Increasing the Settlement of Industrial Conflicts by Applying the K-Means Algorithm

Ucik Faruqo Heni, Kristophorus Hadiono

Universitas Stikubank, Semarang, 50241, Indonesia

Abstract

Disputes over the execution of the contract between workers and employers sometimes occur. The settlement of conflicts that exceed the time according to Law no. 2/2004 will be handled by the fields of industrial relations, worker welfare, and social security for workers from the City/Regency Manpower and Transmigration Office. Due to the limited number of mediators, a particular method is needed to increase the number of dispute resolutions available. The dispute data used in this study was from the year 2020 and was obtained from the Semarang City Manpower and Transmigration Office. Data processing is carried out by applying the steps in the CRISP-DM method, including applying the K-Means algorithm. The clustering process based on K-Means Algorithm is started after the data cleaning process and then grouped into 3 parts based on the amount of mediation time and the number of workers. 5 groups of data were obtained after being processed using the k-means algorithm based on the ease of case resolution, the number of workers involved, and the type of dispute problem. The first order for ease of case resolution is held by data groups 0 and 4, then followed by data groups 2 and 3 in second place. The last position is in data group 1

Keywords: clustering, k-means algorithm, dispute settlement

1. Introduction

Disputes in the implementation of employment contracts between workers and employers can occur. The causes of the dispute, according to law no. 2/2004 are (1) conflict of rights; (2) conflict of interest; (3) termination of employment; and (4) conflicts between trade unions in one company. According to the law, the resolution of disputes is usually handled by the unit of Industrial Relations, Workers' Welfare, and Labor Social Security (this unit is often referred to as HI). Officers in the HI unit in carrying out their duties are always guided by applicable laws, one of which is Law no. 2/2004.

In article 3 paragraph 3 of Law no. 2/2004 states that the duration time for resolving a dispute through bipartite is 30 days (1 month). If one of the disputing parties refuses to negotiate or the negotiations do not reach an agreement, then the bipartite negotiation is considered a failure. The failed dispute resolution can be reported to the relevant agency in the local manpower sector and of course accompanied by evidence of the negotiation efforts that have been carried out.

Disputes that are reported to the HI unit often cannot be resolved within the timeframe specified by existing regulations. This creates problems if a new dispute is reported to the HI unit, and dispute resolution cannot be handled quickly. So that the possibility of accumulation of dispute resolution problems is increased. In order to avoid the accumulation of dispute resolution, it is proposed to apply the K-Means Clustering algorithm. The goal is for the HI unit to obtain accurate data in determining the priority of resolving disputes.

It is a common knowledge that the K-means clustering method is the popular method to do clustering. The K-means clustering method is a clustering method that is based on th

e similarity between items in the same group and the difference between items in differen t groups [1]. By applying the k-means clustering method, it is hoped that groups/clusters will be formed which can be used to determine the priorities for dispute resolution proble ms.

In a study, the application of the k-means cluster method was carried out to classify fl oods that occurred in urban areas. In addition to the k-means cluster method, the entropy weight increase method was applied to the grouping of urban areas affected by flooding. The application of these methods is intended to evaluate the risk of flooding that occurs i n urban areas [2]. The k-means cluster method was also applied in another study to meas ure the performance of students so that students who have similar performance can be pla ced in the same group [3].

In addition to the k-means cluster method, other methods are needed for data processi ng in order to obtain the required information. One of them is the CRISP-DM method, w hich is used by [4] to segment the academic quality of students. This research is similar t o the research conducted by [3] in grouping students with the same performance into the same group. However, the researcher [3] did not use the CRISP-DM method when the re search was conducted. Utilization of the CRISP-DM method aims to make large data min ing processes faster, cheaper, and reliable, and can be managed and the process repeated [5], [6].

The k-means cluster method can be combined with the k-nearest neighbor method as part of the process of making a film recommendation system [7]. In this study, the film d ataset was taken from Kaggle which was then used as input for the k-means, k-nearest ne ighbor, collaborative filtering, and content-based filtering cluster methods. These method s were tested with the same data, and the results of the tests carried out will be used as th e introduction stage of the machine learning concept and recommendation system. The st udy concluded that the proposed model gives better results where when the number of cl usters decreases, the RMSE (Root Mean Squared Error) value also decreases.

