Analysis of Drug User Pattern Grouping Based on Social and Psychological Factors Using the K-Modes Method

Besse Helmi Mustawinar[1*], Irenita[2], Yuliani[3], Nirma Wulandari[4], Hasni Hirman[5] ,

[1,2,3]*Department of Mathematics, Faculty of Sciences, Universitas Cokroaminoto Palopo*

Received: **18 October 2025**

Accepted: **09 December 2025**

Correspondent Email:
emm.emm92@gmail.com.

Abstract:
Drug abuse in Indonesia remains a complex social problem with multidimensional impacts on the health, economy, and social stability of the community. South Sulawesi, particularly Palopo City, is an area with increasing rates of drug abuse from year to year. Prevention efforts are often hindered by a lack of information regarding user behaviour patterns, which vary based on social and psychological factors. This study aims to group drug user patterns using the K-Modes algorithm and determine the optimal number of clusters using the Silhouette Coefficient method. The research data is secondary categorical data from the BNNK Palopo City in 2024, with variables including age, education, occupation, social environment, and duration of drug use. The analysis using R software with the klaR and cluster packages. The results showed that the highest Silhouette value was obtained in five clusters with an average silhouette score of 0.3363, which represents the five main types of drug users based on social and psychological characteristics. Cluster 4 has the highest drug user pattern based on the k-modes analysis. Cluster 4 represents the group of Young Adults Exposed to Strong Social Environments with low educational backgrounds and precarious employment. This permissive social environment makes them susceptible to drug exposure. Interventions that can be provided to users focus on alternative education and the develop of life skills.
Keywords: *K-Modes, Silhouette Method, Clustering, Drugs Abuse, Social and Psychological Factors*

1. PENDAHULUAN

Drug abuse in Indonesia is a national issue that requires special attention to prevent its use. According to the National Narcotics Agency (BNN, 2024), the prevalence of drug abuse in Indonesia has reached 1.73%, equivalent to approximately 3.3 million people. This figure indicates a significant upward trend, particularly among teenagers and young adults. This phenomenon not only impacts individuals' physical and mental health but also has broad implications for social, economic, and national security dimensions.

In South Sulawesi Province, similar conditions also show an upward trend. According to the Badan Narkotika Nasional Provinsi (BNNP) Sulawesi Selatan report, this province ranks fifth nationally in drug abuse prevalence, thus categorising it as a "drug emergency" (Hasanuddin, 2024). One of the areas with the most significant increase in cases is Palopo City. Data from Polres Palopo shows that between January and September 2025, at least 63 cases of drug abuse were recorded. The majority of perpetrators are from the productive age group, namely late adolescence and young adulthood (Nandini et al., 2025). This situation indicates that drug abuse is not simply an individual issue, but rather a complex, multidimensional social phenomenon rooted in the social dynamics of society.

Several previous studies have confirmed that social and psychological factors significantly contribute to an individual's tendency to use drugs. A youthful society, priced correct perception, easy availability of substance, peer pressure, emotional stress or imbalance and low-income family support are also related to drug abuse behaviour (Deep et al., 2024; Hisyam et al., 2025). The intricate nature of these factors necessitates an analysis that is both descriptive and capable of detecting subtle patterns between social, mental, psychological, and economic variables.

Considering this, data mining is a promising alternative to the exploration of hidden patterns under the complex features of drug abuse. Clustering, as one of the relevant approaches, enables research to identify structures within data without prior class labels. Clustering methods, however, are exploratory and search for homogeneous groups based on proximity relations rather than making predictions. Compared to predictive classification approaches, they can discover homogeneous regions based on standard features that have not been previously identified in the segmented space (Xu & Ye, 2023). Thus, it is very appropriate for studying the social fragmentation of drug users according to demographic, behavioural and psychosocial characteristics.

