Islamic State University , Malang, Indonesia (email 18650076@student.uin-malang.ac.id )

Yusuf Ansori, Author is the Informatic Engineering Departement of Maulana Malik Ibrahim Islamic State University, Malang, Indonesia (e-mail: 18650041@student.uin-malang.ac.id ).

Khadijah Fahmi H.H., Author is the Informatic Engineering Departement of Maulana Malik Ibrahim Islamic State University, Malang, Indonesia (email khadijah.holle@uin-malang.ac.id )

fundamental objective of the policy of the program is to reduce the unemployment rate from 9-10 percent of the population to less than 6 percent and reduce the number of poor people from 15,97 percent to 8,2 percent within five years of his administration [3].

Clustering is the unsupervised classification of patterns (Observations, data items, or feature vectors) into groups (clusters). It is important to understand the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification) [4]. Clustering has been effectively applied in a variety of engineering and scientific fields such as psychology, biology, medicine, computer vision, communications, and remote sensing. Clustering algorithms can be broadly classified into hierarchical and partitional algorithms based on the structure of abstraction [5].

K-Means is a non-hierarchical data clustering method that attempts to partition existing data into groups/clusters so that data that has the same characteristics are grouped into one group and data that has different characteristics are grouped into the other groups. The purpose of clustering is to minimize the objective function set in the clustering process, which generally tries to minimize variations within a cluster and maximize variations between clusters [6].

The K-Means algorithm is one of the most popular data mining algorithms, so it is widely used in processing data. Such as using the K-Means algorithm to cluster corporate bond data, the result is that the K-Means method is more suitable than the FCM method [7]. Analyzing the distribution of UMKM in Malang city has concluded that UMKM in Malang city consists of 3 clusters when analyzed using the K-Means method [8]. mining web user data using K-Means shows that the K-Means algorithm is feasible and has the scalability to such data [9]. Clustering sales data using the K-Means method helps store management in managing stock in stores. This is very useful both for the shop owner and the buyers [10]. Besides nominal and numeric data, clustering using the K-Means can also be done in text data (text mining). The K-Means method can be used to improve the level of accuracy in predicting the classification of document data on the theme of student final assignments [11].

## II. RESEARCH METHOD

Cross Industry Standard Process Model for Data Mining (CRISP-DM) method can be applied to dataset research on the number of villages/sub-district according to the presence and type of small and micro industries using the K-Means algorithm. This method has six phases which are Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment [12]. The life cycle of CRISP-DM is presented in Figure. 1.

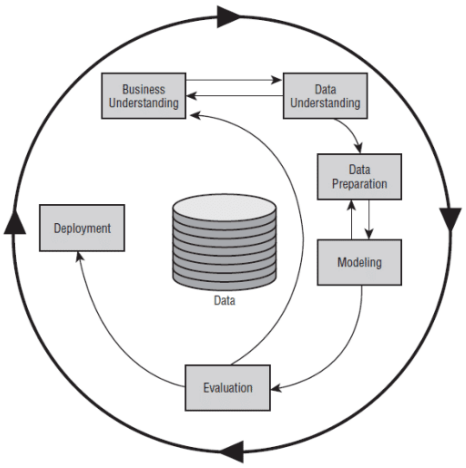


Fig. 1. CRISP-DM Method Phases

1. Business Understanding
   Business understanding aims at gathering as much information as possible before starting the research [26]. It's important to know the use of dataset to determine the potential of UMKM in each province in Indonesia. The goal will lead to policies that can be carried out on UMKM in Indonesia so that they can continue to grow effectively.
2. Data Understanding.
   Data understanding consists of data formatting, description, exploration, and data quality verification. Usually, data analysts return to business understanding to reconsider the aims of data [13]. The data was collected in 2018 provided by the Central Bureau of Statistics (Badan Pusat Statistik). The data consists of the distribution of UMKM in every province in Indonesia. The data has attributes, namely the province, the leather industry, the wood industry, the metal industry, the woven industry, the pottery industry, the fabric industry, the food and beverage industry, and other industries. Table 1 show the data attributes

