EPI International Journal of Engineering
Volume 5, Number 2, February 2022, pp. 21-25
DOI: 10.25042/epi-ije.022022.04

*p*ISSN 2615-5109
*e*ISSN 2621-0541

Clustering Mining Equipment Productivity Data using K-Means Algorithm

Muhammad Sandi Arista Ikhsan Yahmid[a,*], Tony Chen[b], Muhammad Emirat Millenium Try[c], Karno Nugroho Silangin[d], Nurul Alifia Putri[e], Aryanti Virtanti Anas[f]

[a]Department of Mining Engineering, Faculty of Engineering, Hasanuddin University, Indonesia. Email: sandiarista85@gmail.com
[b]Department of Mining Engineering, Faculty of Engineering, Hasanuddin University, Indonesia. Email: tonychena33@gmail.com
[c]Department of Informatics Engineering, Faculty of Engineering, Hasanuddin University, Indonesia. Email: emirat.millenium@yahoo.co.id
[d]Department of Mining Engineering, Faculty of Engineering, Hasanuddin University, Indonesia. Email: karnonugroho.kn@gmail.com
[e]Department of Mining Engineering, Faculty of Engineering, Hasanuddin University, Indonesia. Email: nurulalifia29@gmail.com
[f]Department of Mining Engineering, Faculty of Engineering, Hasanuddin University, Indonesia. Email: virtanti@gmail.com

Abstract

Productivity and efficiency of mining equipment are among the most important factors contributing to unit mining cost. Therefore, knowing the condition of the machines on mining equipment is mandatory for mining supervisors. In analyzing the condition of mining equipment in large quantities requires a lot of energy and time. Based on these problems, a classification model is needed that can categorize mining equipment based on the performance and condition of the equipment. Clustering production data to determine the performance of mining equipment is important because it increases the productivity of production activities and reduces company losses. The purpose of this study is to obtain information on variables that affect the productivity of mining equipment and divide mining equipment categories based on production performance with data mining clustering techniques. The method in this study refers to the CRISP-DM with several adjustment using K-Means algorithm. The tools used are Jupyter Notebook with Python programming language. Variables used to cluster the mining equipment are Distance, Rites, and HaulDuration1 (working duration). There are 3 clusters formed based on the data used. Comparison of data on cluster with less production and less working duration can reach 58.50%. It shows that damaged equipment (less working duration) greatly affect the production of mining equipment as a whole.

Keywords: Data mining; mining equipment; machine learning; data clustering; K-Means algorithm

1. Introduction

There are many factors that affect productivity in mining production activities. A study of the digging and loading equipment can be carried out by monitoring the conditions in the field and the factors that affect the production capability of the mechanical equipment. The productivity of heavy equipment that is less than the target can be detrimental to the company [1], [2]. The productivity of the equipment can be seen from the ability of the equipment in its use. The factors that affect the productivity of equipment are the nature of the material, where the harder the type of material, the smaller the production of digging and loading equipment [3]. Productivity and efficiency of mining equipment are among the most important factors contributing to unit mining cost [4]. Therefore, knowing the condition of the machines on mining equipment is mandatory for mining supervisors.

However, one of the challenges of determining the performance of mining equipment requires consideration of quite a number of variables. In analyzing the condition of mining equipment in large quantities requires a lot of energy and time. Based on these problems, a classification model is needed that can categorize mining equipment based on the performance and condition of the equipment.

Clustering production data to determine the performance of mining equipment is important because it increases the productivity of production activities and reduces company losses. Related research, K-Means is a fairly simple clustering algorithm that partitions the dataset into several k clusters. The algorithm is quite easy to implement and run, relatively fast, easy to customize and widely used [5]. The main principle of this technique is to arrange *k* partitions/centroids/means from a set of data. The K-Means algorithm starts with the formation of a cluster partition at the beginning and then iteratively improves the cluster partition until there is no significant change in the cluster partition [6]. Data Mining (DM) is a series of processes to explore added value from a data set in the form of knowledge that has not been known

*Corresponding author.
*Jalan Poros Malino km. 6, Bontomarannu
Gowa, Indonesia, 92171*

manually. Several techniques that are often mentioned in the DM literature include clustering, classification, association rule mining, neural networks, and genetic algorithms [7].