In performing customer segmentation, three (3) clustering algorithms were used to obt ain groups of customers who have the same behavior in the same segment and with differ ent patterns in different segments. This research was conducted with the aim of getting a group of customers who have high purchasing characteristics as well as frequent custome rs. Also, customers who have high purchasing characteristics even though they visit occa sionally [8]. The results of this study provide 5 groups that can be referred to as customer groups with the characteristics of careless, careful, standard customers, target, and reason able.

Another study using the k-means algorithm was carried out with the aim of making th e potential level of rubber plantation farmers [9]. The research data was taken from the C entral Statistics Agency (BPS) of North Sumatra Province. The results of this study obtai ned 3 cluster groups (high, medium, and low), where the high group for the potential of r ubber plantations obtained 1 cluster. The middle group for potential rubber plantations o btained 6 clusters, and the low group for potential rubber plantations obtained 19 clusters.

Utilization of the k-means algorithm for clustering also occurs within the scope of the feasibility area for planting corn [10]. The Agriculture Service of South Lampung Regen cy found that several areas were potential areas for maize. Therefore, it is necessary to gr oup corn-producing areas to find out which areas produce large and small amounts of cor n. The distribution of the harvest is usually done based on the name of the corn-producin g district. The research was conducted by looking for common characteristics which wer e then grouped into one cluster, other characteristics were grouped into other clusters. dat a in one cluster has a small degree of variation.

Bipolar disorder (BD) is a disease that requires serious treatment. In terms of the bipol ar disorder treatment, is needed according to the level of symptoms experienced by the p atient. With the application of the k-means algorithm, this study resulted in 5 (five) cluste r proposals. The cluster consists of groups based on parameters (1) the number of hospita lizations and suicide attempts; (2) co-existing personality disorders; (3) body mass index; (4) metabolic syndrome; (5) the number of comorbid physical illnesses; (6) cognitive fun ction; (7) become permanently disabled due to BD; (8) leisure and global functions; and (9) patients' perceptions of their mental health and functioning. The results of this study p rovide a suggestion for the management of patients with bipolar disorder. The treatment i s grouped into 5 clusters based on the severity of bipolar disorder experienced by the pati ent [11].

Increasing crop yields is something that farmers really want. An increase in yields can be done by, one them, predicting crop yields. Predicting crop yields can be done by modi fying the planting process and care throughout the growing season. These modification a ctions have the aim of increasing crop yields more than before. With the help of UAV (U nmanned Aerial Vehicle), rice fields pictures can be taken. Then, the color of the rice pla nt is processed using the k-means algorithm, the graph-cut (KCG) algorithm, and the lab color space is used to group the rice fields. The results show that the proposed method ca n classify areas with a relative error rate of 6% - 33% better than before (by 1% - 31%) [12].

From several scientific articles that have been discussed, a common thread can be dra wn that the k-means algorithm is a popular algorithm for grouping/clustering. The results from previous research show that the k-means algorithm is useful for grouping objects ba sed on similar features or performance. Thus, this method can also be used to group disp ute issues in the HI unit to increase the chances of resolving disputes.

2. Material and Methods

The research method used to conduct this research is as follows. The researcher interv iews the HI unit officers and doing literature studies on the k-means algorithm first. The data used for this research is the settlement of industrial relations problems data in 2020. The data was obtained from the Semarang City Manpower and Transmigration Office. T he initial process after the data is obtained is the identifying the parameters that will be u sed in conducting the clustering process with the k-means algorithm. Furthermore, the da ta processing is carried out by following the stages of the process in the CRISP-DM (Cro ss Industry Standard Process for Data Mining) method. The K-Means clustering will be c alculated with formula (1).

$$d_{euclidean}(x, y) = \sqrt{\square} \quad (1)$$

3. Result and Discussion

This part will describe how to increase the settlement dispute. It consists of several ste ps like this:

1. Identification and Data Cleaning

The dataset can be seen that have 1302 rows, 13 columns, and 3 data types. Some of the columns contain different amounts of data. The initial analysis was done, be cause of the dataset conditions. The first step was data cleaning, which is done by deleting rows that have no-value. This action was chosen because the comparison number of missing values per attribute to the total data was small (below 10%). T he NEGOTIATION_QUANTITY attribute has 0.4% missing values and the PRO

BLEM_TYPE attribute has 0.07%. After deleting the no-value columns in rows, t he total row was changed into 1295 row of data.