Table 1. Data attributes

| Provinces | Leather | Wood | Metal | Woven | … | Other |
|---|---|---|---|---|---|---|
| Aceh | 25 | 1146 | 320 | 504 | … | 310 |
| Sumatera Utara | 87 | 1225 | 370 | 1004 | … | 524 |
| Sumatera Barat | 120 | 944 | 279 | 582 | … | 299 |
| … | … | … | … | … | … | … |
| Papua | 19 | 238 | 30 | 104 | … | 44 |

3. Data Preparation
   Data preparation to clean the data from missing values, and noise data. The numeric value that has more than three digits was detected as text value. The missing value was being reduced in order to obtain accurate results [14]. The numeric data that has problems was also fixed.
4. Modelling
   In this phase, the modeling is done using the K-means algorithm in RapidMiner. RapidMiner allows users to do data mining using visual code,

therefore the research time will be faster than manual. The flow diagram of the k-means algorithm is shown in Figure. 2

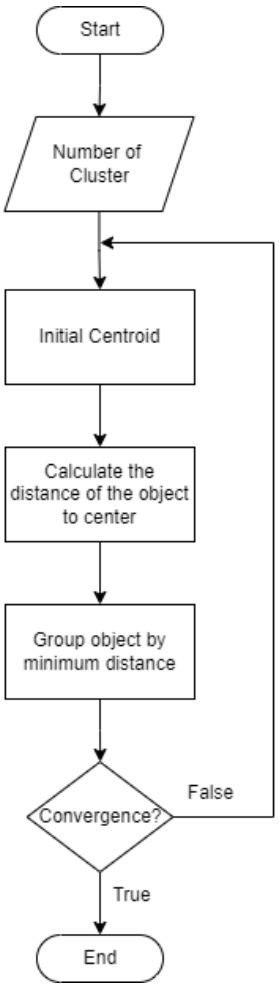


Fig. 2. Flowchart k-means algorithm

The k value used is five which means the data is grouped into five clusters. Then, the initial of centroid is done randomly. Calculate the distance of each data to the center of the cluster using the Euclidean Distance. The equation is written as below [15].

$$d(x_i, y_i) = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2} \quad (1)$$

Calculations produce the same class based on the nearest cluster. Then recalculate the center of cluster to determine the next iteration based on the average members in the cluster until no changing in the value at the center of the cluster, so data modeling using K-Means is complete.

5. Evaluation

The evaluation phase is carried out to maintain the results of the modelling phase to keep in line with the goals in the business understanding phase. The evaluation of this research is about the process of understanding the graphs to make them into useful knowledge for making decisions.

6. Deployment

This phase is the last of phases in CRISP-DM. The knowledge obtained from the evaluation provides an overview of how a policy will be carried out to increase the growth of UMKM in Indonesia. The results of the research are expected to be easily understood by others and can be implemented in making policy related to UMKM.

## III. RESULTS AND ANALYSIS

1. K Value

Determining the appropriate value of k serves to get a good number of clusters. The smaller the Davies-Bouldin Index (DBI) value, the better the clustering of the data. Therefore, the data were tested using different k values to see the smallest DBI value. The results of these trials are shown in Table 2.

Table 2. Davies-Bouldin Index On Different K

| K Value | DBI Value |
|---|---|
| 3 | 0.175 |
| 4 | 0.423 |
| 5 | 0.512 |
| 6 | 0.518 |
| 7 | 0.419 |
| 8 | 0.512 |

The results from table 2 shows that the smallest DBI is obtained at k=3, and then k=7, k=4, k=5 and k=8 got the same result, and the last is k=6. In conclusion, k value that suits all of the data based on its smallest DBI is 3.