Clustering is one of the sub categories of data mining and is a process where the same sample is divided into groups called clusters. Each cluster includes a sample where the members are similar to each other and different from the available samples from other groups [8]. Cluster analysis is a multivariate technique that has the main goal of grouping objects based on their characteristics. Cluster analysis classifies objects so that each object with the closest similarity to another object is in the same cluster [9]. The K-Means algorithm is one of the partitional algorithms, because K-Means is based on determining the initial number of groups by defining the initial centroid value [10]. The K–means algorithm will group data items in a dataset into a cluster based on the closest distance to the randomly selected initial centroid value which is the starting center point, the distance with all data will be calculated using the Euclidean Distance formula [11].

The purpose of this study is to obtain information on variables that affect the productivity of mining equipment and divide mining equipment categories based on production performance with data mining clustering techniques. The model created is expected to help companies know the performance of equipment quickly based on the cluster group of the equipment at a time.

2. Theoretical Basis

Artificial intelligence (AI) is a method used to solve a problem by imitating the capabilities of living things into a computer program. One branch of AI technique is machine learning (ML) which tries to imitate how human processes or intelligent creatures learn and generalize. The hallmark of ML is the existence of a training process or training. Therefore, ML requires data to be learned which is called training data. There are at least two main functions in ML, namely prediction and classification. The prediction or regression function is used by the machine to estimate the output of an input data based on the data that has been studied in training. While classification is a method in ML used by machines to sort or classify objects based on certain characteristics [12].

Data mining is the process of extracting previously unknown information and patterns from large amounts of data. Data mining uses artificial intelligence, statistics, mathematics and machine learning in the process of extracting information for decision making [13]. K-Means is a machine learning algorithm that is often used in data mining. K-Means groups data into several clusters that have the maximum similarity between data in one cluster. Determination of the most optimal number of clusters can be done using the elbow method [14].

K-Means Clustering algorithm is one of the clustering methods by partitioning from set data into cluster *K*. It is a distance-based clustering algorithm that divides data into a number of clusters in numerical attributes [15].

1. Determine the number of clusters K and the number of maximum iterations.
2. Perform the initialization process K midpoint cluster, then the equation of centroid count feature [15]:

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \quad (1)$$

   Equation 1 is done as much as *p* dimensions from *i* = 1 to $i = p$

3. Connect any observation data to the nearest cluster. Euclidean distance spacing measurements can be found using Eq. 2 [15].

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2)$$

4. Reallocation of data to each group based on comparison of distance between data with each group's centroid [15].

$$a_{ij} \begin{cases} 1 & d = min\{D(x_i, c_j) \\ 0 & otherwise \end{cases} \quad (3)$$

5. Recalculate the cluster midpoint position.
   a_{ij} is the value of the membership of point x_i to the center of the group c_1, d is the shortest distance from the data x_i to the group K after being compared, and c_1 is the center of the group to 1. The objective function used by this method is based on the distance and the value of the data membership in the group. The objective function can be determined using Eq. 4 [15].

$$J = \sum_{i=1}^{n}\sum_{l=1}^{k} a_{ij} D(x_i, c_1)^2 \quad (4)$$

   n is the amount of data, *k* is the number of groups, a_{i1} is the membership value of the data point x_i to the *c1* group followed a has a value of 0 or 1. If the data is a member of a group,
   the value a_{i1} = 1. If not, the value a_{i1} = 0.

6. If there is a change in the cluster midpoint position or number of iterations < the maximum number of iterations, return to step 3. If not, then return the clustering result [15].

Determination of the optimal number of clusters can use the Elbow method. The Elbow method provides ideas by choosing the cluster value and then adding the cluster value to be used as a data model in determining the best cluster. This method will produce information in determining the best number of clusters by looking at the percentage of comparison results between the number of clusters that form an angle at a point [16].

Determination of the optimal number of clusters using the elbow method is done by looking at the SSE (Sum of Squared Error) value. SSE is shown as in Eq. 5 and Eq. 6 [17].

$$SSE = \sum_{i=1}^{K}\sum_{xj \in Ci}^{N_i} |X_j - M_i|^2 \quad (5)$$

$$M_i = \frac{\sum_{i=1}^{n} X_i}{n} \quad (6)$$

where,
K = Number of clusters
X = {$x_1, x_2, \ldots, x_i, \ldots, x_n$}
C = {$C_1, C_2, \ldots, C_i, \ldots, C_n$}
M = Centroid of cluster (C_i)

Table1. Types and sources of research data

No.	Data Type	Time span	Amount of data
1	Mining equipment production data recorded every shift (Secondary)	2 months	15,427 rows and 62 columns data
2	Mining equipment incident status data (Secondary)	2 months	120,876 rows and 23 columns data

3. Method

The data used in the form of secondary data derived from the research of Mining Engineering students at Hasanuddin University in one of the coal mining companies. The data consists of 2 different files, namely production data and mining equipment incident status data. The time span of this data is two months with details which can be seen in Table 1.

