Analysis Of Drug Data Mining With Clustering Technique Using K-Means Algorithm

April Lia Hananto 1, Priati Assiroj 2, Bayu Priyatna 3, Nurhayati 4, Ahmad Fauzi 5, Aviv Yuniar Rahman 6, Shofa Shofiah Hilabi 7

1, Faculty of Computing Universiti Teknologi Malaysia Skudai Johor, Malaysia
2,3,5,7 Faculty of Technology and Computer Science, Universitas Buana Perjuangan, Karawang, Indonesia
4 Master of Computer Science, Universitas Budi Luhur, Jakarta, Indonesia
6 Department of Informatics Engineering, Universitas Widyagama, Malang, Indonesia

1E-mail: aprilia@ubpkarawang.ac.id, hananto1983@graduate.utm.my
2E-mail: priati@ubpkarawang.ac.id
3E-mail: bayu.priyatna@ubpkarawang.ac.id
4E-mail: nurhayati@gmail.com
5E-mail: afauzi@ubpkarawang.ac.id
6E-mail: aviv@widyagama.ac.id
7E-mail: shofa.hilabi@ubpkarawang.ac.id

Abstract. Data processing is very important in the development of information technology. Almost all fields of work have information data. Data can be used to help analysis in work. At present, health information data is very important to be processed in order to help medical personnel to make decisions. So that the results of the right decision to help patients. Lately, drug data has been misused for information eliminating a depressed patient without a doctor's prescription with a total data of 53766. The results shown are very large. So it requires very much attention from the government. As a result of the deviation of information and applied to the patient will result in death. Therefore, research needs to be conducted to group data on drug data. The source of research data is obtained from the UCI Machine Learning Repository Education website. The method proposed in this research is data mining. This solution can help researchers in the analysis of these data. One technique in data mining with clustering is using the K-means algorithm. The variables used are drug name, condition, useful count. The first research results can classify three categories consisting of using the highest drugs, using medium drugs and using lace drugs. Then the accuracy of the data is obtained with condition 99.45% valid records 53471, drug name 100% with valid records 53766, useful count 100% with valid records 53766.

Keyword: Drug Data, Data Mining, Clustering Technique, K-Means Algorithm.

1. Introduction

1 To whom any correspondence should be addressed.
2 To whom any correspondence should be addressed.
Human Technology and information are increasingly rapid, making uncountable data information in human life, clearly observed in the field of large-scale data processing in data storage. This has become a big attraction for companies and organizations, both public and private, to have sufficient data storage capacity. The ability of technology and information to collect data holds various types of data that are very large in number, generally supporting in terms of processing internal and external data in transactions of companies and services [1][2] that are supported and managed by information technology.

Utilization of data in a company to support decision making is not enough in the operating system alone, and it is necessary to analyze company data to get the right and accurate study results. This is an attraction in the use of knowledge that can solve the problem of large amounts of data into information. Data mining can be done with an application such as WEKA, SPSS Clementine, Mathlab, Data miner, Orange canvas, Data engine, DB Miner, Web Miner, and Rapid mine which can facilitate data mining [3][4]. According to Witten, Frank & Hall, 2011. "The Waikato Environment for Knowledge Analysis (WEKA) is an association of complete and implemented State-of-the-art learning and algorithms in data mining." Statistical Product and Service Solution (SPSS) was chosen as the application used for data processing with analytical procedures used in the business field, starting from a single level [5].

Data mining drug analysis is an observation of the data drug obtained from the UCI Machine Learning [6][7] Repository Education website, which is a substantial amount of data and has been processed using classification techniques. Therefore, by utilizing data mining analysis, this is used by the author as a solution to handle information hidden in the data buildup. The author takes the completion of mining drug data analysis case studies using Clementine SPSS application using the Clustering Technique method using the K-Means algorithm [8][9].

2. Literature Review

2.1. Data Mining

Data mining is defined as a process to determine the relationship of new patterns and trends that are meaningful by filtering, using pattern recognition techniques such as statistical and mathematical techniques [10][11]. Data mining is a term used to describe the discovery of knowledge databases. Data mining is a process that uses analytical methods [12], mathematics, artificial intelligence, and machine learning to extract and identify useful information and detailed knowledge from various large databases [13].

Clustering refers to grouping notes, observations, or cases into classes of similar objects. Clustering is a collection of records that are similar to each other, and different from files in other groups. Clustering is different from classification in that there are no target variables for grouping. The grouping task does not try to classify, estimate, or predict the value of the target variable. Grouping algorithms seek to segment all data sets into relatively homogeneous subgroups or clusters, where the similarity of records in the group is maximized and their similarity to file outside the cluster is minimized [11].

