Using K-means Techniques for Clustering Depressed Patients in Mahasarakham Province

Supakron Srisanga[1], Arisara Pahdungcharoen[2], Nanthicha Wonghong[3], Kanjana Hinthaw[4], Anupong Sukprasert[5] and Warawut Narkbunnum[6*]

Mahasarakham Business School, Mahasarakham University[1,2,3,4,5,6]

E-Mail: warawut.n@acc.msu.ac.th[6]

(Received: November 13, 2023; Revised: December 2, 2023; Accepted: December 7, 2023)

ABSTRACT

This study used advanced data mining techniques to examine the patterns of depression and service use among patients in Mahasarakham Province. The focus was to understand how people with depression access and use medical services in this area. Data collected from the Mahasarakham Provincial Public Health Office were meticulously analyzed using the K-means clustering method. This analysis involved a dataset consisting of five key variables used in the clustering process. This study successfully identified six unique clusters of patients, each showing different symptoms and degrees of depression. These findings provide essential insights into the diverse manifestations of depression within Mahasarakham's patient population, contributing valuable information for healthcare providers and policymakers to optimize service delivery and treatment approaches.

Keywords : Depression, Data Mining Technique, Clustering, K-means clustering

Introduction

Depression is a psychiatric illness characterized by diminished emotional responsiveness to a range of stimuli in individuals with schizophrenia. Individuals who possess a diminished awareness of their happiness tend to exhibit a decreased inclination toward actively engaging in life experiences. Numerous factors contribute to the development of depression. Whether depression arises from genetic factors or external societal influences, according to the World Health Organization, the global prevalence of depression stands at 4.4% at present. Nevertheless, studies have indicated a significantly higher occurrence of depression among medical students, ranging from 11.5% to 48.2%, which is approximately 3 to 10 times greater than the general population [1]. Depression has been recognized as a significant public health concern in Thailand, affecting the entire population. It has also been documented to manifest concurrently with a range of physical health conditions, with a higher prevalence among women and the elderly [2]. This research employed data mining theory to examine a representative sample of patients diagnosed with depression in Mahasarakham Province. The objective was to identify inherent patterns or structures among depression patients in Mahasarakham using the principle of the clustering technique. The primary objective was to investigate the clustering patterns of depressed patients in the province through the application of K-means clustering. The utilization of clustering techniques to determine the age group This study aims to investigate the prevalence of depressive symptoms among different groups in Mahasarakham Province, with a focus on district and gender as the primary variables of interest.

1. Objective

1.1 To identify the optimal value of k for clustering using the k-means clustering technique

1.2 To explore Mahasarakham depressive symptoms further by k-mean clustering

2. Related work

Data Mining [3] involves a diverse array of methodologies aimed at obtaining meaningful insights from extensive databases. The techniques encompassed in this list include association rule mining, classification, clustering, regression analysis, time series analysis, anomaly detection, text mining, neural networks, dimensionality reduction, ensemble approaches, and other related methodologies. The utilization of data mining is of crucial significance for supporting informed decision-making processes and uncovering latent patterns within datasets across diverse domains, such as business [4,5] and healthcare [6,7] The selection of a particular methodology is contingent upon the inherent characteristics of the data as well as the specific objectives of the research. Data clustering [8] is a widely employed data mining approach that aims to categorize data points into clusters or groups based on their intrinsic features. The K-means technique is one of the most utilized clustering algorithms. The following is a comprehensive exposition of the concept of data clustering utilizing the K-means algorithm.

K-Means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into distinct groups or clusters based on [9] The K-means method is a partitioning technique that divides data into K clusters, where K is a predetermined number that the user selects. The Euclidean distance is employed for the computation of the distances between each data point and the initially assigned centroids. The Euclidean distance, also known as the Euclidean metric, refers to the direct line connecting two elements or the shortest distance between two items. It is widely recognized as the easiest method for measuring the distance

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

The primary concept underlying the K-means algorithm is to reduce the intra-cluster variance while simultaneously maximizing the inter-cluster variance. Figure 1 illustrates the operational procedure of the K-Means algorithm.

