Imbalanced Data Analysis of Adolescent Risk Behavior of Drug Abuse using Random Forest

Ismaini Zain[1*], Kartika Fithiasari[2], Erma Oktania Permatasari[3], Tyas Ajeng Nastiti[4], Mardyono[5], Nilam Novita Sari[6], Resti Pujihasvuty[7], Sri Lilestina Nasution[8]
{ismaini_z@statistika.its.ac.id[1*], kartika_f@statistika.its.ac.id[2], erma.oktania@gmail.com[3]}

Department of Statistics, Faculty of Science and Data Analystics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia[1,2,3,6], Department of Visual Communication Design, Universitas International Semen Indonesia, Gresik, Indonesia[4], National Population and Family Planning Board, East Java, Indonesia[5]

Abstract. Adolescence represents a period of self-searching and vulnerability to fall into risky behavior such as drug abuse. In Indonesia, the case of drug abuse by adolescents is high. Therefore, to know the factors behind it can be done using classification such as random forest. The data used in this research were adolescent risk behavior of drug abuse based on SKAP. The percentage of drug abuse among adolescents are 4.1% shows that there is an imbalanced class in the data. It is necessary to handle the imbalanced data by applying the SMOTE-N. This study will classify the adolescent risk behavior of drug abuse using random forest combine with SMOTE-N to handle the imbalanced class. The results show that the model using SMOTE-N is better because it can increase specificity and g-means. The variables affect the classification of drug abuse among adolescents are the age, sex, and psychology consequence.

Keywords: Adolescent Risk Behavior, Drug Abuse, Imbalanced Data, Random Forest, SMOTE-N

1 Introduction

Adolescence is when an individual is no longer a child but has not yet become an adult. During this time, an individual begins to experience physical and psychological changes internally while also changing social expectations and perceptions. Adolescence is a period of self-searching. In this period, they are beginning to experience changes in lifestyle and relationships. At this time, adolescents are vulnerable to being influenced by their surroundings. The wrong surrounding can cause teenagers to fall into risky behavior. Risky behavior is defined as any conscious or non-conscious behavior with adverse social effects, such as drug abuse [1,2].

Drug abuse refers to the usage of drugs in excessive quantities or for purposes for which it was not intended, leading to significant distress, allowing the person to suffer from tolerance [3]. Different drugs can be abused, including illegal drugs such as heroin, prescription medicines like pain killers, and other medicines that can be bought off the supermarket, such as cough mixers [3]. Three factors affect adolescent risk behavior. First is the predisposing or motivating factors coming from the individual such as knowledge, age, gender, education, etc. The second is the enabling factor that allows or encourages such behavior to be carried out,

AECon 2020, December 19-20, Purwokerto, Indonesia

DOI 10.4108/eai.19-12-2020.2309145

such as health resources, economic status, access to information media, domicile location (urban or rural), etc.

The last is the reinforcing factor determined by other people, including family, friends, teachers, etc. The consequent reasons for drug abuse among adolescents are the absence of communication between parents and children, ease of access to drugs, depression, socialization problems, experimentation, peer pressure, confidence problems, etc. [4]. In 2015, [5] discuss about the factors affect the adolescent drug abuse behavior and the results show that ease of access to drugs and friends are the factors affect the drug abuse behavior. Knowledge and attitudes on drug abuse have been discussing by [6], with the results showing that the higher the adolescents' knowledge, the fewer adolescents will be involved in drug abuse. The lack of the knowledge of the drug

The lower the adolescents' knowledge, the higher adolescents will be involved in drug abuse and it can cause the number of drug abuse will increase. In Indonesia, in 2017, the number of people between 10 and 59 who abused drugs is about 3.37 million. Meanwhile, the number of students who abused drugs in 2018 is 2.29 million. The age range of 15-35 years old is the group that is prone to drug abuse [7]. Based on SKAP, in East Java in 2019, the number of adolescents who consume drugs is 191 or 4.1%. Although the number of adolescents who consume drugs is small, but it will have negative impact. Therefore, analysis is needed to investigate the factors behind it. Classification can be used to find out the factors to predict adolescent risk behavior of drug consumption.

