The Open Public Health Journal

Content list available at: https://openpublichealthjournal.com

RESEARCH ARTICLE

Imbalanced Data Analysis of Adolescent Unintended Pregnancy and Pre-marital Sex using Univariate and Bivariate Random Forest

Ismaini Zain[1,*], Agnes Tuti Rumiati[1], Erma Oktania Permatasari[1] and Nilam Novita Sari[1]

[1]*Department of Statistics, Faculty of Science and Data Analytics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia*

Abstract:

Background:

Unintended pregnancy and pre-marital sex are common in adolescents due to the lack of understanding about Adolescent Sexual and Reproductive Health (ASRH), family planning, and/or birth control.

Aims:

The purpose of this study was to classify unintended pregnancy and pre-marital sex among adolescents in East Java to find the factors behind them.

Methods:

The data used in this research included women of childbearing age based on SKAP by BKKBN in 2019. This study employed SMOTE-NC as a method for handling imbalanced data and random forest as a classification method.

Results:

The result of this study showed that the most influential variable was education and domicile for unintended pregnancy and pre-marital sex, respectively.

Conclusion:

The most influential variable in the classification of unintended pregnancy and pre-marital sex was age was education and domicile for unintended pregnancy and pre-marital sex.

Keywords: Unintended pregnancy, Pre-marital sex, Classification, Imbalanced data, Random forest, Education.

1. INTRODUCTION

Pre-marital sex and unintended pregnancy often happen, especially during adolescence, and this often leads to negative impacts on the individual's physical and mental health. Pre-marital sex is often associated with adolescent sexual and reproductive health (ASRH). Sexual and reproductive health (SRH) is defined as a condition where one's physical, mental and social conditions are completely healthy, including the health of the reproductive system and its functions and processes [1]. ASRH is one of the important aspects that need to be known and understood by an adolescent. Many consequences arise by neglecting ASRH, such as pre-marital sex and early pregnancy for girls, which can be physically risk.

Adolescents face a higher risk of complications and death as a result of pregnancy than older women because their bodies have not reached the physical maturity to cope with the pregnancy [2, 3]. Pregnancies among female adolescents are more likely to be unintended. They do not want to be pregnant, but they do not have contraceptive knowledge [4]. Moreover, it has been observed that a low rate of contraceptive knowledge results in a lack of family planning [5]. One major consequence due to the lack of family planning is unwanted pregnancy. Unintended pregnancies and childbirth may lead to complications, which is the leading cause of death in girls aged 15-19 years. It has been estimated that out of 10.2 million unintended pregnancy, as many as 5.6 million girls undergo abortions, 3.9 million of which are unsafe [5]. Unintended pregnancy in adolescents is estimated to increase, ranging from 150,000 to 200,000 cases every year [6]. In 2017, East Java

* Address correspondence to this author at the Department of Statistics, Faculty of Science and Data Analytics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia; Tel: +628123036678;
E-mails: ismaini_z@statistika.its.ac.id, researchcation.id@gmail.com

had the highest rate of pregnancy cases out of wedlock among junior and senior high school students [7].

The factors for unintended pregnancy have been discussed previously [8], with results showing a significant relationship among unintended pregnancy, age, and education. Meanwhile, concerning pre-marital sex, a study [9] showed that age and education are factors that influence pre-marital sex.

In the SKAP, the percentage of adolescents who had unintended pregnancies in East Java was reported to be 1.84%, and the percentage of adolescents who had pre-marital sex in East Java was 0.4%. An imbalanced condition has been reported while comparing the number of adolescents with and without unintended pregnancy versus the numbers for adolescents who have and did not have pre-marital sex. Hence, the imbalanced data needs to be handled using SMOTE to get unbiased results. SMOTE is a method used for numeric datasets that creates synthetic instances in minority classes [10]. For nominal features, SMOTE can be extended to SMOTE-N. The classification of imbalanced data using SMOTE-N was done by Zain *et al* and Fithriasari *et al*., indicating that SMOTE-N has the best performance because it increases the specificity and G-mean [11, 12]. In this paper, we used SMOTE-NC to handle a mix of continuous and nominal attributes that create synthetic instances in minority classes. In 2021, the classification using the random forest as the base classifier and SMOTE-NC, TL, and SMOTE-NC+TL as the imbalanced methods was done. The results showed that the combination of random forest and SMOTE-NC is better than the combination of random forest and SMOTE-NC+TL; however, it is not better than the combination of random forest and TL for extreme cases in imbalanced data [13, 14].

In this study, unintended pregnancy and pre-marital sex are used as dependent variables. It is different from previous research that has only analyzed unintended pregnancy and pre-marital sex separately. However, in this study, an analysis is carried out using both variables, unintended pregnancy and pre-marital sex, simultaneously to determine the underlying factors using random forest classification.

