SITEKIN: Jurnal Sains, Teknologi dan Industri, Vol. 22, No. 1, December 2024, pp.75 - 82
ISSN 2407-0939 print/ISSN 2721-2041 online

Use of Data Mining Technology to Identify Narcotics Distribution Patterns

Nirwan Moningka[1], Kusrini[1]
[1]Master of Informatics, Amikom Yogyakarta University
Jl. Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta, Indonesia 55283
*Email: nirwan@students.amikom.ac.id, kusrini@amikom.ac.id

ABSTRACT

The abuse of narcotics has become one of the significant social and health problems in various countries worldwide. Conventional methods relying on manual analysis or traditional approaches may not be effective enough in addressing this challenge. Therefore, a more sophisticated and efficient approach is needed to tackle this issue. Data mining uses techniques from statistics, machine learning, and pattern recognition to extract valuable information from large data sets. This research employs data collection methods from the Narcotics Investigation Directorate of the Maluku Regional Police from 2021 to 2023. This data includes profiles of narcotics users, such as the age of the perpetrators, gender, last education level, occupation, location of arrest, and type of narcotic. The aim is to identify the patterns of narcotic distribution in the Maluku Province using data mining techniques, namely the Apriori algorithm, Naive Bayes, Random Forest, and Support Vector Machine (SVM). The exclusion of the age variable was a correct decision, as it resulted in an increase in accuracy. This increase is likely due to the high variation in the age variable. The accuracy improvement was more evident in the Random Forest algorithm compared to Naive Bayes and SVM. Random Forest achieved satisfactory results with an accuracy of 0.96. This indicates that Random Forest is a good algorithm for predicting narcotics user data. These results suggest that the pattern of narcotics distribution is associated with specific factors, including the male gender, the highest level of education being high school, a self-employed occupation, arrest locations on public roads, and the type of narcotic being Sabu.

Keywords: *Apriori Algorithm, Machine Learning, Naïve Bayes, Random Forest, SVM.*

Introduction

The misuse of narcotics has become one of the significant social and health issues in various countries around the world. In Indonesia, this problem cannot be ignored either [1][2]. The impact is very damaging, not only for the individuals involved in the misuse, but also for society as a whole [3][4]. The misuse of narcotics has led to an increase in crime rates, damaged public health, harmed social relationships, and even caused deaths. [5][6].

The Maluku Province is one of the provinces in Indonesia, consisting of a cluster of islands scattered across the eastern part of the country. Covering an area of approximately 712,480 km², with 85% being sea and 15% land, Maluku faces significant challenges in terms of supervision and law enforcement. These challenges are particularly pronounced in efforts to prevent and combat narcotics trafficking and abuse, given its vast and dispersed geographic layout.

Geographically, Maluku consists of 11 regencies/cities, with a population spread across various large and small islands as shown in Figure 1. The geographic conditions present challenges in accessibility and mobility for law enforcement officers, including in identifying and addressing the spread of narcotics. The province has abundant natural resources and a strategic geographic location, making it a potential target for narcotics syndicates to expand their distribution networks. Although the number of drug abuse cases in Maluku Province may not be as high as in other metropolitan areas in Indonesia, the impact remains significant and requires serious attention.