There are many methods for analyzing data. One of the most popular is the Cross Industry Standard Process for Data Mining (CRISP-DM). The method in this study refers to the CRISP-DM with several adjustment. The tools used are Jupyter Notebook with Python programming language.

Data processing stages generally take up most of the time. Very large amounts of data can be analyzed further to get more information. Data Mining is a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract information. After further data exploration, an anomaly was found in the data used, so that the anomaly data was cleaned. Then visualization of correlation data between several variables and the production of mining equipment is carried out. This visualization aims to better understand the relationship of variables that influence the production of mining equipment.

The last stage is done with one of the machine learning techniques, namely clustering with the K-Means algorithm. The K-Means algorithm will divide the data based on several influential variables given. Determination of the optimal number of clusters is done by the Elbow method. The Elbow method is done by looking at the point where the decrease in inertia (how far apart each sample is in a cluster) is no longer significant. This division of equipment categories will be useful as recommendations for equipment performance on a shift.

4. Research Discussion

The results of the first data mining process obtained ineffective data that actually did not need to be entered into the data recording. The data in question is data that has only one unique data and data that has the same information purpose as data in other columns. The data is discarded so that it can reduce data storage and data recording becomes faster.

During the visualization process, an anomaly was found in the data. It can be seen in Fig. 2, that many plots of data are gathered at long distances and the production is much higher than the other data. After being traced, it turns out that all of these data belong to the HBM999 type of equipment. It is assumed that this HBM999 represents many devices that do not yet have an ID so that they are combined into type HBM999. This data is then removed from the analysis and re-visualization is carried out.

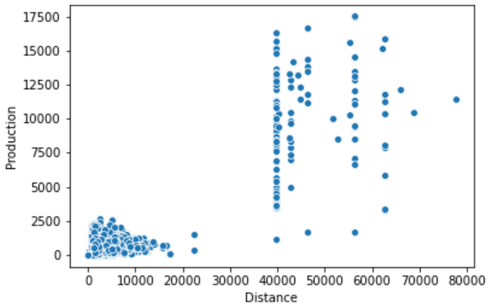

Figure 2. Anomalies in the distance and production variable data plots

The data visualization process is carried out for each type of variable that is considered to affect the production of equipment. The variables analyzed included Shift, Distance, Rites, Capacity, HaulDuration1 (duration of work), HaulDuration2 (duration of delay), HaulDuration3 (duration of standby) and HaulDuration4 (duration of maintenance). Visualization is made by scatter plot for each variable pair with production and heatmap to display correlation values. Figure 3 shows the difference in the average productivity of each equipment on the morning shift and night shift not much different. Even so, the average productivity on the morning shift is slightly higher than the night shift.

Figure 4 shows the pattern of equipment production with respect to distance, rotation and equipment capacity. The farther the distance of conveyance of the equipment, the production tends to be smaller. On the other hand, for large rites and capacities, the production tends to be greater.

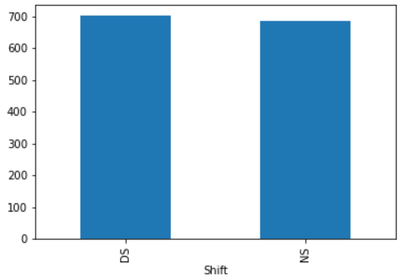

Figure 3. Production graph against shift time

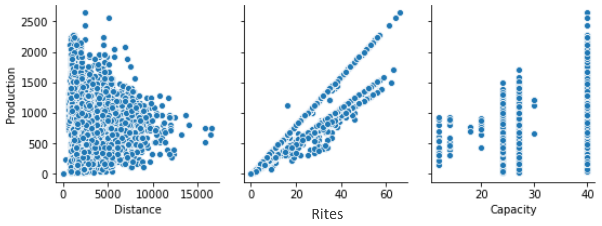

Figure 4. Plot production data against distance, rites and equipment capacity

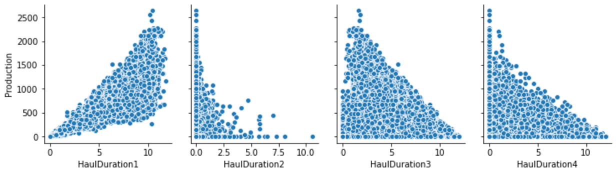

Figure 5. Plot production data against equipment activity duration

Figure 5 shows scatter plots of production data against equipment activity duration. *HaulDuration1* (working duration of equipment) tends to be directly proportional to its production. Meanwhile, HaulDuration2 (snooze duration), HaulDuration3 (standby duration) and HaulDuration4 (maintenance duration) tend to be inversely proportional to their production. Figure 6 shows the correlation between variables. It can be seen that Rites and HaulDuration1 have a high positive correlation to production, while HaulDuration2, HaulDuration3 and HaulDuration4 have a negative correlation with production.