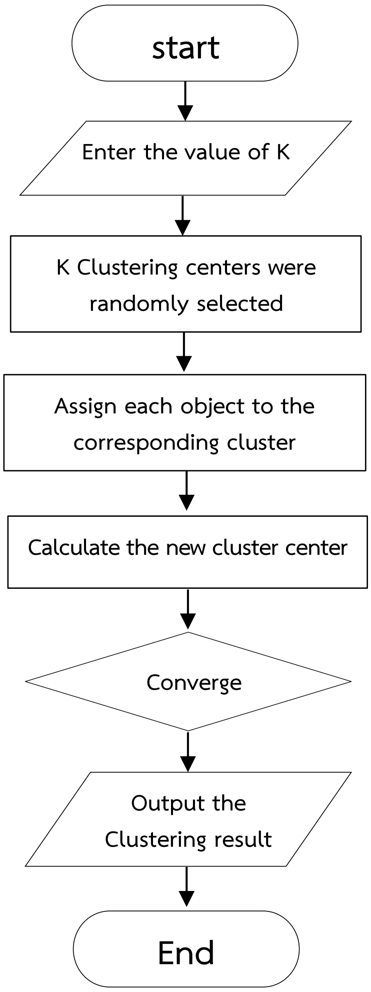
start
Enter the value of K
K Clustering centers were randomly selected
Assign each object to the corresponding cluster
Calculate the new cluster center
Converge
Output the Clustering result
End

Figure 1 flowchart of the traditional k-means method

Figure 1 displays the flowchart of the traditional k-means method. The time complexity of the K-means algorithm is represented as O (nkfl), where k represents the number of clusters, n represents the number of items in the dataset (X), f signifies the number of features (dimensions) of each object (xi), and l is the number of iterations performed by the algorithm. [11]

This study uses data mining and K-means clustering techniques to evaluate depression patterns in Mahasarakham Province. It is based on prior research in the field of mental health. This corresponds to research on mental clustering for the comprehension of depressive symptoms [12,14] Furthermore, this study enhances the research conducted regarding the use of healthcare services, thus supporting the customized medicine approach, and highlighting the significance of demographic aspects in mental health.

Research Methodology

1. Research proposed

The initial stage of this research endeavor involves conducting a comprehensive search and gathering relevant material. The data utilized in this study was acquired from the Mahasarakham Provincial Public Health Office, with due consideration to the privacy and confidentiality of the individuals involved. Upon acquiring the data, the researcher proceeded to preprocess the data to render it appropriate for utilization in clustering methodologies. The present study employs the K-means clustering technique to investigate the association among individuals diagnosed with depression in Mahasarakham province. The proposed workflow is shown in Figure 2.

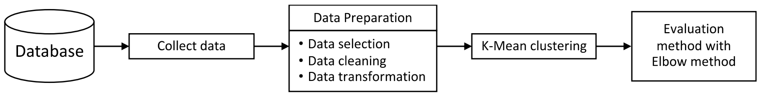

Figure 2 The typical system’s workflow

2. Cross Standard Process for Data Mining

The implementation of data mining methods involves the utilization of the Standard Process in Data Mining (CRISP-DM), as illustrated in Figure 3.

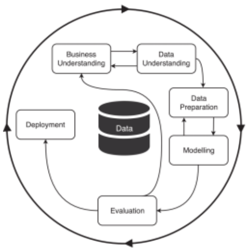

Figure 3 The CRISP-DM process model of data mining [15].

Figure 3 illustrates the sequential arrangement of the six phases of CRISP-DM within a conventional data mining application. During the early 2000s, several process models and approaches were established, drawing upon CRISP-DM as a foundational framework.

2.1 Business Understanding: According to data from depression access to services (Workload) in Mahasarakham province in fiscal year 2022, the number of patients who came for diagnostic and treatment services was 13,780 [16] In considering this issue, the researcher conducted a study aimed at categorizing people who have depression in the province of Mahasarakham, to optimize the grouping of distinct datasets.

2.2 Data Understanding: The researcher has obtained a dataset relating to people who were diagnosed with depression in Mahasarakham Province. This dataset excludes any personally identifiable information such as first names, last names, or national identification numbers, hence precluding the possibility of individual identification. The data utilized for this study was acquired from the Mahasarakham Provincial Public Health Office and subsequently organized in an Excel spreadsheet for data categorization, that displays a collection of 10 attributes, encompassing features such as HOSCODE, HOSNAME, Date of birth, gender, Nation, TYPEAREA, HOSP_DX, DIAGCODE, 2Q, and 9Q assessment score. The 2-Question (2Q) and 9-Question (9Q) assessments are concise tools used for diagnosing depression [17] and assess a broader range of symptoms like sleep disturbances, energy levels, and feelings of worthlessness, aiding in both diagnosis and severity assessment.

2.3 Data Preparation

The process described relates to the preparation of data for analysis using data mining techniques. The data preparation process consists of 3 processes.