Classification is the process of finding a model that describes data classes, and the model is used to predict the class label of objects for which the class label is unknown [8]. One of the popular classification methods is a tree-based classification, of which Random forest is one of them. A random forest is a collection of decision trees. The individual decision trees are generated using a random selection of attributes at each node to determine the split. During the classification, each tree votes, and the most popular class is returned [8]. Random forest is appealing because of the additional features they provide, such as measuring variable importance [9]. The variable importance is used to measure the candidate predictor variables that most influence is predicting the response variable. MDG (Mean Decrease Gini) is one of the essential measures. MDG calculated by adding up all the node impurity (Gini index for classification) decrease from splitting variable and then averaged over all the trees [10]. Apart from measuring variable importance, random forest also computes the missing value, detect outliers, give class weighting, etc. [8].

Although random forest has many advantages and provides additional features, if working on imbalanced data, the random forest will produce a higher accuracy for the majority class than for the minority class. The percentage of an adolescent who consumed drugs is lower than not consume the drug. This condition shows the class imbalance in the data. Imbalanced data is a condition where one of the classes (majority class) contains a much larger number of data than the other classes (minority class) [11].

The imbalanced classes can cause a bias in the majority classes because the data from the minority class tend to be misclassified [11]. Several methods have been proposed to learn from imbalanced data. The data-level approach or resampling methods are the most popular approach to handle imbalanced data [11]. It generates new training datasets to make the class distribution more balanced [12]. Data-level approaches They are divided into over-sampling, under-sampling, and hybrids sampling. SMOTE is one of the most popular over-sampling methods that overcome random over-sampling [13]. It does this by creating "synthetic" data rather than over-sampling with replacement in the minority class. While SMOTE is used for numeric datasets, SMOTE-N is used to handle datasets with nominal features [13]. The

classification of imbalanced data using SMOTE-N and logistic regression was done by [14]. The results show that the model using SMOTE-N has the highest AUC than the model without SMOTE-N, which means that SMOTE-N can increase the model's accuracy.

This study aims to classify imbalanced data on adolescent risk behavior of drug abuse using random forest and combined with SMOTE-N to handle the imbalanced classes in the data, which can help identify the factors that affect adolescent risk behavior of drug abuse.

2 Methods

2.1 Data

The data used in this research were the adolescent risk behaviors based on SKAP (Survei Kinerja dan Akuntabilitas Program KKBPK) of East Java in 2019 by BKKBN.

2.2 Research Variables

The variables used in this research are as follows:

Y_1 = drug consumption (0 = No, 1 = Yes)

X_1 = age (0 = < 19 years old, 1 = $\geq$ 19 years old)

X_2 = sex (0 = Male, 1 = Female)

X_3 = education (0 = did not go to school or has completed either elementary or junior high school, 1 = completed either senior high school or college education)

X_4 = domicile (0 = urban, 1 = rural)

X_5 = knowledge of drugs (0 = no, 1 = yes)

X_6 = knowledge of the physical consequences of the drug (0 = no, one = yes)

X_7 = knowledge of the psychological consequences of the drug (0 = no, one = yes)

X_8 = knowledge of the socioeconomic consequences of the drug (0 = no, one = yes)

X_9 = knowledge of adolescent sexual and reproductive health (ASRH) (0 = no, 1 = yes)

2.2 Research Design

The classification process for imbalanced data of adolescent risk behavior of drug abuse starts with partitioning the adolescent risk behavior of drug abuse into training and testing data using 5-fold cross-validation. Every fold will have the training and testing data. SMOTE-N will be applied in the training data, and then the training data will be classified using random forest. The model obtained will be tested using training data, and the performance of the model will be evaluated using accuracy, sensitivity, specificity, and G-means.

3 Results and Discussion

In this paper, we classify the adolescent risk behavior of drug abuse that has binary classes. Figure 1 shows the class distribution of the drug consumption variable. The total data are 5300, of which the minority class (consuming drug) has 192 data, and the majority class (not consuming drug) has 5108 data.