Based on the relationship between variables and their relation to production, only a few variables are taken that are used in the clustering stage. the variables used are Distance, Rites and HaulDuration1. These three variables are used to see the production performance of mining equipment.

The last step is data clustering using the K-Means machine learning method. Determination of the optimal number of clusters is done by the Elbow method as shown in Fig. 7. It was found that the change in the inertia value was no longer significant at the point k=3. So it was determined to divide the data into 3 clusters.

The results of the K-Means clustering can be seen in Fig. 8. Clusters with label 0 (blue) indicate mining equipment with very low production even with the distance between the location where the material is taken to disposal is relatively close. Label 1 (orange) shows mining equipment with medium production. Most of the equipment that has long distances is also included in this cluster although its production is classified as less. Label 2 (green) indicates mining equipment with good to very high production performance. Comparison of data labelled 0 to the overall data reached 24.65%.

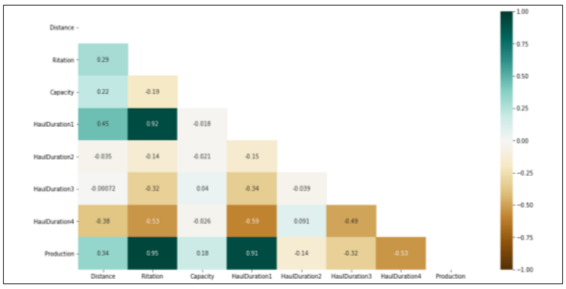

Figure 6. Heatmap of correlation between variables

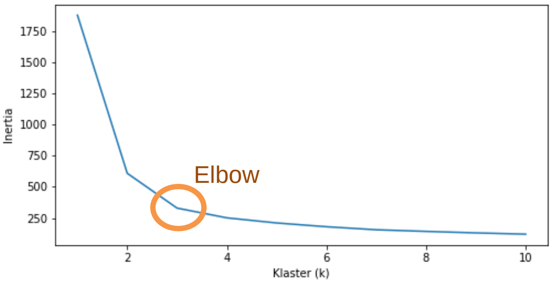

Figure 7. Clustering with elbow

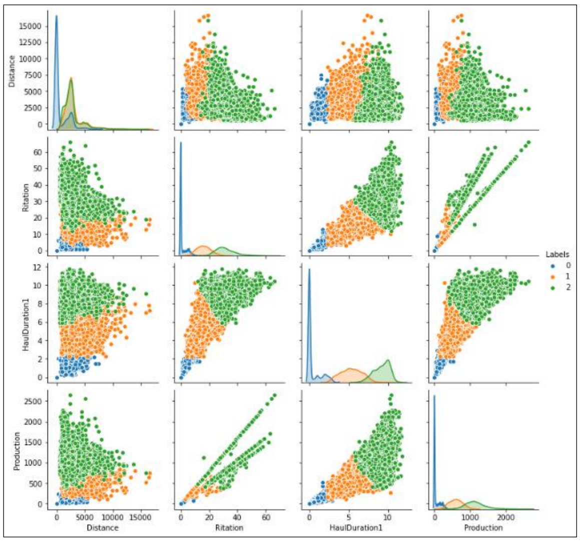

Figure 8. Mining equipment data plot along with its cluster

Based on Fig. 8, it can be seen that the duration of the equipment work greatly affects its production. HaulDuration1 shows how long the equipment has been running while HaulDuration4 shows how long the equipment has been down. The greater the time the equipment works, the production tends to be even greater. On the other hand, the greater the HaulDuration4 time, the smaller the production will tend to be. Figure 8 shows the correlation value for HaulDuration1 with production reaching 0.91 while the correlation value for HaulDuration4 with production is -0.53.