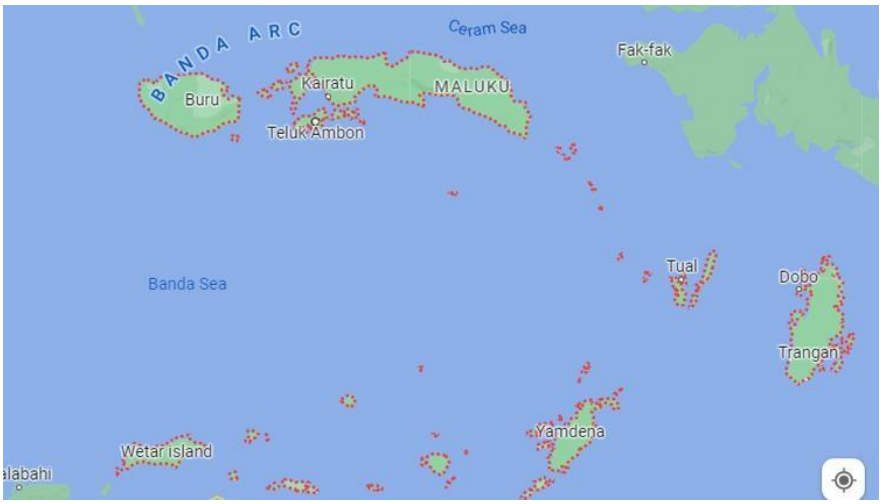

Figure 1. Map of Maluku Province

The Narcotics Investigation Directorate, hereinafter referred to as Ditresnarkoba, is an implementing unit responsible for narcotics investigations at the regional police level under the Provincial Police Chief. From 2021 to 2023, Ditresnarkoba has arrested at least 572 drug users. The users are categorized based on criteria such as age, gender, last education level, occupation, arrest location, and type of narcotic. Conventional methods relying solely on manual analysis or traditional approaches may not be sufficiently effective in tackling these complex challenges. Therefore, it is essential to adopt a more advanced and efficient approach to effectively address these issues.

Computer science has become one of the most influential fields in various aspects of life, including in addressing complex social and security issues. One significant contribution of computer science is in the field of data mining, which enables deep and sophisticated analysis of large datasets to uncover patterns, relationships, and trends that cannot be identified with traditional analytical methods. This capability is crucial for making informed decisions and improving problem-solving strategies in various domains [7]. Data mining has become an exceptionally powerful tool for understanding complex phenomena, including the widespread distribution and impact of narcotics [8].

Data mining employs techniques from statistics, machine learning, and pattern recognition to extract valuable and actionable information from large and complex data sets [9][10]. In this context, Data Mining technology offers significant potential in effectively supporting efforts to prevent and address drug abuse [11]. The Apriori algorithm is a data mining algorithm that utilizes association methods to identify item sets that frequently occur together in large datasets. Its main goal is to uncover patterns of variables that often appear in conjunction with one another, thereby providing valuable insights into the relationships between different items within the data [8]. The Apriori algorithm is used to find association rules in data. It is highly effective for identifying relationships between various attributes in narcotic distribution data, such as types of narcotics that are frequently found together or locations that often become distribution points. This can help in understanding patterns and correlations that may aid in targeted interventions and preventive measures.

Naïve Bayes is a statistical algorithm that applies probability theory to predict future outcomes based on observations from previous experiences, leveraging the expertise and knowledge of domain experts and practitioners to make informed predictions [12]. Naive Bayes is a simple, yet effective classification algorithm based on Bayes' theorem. Despite its simplicity, this algorithm can be used to predict the future occurrence of drug distribution events with a reasonable degree of accuracy based on historical data. Random Forest is an ensemble learning algorithm that combines numerous decision trees to enhance prediction accuracy. This algorithm is adept at managing data with a large number of diverse variables, making it particularly effective for analyzing complex patterns in drug distribution and other intricate datasets [13].

Support Vector Machine (SVM) is a powerful classification method that works by identifying the optimal hyperplane to separate different classes of data with the maximum margin. SVM has proven to be

highly efficient in high-dimensional spaces and is particularly effective for applications that require accurate pattern recognition and classification, such as image recognition and text categorization [14]. However, despite the significant potential of Data Mining technology in addressing drug abuse issues, its application in this context remains limited. Several challenges are encountered, including high data complexity, large data volumes, and the diversity of drug types along with various related factors such as socio- economic conditions, geographical distribution, and patterns of abuse. Therefore, comprehensive, and in-depth research is essential to develop more effective methods and strategies for leveraging Data Mining technology to identify and analyze patterns of drug distribution and abuse in Maluku Province. This includes refining algorithms, improving data integration, and addressing privacy and ethical considerations.

Another study conducted by Gesang Bekti Setyo Nugroho in 2021 aimed to design a decision support system to provide recommendations in the rehabilitation assessment process. This research employed the Random Forest algorithm, but the results were suboptimal due to the insufficient amount of available data and the imbalance in the data used for predictions. These factors caused Random Forest's performance in this study to fall short of its full potential, highlighting the need for more balanced and comprehensive data for improved accuracy [15].