2.2. K-Means Clustering Algorithm

Clustering problems arise in many different applications, including data mining and knowledge discovery [14]. K-means plays an essential role in various fields of data mining [15]. K-means clustering algorithm, which we call the filtering algorithm. This algorithm is easy to implement, requiring a KD-tree as the only primary data structure [15].

3. Research Methodology

The method used refers to the Cross-Industrial Process Model for Data Mining (CRISP-DM). The research method used, as shown in Figure 1:
4. Results and Discussion

4.1. Business Understanding Phase
1. Determine the purpose of the analysis in detail in the overall health unit. Aiming to analyze the data drug consumed by the community with the results of mental changes after using it, 3 cluster clusters will be based on cluster variables.
2. Data is obtained from the Universal Learning and Intelligent Systems (UCI) Machine Learning Repository website: http://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29#. The number of records of the information is 53767 data with seven variables. At this stage, an understanding of the number of records of the data is carried out with 53767 data with seven variables. At this stage, an understanding of research needs is carried out in detail and translating them into data mining objectives.
3. *Prepare an initial strategy for achieving goals. The initial goal in achieving goals is to search for data on the internet.*

4.2. Data Understanding Phase
Understanding the Drug set data obtained on the internet in the form of a .tsv file which is then changed to a spreadsheet format.
1. Data collection
   The data used in the study is the Drug dataset, which is a data of patients who experience severe depression and then apply hard drugs without doctor's advice and exceed the recommended dosage in general.
2. Understand further data from initial knowledge search using analysis of data drug investigations.
3. Select a small portion of data to be analyzed in small data groups that contain analysis patterns. The variables used are drug names, conditions, useful contracts. The following are some sample samples that can be seen in Table 1:
Table 1. Samples of Drug Data Group Data Variables From 53766 Data.

| Drug Name     | Condition       | UsefulCount |
|---------------|-----------------|-------------|
| Mirtazapine   | Depression      | 22          |
| Masalamine    | Crohn’s         | 17          |
| Bactrim       | Urinary         | 3           |
| Contrave      | Wight Los       | 35          |
| Cyclafem 1/35 | Birth Control   | 4           |
| Ziclara       | Migraine        | 13          |
| Copper        | Opiate          | 1           |
| Amitriptyline | Birth Control   | 32          |
| Methadone     | Hot Flashes     | 21          |
| Levora        | Vaginal Yeast Infection | 3 |
| Paroxetine    | Wight Loss      | 17          |
| Miconazole    | Disorde         | 7           |
| Belviq        | Insomnia        | 57          |
| Seroquel      | Narcolepsy      | 19          |
| Ambien        | Insomnia        | 44          |
| Nuvigil       | Acne            | 14          |
| Chantix       | Smoking Cessation| 26         |
| Microgestin   | Acne            | 1           |
| Klonopin      | Bipolar Disorde | 24          |
| Ciprofloxacin | Urinary Tract Infection | 9 |

1. The results of the evaluation of data contained duplicate data with an empty value found with a missing value. Perform preprocessing and cleansing to handle incomplete data and avoid interference in processing research data. The results of the test are seen in Figure 2:

![Figure 2. Sample Data That Duplicates Missing Value Data](image-url)
4.3. Data Preparation Phase

Prepare the initial data to be used for analysis purposes throughout the next stage. Intensively carried out includes all activities in building a dataset of drugs that will be applied to the modeling which starts from the initial data in the form of a dataset which will then be carried out processing data modeling using the k-means cluster method.

Using three variables to analyze. The variables used are the drug name, condition, and usefulcont. Prepare the initial data that is ready to go to the modeling stage by ensuring that the information is feasible for data mining processing can be seen in Figure 3:

![Figure 3. Data Ready to Be Processed](image)

4.4. Modeling Phase

The stages directly involve data mining modeling and determining the algorithm to be performed. The algorithm used is the K-Means algorithm with Clustering technique. It can be understood for the modeling stage, learning the number of clusters, taking the cluster center on the data to be analyzed, looking for the center on the data studied, in the process using tools. The results there is a data cleaning missing value, which means that the data is double and the cost is empty, then there will be data cleaning.

![Figure 4. Clustering Technique of K-Means Algorithm in Clementine SPSS](image)
4.5. Evaluation Phase

This stage is the evaluation stage, which has an analysis model that meets the objectives at the initial step following the purposes.