Comparison of data in cluster 0 to the overall data reached 24.65%. In addition, data in cluster 1 shows equipment whose production is still not optimal with a percentage reaching 33.85% of the total data. This shows that damaged equipment with less duration of work greatly affects the production of mining equipment as a whole. Product productivity can be increased by anticipating potential breakdowns early with a predictive maintenance program. In addition, mining equipment management can be maximized because potential damage can be detected early.

5. Conclusions

Based on the results of the analysis of mining equipment productivity with data mining techniques is found that the working duration of the equipment greatly affects its production. Correlation value between equipment work duration and its production is 0.9, while the correlation value between maintenance duration and its production is -0.53. Comparison of data on cluster with less production can reach 58.50% (total of clusters 0 and 1). It shows that damaged equipment (less working duration) greatly affect the production of mining equipment as a whole.

The K-Means model created in this study still cannot be said to be optimal for needs at the industrial level. This is because the data used is still raw and limited in terms of quantity and complexity. Further pre-processing and comparison with other models are required. However, this model can sufficiently demonstrate the production performance of mining equipment with the three most influential variables.

Acknowledgements

A big thank you for Ministry of Education and Culture Republic of Indonesia for providing a forum for Indonesian students to be creative and work through the Student Creativity Program (PKM), especially in PKM-RE activities, because of this program this research can be carried out.

References

[1] D. Sujatmiko, "Analisis Produktivitas Alat Berat Studi Kasus Proyek Pembangunan PLTU Talaud 2 x 3 MW Sulawesi Utara," Universitas Gadjah Mada, 2013.

[2] A. T. Tenriajeng, "Pemindahan Tanah Mekanis," Universitas Gunadarma, 2003.

[3] W. Hartono, Pemindahan Tanah Mekanik (Alat-alat Berat). Surakarta, Indonesia: Lembaga Pengembangan Interscience Publication, 2005.

[4] J. Paraszczak, "Understanding and Assessment of Mining Equipment Effectiveness," Trans. Inst. Min. Metall. Sect. A Min. Technol., vol. 114, no. 3, pp. 147–151, 2005.

[5] X. Wu and V. Kumar, The top ten algorithms in data mining. London, United Kingdom: CRC Press, 2009.

[6] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning and Tool, 3th Editio. Burlington, Vermont, USA: Morgan Kaufmann Publisher, 2011.

[7] Lindawati, "Data mining dengan teknik clustering dalam pengklasifikasian data mahasiswa studi kasus prediksi lama studi mahasiswa Universitas Bina Nusantara," Semin. Nas. Inform. 2008 (semnasIF 2008), vol. 1, no. 5, pp. 174–180, 2008.

[8] Ediyanto, N. M. Mara, and N. Satyahadewi, "Pengklasifikasian karakteristik dengan metode K-Means Cluster Analysis," Bul. Ilm. Mat. Stat. dan Ter., vol. 2, no. 2, pp. 133–136, 2013.

[9] S. G. Farhac, R. Yasin, and R. K. Seyyed, "Combining Clustering Algorithhms for Provide Marketing Policy in Electronic Stores," Int. J. Program. Lang. Appl., vol. 4, no. 1, 2014.

[10] T. S. Madhulatha, "An overview on clustering methods," IOSR J. Eng., vol. 2, no. 4, pp. 719–725, 2012.

[11] A. Agrawal and H. Gupta, "Global K-Means (gkm) clustering algorithm: A survey," Int. J. Comput. Appl., vol. 79, no. 2, pp. 20–24, 2013.

[12] A. Ahmad, "Mengenal Artificial Intelligence, Machine Learning, Neural Network, dan Deep Learning," J. Teknol. Indones., 2017.

[13] P. Bhatia, Data Mining and Warehousing Principles and Practical Techniques, Edisi ke-1. United Kingdom: Cambridge University Press, 2019.

[14] A. Winarta and W. J. Kurniawan., "Optimasi Cluster K-Means Menggunakan Metode Elbow pada Data Pengguna Narkoba dengan Pemrograman Python," J. Tek. Inform. Kaputama, vol. 5, no. 1, pp. 113–119, 2021.

[15] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster," in IOP Conf. Series: Materials Science and Engineering 336, p. 2018.

[16] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," Int. J. Comput. Appl., vol. 105, no. 9, pp. 975–8887, 2014.

[17] T. Thinsungnoen, N. Kaoungku, P. Durongdumronchai, K. Kerdprasop, and N. Kerdprasop., "The Clustering Validity with Silhouette and Sum of Squared Errors," in Proceedings of the 3rd International Conference on Industrial Application Engineering 2015, 2015, pp. 44–51.