To analyse categorical data, the K-Modes algorithm is the most suitable method. The K-Modes algorithm is a development of the K-Means algorithm for handling non-numeric or categorical

Available online at https://conference.lppm.unila.ac.id/index.php/icsiger
DOI: https://doi.org/10.23960/icsiger

data. K-Modes explains that the distance between data is calculated by measuring the similarity of the mode value or the highest frequency appearing in the data (Dorman et al., 2022). Therefore, grouping data using k-modes allows for the formation of clusters based on similar attribute categories, such as occupation, education level, economic status, and motivation. Therefore, it is hoped that the K-Modes algorithm can be applied to the drug user dataset in Palopo City to help identify social and psychological patterns that indicate drug abuse.

The effectiveness of the K-Modes algorithm has been demonstrated through various studies in the social sciences. Zamaninasab et al. (2018) applied the Evidence Accumulation (EA) method to the K-Modes algorithm to analyse patterns among injecting drug users. Their research results showed that the combination of Naïve Bayes and K-Modes EA was able to produce more stable and representative segmentation. Meanwhile, Tolner et al. (2021) demonstrated that the K-Modes algorithm has a higher level of stability than K-Means when applied to social data with many categorical variables. Another study by Soleimanian and Haggi (2020) also found that the K-Modes algorithm is effective in clustering crime data based on similar social attributes of perpetrators, making it relevant for use in the context of crime and drug abuse.

To assess the validity of the clustering results, this study used the Silhouette Coefficient. This measure considers the average distance between members within the same cluster and the distance between different clusters. Silhouette values range from -1 to 1, with values closer to 1 indicating that the object is in the correct cluster. Conversely, values closer to -1 indicate a potential mismatch in cluster grouping (Faizan et al., 2020). The use of the Silhouette Coefficient is commonly applied in data mining-based social research because it provides an objective measure of the degree of homogeneity and separation between clusters (Jung & Chung, 2021).

By applying the K-Modes algorithm using the Silhouette Coefficient method, this study is expected to identify patterns of drug users in Palopo City based on social and psychological factors. This analysis has the potential to make a significant contribution to mapping high-risk drug user groups based on real-world data. The results of this study can serve as a basis for the BNNK Palopo to formulate policies related to more effective prevention and rehabilitation strategies. A data mining-based approach not only enriches social research methodology but also opens up opportunities for the development of evidence-based public policies. Thus, the application of data mining in social studies on drug abuse is a strategic step towards more systematic and sustainable prevention efforts.

2. MATERIALS AND METHODS

This research is a quantitative study that applies data mining methods using unsupervised learning (clustering) techniques. This approach was used to identify patterns and characteristics of drug users based on social and psychological factors without predetermined class labels (Alpenia & Kariyam, 2022). The main objective of this study was to group drug user data into several homogeneous clusters using the K-Modes algorithm, then evaluate the quality of the clustering results using the Silhouette Coefficient method. The variables used in this study are nominal categorical, so the K-Modes method was chosen because it is more appropriate for data that does not have numerical values. The analysis tool used was the R software.

The data used is secondary data obtained from the BNNK Palopo in 2024. The dataset consists of 5 categorical attributes relevant to the social and psychological aspects of drug users, which can be observed in the following table:

Table 1: Research Variables and Their Descriptions

Variable	Desciption
Age	Teenager (14 – 18 y.o)
	Young Adult (19 – 25 y.o)
	Adult (26 – 59 y.o)
Occupation	Student
	Self-Employed
	Civil Servant
	Jobless
Academic Level	High
	Low
Duration of Drug Abuse	Moderate
	Low-Term
Sosio-Environment	Friends
	Companions
	Family
	Society

The research procedure was carried out through several systematic stages as follows:

1. Literature Review
   Conducting a literature review of previous research on drug abuse, the social and psychological factors that influence it, and the application of data mining algorithms, particularly clustering and the K-Modes Algorithm. The literature review also includes an understanding of clustering evaluation methods, such as the Silhouette Coefficient.
2. Data Collection
   Data was collected from official reports from the Palopo City National Narcotics Agency (BNN), and attributes relevant to the research objectives were selected.
3. Data Preprocessing: Data transformation was performed to convert numerical values into categorical data using label encoding techniques, allowing them to be processed by the K-Modes algorithm.
4. Clustering Process with K-Modes

Available online at https://conference.lppm.unila.ac.id/index.php/icsiger
DOI: https://doi.org/10.23960/icsiger

The K-Modes algorithm was used because it can group categorical data based on the similarity of modes between variables.
The process includes:

a) Determining the initial number of clusters (k),
b) Initialising the initial mode randomly,
c) Calculating the dissimilarity between the data and the cluster mode using the following equation:

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \quad (1)$$

where $\delta(x_j, y_j) = 0$ if $x_j = y_j$ and $\delta(x_j, y_j) = 1$ if $x_j \neq y_j$

d) Updating the mode value based on the most frequent value (mode) in each attribute, and
e) Repeating the process until the mode is stable (converged).