2.3.1 Data Selection: The process of data selection is a crucial technique that prioritizes the identification and inclusion of pertinent data for subsequent data analysis. The patient's date of birth and sex, as well as the name of the service facility, are requested. The assessment score for the diagnostic code (DIAGCODE) 9Q transforms the subsequent stage.

2.3.2 Data Cleaning: The process of data cleaning revealed that, upon conducting the survey and examining the data through exploratory analysis, it was observed that the obtained data did not contain any zero values. This characteristic renders this data segment suitable for analyzing the segmentation data.

2.3.3 Data Transformation: The data followed a process of transformation, as shown in Table 1.

Table 1 The list of data attributes used in research

No.	Name	Data type	Value	Description
1	age	integer	1 = 1 to 12, 2 = 13 to 19, 3 = 20 to 39, 4 = 40 to 59, 5 = > 60	Age (year)
2	gender	Integer	1 = male, 2 = female	sex
3	district	Integer	1 = Phayakkhaphum, 2 = Kantharawichai, 3 = Kosum Phisai, 4 = Mueang Mahasarakham, 5 = Chiang Yuen, 6 = Wapi Pathum, 7 = Na Chueak, 8 = Borabue, 9 = Na Dun, 10 = Kae Dam	district
4	DIAGCODE	Integer	1 = (F32.0, F32.1, F32.2, F32.3, F32.8 and F32.9) 2 = (F33.0, F33.1, F33.2, F33.3, F33.4, F33.8 and F33.9) 3 = (F34.0, F34.1, F34.8 and F34.9) 4 = (F38.0, F38.1 and F38.8) 5 = (F39.0)	Diagnosis code
5	9Q	integer	1 = < 7 score, 2 = 8 – 12 score, 3 = 13 – 18 score and 4 = > 19 score	9Q assessment score

Table 1 depicts the subsequent manner: 1) The patient's date of birth was transformed into AGE and subsequently classified into five age groups. 2) The concept of gender was classified into two distinct categories. 3) The service area was designated and categorized into

10 districts based on their respective names. 4) The diagnosis code (DIAGCODE) was categorized into five distinct groups. 5) The 9Q evaluation score data was classified into four groups.

2.4 Modeling

This study presents an algorithm for constructing models utilizing clustering techniques, namely the K-means Clustering method, within the RapidMiner Studio Version 10 software. The objective of this algorithm is to generate models that can be employed for patient segmentation. The present study aims to investigate the prevalence of depression in Mahasarakham province through the utilization of an optimal K value. This will be achieved by determining the K value and calculating the average value within the centroid distance.

Results

Based on the upward trajectory of depression indicated in the Introduction section. The findings of the present study highlight the significance of individualized intervention and assistance for different subcategories of depression-affected populations. The first objective of the study, which aimed at identifying the optimal k value using the k-means clustering technique, the analysis revealed that the most suitable k value is 6, as illustrated in Table 2.

Table 2 Average Within Centroid Distance of each group (K)

K	Average Within Centroid Distance	K	Average Within Centroid Distance
2	2.383	9	0.576
3	1.507	10	0.536
4	1.378	11	0.552
5	1.197	12	0.517
6 *	0.729	13	0.499
7	0.694	14	0.386
8	0.639	15	0.379

* Optimal K for K-Means

Following the information provided in Table 2 , the K-means clustering technique was employed to cluster the data. The parameter K was set to a value of 15, and the average “Average Within Centroid Distance” was calculated. This calculation was performed to visualize the data distribution through the creation of a scatter plot. To determine the optimal k value, a plot is utilized to identify the elbow point [18]

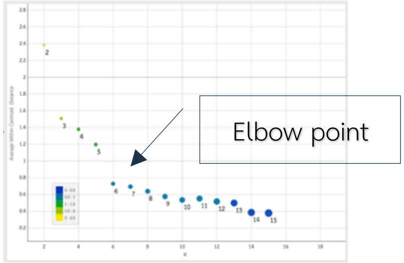

Figure 4 Scatter Plot Technique

The scatter plot technique reveals in Figure 4 that the point with the highest k value is characterized by a strong peak. The graphic illustrates that the value of k is 6. The optimal approach for partitioning data of individuals diagnosed with depression in Mahasarakham Province. To achieve objective number 2, the aim is to conduct a more in-depth investigation into depression symptoms in Mahasarakham using the k-means clustering technique. This can be seen by referring to the heatmap illustrated in Figure 5.