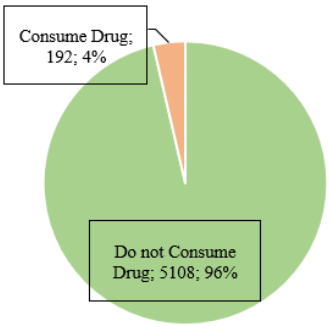

Fig. 1. Percentage of Adolescent Risk Behavior of Drug Abuse

The adolescent risk behavior of drug abuse data is partitioned into 5-fold cross-validation. Every fold will contain about 4240 training data and 1060 testing data. The training data consist of about 4086 data in the majority class and about 154 data in the minority class. In comparison, the testing data will have about 1022 data in the majority class and about 38 data in the minority class. This condition shows that there is a class imbalance in the data. Hence, it needs to handle the imbalanced data using SMOTE-N in the training data. After SMOTE-N is applied, the data in the minority class increases. The data distribution can be seen in Table 1.

Table 1. Class Distribution

	Original Data		**SMOTE-N**	
	Majority	Minority	Majority	Minority
Fold 1	4086	154	4086	4085
Fold 2	4086	153	4086	4086
Fold 3	4086	154	4086	4086
Fold 4	4087	153	4087	4087
Fold 5	4087	154	4087	4086

The performance of random forest without handling the imbalanced data and with handling the imbalanced data using SMOTE-N are shown as below.

Table 2. Performance of Original Data

	Accuracy	**Sensitivity**	**Specificity**	**G-Means**
Fold 1	0.964	0.964	0.000	0.000

	Accuracy	Sensitivity	Specificity	G-Means
Fold 2	0.963	0.963	0.000	0.000
Fold 3	0.964	0.964	0.000	0.000
Fold 4	0.963	0.963	0.000	0.000
Fold 5	0.964	0.964	0.000	0.000
Average	**0.964**	**0.964**	**0.000**	**0.000**

From Table 2, it can be seen that the average accuracy is 96.4%, and the average sensitivity is 96.4%. Although the accuracy and sensitivity are high, the specificity is 0%. Without handling the imbalanced data, the model cannot correctly classify the data in the minority class or adolescents using drugs.

Table 3. Performance of SMOTE-N

	Accuracy	Sensitivity	Specificity	G-Means
Fold 1	0.722	0.995	0.105	0.323
Fold 2	0.687	0.990	0.090	0.298
Fold 3	0.765	0.979	0.083	0.285
Fold 4	0.662	0.990	0.084	0.288
Fold 5	0.683	0.982	0.072	0.266
Average	**0.704**	**0.987**	**0.087**	**0.292**

Table 3 shows the average accuracy of the model with SMOTE-N is 72.5%, or in other words, the overall adolescent risk behavior of drug abuse was correctly classified as much as 72.5%. The model with SMOTE-N obtained high sensitivity and increased specificity and g-means. Thus, the SMOTE-N model can capture the data in minority class or the adolescents who are using drugs.

Table 4. Performance Comparison

	Accuracy	Sensitivity	Specificity	G-Means
Original	0.964	0.964	0.000	0.000
SMOTE-N	0.704	0.987	0.087	0.292

Table 4 shows the performance comparison of random forest without handling the imbalanced data and using SMOTE-N. The accuracy of the model with SMOTE-N is lower than the accuracy of the model with original data, but the former's sensitivity is higher. Using SMOTE-N can increase the model's specificity from 0% to be 8.7% and increase the G-means significantly to 29.2%.

Figure 3 shows the variables importance in the model that uses SMOTE-N. The higher the MDG, the more influential the variable is. Figure 3 shows that the age variable has the highest MDG of 446 compared to the other variables. It means that age is the most crucial variable in determining adolescent risk behavior of drug abuse. The next most important variables are sex and psychological consequence with MDG as much as 304 and 208. The least essential variable is ASRH, with the lowest MDG, which is 40.