2. Split the dataset into 3 part of data
   The dataset will be divided into 3 data frames, namely data1, data2, and data3. Th e first group was form based on the NEGOTIATION_QUANTITY attribute. The r eason to form the first cluster was based on the rule that the case was considered d one in time if the case settlement was done with 4-time negotiation. Based on that, it is necessary to look for cluster values that have members below 4 or close to 4. The second cluster was formed based on the number of workers involved. Judging from the boxplot result, the MAN_POWER_QUANTITY attribute has a high data range, and the quartile values tend to be downwards. However, the MAN_POWE R_QUANTITY attribute was quite influential on the outcome of the settlement so it cannot be removed. Based on this fact, It needs to be grouped first through the s econd cluster. The third cluster is an integration between the first and the second c luster. The third cluster will be the final result to increase the settlement of the dis pute. Here is the description of the first, second, and the third cluster process.
   a. The First Cluster
      The parameter used to form the first cluster was the NEGOTIATION_QU ANTITY attribute (column). As mentioned in the previous step, the first cl uster is shaped based on the number of negotiations or tramped on the easi est to the highest endeavor part of the settlement case. Before doing the k-means clustering, there needed to determine the optimal number. The simp le way to identify the optimal number was by implementing the Elbow met hod. The Elbow Method is a technic for determining the k value before im plementing the k-means clustering method. Figure 1 was the result of the E lbow Method for the Optimal k value.

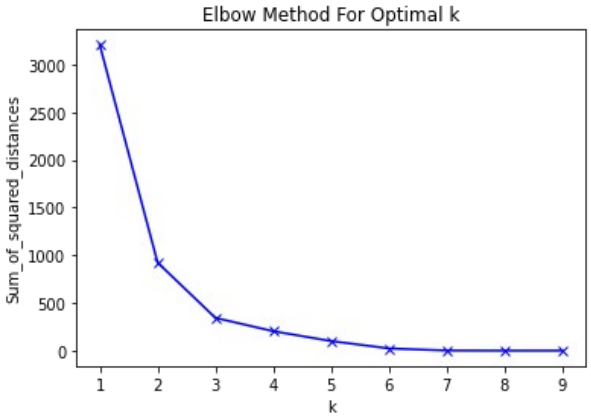

FIGURE 1. ELBOW METHOD RESULT FOR THE FIRST CLUSTER.

The Elbow method results in Fig. 1 showed the extreme slop was in betwe en numbers 2 and 3. From the elbow method result which gives two numb ers, it is needed to implement the Silhouette score method. The Silhouette score method is used to make sure the quality of the clustering created by t he k-means algorithm. If the silhouette score result comes near to 1 (one) means that the data point is very near within the cluster center and it indica tes the cluster number is the optimal one. The average silhouette score for 2 clusters is 0.6574, and for 3 clusters is 0.7171; which can be concluded t hat the optimal number of clusters is 3.
After this step, the k-means clustering was implemented and the results sh owed in Table 1.

TABLE 1. THE FIRST CLUSTER RESULT

Cluster Label	Number of Negotiation	Category
0	1, 2	Easy
1	5, 6, 7, 8	Difficult
2	3, 4	Average

b. The Second Cluster
The second cluster was made based on the MAN_POWER_QUANTITY a ttribute. The grouping decision for the second cluster was based on the nu mber of workers that involved in the dispute case. The forming process of t he second cluster are the same with the first cluster. The first step was impl ement the elbow method, and it can be seen in Figure 2.

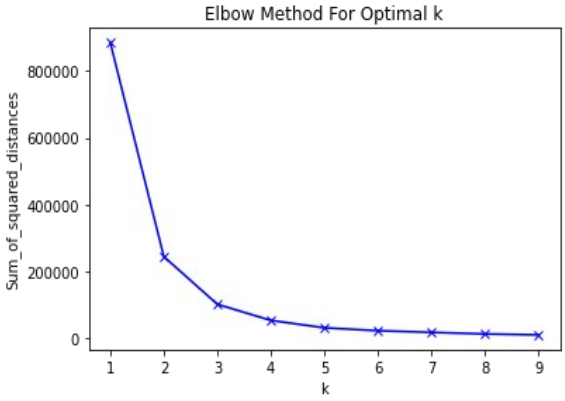

FIGURE 2. ELBOW METHOD RESULT FOR THE SECOND CLUSTER

From Fig. 2, it gives information that the extreme slop is between 2 and 3. Then, the silhouette score method need to be applied. The silhouette score for 2 cluster was 0.6194 and for 3 cluster was 0.6265. This result can be co ncluded that the optimal cluster number for the second cluster is 3. The ai m of the second cluster that chose the MAN_POWER_QUANTITY attrib ute was to see how many of the workers were involved in the dispute when the dispute case was reported to the HI unit. The result of the k-means clus tering is presented in Table 2.