In a previous study conducted by Agus Setiawan in 2022, the aim was to predict the number of narcotics cases that would occur in future years based on the historical data of cases from previous years. The research utilized various methods, including Linear Regression, Neural Networks, and Support Vector Machines, to achieve this goal. The findings revealed that the Support Vector Machine (SVM) algorithm provided the most accurate predictions for managing narcotics cases in 2021 compared to the other methods. Future research is anticipated to be enhanced by integrating classification algorithms, which would help in identifying the most used types of narcotics and further improving the accuracy of predictions [16].

Based on previous research, this study aims to comprehensively identify and analyze drug distribution patterns in the Maluku Province using advanced data mining techniques, specifically the Apriori algorithm, Naive Bayes, Random Forest, and Support Vector Machine (SVM).

Research Methods

Datasets

This research employs data collection methods from the Narcotics Investigation Directorate of the Maluku Regional Police covering the period from 2021 to 2023. The data comprises profiles of drug users, including details such as the age of the offenders, gender, highest level of education, occupation, location of arrest, and types of narcotics. This information is detailed in Table *1*.

Table 1. Narcotics user data

NUMBER	AGE	GENDER	LAST EDUCATION	WORK	CAPTURE LOCATION	TYPES OF NARCOTICS
1	44	MALE	SENIOR HIGH SCHOOL	SELF-EMPLOYED	PUBLIC ROAD	SABU
2	41	MALE	SENIOR HIGH SCHOOL	NOT YET WORKING	MARKET	GANJA
3	33	FEMALE	COLLEGE	EMPLOYEE	PARKING AREA	SINTETIS
4	26	MALE	SENIOR HIGH SCHOOL	NOT YET WORKING	LODGING	GANJA
5	35	MALE	SENIOR HIGH SCHOOL	SELF-EMPLOYED	PUBLIC ROAD	SABU

Next, the data will be validated and preprocessed. Data validation is performed to ensure that the data used is accurate, complete, and meets the specified criteria before being used for further analysis. Data preprocessing is carried out to transform the data into a form that is more suitable for analysis and to ensure its compatibility with the analytical methods to be applied.

Apriori Algorithm

This research uses the Apriori Algorithm to identify item sets in a dataset of drug users. The application of the Apriori Algorithm is anticipated to produce item sets that will be utilized for predictive analysis. The support value is derived from the following equation (1).

$$\text{Suppprt (x)} = \frac{Lots\ of\ items\ (x)}{N} \quad (1)$$

Where:
Support (X) = The percentage of cases for a particular combination of items
Item (X) = The amount of data that has value X
N = The total amount of all data

By using machine learning, the apriori algorithm is is carried out by:

from mlxtend.frequent_patterns import apriori, association_rules (to retrieve the apriori and association_rules functions from the mlxtend library)
itemset = apriori(transaksi_df, min_support=0.1, use_colnames=True) (to carry out the apriori algorithm on the df dataset with a minimum support of 0.1)

The support value ranges from zero (0) to one (1). Combinations of variables with high support values will be used to make predictions using Naïve Bayes, Random Forest, and Support Vector Machine (SVM) algorithms, ensuring more accurate and reliable outcomes.

Naïve Bayes Algorithm

The Bayes' Theorem operates according to the principle of conditional probability. Conditional probability represents the likelihood or chance of an event occurring, given that another related event has already occurred. The calculation of conditional probability can be done using the following equation (2).

$$P(A|B) = \frac{P(B|A)\times P(A)}{P(B)} \quad (2)$$

Where:
P(A|B) = A's conditional probability given by B
P(B|A) = B's conditional probability given by A
P(A) = Probability of event A
P(B) = Probability of event B

By using machine learning, the Naïve Bayes algorithm is carried out by:

from sklearn.naive_bayes import GaussianNB (to build and evaluate a gaussian naïve bayes model from the sklearn library)
model = GaussianNB() (for the Initiation of the Gaussian Naïve Bayes model)
model.fit(X_train, y_train) (for training models)
y_pred = model.predict(X_test) (to make predictions)

This algorithm is very fast and efficient, especially for large datasets with relatively independent features.