5. Cluster evaluation using the Silhouette Coefficient method.
This method measures internal consistency between cluster members and is calculated using the equation:

$$s_i = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

Where: $a(i)$ is the average distance between objects i and all other objects in the same cluster (cohesion), $b(i)$ and is the minimum average distance between objects i and all other objects in different clusters (separation).
The silhouette value is in the range of -1 and 1 with the following interpretations:
i) $s_i \approx 1$ means the object i is well clustered,
ii) $s_i \approx 0$ means the object i is on the boundary between two clusters
iii) $s_i \approx -1$, which means the object i is likely in the wrong cluster.
The highest average silhouette value is used to determine the optimal number of clusters.
6. Clustering results are visualised using a ggplot2-based data profile dashboard in R. Interpretation is performed by examining the dominant characteristics of each cluster.

3. Results and Discussions

The initial step in the analysis process is data cleaning, which is performed to ensure data integrity and quality before the modelling stage. This process includes correcting structural errors, such as duplicate and inconsistent entries, and handling missing values through imputation or removing cases with incomplete information. After cleaning, the data set suitable for analysis was 74 respondents registered as drug users in 2024. This dataset was then used as the basis for clustering social characteristics and drug use behaviours.

The next stage is data transformation, which converts numeric variables into categories to accommodate the characteristics of the K-Modes algorithm, which can only process categorical data. This transformation process includes grouping variables such as age into "teenager," "young adult," and "adult," and grouping education levels into "low," "middle," and "high." This transformation aims to facilitate the interpretation of the clustering results and ensure that the data type matches the algorithm's requirements.

```
> str(data_responden)
'data.frame':   74 obs. of  5 variables:
 $ Age                     : Factor w/ 3 levels "Adult","Teenager",..: 3 1 3 1
 $ Occupation              : Factor w/ 4 levels " Civil Servant",..: 3 2 2 2 4
 $ Academic.Level          : Factor w/ 2 levels " High"," Low": 2 2 2 2 2 1 1
 $ Duration.of.Drug.Abuse  : Factor w/ 2 levels " Long-Term"," Moderate": 2 1
 $ Socio.environment.Effects: Factor w/ 4 levels " Companions",..: 3 3 3 3 4 3
```

Figure 1. Data Type of Dataset

After the transformation is carried out, all variables in the dataset are converted to factor type using the R software. Checking the data type is done using the str function in R, as shown in Figure 1. Based on the results of the data structure display, all variables have been confirmed to be categorical (factor) type. Hence, the dataset is ready to be used in the clustering process.

The next step is to determine the optimal number of clusters using the Silhouette method. This method is used to evaluate clustering quality by measuring the degree of closeness between objects within a cluster compared to the distance between objects within other clusters. Silhouette coefficient values range from -1 to 1, with values closer to 1 indicating a strong fit within the cluster, while negative values indicate a clustering error.

In this study, the silhouette calculation was performed using the Gower distance matrix, as this metric is most appropriate for categorical data with mixed variables or nominal scales. Gower distance calculates similarity between respondents based on the proportion of identical attributes, providing a more accurate measure of similarity than the Euclidean distance commonly used for numerical data. Using the Gower distance also ensures that each attribute contributes equally to determining the closeness between objects within a cluster.