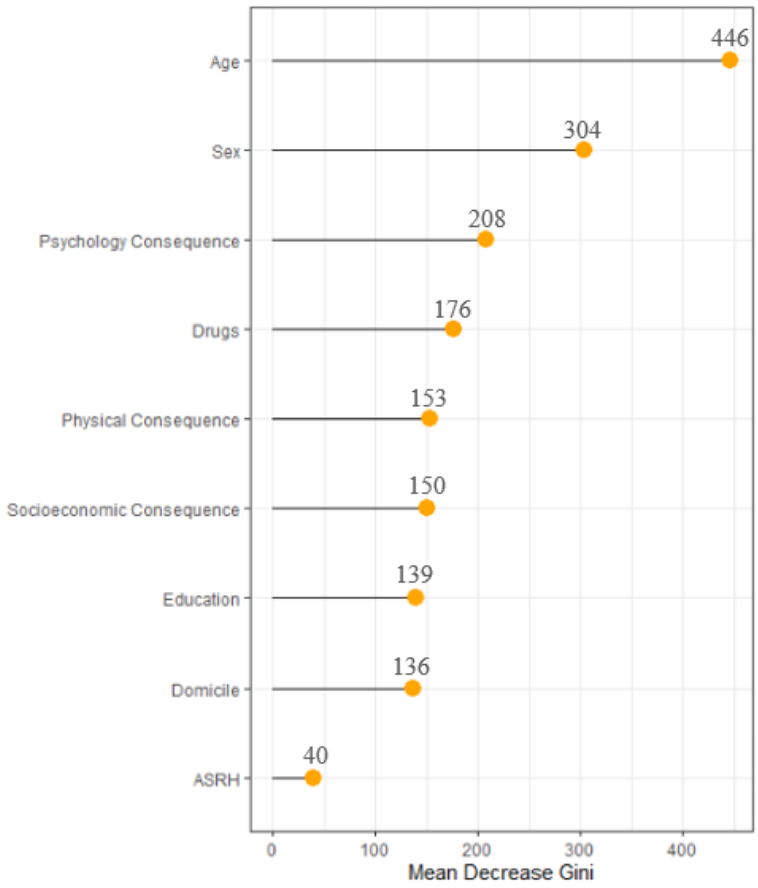

Fig. 2. Mean Decrease Gini

This study shows that age is the most crucial variable in determining adolescent risk behavior of drug abuse. This result agrees with the research done by [15] and [16], which found out that age affects the use of drugs such as narcotics, alcohol, psychotropic substances, and addictive substances. In this study, age is in continuous attribute, referring to research in Iran, which found that adolescents who consume drugs and alcohol are about 17-18 years old. Another research in Indonesia shows that drug abuse behavior tends to be exhibited by adolescents in the group of 20-24 years old rather than the group of 15-19 years old.

The next most crucial variable is sex. This result is congruent with the research done by [17], which found that most people who abuse drugs are male. [18] also showed that gender affects drug abuse and alcoholism. In Brazil, most of the alcoholism are male adolescents. In Indonesia, the same results were found in how gender affects drug abuse, and that male adolescents tend to be more vulnerable to drug abuse [16].

The knowledge of the psychological consequence of drug abuse is the third most crucial variable. The risk behavior faced by adolescents is related to psychology. The knowledge of the psychological consequence of drug abuse is an essential factor in drug abuse behavior [19]. In this study, there are six indicators of psychological consequences: brutal behavior,

delusional behavior, difficulty in concentrating, anxiety, self-harm, and suicidal thoughts. This condition is unexpected and can encourage adolescents to engage in negative behavior, including drug abuse.

4 Conclusion

The original data model has the highest accuracy, 96.4%, but the specificity and the G-means are 0. It shows that the model without handling the imbalanced data cannot capture the data in the minority class. While the SMOTE-N model for handling the imbalance data obtained a lower accuracy, it also has higher sensitivity, specificity, and G-means than the original data model. Hence, the best model is the model with SMOTE-N. From the best model obtained that the variables affect the classfication of the adolescent risk behavior of drug abuse are age, sex, and psychological consequence.

Acknowledgments. This research was funded by BKKBN. The authors thanks to BKKBN for funding which support this research and all individuals associated with this research work.

References

[1] Trimpop RM. Advances in the psychology. Amsterdam: 1994. https://doi.org/10.1111/apps.12164.