Random Forest Algorithm

The Random Forest process begins by constructing multiple decision trees independently, where each tree is trained on a random subset of the training data obtained through a technique called bootstrapping (sampling with replacement). During the training of each tree, only a random subset of features is considered for each split, which adds diversity among the trees and helps to ensure that the model does not rely too heavily on any single feature. Once all the trees are trained, Random Forest combines the predictions from all the trees through majority voting (for classification) or averaging (for regression) to provide a final result that is more stable and accurate. This approach significantly reduces the risk of overfitting compared to using just a single decision tree, resulting in a more robust and reliable model.

By using machine learning, the Random Forest algorithm is carried out by:

from sklearn.ensemble import RandomForestClassifier (to build and train models from the sklearn library).
model = RandomForestClassifier(random_state=42) (to initiate and train a random forest model)
model.fit(X_train, y_train).
y_pred = model.predict(X_test) (for Predictions using the testing set).

Algoritma Support Vector Machine

If the data is linearly separable, SVM tries to find the hyperplane that separates the data with the maximum margin. With the equation:

$$Minimize = \frac{1}{2}||w|| \quad (3)$$

$$yi(w.xi + b) \geq 1 \quad (4)$$

Dimana:
w = weight vector
b = refraction
xi = feature vector of data point i
yi = class label (1 or -1)

By using machine learning, the Support Vector Machine (SVM) algorithm is carried out by:

from sklearn.svm import SVC (to build an SVM model from the sklearn library)
svm_model = SVC(kernel='linear') (for creating linear SVM models)
svm_model.fit(X_train, y_train)
y_pred = svm_model.predict(X_test) (for evaluation of prediction models)

Results and Discussion

Testing of the naïve Bayes, random forest, and Support Vector Machine algorithms using 70% training data and 30% testing data on a narcotics user dataset, without analyzing the results of the apriori algorithm, produced the following performance metrics: accuracy, precision, recall, and F1-Score:

Table 2. The test results use 6 variables.

Method	Accuracy	Types of Narcotics	Precision	Recall	F1-Score	Support
Naïve Bayes	0.76	Ganja	0.56	1.00	0.72	47
		Sabu	0.95	0.72	0.82	98
		Sintetis	1.00	0.48	0.65	27
Random Forest	0.95	Ganja	0.95	0.87	0.91	47
		Sabu	0.98	1.00	0.99	98
		Sintetis	0.86	0.93	0.89	27
SVM	0.80	Ganja	0.70	0.55	0.62	47
		Sabu	0.83	0.90	0.86	98
		Sintetis	0.83	0.89	0.86	27

The use of the apriori algorithm on narcotics user datasets produces various support results that will be utilized for making predictions. The itemsets derived from the apriori algorithm consist of one (1) variable up to five (5) variables. This study will employ combinations of three (3) to five (5) variables, which include

primary variables with the highest support values as illustrated in Table *3*. Consequently, the itemsets that will be used in this analysis are as follows:

Table 3. Results of using the a priori algorithm

Variabel	Item 1	Item 2	Item 3	Item 4	Item 5	Support
3 Variabel	Male	Sabu	Senior High School			0.573426573
4 Variabel	Male	Sabu	Senior High School	Settlement		0.326923077
5 Variabel	Senior High School	Public road	Male	Sabu	Self-employed	0.241258741

The first test was conducted using 3 variables: Male representing the gender variable, Sabu representing the primary variable, and Senior High School representing the last education variable. Testing the naïve Bayes, random forest, and Support Vector Machine algorithms with 70% training data and 30% testing data on these three (3) variables yielded the following results:

Table 4. The test uses 3 variables.

Method	Accuracy	Types of Narcotics	Precision	Recall	F1-Score	Support
Naïve Bayes	0.38	Ganja	0.31	1.00	0.48	47
		Sabu	0.56	0.05	0.09	98
		Sintetis	1.00	0.48	0.65	27
Random Forest	0.65	Ganja	0.00	0.00	0.00	47
		Sabu	0.62	1.00	0.76	98
		Sintetis	1.00	0.48	0.65	27
SVM	0.65	Ganja	0.00	0.00	0.00	47
		Sabu	0.62	1.00	0.76	98
		Sintetis	1.00	0.48	0.65	27

The second test was conducted using 4 variables: Male representing the gender variable, Sabu representing the primary variable, Senior High School representing the highest education variable, and Settlement representing the arrest location variable. Testing the Naïve Bayes, Random Forest, and Support Vector Machine algorithms using 70% training data and 30% testing data on these four (4) variables yielded the following results:

Table 5. The test uses 4 variables.

