I.J. Information Engineering and Electronic Business, 2020, 4, 15-20
Published Online August 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2020.04.02

Descriptive Modeling Uses K-Means Clustering for Employee Presence Mapping

Warnia Nengsih
Information Technology of Department, Pekanbaru,28265 Indonesia
Email: warnia@pcr.ac.id

Muhammad Mahrus Zain
Information Technology of Department, Pekanbaru,28265 Indonesia
Email: mahrus@pcr.ac.id

Received: 21 August 2019; Accepted: 07 May 2020; Published: 08 August 2020

Abstract: Human resource is valuable asset for an agency. The success of an institution is not only determined by the quality of its human resources, but also by the level of discipline. The discipline of an employee in an institution can be seen and measured by the level of attendance in doing a job, because the level of attendance is one of the factors that determine productivity. The current problem is the management level of the company that has difficulty in monitoring and controlling the employee attendance data. There needs to be a mapping and grouping to find out patterns of absence. Mapping or patterns that are obtained help management levels to monitor employees, take approaches and take action so as to improve employee discipline. In this study, it was used descriptive modeling with the implementation of the k-means clustering method. The results of the mapping obtained help the management level in controlling and monitoring as a reference for the next policy maker.

Index Terms: Descriptive Modelling, K-Means, Clustering

1. Introduction

The role of management level is very large in implementing employee work discipline. In addition, the role of the management level also determines whether the implementation of the discipline is appropriate and in line with the company's vision and mission. The weakness of the implementation of discipline so far is the lack of supervision of the management level on employee discipline development. Though work discipline plays an important role for the continuity of the work of the organization. With high work discipline from employees, it will have a positive impact on the achievement of work effectiveness and efficiency, which of course will be directly proportional to work productivity.

In relation to the level of attendance, provides a measure or criterion of employee discipline as follows: "When the absenteeism or absenteeism per month reaches 2-3%, it is said that the employee has high discipline. When the absentee level reaches 15-20% per month, it is said that employee discipline is low, and if it is between the two conditions above, then the discipline level of employees can be said to be moderate. " [1]

At present almost all companies have used media or technology for employee attendance. All employee attendance data is stored in a database. The data and information can be seen on certain URLs, so that the management level or the employee concerned can see for themselves the presence information. But the problem is that the management level has difficulties in monitoring or controlling and also in getting patterns or knowledge and mapping from the data and information. [2]. There needs to be a grouping into the categories of whether an employee falls into the cluster of absenteeism and severe delays, absenteeism and quite late delays and absenteeism and delays that are still in the normal level. The pattern and knowledge gained will help the management level to monitor, approach and take action so that it can improve work discipline.

2. Research Method

The stages to be carried out in this study are:

A. Justification/Business Case Assessment

At this stage the specification of user requirements will be produced which is in the form of data, information, procedures and information system limitations.

B. Design and Building System

The focus point of the design is towards grouping employees and mapping and visualizing employee data and the resulting pattern.

C. Implementation of the K-Means Clustering Method

The number of clusters used is 3 *clusters* namely Delays and absences of employees' severe category, Delays and absences of employees' moderate category, Delays and absences of employees' Normal category.

D. Implementation of the K-Means Clustering

K-means algorithm uses input in the form of parameters, number k, and set of data sets of objects to be included in the k class / cluster so that the *intracluster* similarity is higher while the inter-cluster similarity is lower. Cluster similarity is measured based on the average value of all objects in the cluster, which can be seen as a cluster center [3].

This algorithm will select a number of objects from several objects that are in a data set. Each selected object represents the average value of a number of clusters. Then the remaining objects, each of which will be assigned to a cluster that has been determined based on its similarity with the average value of each cluster. After that, a change in the average value of the cluster is already made into several cluster members. The three steps above are done iteratively (repeated) until there is no change in the average value and all data has been distributed to each of the existing clusters.

The data in K-means are classified in advance into K clusters to define the k-centroid value of each cluster. The Euclidean distance between an object and all the nearby centroid is calculated as per the formula (1)

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

||Xi(j) – Cj ||2 is the nearest distance measure between a data point xij and the Centroid Cj. To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids It halts creating and optimizing clusters when either [4]:

-The centroids have stabilized, there is no change in their values because the clustering has been successful.
-The defined number of iterations has been achieved.

E. Descriptive Modelling

Descriptive analytics recaps and transforms data into expressive information for reporting and one-to one care but also allows for thorough examination to answer questions such as "what has occurred?" and "what is presently bang up to-date?" [5] Descriptive analytics as control panel applications that support development implementation in sales and procedures administration, allowing for real-time tracking [6]. Summarization can be observed as squeezing a given set of dealings into a smaller set of designs while recollecting the supreme likely information. Summarization is a common and authoritative though often time-consuming method to examining large datasets [7-9].