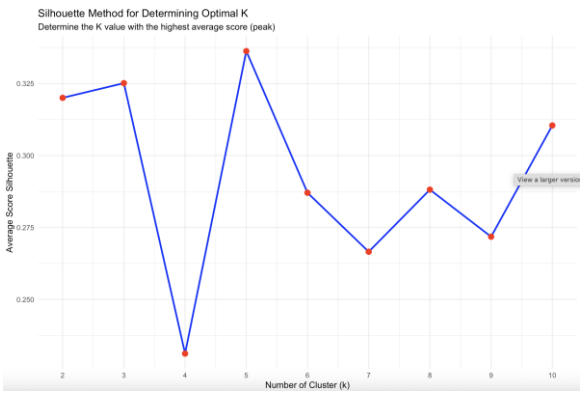

Figure 2. Plot of Silhouette Methods

Figure 2 shows that the highest average silhouette score was obtained with five clusters, with a value of 0.3363. This value indicates that forming five clusters (K = 5) is the optimal number, as it achieves the best balance between homogeneity within clusters and differences between clusters. Although the silhouette score falls within the

Available online at https://conference.lppm.unila.ac.id/index.php/icsiger
DOI: https://doi.org/10.23960/icsiger

moderate range, this result remains acceptable in the context of categorical social data with high diversity. Thus, the K-Modes algorithm is considered capable of forming a clustering structure that is sufficiently representative of the social characteristics of drug users.

Based on these results, the following clustering process was carried out using the K-Modes algorithm with an optimal number of five clusters. This clustering aims to identify similar patterns among respondents based on categorical variables such as gender, education level, employment status, and motivation for use. The final results of the clustering process are presented in Table 2, which displays the distribution of members and the main characteristics of each drug user cluster.

Table 2. Number of clusters using k-modes

Cluster	Frequent	Age	Occupation	Academic Level	Duration of Drug Abuse	Socio-Environment Effects
1	5	Teenager	Student	High	Moderate	Society
2	11	Young Adult	Jobless	High	Long-Term	Companions
3	18	Adult	Self-Employed	High	Long-Term	Companions
4	27	Adult	Self-Employed	Low	Moderate	Friends
5	13	Young Adult	Jobless	High	Moderate	Friends

Based on the clustering results shown in Table 2, five main clusters were identified, representing the social patterns and characteristics of drug users in Palopo City. Each cluster has distinct characteristics based on age, type of employment, education level, duration of abuse, and social environmental influences.

Cluster 4, with the most significant number of members —27 individuals —can be identified as having the highest level of abuse among all clusters. The main characteristics of this cluster are that adults dominate it, are self-employed, have a low level of education, and exhibit moderate duration of drug abuse. Furthermore, the most influential social environmental factor in this group comes from their circle of friends. This pattern suggests that adults with unstable economic status and a permissive social environment toward drug use are at higher risk of engaging in substance abuse.

These findings suggest that the BNNK Palopo should prioritise attention to individuals in Cluster 4, particularly through community-based and social environment-based intervention programs. Possible approaches include raising awareness of the dangers of drugs through peer groups, counselling in informal workplaces, and fostering productive economic practices for low-educated adults. These strategies are expected to reduce dependency rates and prevent the spread of drug abuse among productive members of society.

Meanwhile, other clusters exhibit varying characteristics that also provide important insights for prevention policies. Cluster 1, for example, consists of young students with a high level of education and moderate duration of use, where the primary influence comes from the community environment. This situation suggests that drug abuse among students continues to occur due to broader social environmental pressures, such as exposure to lifestyle and media influences.

Clusters 2 and 5 are dominated by jobless young adults -with a high level of education- and are primarily influenced by peers (friends or companions). The user patterns of these clusters indicate a strong link between unemployment and drug abuse. Cluster 3 describes the characteristics of a group of self-employed adults. This group also has a high level of education and long-term drug abuse. This may indicate a link between drug abuse as a way to relieve work pressure.

Overall, the results of this clustering demonstrate that drug abuse in Palopo City is not only influenced by individual factors but is also closely linked to social and economic conditions, as well as the influence of the peer group. These findings align with research by Deep et al. (2024) and Hisyam et al. (2025), which confirmed that social pressure and economic status play a significant role in the tendency for drug abuse behaviour among productive-age groups.