Method	Accuracy	Types of Narcotics	Precision	Recall	F1-Score	Support
Naïve Bayes	0.38	Ganja	0.31	1.00	0.48	47
		Sabu	0.56	0.05	0.09	98
		Sintetis	1.00	0.48	0.65	27
Random Forest	0.96	Ganja	1.00	0.85	0.92	47
		Sabu	0.98	1.00	0.99	98
		Sintetis	0.84	1.00	0.92	27
SVM	0.71	Ganja	0.00	0.00	0.00	47
		Sabu	0.69	1.00	0.82	98
		Sintetis	0.83	0.89	0.86	27

The third test was conducted using 5 variables, where: Male represents the gender variable, Sabu represents the primary variable, Senior High School represents the highest education variable, Public road represents the arrest location variable, and Self-employed represents the occupation variable. Testing of the Naive Bayes, Random Forest, and Support Vector Machine algorithms using 70% training data and 30% testing data on these five (5) variables yielded the following results:

Table 6. The test uses 5 variables.

Method	Accuracy	Types of Narcotics	Precision	Recall	F1-Score	Support
Naïve Bayes	0.76	Ganja	0.56	1.00	0.72	47
		Sabu	0.95	0.72	0.82	98
		Sintetis	1.00	0.48	0.65	27
Random Forest	0.96	Ganja	1.00	0.85	0.92	47
		Sabu	0.98	1.00	0.99	98
		Sintetis	0.84	1.00	0.92	27
SVM	0.80	Ganja	0.60	0.89	0.72	47
		Sabu	0.97	0.72	0.83	98
		Sintetis	0.83	0.89	0.86	27

The selection of variables in prediction models significantly affects their performance. For instance, the Naïve Bayes algorithm shows a decrease in accuracy when using 3 and 4 variables, but it maintains its accuracy with the use of 5 variables. On the other hand, the Random Forest algorithm experiences a drop in accuracy with the use of 3 variables but demonstrates an improvement with the use of 4 and 5 variables, with accuracy slightly higher than when using 6 variables. Support Vector Machine (SVM) also shows a decrease in accuracy with 3 variables but improves with 4 variables and continues to improve further with the use of 5 variables.

The decision to exclude the age variable proves to be beneficial as it results in an enhancement in accuracy. This improvement is likely due to the substantial variation in the age variable, which can introduce noise into the model. The increase in accuracy is more pronounced with the Random Forest algorithm compared to Naïve Bayes and SVM.

Among the algorithms tested, Random Forest delivers the most satisfactory results, achieving an impressive accuracy rate of 0.96.

Conclusion

The application of the apriori algorithm proves to be beneficial in enhancing the accuracy of the results. It is important to note that using fewer secondary variables does not necessarily guarantee high accuracy values, despite potentially having better support values. In our study, the random forest algorithm demonstrated superior accuracy with a value of 0.96, in contrast to the naïve bayes algorithm, which achieved an accuracy of 0.76, and the support vector machine (SVM) algorithm, which recorded an accuracy of 0.80. These results suggest that the pattern of narcotics distribution is closely associated with specific factors, including the male gender, the highest level of education being high school, a self-employed occupation, arrest locations on public roads, and the type of narcotic being Sabu.

Future research could benefit from incorporating additional variables such as the specific arrest region, longitude, and latitude. Furthermore, presenting the data through a geographic information system (GIS) could provide a more comprehensive visualization of narcotics distribution patterns. This approach would not only enhance the analysis but also offer valuable insights into the geographical spread of narcotics, potentially informing more effective intervention and prevention strategies.

References

[1] R. D. Situmorang, Sumarno, and N. Hidayati, “Penerapan Data Mining dalam Klasifikasi Pencegahan Narkoba Menggunakan Algoritma Naïve Bayes di BNN Kota Pematangsiantar,” *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 4, pp. 295–302, Sep. 2022.