3. System Design

The Fig.1 is the justification flow of user needs

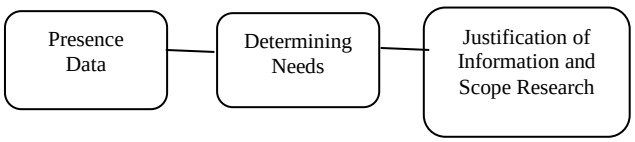

Fig. 1. Justification Flow of User Needs

The data used is employee attendance data of a company. The next step is to specify the user's needs. The user in this case is the level of company management that has authority and access related to the patterns and knowledge obtained from processing the data. Determine the procedures and limitations and scope of the research undertaken.

Fig 2 shows, this algorithm is performed in following steps [10]

1. The initial centroids are prepared by placing k number of points containing the objects that are to be clustered.
2. The nearest Centroid is the group of each object moved.
3. The recalculation of k- centroids is performed in the case of all the objects that has been allotted a group.
4. All the procedure of allocation of centroids are repeated until there remains to movable centroids. This will result in the formation of groups from which the reference metric can be minimized.

5. K-Means clustering intends to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean [8]. This method produces exactly *k* different clusters of greatest possible distinction. The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data. [9,11]

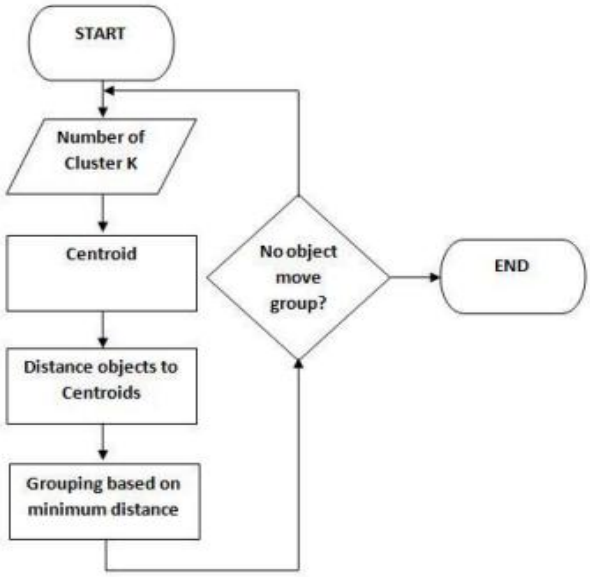

Fig. 2. Flowchart of K-means Algorithm

4. Results and Discussion

Preprocessing data from this research process is illustrated in the following workflow on Fig 3:

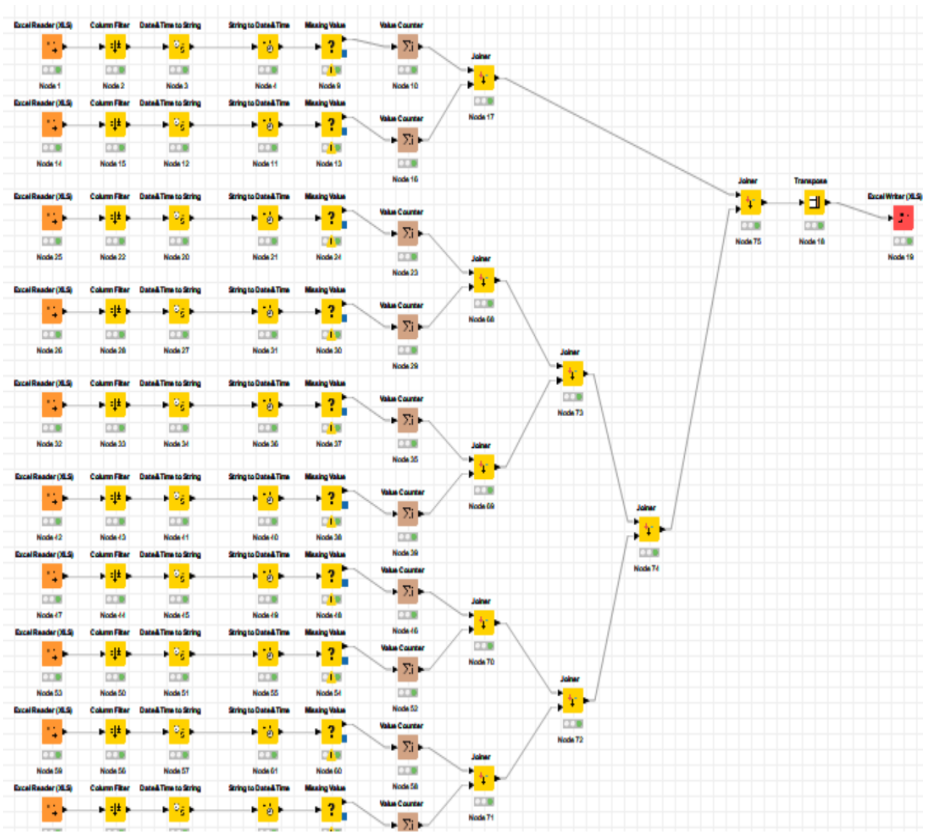

Fig. 3. Preprocessing Data

The join process is shown in Fig. 3. Data processed is sourced from employee data from a university. There are several stages of data preprocessing that are carried out starting from cleaning data, data selection, integration and transformation data. In the Selection data section there is a join process for several sources. Join the results of several files to determine the attributes used in the study.

From Fig. 4 shows, it can be seen that the join is done between two different tables, namely:

1. A and B, C and D, E and F, G and H, I and J
2. Results of C & D join E & F, Results of G & H, Results join of I & J
3. Results of C, D, E, F join G, H, I, J
4. Results of C, D, E, F, G, H, I, J join A, B

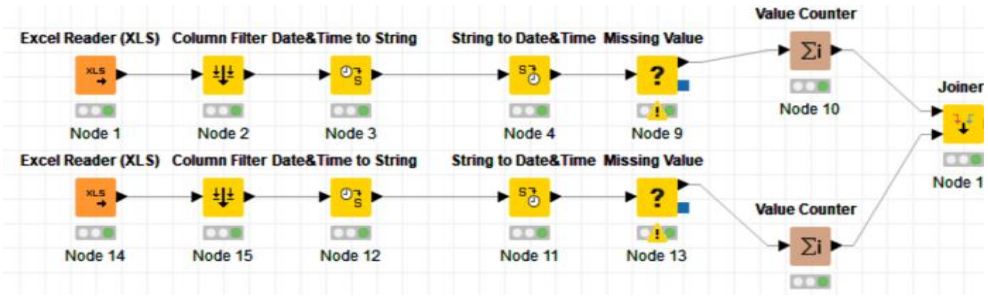

Fig.4. Join Process

Fig 5 shows, at the repository node, connect K-Means with Excel Reader.

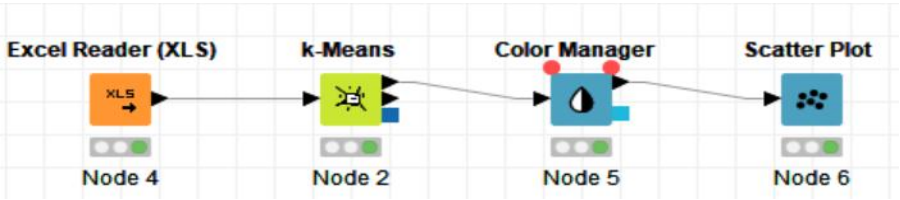

Fig. 5. Visualization Result

Fig 6 shows, from the results of the research that has been done, it was found that in the presence of clusters it can be seen the pattern of employee absence based on groups of severe, moderate and normal absence. The grouping is as follows:

Cluster 0: Cluster of severe absence
Cluster 1: Cluster of normal absence
Cluster 2: Cluster of moderate absence

Data processed is sourced from employee data from a university. There are several stages of data preprocessing that are carried out starting from cleaning data, data selection, integration and transformation data. In the Selection data section there is a join process for several sources

5. Conclusions

From the results of the research that has been done, it was found that in the presence of clusters it can be seen the pattern of employee absence based on groups of severe, moderate and normal absence. The grouping is as follows cluster of severe absence, cluster of normal absence and cluster of moderate absence.