TABLE 2. THE SECOND CLUSTER RESULT

Label	Number of Involved Workers	Category
0	< = 25	low
1	> 60	substansial
2	26 < x <= 60	moderate

3. Combining the first and the second cluster
When combining the first and the second cluster, not all attribute was used. The ad ded attribute that used in the combining cluster besides the attribute from the first and the second cluster were the EASY_SETTLEMENT_OF_CASE_BASED _ON_THE_NUMBER_OF_WORKERS _INVOLVED and PROBLEM_TYPE at tribute.
4. Testing the Cluster
In this part, the process for the third cluster is the same as the first and the second cluster. The first process for the third cluster is the implementation of the Elbow Method to determine the optimal value for the k in the k-means clustering algorith m. The elbow method resulting only one point where the extreme slop occurs, whi

ch is in the number 5 (Fig. 3). The silhouette score for this point is 0.7984 and this value is closer to 1.

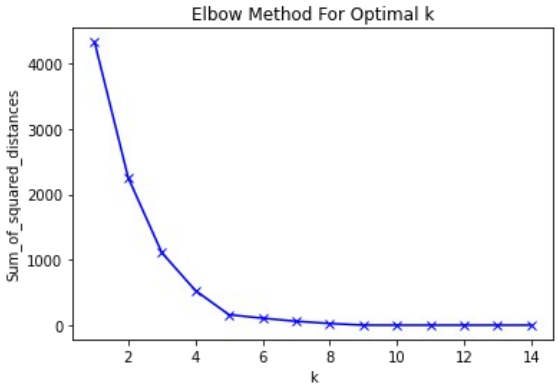

FIGURE 3. THE ELBOW METHOD RESULT FOR THE THIRD CLUSTER

5. Clustering Result

This part will explain the final step of the last step from the CRISP-DM which is e valuation. The evaluation was done by looking at the result of the model as showe d in Figure 4.

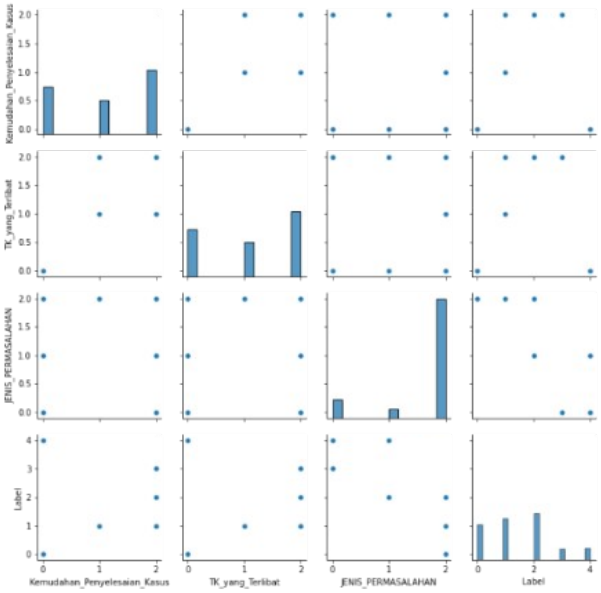

FIGURE 4. VISUALIZATION IMAGE USING PAIRPLOT FOR THE FINAL RESULT OF THE T HIRD CLUSTERING.

After the k-means clustering was implemented, the Figure 4 shows the results. The cl uster is labeled as 0, 1, 2, 3, 4; thus the final number of cluster is 5. Table 3 gives the deta il of the final number of cluster including the attribute that used for clustering.