To provide a deeper understanding, the results of this clustering are visualised in the form of a data profile dashboard. The dashboard displays visual comparisons between clusters based on key variables, such as age, education level, occupation, and social environmental influences. Through this visualisation, BNNK Palopo can easily identify the most at-risk target groups and design evidence-based prevention strategies.

Available online at https://conference.lppm.unila.ac.id/index.php/icsiger
DOI: https://doi.org/10.23960/icsiger

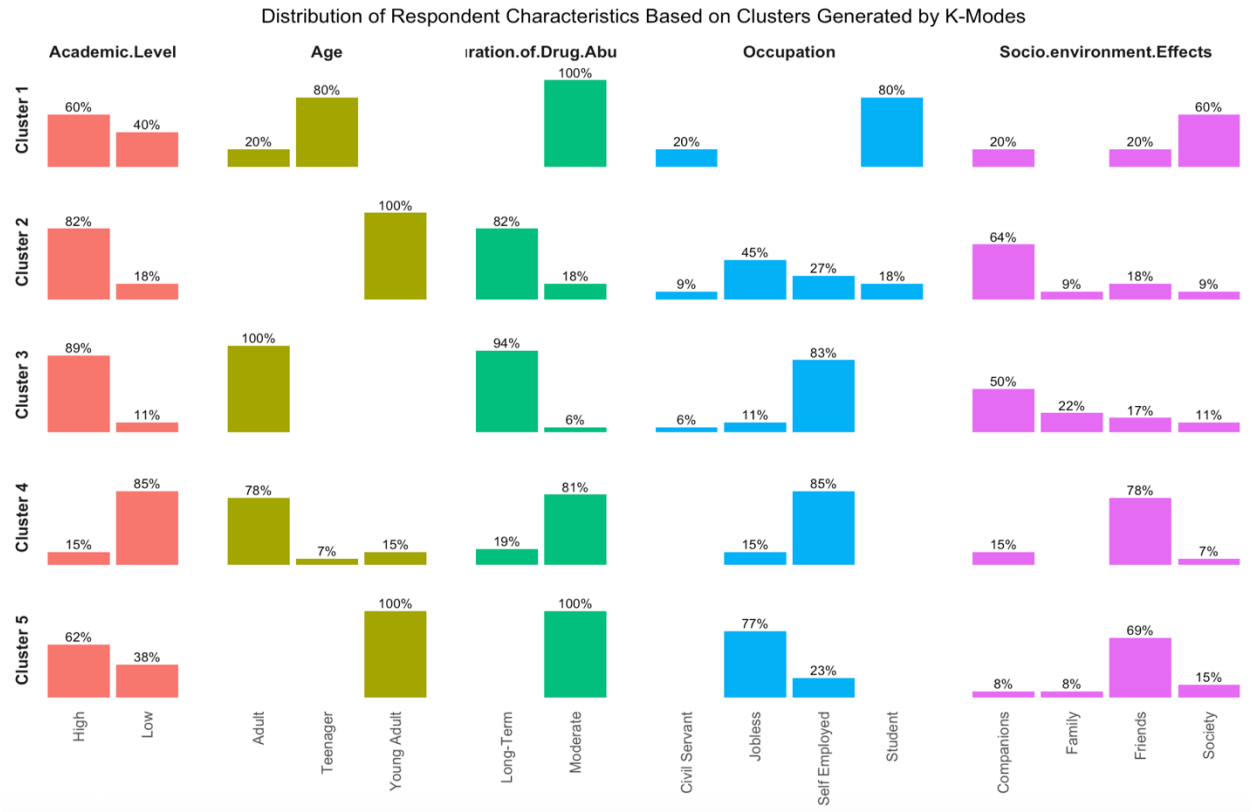

Figure 3. Profil Dachboard for Each Cluster

Based on Figure 3, which displays the data profile dashboard resulting from the clustering, it can be interpreted that each drug user cluster has distinct social and psychological characteristics. This visualisation confirms the findings of the previous K-Modes analysis, which found that a combination of factors, such as age, employment status, education level, duration of use, and the intensity of social influence, influences drug abuse behaviour in Palopo City. The characteristics and policy implications of each cluster are explained in detail as follows:

Cluster 1 – Experimental Teenagers

This cluster represents the age group of teenagers who use drugs in the early stages as part of their teenage experimentation. The primary influence comes from the broader social environment, such as schools, communities, and social media, which encourage experimentation without full awareness of the dangers of drug abuse and its future impact. Use in this group is generally temporary, as a means of proving themselves within their social circle, and they do not yet show signs of severe dependence. Therefore, recommended intervention strategies include preventive education programs and anti-drug campaigns that emphasise self-awareness, strengthening moral values, and refusal skills training in schools and youth communities.