2. Clustering

a. The Pottery Industries

The distribution of Pottery attribute is shown in Fig. 3

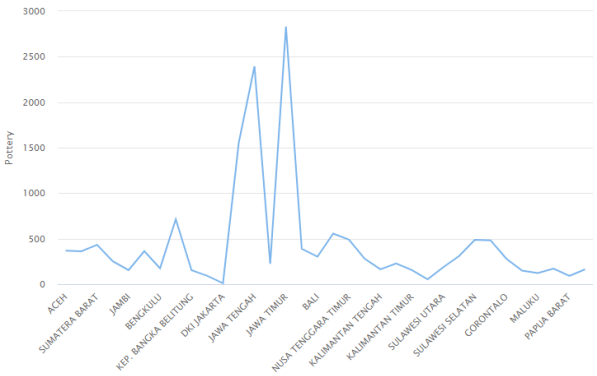


Fig. 3. Pottery attribute

b. The Woven Industries

The distribution of Woven attribute is shown in Fig. 4

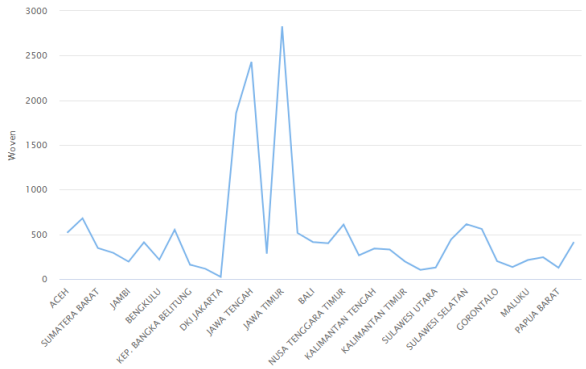


Fig. 4.Woven attribute

c. The Wood Industries

The distribution of Wood attribute is shown in Fig. 5

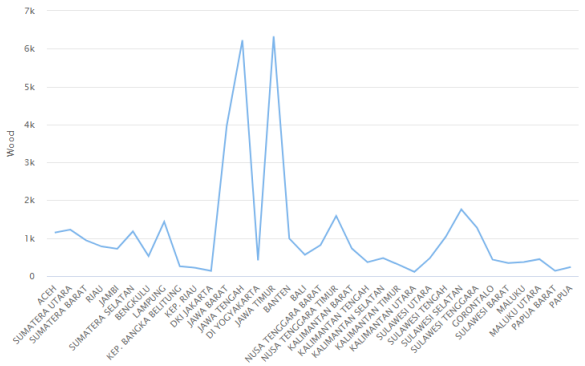


Fig. 5. Wood attribute

d. The Leather Industries

The distribution of Leather attribute is shown in Fig. 6

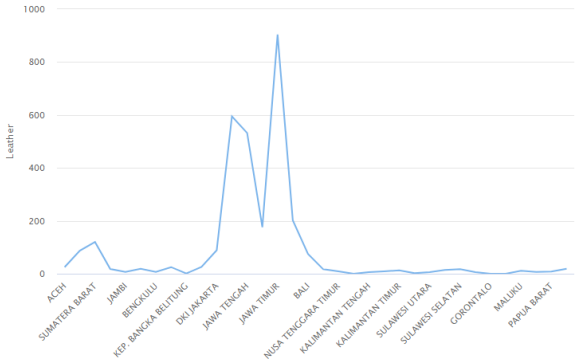


Fig. 6. Leather attribute

e. The Fabric Industries

The distribution of fabric attribute is shown in Fig. 7

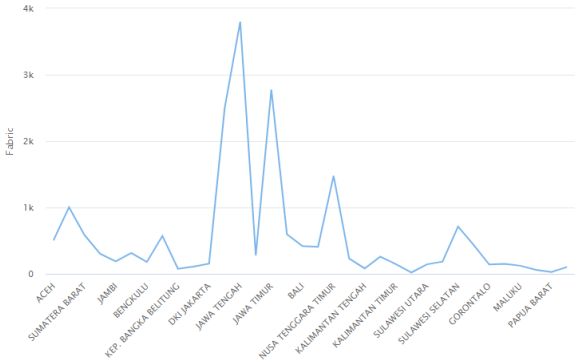


Fig. 7. Fabric attribute

f. The Metal Industries

The distribution of metal attribute is shown in Fig. 8

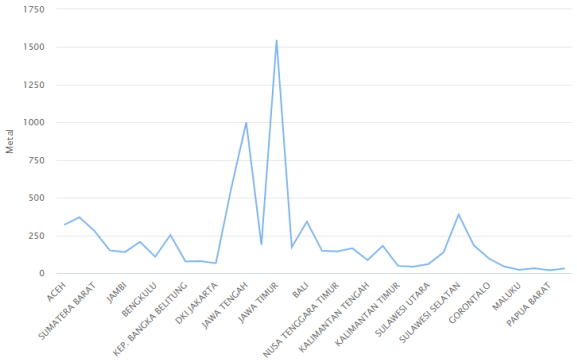


Fig. 8. Metal attribute

g. The Food and Beverage Industries

The distribution of metal attribute is shown in Fig.

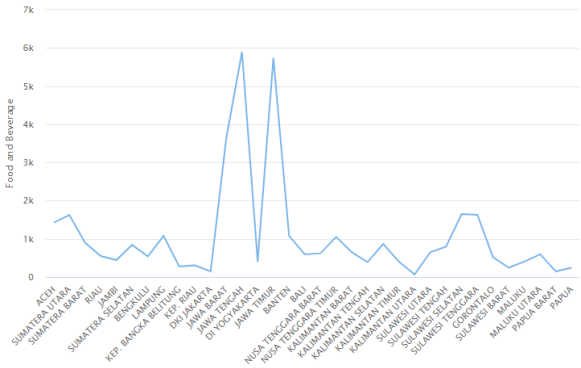


Fig. 9. Food and Beverage attribute

9

h. The Other Industries

The distribution of metal attribute is shown in Fig. 10

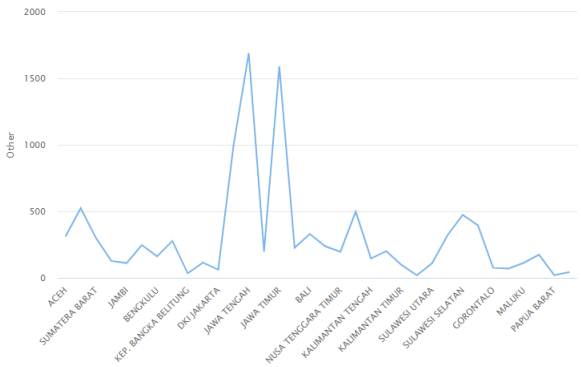


Fig. 10. Other attribute

The prepared data was clustered into three clusters using the k-means algorithm. Data centroid as shown in Table 3

Table 3. Centroid table