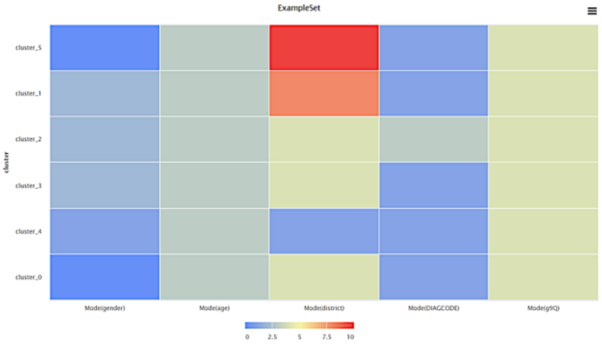

Figure 5 The prevalence of individuals diagnosed with depression within each respective data group (Heatmap)

Figure 5 illustrates a graphical representation of the frequency distribution (mode) of data of depressed patients in Mahasarakham Province, categorized by category. Heatmap charts employ color intensity as a means of representing data segments, rather than relying on frequency. Color-coded heatmaps illustrate the correlation between variables and clusters; areas displaying higher correlations are indicated by warm or red colors. However, in cases where the regions with minimal correlation present cold or blue colors. For example, the correlation between cluster_5 and the district with the highest correlation, heatmaps could show clear clustering of patients based on their symptom profiles. Each cluster would represent a group of patients with similar symptom patterns, indicating different subtypes or severities of depression.

Based on the findings from the k-means clustering model in the study of depression in Mahasarakham Province, the researchers identified six distinct clusters of individuals with depressive conditions, each exhibiting unique characteristics. These clusters provided valuable insights into the intricate nature of depression in the region:

Cluster_0: Comprised mostly of residents aged 20 to 39 from Mueang Mahasarakham District, with an unspecified gender distribution. This group showed severe depression symptoms and high 9Q scores. Cluster_1: Mainly females aged 20 to 39 from the Borabue District, consistently diagnosed with depression and exhibiting elevated 9Q scores. Cluster_2: Females from Mueang Mahasarakham District, also aged 20 to 39, indicated long-term emotional disorders with high 9Q scores despite issues with diagnostic codes. Cluster_3: Similar to Cluster_1 and 2, primarily composed of women aged 20 to 39 from Mueang Mahasarakham District, characterized by a depression diagnosis and persistent high 9Q scores. Cluster_4: Mostly males aged 20 to 39 from the Phayakkhaphum District, marked by a high prevalence of depression and elevated 9Q scores. Cluster_5: Mirrored the gender distribution of Cluster_0, with residents aged 20 to 39 in the Kae Dam District and high 9Q scores indicative of depression.

Discussion

The investigation of depression in Mahasarakham Province employed the k-means clustering approach to detect six unique clusters, thereby providing valuable insights into demographic patterns and the intensity of depression. This strategy is advantageous because it focuses on specific subgroups for effective intervention and provides detailed demographic information. However, it has certain limitations, such as limited applicability and problems with diagnostic codes. The significant discoveries include the aggregation of specific age and gender cohorts within different clusters, implying that sociocultural forces affect mental well-being, as well as regional discrepancies indicating possible environmental or socioeconomic factors contributing to mental health inequalities. These findings highlight the need for customized mental health solutions and highlight areas that require further research and policy development.

It is crucial to emphasize that the study placed a high priority on ensuring data anonymity and confidentiality, avoiding the inclusion of personally identifiable information. This approach was undertaken to uphold ethical standards and protect privacy. This research study has contributed to a nuanced and comprehensive comprehension of the demographic and diagnostic complexities associated with depression in the province of Mahasarakham. The insights can be regarded as a valuable resource for mental health professionals and policymakers who are seeking to effectively tackle the complex challenges presented by depressive conditions in the specified region. Additionally, it is emphasized that customized interventions and assistance are crucial for specific subgroups within the demographic affected by depression.

Suggestion

To determine whether the patterns discovered in Mahasarakham Province are universally applicable in other areas, further research should prioritize broaden the study's geographical coverage. This would make it possible to apply the findings to a wider range of situations. In addition, a thorough examination of the effects of socio-economic factors and an understanding of the importance of environmental influences on the identified clusters could provide a deeper understanding of the complex nature of depression. In addition, it would be highly beneficial to conduct longitudinal studies to monitor the evolution of depressed symptoms within these clusters over time. In addition, qualitative approaches can be used to gain insight into the individual experiences of people within each cluster, which could enhance the quantitative data. Implementing and assessing focused intervention tactics that align with cluster attributes discovered in this investigation would be a pivotal measure in enhancing mental health care outcomes.