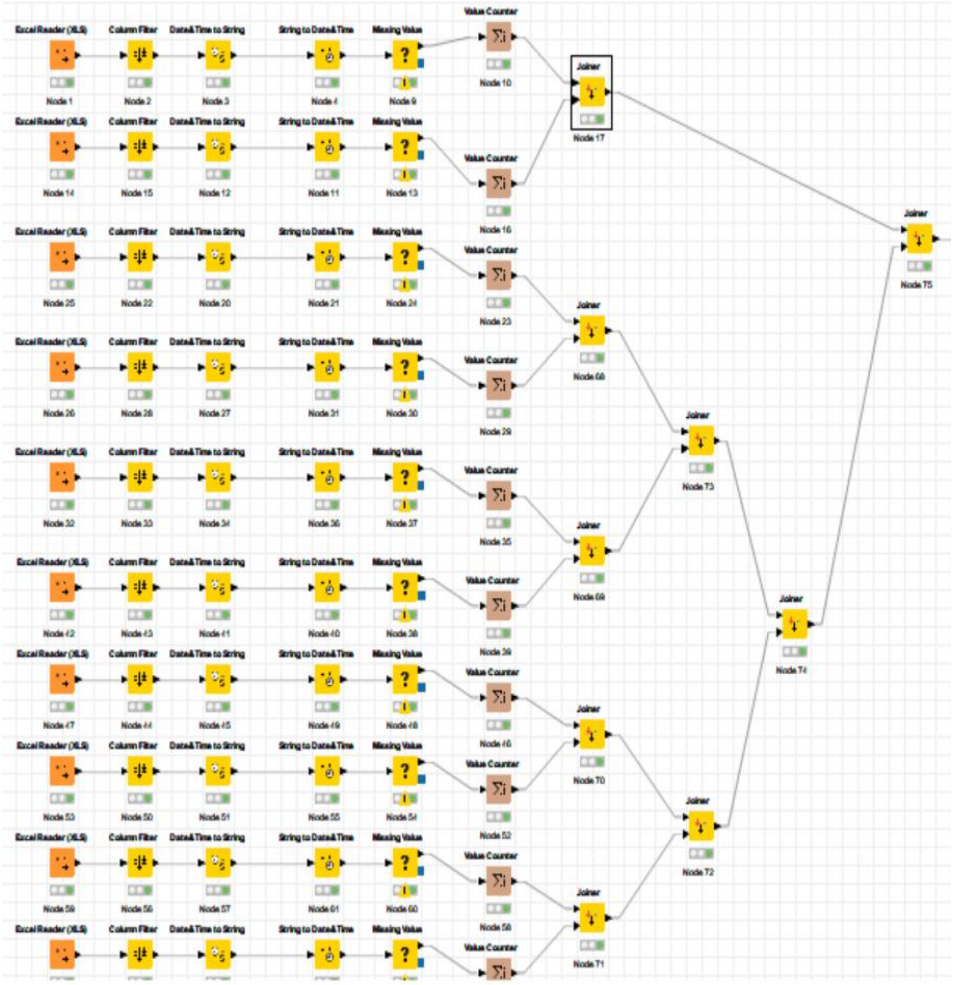

Fig.6. K-Means method

References

[1] MacQueen, J. B., "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. March 2013, pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.

[2] Madhuri V. Joseph, "Significance of Data Warehousing and Data Mining in Business Applications", International journal of Soft Computing and Engineering, March 2013, Vol No:3,Issue no:.

[3] C. Zhang, and Z. Fang, "An improved k-means clustering algorithm", Journal of Information & Computational Science, 10(1), 2013, 193-199

[4] Fahad, A, Alshatri, N., Tari, Z., AlAmri, A., Zomaya, Y., Khalil, I., Foufou, S., Bouras, A, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis," Emerging Topics in Computing, IEEE Transactions on,2014 ,vol.PP, no.99, pp.1,1.

[5] Mortenson, M. J., Doherty, N. F., & Robinson, S. (2014). OperationalresearchfromTaylorism to terabytes:aresearch agenda for the analyticsage. European Journal of OperationalResearch, 583-595.

[6] SAP. SAP HANA Marketplace. Retrievedfrom SAP : http://marketplace.saphana.com SAP. (2014, 05 31). SAP HANA partner race`. Retrievedfrom SAP : http://global.sap.com/germany/campaigns/2 012_inmemory/partner-race/race.epx S

[7] Kwame Boakye Agyapongn, DR.J.B Hayfron-Acquah " An overview of Data Mining Models(Decsriptive and Predictive") International Journal of Software and Hardware Research in Enginering Volume 4 Issue 5 May 2016

[8] Shalini S Singh & N C Chauhan, "K- means v/s K- medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology, 2011.

[9] Anil K. Jain, "Data clustering: 50 years beyond Kmeans", 19th International Conference in Pattern Recognition,2009.

[10] Arora, Deepali, Varshney, "Analysis of K-Means and K-Medoids Algorithm For Big Data", International Conference on Information Security & Privacy (ICISP2015), 2015.

[11] H. Jiawei,M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques", San Francisco California, Morgan Kaufmann Publishers, 2012.

Authors' Profiles

Warnia Nengsih is a senior lecturer. Active in activities related to data science. The focus of research and publications in the field of soft computing and data engineering. Training and certification related to this field.

Muhammad Mahrus Zain is a junior lecture. He is a data scientist, a research focus in the data engineering area. Involved in several soft computing and IT projects

How to cite this paper: Warnia Nengsih, Muhammad Mahrus Zain, " Descriptive Modeling Uses K-Means Clustering for Employee Presence Mapping", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.12, No.4, pp. 15-20, 2020. DOI: 10.5815/ijieeb.2020.04.02