[2] Turner C, McClure R, Pirozzo S. Injury and risk-taking behavior - A systematic review. Accid Anal Prev 2004;36:93–101. https://doi.org/10.1016/S0001-4575(02)00131-8.

[3] Zaman M, Razzaq S, Hassan R, Qureshi J, Ijaz H, Hanif M, et al. Drug abuse among the students. Pakistan J Pharm Res 2015;1:41. https://doi.org/10.22200/pjpr.2015141-47.

[4] Srivastava SK, Barmola KC. Psychopathology in Drug and Alcohol Abuse Adolescents. Glob Vis Publ House 2012:3–18.

[5] Maharti VI. Faktor-faktor yang Berhubungan dengan Perilaku Penyalahgunaan Narkoba pada Remaja Usia 15-19 Tahun di Kecamatan Semarang Utara Kota Semarang. J Kesehat Masy 2015;3:946–53.

[6] Firdaus AM yunanta, Hidayati E. Pengetahuan Dan Sikap Remaja Terhadap Penggunaan Napza Di Sekolah Menengah Atas Di Kota Semarang. J Keperawatan Jiwa 2019;6:1. https://doi.org/10.26714/jkj.6.1.2018.1-7.

[7] Sugihartati R, Susilo D. Acts against drugs and narcotics abuse: Measurement of the effectiveness campaign on Indonesian narcotics regulator Instagram. J Drug Alcohol Res 2019;8. https://doi.org/10.4303/jdar/236079.

[8] Han J, Kamber M, Pei J. Data Mining Techniques, Third Edition 2011:847.

[9] Cutler A, Cutler DR, Stevens JR. Ensemble Machine Learning. Ensemble Mach Learn 2012. https://doi.org/10.1007/978-1-4419-9326-7.

[10] Han H, Guo X, Yu H. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. Proc IEEE Int Conf Softw Eng Serv Sci ICSESS 2016;0:219–24. https://doi.org/10.1109/ICSESS.2016.7883053.

[11] Stefanowski J, Wilk S. Selective Pre-processing of Imbalanced Data for. Data Warehous Knowl Discov (Lecture Notes Comput Sci Ser 5182) 2008:283–92. https://doi.org/10.1007/978-3-540-85836-2.

[12] Zhang H, Wang Z. A normal distribution-based over-sampling approach to imbalanced data classification. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2011;7120 LNAI:83–96. https://doi.org/10.1007/978-3-642-25853-4_7.

[13] Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. Deep synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.
[14] Barro RA, Sulvianti ID, Afendi FM. Penerapan Synthetic Minority Oversampling Technique (Smote) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu. Xplore J Stat 2013;1:1–6. https://doi.org/10.29244/xplore.v1i1.12424.
[15] Bajwa HZ, Al-Turki ASA, Dawas AMK, Behbehani MQ, Al-Mutairi AMA, Al-Mahmoud S, et al. Prevalence and factors associated with the use of illicit substances among male university students in Kuwait. Med Princ Pract 2013;22:458–63. https://doi.org/10.1159/000350609.
[16] Nasution SL, Puspitawati H, Rizkillah R, Puspitasari MD. Pengaruh Pengetahuan Remaja tentang NAPZA dan HIV serta Pengetahuan Orang Tua tentang Program Pembangunan Keluarga terhadap Perilaku Penggunaan NAPZA pada Remaja. J Ilmu Kel Dan Konsum 2019;12:100–13. https://doi.org/10.24156/jikk.2019.12.2.100.
[17] Sitorus RJ, Natalia M. Perilaku seksual beresiko penggunaan narkoba risky sexual behavior of narcotic users. Kesmas J Kesehat Masy Nas 2015;9:348–52.
[18] Madruga CS, Laranjeira R, Caetano R, Pinsky I, Zaleski M, Ferri CP. Use of licit and illicit substances among adolescents in Brazil - A national survey. Addict Behav 2012;37:1171–5. https://doi.org/10.1016/j.addbeh.2012.05.008.
[19] Deković M. Risk and protective factors in the development of problem behavior during adolescence. J Youth Adolesc 1999;28:667–85. https://doi.org/10.1023/A:1021635516758.