References

[1] Phomprasith, S., Karawekpanyawong, N., Pinyopornpanish, K., Jiraporncharoen, W., Maneeton, ... Lawanaskol, S. (2022). Prevalence and Associated Factors of Depression in Medical Students in a Northern Thailand University: A Cross-Sectional Study. *Healthcare,* 10(3), 1-13.https://doi.org/10.3390/HEALTHCARE10030488

[2] World Health Organization. (2023). *Depressive disorder (depression).* Retrieved October 25, 2023, from https://www.who.int/news-room/fact-sheets/detail/depression

[3] Gupta, M. K., & Chandra, P. (2020). A comprehensive survey of data mining. *International Journal of Information Technology (Singapore),* 12(4), 1243–1257. https://doi.org/10.1007/S41870-020-00427-7/TABLES/7

[4] Kolukuluri, M., Keerthana Devi, V., Tejaswini, S. S., & Anusha, K. (2023). Business Intelligence Using Data Mining Techniques And Predictive Analytics. *Journal of Pharmaceutical Negative Results*, 13, 6923–6932. https://doi.org/10.47750/PNR.2022.13.S07.837

[5] Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems With Applications,* 166, 114060. https://doi.org/10.1016/J.ESWA.2020.114060

[6] Laijawala, V., Aachaliya, A., Jatta, H., & Pinjarkar, V. (2020). Mental Health Prediction using Data Mining: A Systematic Review. *SSRN Electronic Journal.* https://doi.org/10.2139/SSRN.3561661

[7] de la Fuente-Tomas, L., Arranz, B., Safont, G., Sierra, P., Sanchez-Autet, M., Garcia-Blanco, A., & Garcia-Portilla, M. P. (2019). Classification of patients with bipolar disorder using k-means clustering. *PLOS ONE,* 14(1), e0210314. https://doi.org/10.1371/JOURNAL.PONE.0210314

[8] Weißer, T., Saßmannshausen, T., Ohrndorf, D., Burggräf, P., & Wagner, J. (2020). A clustering approach for topic filtering within systematic literature reviews. *MethodsX,* 7, 100831. https://doi.org/10.1016/J.MEX.2020.100831

[9] Towards Data Science. (2018). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Retrieved October 25, 2023, from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[10] Ghazal, T. M., Hussain, M. Z., Said, R. A., Nadeem, A., Hasan, M. K., Ahmad, M., Khan, M. A., & Naseem, M. T. (2021). Performances of K-Means Clustering Algorithm with Different Distance Metrics. *Intelligent Automation & Soft Computing*, *30*(2), 735–742. https://doi.org/10.32604/IASC.2021.019067

[11] Ashabi, A., Bin Bin Sahibuddin, S., & Salkhordeh Salkhordeh Haghighi, M. (2020). The systematic review of K-means clustering algorithm. *ACM International Conference Proceeding Series*, 13–18. https://doi.org/10.1145/3447654.3447657

[12] Baghdadi, N. A., Alsayed, S. K., Malki, G. A., Balaha, H. M., & Farghaly Abdelaliem, S. M. (2023). An Analysis of Burnout among Female Nurse Educators in Saudi Arabia Using K-Means Clustering. European Journal of Investigation in Health, Psychology & Education (EJIHPE), 13(1), 33–53. https://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=161434209&site=ehost-live

[13] Sheha, M. A., Mabrouk, M. S., & Sharawy, A. A. (2022). Feature Engineering: Toward Identification of Symptom Clusters of Mental Disorders. *IEEE Access*, 10, 134136–134156. https://doi.org/10.1109/ACCESS.2022.3232075

[14] Zeng, Y., & Cheng, F. (2021). Medical and Health Data Classification Method Based on Machine Learning. *Journal of Healthcare Engineering*. https://doi.org/10.1155/2021/2722854

[15] Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

[16] Depression patients have access to services in Mahasarakham province. (n.d.). *Department of Mental Health*.

[17] Levis, B., Sun, Y., He, C., Wu, Y., Krishnan, A., Bhandari, ... & Thombs, B. D. (2020). Accuracy of the PHQ-2 Alone and in Combination With the PHQ-9 for Screening to Detect Major Depression: Systematic Review and Meta-analysis. *JAMA*, 323(22), 2290–2300. https://doi.org/10.1001/JAMA.2020.6504

[18] Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip Journal on Wireless Communications and Networking,* 2021(1), 1–16. https://doi.org/10.1186/S13638-021-01910-W/FIGURES/6