Cluster 2 – High-Risk Unemployed Young Adults.

This group is predominantly composed of young adults with relatively high levels of education, yet many of them are unemployed. The main risk factors in this cluster are the influence of peers who have long-term drug use, as well as psychological distress due to economic pressures. Given this social background, this cluster is categorised as a high-risk group for dependence. Recommended intervention programs include community-based rehabilitation, job training, and entrepreneurial guidance to increase economic independence while reducing dependence on social groups that foster permissive behaviours.

Cluster 3 – Independent Adults with Social Pressure from Close Friends.

This cluster consists of adults with stable jobs as self-employed individuals who use drugs due to social pressure from their close circle of friends. This group tends to have relatively stable economic circumstances, but the influence of their friendships and social networks tends to encourage drug abuse. This pattern suggests that economic factors do not drive drug use, but relatively strong social ties within their friendship groups. The wrong community will lead individuals to drug abuse. Recommended interventions include a peer-support system approach and community-based interpersonal counselling to build collective awareness and positive social support within their communities.

Cluster 4 – Young Adults Exposed to a Strong Social Environment.

This cluster has the most significant number of members and the highest rates of drug abuse. It is characterised by young adults with low education and precarious employment, living in a

Available online at https://conference.lppm.unila.ac.id/index.php/icsiger
DOI: https://doi.org/10.23960/icsiger

social environment permissive of drug use. External factors, such as loose social norms and minimal social control, significantly influence the behaviour of this group. Therefore, this cluster is a top priority in prevention and response efforts by the Palopo BNNK. Relevant intervention strategies include alternative education programs, life skills training, and socioeconomic empowerment for individuals with low educational backgrounds.

Cluster 5 – Stable Adults Enmeshed in Social Networks of Users.

This cluster represents a group of adults with relatively stable economic conditions who continue to use drugs due to the influence of long-established social networks. This cluster is similar to Cluster 3, but drug abuse is focused within the same friendship group, driven by the psychological needs of these established friendships. This group has gone through a cycle of relapse in their attempts to break free from drug abuse due to their peer group. Recommended interventions for this cluster are group therapy and long-term psychosocial rehabilitation, aimed at building self-awareness, restructuring social relationships, and strengthening family support.

Overall, these interpretations indicate that social and psychological factors play a significant role in shaping patterns of drug use in Palopo City. Effective interventions must be tailored to the unique characteristics of each cluster. Rehabilitation focuses on medical or legal aspects as a drug prevention effort. However, strategies and interventions from a social and psychological perspective are needed to empower individuals and communities to play an active role in preventing drug abuse.

4. CONCLUSIONS

This study successfully clustered drug user patterns based on social and psychological factors using the K-Modes algorithm, validated by the Silhouette method. The analysis revealed that the optimal number of clusters was five, representing a diverse range of user characteristics based on a combination of age, occupation, education level, duration of use, and social environment.

Each cluster exhibited distinct characteristics: from experimental teenagers who tried drugs due to social influences, to chronic adults who had long-term drug abuse due to the impact of close friends or social networks. The most dominant factors distinguishing the clusters were age, employment status, and social environment.

These findings confirm that data mining approaches, specifically the K-Modes algorithm, are practical for analysing categorical data in social contexts such as drug abuse. The clustering results can provide a basis for rehabilitation institutions, government agencies, and social service organisations in designing more targeted, user profile-based interventions.

For future research, the use of larger and more diverse datasets, as well as the integration of hybrid clustering methods or predictive machine learning algorithms, is recommended to enhance the accuracy and generalizability of the clustering results.