1. All models are following the goals according to the initial stage.
2. There are important issues that are not addressed, including data that is double with the results of the results that cause the analysis of the patient's condition is not optimal, which should be accurate, saying 99.45% is correct in the patient's condition.
3. Take decisions related to the use of data mining results. The results of data processing in clustering techniques can be seen in Figure 5:

![Figure 5. Results of Data Processing in Clustering Techniques](image)

Knowledge gained in interpretation is as follows:

1. The first cluster, this cluster is typically high. Having the intention of patients who use the most top drug with a value of 41702 records is a group of useful count data on drug use of 13477 with a standard definition of 10,688
2. The second cluster, this cluster is typically low. Having the intention of patients using the lowest drug with a value of 1520 records is a group of useful count data on drug use of 171,657 with a standard deviation of 66,817
3. The third cluster, this cluster is a medium type. Having intentions of patients using cheap drugs with a value of 10544 records is a group of useful count data on drug use of 64677 with a standard deviation of 20.56. Can be seen in Figure 6:

![Figure 6. Quality of Processing Results In The Form of Diagrams](image)
**Conclusion**

The results of the study, knowledge is obtained in the following interpretations:

1. The first cluster, this cluster is typically high. Having the intention of patients who use the most top drug with a value of 41702 records is a group of useful count data on drug use of 13477 with a standard deviation of 10,688.

2. The second cluster, this cluster is typically low. Having the intention of patients using the lowest drug with a value of 1520 records is a group of useful count data on drug use of 171,657 with a standard deviation of 66,817.

3. The third cluster, this cluster is a medium type. Having intentions of patients using cheap drugs with a value of 10544 records is a group of useful count data on drug use of 64677 with a standard deviation of 20.56.

**Reference**

[1] A. Y. Rahman, M. Sa’adah, F. W. Setiawan, A. L. Hananto, and E. Supriyanto, “Quality of Service Analysis on Go-Jek Mobile App for Students,” pp. 61–65, 2018.

[2] A. L. Hananto and A. Y. Rahman, “User Experience Measurement On Go-Jek Mobile App In Malang City,” *2018 Third Int. Conf. Informatics Comput.*, pp. 1–6, 2018.

[3] P. Shi and J. M. Goodson, “A data mining approach identified salivary biomarkers that discriminate between two obesity measures,” *J. Obes.*, vol. 2019, 2019.
[4] R. Rakhmawati, Irianto, F. D. Murdianto, A. Luthfi, and A. Y. Rahman, “Thermal Optimization on Incubator using Fuzzy Inference System based IoT,” 2019 Int. Conf. Artif. Intell. Inf. Technol., pp. 464–468, 2019.

[5] A. Y. Rahman, M. Sa’adah, F. W. Setiawan, A. L. Hananto, and E. Supriyanto, “Access Level File using Simple Additive Weighting,” pp. 62–66, 2018.

[6] A. Y. Rahman, S. Sumpeno, and M. H. Purnomo, “Area Detection and Matching Using Scale Invariant Feature Transform (SIFT) Method of Stereo Camera,” Proc. - 2017 Int. Conf. Soft Comput. Intell. Syst. Inf. Technol. Build. Intell. Through IOT Big Data, ICSIIT 2017, pp. 66–71, 2017.

[7] A. Y. Rahman, S. Sumpeno, and M. H. Purnomo, “Video Minor Stroke Extraction Using Learning Vector Quantization,” pp. 1–7, 2017.

[8] Y. Marom and D. Feldman, “k-Means Clustering of Lines for Big Data,” pp. 1–29, 2019.

[9] S. M. Abtahi, “Chaotic Dynamic Analysis and Nonlinear Control,” vol. 5, no. 2, 2019.

[10] Y. Sato, K. Izui, T. Yamada, and S. Nishiwaki, “Data mining based on clustering and association rule analysis for knowledge discovery in multiobjective topology optimization,” Expert Syst. Appl., vol. 119, pp. 247–261, 2019.

[11] J. Gola et al., “Advanced microstructure classification by data mining methods,” Comput. Mater. Sci., vol. 148, no. November 2017, pp. 324–335, 2018.

[12] S. Setiawidayat and A. Y. Rahman, “New method for obtaining Peak Value R and the duration of each cycle of Electrocardiogram,” 3rd Int. Conf. Sustain. Inf. Eng. Technol. SIET 2018 - Proc., pp. 77–81, 2018.

[13] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Type 2 diabetes mellitus prediction model based on data mining,” Informatics Med. Unlocked, vol. 10, pp. 100–107, 2018.

[14] P. Yin, M. Pham, A. Oberman, and S. Osher, “Stochastic Backward Euler: An Implicit Gradient Descent Algorithm for k-Means Clustering,” J. Sci. Comput., vol. 77, no. 2, pp. 1133–1146, 2018.

[15] J. Qi, Y. Yu, L. Wang, J. Liu, and Y. Wang, “An effective and efficient hierarchical K-means clustering algorithm,” Int. J. Distrib. Sens. Networks, vol. 13, no. 8, pp. 1–17, 2017.