| Attribute | Cluster_0 | Cluster_1 | Cluster_2 |
|---|---|---|---|
| Pottery | 269.742 | 2612.0 | 1553.0 |
| Woven | 323.484 | 2629.0 | 1855.0 |
| wood | 691.581 | 6275.5 | 3965.0 |
| Leather | 32.935 | 718.0 | 595.0 |
| Fabric | 322.258 | 3288.5 | 2496.0 |
| Metal | 147.290 | 1273.5 | 557.0 |
| Food and Beverage | 680.710 | 5817.5 | 3637.0 |
| other | 200.290 | 1640.0 | 993.0 |

Label cluster_0 is cluster 1, label cluster_1 is cluster 2, and label cluster_2 is cluster 3. After finding the cluster, then analyzing each cluster based on the proximity between the centroid and the dataset. The distribution of cluster based on provinces shown in table 4.

Table 4. Distribution of Cluster

| Cluster | Frequency (province) | Average of centorid |
|---|---|---|
| Cluster_0 | 31 | 333.54 |
| Cluster_1 | 2 | 3031.75 |
| Cluster_2 | 1 | 1956.375 |

Cluster 1 has 31 provinces, cluster 2 has 2 provinces, and cluster 3 has 1 province.

Province in cluster 3 is Jawa Barat, and cluster 2 are Jawa Tengah dan Jawa Timur, cluster 1 are the provinces besides all of them.

Cluster 1 is a cluster that must be considered because it has a fairly low average compared to the other two clusters. The number of UMKM in cluster 1 is still less than in other clusters, this can be an evaluation for the government to pay more attention to the growth of UMKM in these provinces.