ACKNOWLEDGMENTS

The authors would like to thank the BNNK Palopo for their support, cooperation, and assistance throughout the research process. The availability of accurate data and ease of coordination with the BNNK Palopo were key factors in ensuring the smooth running of the research, particularly in understanding the dynamics of the social environment and efforts to prevent drug abuse.

REFERENCES

Alphenia, S., & Kariyam, K. (2022, December). Clustering of Travel Insurance Cases with K-Modes Algorithm. In Proceeding of The International Conference on Natural Sciences, Mathematics, Applications, Research, and Technology (Vol. 2, pp. 1–9).

Badan Narkotika Nasional. (2025). PERINGATAN HANI 2025: MEMUTUS RANTAI PEREDARAN GELAP NARKOBA MELALUI PENCEGAHAN, REHABILITASI, DAN PEMBERANTASAN MENUJU INDONESIA EMAS 2045. Diakses pada 9 November 2025 dari https://bnn.go.id/peringatan-hani-2025-memutus-rantai-peredaran-gelap-narkoba-melalui-pencegahan-rehabilitasi-dan-pemberantasan-menuju-indonesia-emas-2045/

Deep, P. D., Ghosh, N., Gaither, C., & Rahaman, M. S. (2024). The factors affecting substance use and the most effective mental health interventions in adolescents and young adults. Psychoactives, 3(4), 461–475. https://doi.org/10.3390/psychoactives3040028

Dorman, K. S., & Maitra, R. (2022). An efficient k-modes algorithm for clustering categorical datasets. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15(1), 83–97. https://doi.org/10.1002/sam.11546

Faizan, M., Zuhairi, M. F., Ismail, S., & Sultan, S. (2020). Applications of clustering techniques in data mining: a comparative study. International Journal of Advanced Computer Science and Applications, 11(12).

Hasanuddin, Muhammad. (2024). Sulawesi

Available online at https://conference.lppm.unila.ac.id/index.php/icsiger
DOI: https://doi.org/10.23960/icsiger

Selatan darurat narkoba peringkat lima se-Indonesia. Diakses pada 9 November 2025 dari https://riau.antaranews.com/berita/399486/sulawesi-selatan-darurat-narkoba-peringkat-lima-se-indonesia

Hisyam, C. J., Aditya, A., Az-Zahra, D., Galih, D., Sholihah, F., Nafilata, M., & El Sahla, S. (2025). KETERASINGAN SOSIAL SEBAGAI FAKTOR PEMICU PENYALAHGUNAAN NARKOBA DI KALANGAN REMAJA. JURNAL ILMIAH RESEARCH AND DEVELOPMENT STUDENT, 3(1), 122-135. https://doi.org/10.59024/jis.v3i1.1067

Jung, H., & Chung, K. (2021). Social mining-based clustering process for big-data integration. Journal of Ambient Intelligence and Humanized Computing, 12(1), 589–600.

Nandini, Andi Bunayya., Ibrahim, S. (2025). 9 Bulan 63 Kasus Narkoba di Palopo, 12 Ditahan. Diakses pada 9 November 2025 dari https://makassar.tribunnews.com/palopo/1814800/9-bulan-63-kasus-narkoba-di-palopo-12-tersangka-ditahan.

Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. Social Science Research, 110, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817

Soleimanian Gharehchopogh, F., & Haggi, S. (2020). An Optimization K-modes clustering algorithm with elephant herding optimization algorithm for crime clustering. Journal of Advances in Computer Engineering and Technology, 6(2), 79–90.

Tolner, F., Fegyverneki, S., Eigner, G., & Barta, B. (2021, September). Clustering based on Preferences with K-modes using Categorical Variables. In 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY) (pp. 55–60). IEEE.

Zamaninasab, Z., Sharifi, H., & Bahrampour, A. (2018). Naïve Bayes evidence accumulation K-modes clustering: A new method for classifying binary data and its application on real data of injecting drug users. Journal of Biostatistics and Epidemiology, 4(2), 72–78.

Available online at https://conference.lppm.unila.ac.id/index.php/icsiger
DOI: https://doi.org/10.23960/icsiger