TABLE 3. THE FINAL CLUSTER RESULT

Cluster label	Ease Of Dispute Resolution	Number of Involved Workers	Dispute Category
0	Easy	Low	Layoffs
1	Difficult, Average	Substansial, Moderate	Layoffs
2	Average	Moderate	Interest, Layoffs
3	Average	Moderate	Interest
4	Easy	Moderate	Interest, Rights

4. Acknowledgement

Based on the research results from the 2020 dataset, it can be assured that:

1. The optimal value of k for the first cluster is 3, and the second cluster is 3.

2. After combining the first cluster and the second cluster, the optimal value of k for the combined cluster is 5 with the silhouette score of 0.7984.
3. The third cluster gives 5 levels of a grouping which is labeled as 0 to 4. The third cluster can be categorized based on Ease Of Dispute Resolution, Number of Involved Workers, and Dispute Category level. This result can be seen in Table 3.
4. From Table 3, it can be drawn that clusters 0 and 4 were the easiest cluster thus they can be handled as the first priority. For the second priority was clusters 2 and 3 which are the moderate case that needs to be solved. The last priority was cluster 1, the most difficult one among the other cluster.

In the closing conclusion, the research results are 5 clusters namely 0, 1, 2, 3, and 4. The HI unit can work out the settlement cases on clusters 0 and 4 which are the number one priority to be settled. Then cases in clusters 2 and 3 as the second priority to be settled which are the moderate cases and the last one is the difficult cases in cluster 1. This research still can be updated and expanded in order to get the best way of dispute settlement clustering. The next step for future research is implementing the K-Nearest Neighborhood method which might give better results.

5. Reference list

[1] 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," IOP Conf. Ser. Mater. Sci. Eng., vol. 336, p. 012017, Apr. 2018, doi: 10.1088/1757-899X/336/1/012017.

[2] H. Xu, C. Ma, J. Lian, K. Xu, and E. Chaima, "Urban flooding risk assessment based on an integrated k-means cluster algorithm and improved entropy weight method in the region of Haikou, China," J. Hydrol., vol. 563, pp. 975–986, Aug. 2018, doi: 10.1016/j.jhydrol.2018.06.060.

[3] R. Vankayalapati, K. B. Ghutugade, R. Vannapuram, and B. P. S. Prasanna, "K-Means Algorithm for Clustering of Learners Performance Levels Using Machine Learning Techniques," Rev. Intell. Artif., vol. 35, no. 1, pp. 99–104, Feb. 2021, doi: 10.18280/ria.350112.

[4] Y. Suhanda, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," J. Teknol. Inform. Dan Komput., vol. 6, no. 2, pp. 12–20, Sep. 2020, doi: 10.37012/jtik.v6i2.299.

[5] R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min., no. 24959, pp. 29–39, 2000, doi: 10.1.1.198.5133.

[6] F. Martinez-Plumed et al., "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," IEEE Trans. Knowl. Data Eng., vol. 33, no. 8, pp. 3048–3061, Aug. 2021, doi: 10.1109/TKDE.2019.2962680.

[7] R. Ahuja, A. Solanki, and A. Nayyar, "Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor," in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, Jan. 2019, pp. 263–268. doi: 10.1109/CONFLUENCE.2019.8776969.

[8] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer Segmentation using K-means Clustering," in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, Dec. 2018, pp. 135–139. doi: 10.1109/CTEMS.2018.8769171.

[9] P. Alkhairi and A. P. Windarto, "Penerapan K-Means Cluster Pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara," Semin. Nas. Teknol. Komput. Sains SAINTEKS,

vol. 1, no. 1, Art. no. 1, Feb. 2019, Accessed: Jun. 23, 2022. [Online]. Available: https://prosiding.seminar-id.com/index.php/sainteks/article/view/228

[10] A. A. Aldino, D. Darwis, A. T. Prastowo, and C. Sujana, "Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency," J. Phys. Conf. Ser., vol. 1751, no. 1, p. 012038, Jan. 2021, doi: 10.1088/1742-6596/1751/1/012038.

[11] L. de la Fuente-Tomas et al., "Classification of patients with bipolar disorder using k-means clustering," PLOS ONE, vol. 14, no. 1, p. e0210314, Jan. 2019, doi: 10.1371/journal.pone.0210314.

[12] M. N. Reza, I. S. Na, S. W. Baek, and K.-H. Lee, "Rice yield estimation based on K-means clustering with graph-cut segmentation using low-altitude UAV images," Biosyst. Eng., vol. 177, pp. 109–121, Jan. 2019, doi: 10.1016/j.biosystemseng.2018.09.014.