[2] I. Sari, R. Kosasih, and D. Indarti, "Clustering and Topic Modeling of Verdicts of Narcotics Cases Using Machine Learning," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 27, no. 6, pp. 1168–1174, Aug. 2023.

[3] C. M. Simamora, H. F. Kennedy, S. Nurhuda, M. Agustiawan, M. Yogi Prawira, and R. Siregar, "Penyalahgunaan Narkoba Pada Remaja Ditinjau Dari Teori Asosiasi Diferensial," *EKOMA : Jurnal Ekonomi*, vol. 3, no. 3, pp. 811–817, Mar. 2024.

[4] I. Amal and R. A. Putri, "Clustering Pecandu Narkoba Menggunakan Algoritma K-Means Clustering," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 5, no. 2, pp. 434–443, Dec. 2023.

[5] A. Winarta and W. J. Kurniawan, "Optimasi Cluster K-Means Menggunakan Metode Elbow pada Data Pengguna Narkoba dengan Pemrograman Python," *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 5, no. 1, pp. 113–119, Jan. 2021.

[6] B. P. Tomasouw and Y. A. Lesnussa, "Deteksi Penyalahgunaan Narkoba dengan Metode Twin Bounded SVM," *Jurnal Ilmu Matematika dan Terapan*, vol. 15, no. 4, pp. 753–760, Dec. 2021.

[7] B. L. Hasibuan, Sofiah, and E. Yolanda, "Pengklasifikasian Data Pasien Tes Urine Dengan Metode Clustering Pada Kantor Badan Narkotika Nasional Provinsi Sumut (BNNP SUMUT)," *JUKI : Jurnal Komputer dan Informatika*, vol. 4, no. 2, pp. 183–193, Nov. 2022.

[8] N. D. Sari and S. Khoiriah, "Penerapan Metode Asosiasi Pada Toko Afifa Dengan Algoritma Apriori," *INSTINK (Jurnal Inovasi Pendidikan, Teknologi Informasi & Komputer)Teknologi Informasi & Komputer)*, vol. 1, no. 1, pp. 8–17, Apr. 2022.

[9] W. Ginting, "Pengelompokan Data Pasien Test Urine dengan Metode Clustering pada Kantor Badan Narkotika Nasional," *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 5, no. 2, pp. 327–338, Jul. 2021.

[10] E. Yolanda and Suhardi, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Data Pasien Rehabilitasi Narkoba," *KILIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 1, pp. 182–191, Aug. 2023.

[11] S. S. M. Ajibade, O. J. Oyebode, J. P. Dayupay, N. G. Gido, A. C. Tabuena, and O. K. T. Kilag, "Data Classification Technique for Assessing Drug Use in Adolescents in Secondary Education," *J Pharm Negat Results*, vol. 13, no. 4, pp. 971–977, 2022.

[12] D. Setiadi and R. Syahri, "Penerapan Algoritma Naïve Bayes pada Sistem Prediksi Pengguna Narkoba di Kota Pagar Alam," *JUTIM: Jurnal Teknik Informatika Musiwaras*, vol. 7, no. 1, pp. 1– 10, Jun. 2022.

[13] U. Azmi, Hendrick, and Humaira, "Pendeteksian Aroma Ganja Kering Menggunakan Algortima Random Forest," *Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 4, no. 1, pp. 28–33, Mar. 2023.

[14] R. Dasmasela, B. P. Tomasouw, and Z. A. Leleury, "Penerapan Metode Support Vector Machine (SVM) untuk Mendeteksi Penyalahgunaan Narkoba," *PARAMETER: Jurnal Matematika, Statistika dan Terapannya*, vol. 1, no. 2, pp. 111–122, Oct. 2022.

[15] G. B. S. Nugroho, D. Rolliawati, and A. Yusuf, "Sistem Pendukung Keputusan Asesmen Rehabilitasi Narkotika Menggunakan Metode Random Forest Penulis Korespondensi," *Jurnal Sistem Informasi dan Teknologi*, vol. 4, no. 1, pp. 29–42, Jun. 2021.

[16] T. A. Setiawan, A. Ilyas, and Arochman, "Komparasi Model Prediksi Penanganan Kasus Narkotika," *Journal of Informatic and Computer Technology*, vol. 17, no. 1, pp. 42–48, Apr. 2022.