While in cluster 2 and cluster 3, the number of UMKM is more than cluster 1, especially in the fabric industry, the wood industry and food and beverage industry. These industries play important role in the development of UMKM in Indonesia.

## IV. CONCLUSION

Clustering on the growth of Indonesia's UMKM can be done by using the K-Means algorithm. Because the research was done using RapidMiner, data processing can be done quickly. The best DBI values 0.175 using k=3. Cluster 1 has 31 provinces, cluster 2 has 2 provinces, and cluster 3 has 1 province. The provinces with the highest UMKM growth are Jawa Tengah dan Jawa Timur. They plays important role in the development of UMKM in Indonesia especially in Fabric industry, the wood industry and food and beverage industry.

## REFERENCES

[1] Marijan, Kacung. "Mengembangkan industri kecil menengah melalui pendekatan kluster." *dalam Jurnal INSAN* 7, no. 3, 2005.
[2] Chaerani, Diah, Melda Noereast Talytha, Tomy Perdana, Endang Rusyaman, and Nurul Gusriani. "Pemetaan Usaha Mikro Kecil Menengah (UMKM) Pada Masa Pandemi Covid-19 Menggunakan Analisis Media Sosial Dalam Upaya Peningkatan Pendapatan". *Dharmakarya* 9, no. 4 (2020): 275-282.
[3] Prasetyo, P. Eko. "Peran usaha mikro kecil dan menengah (umkm) dalam kebijakan penanggulangan kemiskinan dan pengangguran." *Akmenika Upy* 2, no. 1, pp. 1-13, 2008.
[4] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." *ACM computing surveys (CSUR)* 31, no. 3, pp. 264-323, 1999.
[5] Krishna, K., and M. Narasimha Murty. "Genetic K-means algorithm." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 29, no. 3, pp. 433-439, 1999.
[6] Agusta, Yudi. "K-Means–Penerapan, Permasalahan dan Metode Terkait." *Jurnal Sistem dan Informatika* 3, no. 1, pp.47-60, 2007.
[7] Ningrat, Desy Rahmawati, I. Maruddani Di Asih, and Triastuti Wuryandari. "Analisis Cluster Dengan Algoritma K-Means Dan Fuzzy C-Means Clustering Untuk Pengelompokan Data Obligasi Korporasi." *Jurnal Gaussian* 5, no. 4, pp. 641-650, 2016.
[8] Puntoriza, Puntoriza, and Charitas Fibriani. "Analisis Persebaran UMKM Kota Malang Menggunakan Cluster K-means." *JOINS (Journal of Information System)* 5, no. 1, pp. 86-94, 2020.
[9] Xu, JinHua, and Hong Liu. "Web user clustering analysis based on KMeans algorithm." In *2010 International Conference on Information, Networking and Automation (ICINA)*. 2010.
[10] Indriyani, Fintri, and Eni Irfiani. "Clustering Data Penjualan pada Toko Perlengkapan Outdoor Menggunakan Metode K-Means." *JUITA: Jurnal Informatika* 7, no. 2, pp. 109-113, 2019.
[11] Somantri, Oman, Slamet Wiyono, and Dairoh Dairoh. "Metode k-means untuk optimasi klasifikasi tema tugas akhir mahasiswa menggunakan support vector machine (SVM)." *Scientific Journal of Informatics* 3, no. 1, pp. 34-35, 2016.
[12] Wirth, Rüdiger, and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining." In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. London, UK: Springer-Verlag, 2000.
[13] Bosnjak, Zita, Olivera Grljevic, and Sasa Bosnjak. "CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data." In *2009 5th International Symposium on Applied Computational Intelligence and Informatics*, pp. 509-514. IEEE, 2009.
[14] Nuraeni, Fitri, N. Nelis Febriani SM, Lina Listiani, and Eka Rahmawati. "Implementation of K-Means Algorithm with

Distance of Euclidean Proximity in Clustering Cases of Violence against Women and Children." In *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)*, vol. 1, pp. 162-167. IEEE, 2019.