Random forest is a method that overcomes the limitations of the decision tree. Random forest is a classification method consisting of a collection of trees; all the trees are then voted on and averaged out [15]. The classification process in the random forest begins with breaking the existing sample data into a random decision tree. After the tree is formed, voting will be carried out in each class from the sample data. Next, the votes are collected from each class, and the class with the most votes is selected [16, 17]. By using the random forest in the data classification, it will produce the best performance [17].

Apart from having good performance, random forest is appealing because it can measure the most important variables in predicting the response variable using MDA (Mean Decrease Accuracy) or MDG (Mean Decrease Gini).

The higher the MDG and MDA, the more important the variables are [18].

This study has several aims: (1) to show the imbalance of adolescent behavior data on unwanted pregnancy and premarital sex, (2) to analyze the relationship between unintended pregnancy and pre-marital sex with age, education, domicile, knowledge of contraceptives, and knowledge of ASRH, and (3) to classify unintended pregnancy and pre-marital sex among adolescents in East Java univariately and bivariately using the random forest to find the factors that affect unintended pregnancies and pre-marital sex among adolescents.

2. MATERIALS AND METHODS

2.1. Research Design and Unit of Analysis

This study is a cross-sectional study. The data used is secondary data in the form of raw data from SKAP (Program Performance and Accountability Survey) for East Java in 2019 conducted by BKKBN (National Population and Family Planning Agency). SKAP is a survey conducted by the BKKBN in order to measure the success of population development programs, family planning, and family development.

In this study, women aged 15-18 years were selected, totaling 1815 women of childbearing age and adolescents. The adolescents included in this study were residents of East Java.

2.2. Research Variables

The variables used in this research are as follows:

Y_1 = unintended pregnancy (0 = No, 1 = Yes)

Y_2 = pre-marital sex (0 = No, 1 = Yes)

X_1 = age

X_2 = 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_3 = domicile (0 = rural, 1 = urban)

X_4 = knowledge of contraceptives (0 = no, 1 = yes)

X_5 = knowledge of adolescent sexual and reproductive health (ASRH) (0 = no, 1 = yes)

2.3. Analysis of Statistics

Age, education, domicile, knowledge of contraceptives, and ASRH were used to determine the adolescent's unintended pregnancy and pre-marital sexual behavior. Chi-square analysis was used to measure the relationship among unintended pregnancy and pre-marital sex with age, education, domicile, knowledge of contraceptives, and knowledge of ASRH. When the relationship between the two variables was stronger, the deviation between the observed and expected frequencies was higher, indicating a stronger relationship between the two variables [19].

After that, SMOTE-NC was applied to handle the imbalanced class in the data and make the data class balanced by creating "synthetic" examples rather than over-sampling with replacement [10]. Random forest was then used to classify unintended pregnancy and pre-marital sex among adolescents in East Java univariately and bivariately using random forest to find the factors that affect unintended pregnancy and pre-

marital sex among adolescents. Random forest classification in this study used parameter tuning with the m parameter tuned to 2 and the number of trees being 50, 100, 250, and 500 trees, respectively, to make several models. Accuracy (overall accuracy), macro sensitivity, macro specificity, and macro G-means (geometric mean of sensitivity and specificity) were used to evaluate the model. Higher accuracy, macro sensitivity, macro specificity, and macro G-means indicate a better model. Afterward, the models were compared to get the best model. The best model was used to classify the data and was interpreted using MDA to find the important variables in classifying unintended pregnancy and pre-marital sex. The higher the MDA, the more important the variable was in classifying unintended pregnancy and pre-marital sex.

3. RESULTS AND DISCUSSION

3.1. Imbalance Data of Adolescent Behavior Regarding Unintended Pregnancy and Premarital Sex

The class distribution of the unintended pregnancy variable and pre-marital sex is shown in Fig. (**1**). It shows that the number of women who had unintended pregnancies was 34, while the number of women who did not have unintended pregnancies was 1815, and the number of women who had pre-marital sex was 102 out of 1849.
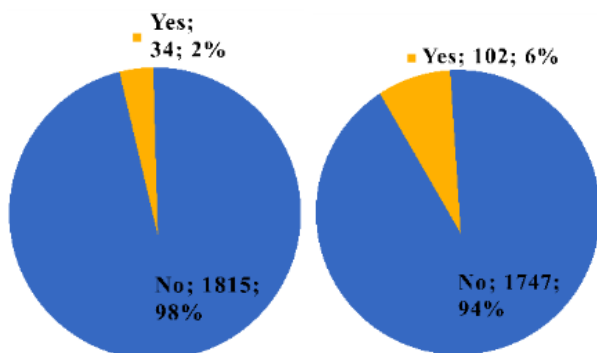

Fig. (1). Percentage of dependent variable.
a) Percentage of Unintended Pregnancy; b) Percentage of Pre-marital Sex

3.2. The Relationship between Dependent Variables (Unintended Pregnancy and Pre-marital Sex) and Independent Variables

The Chi-Square test of independence was used to determine whether unintended pregnancy and pre-marital sex were independent of the variables age, education, domicile, knowledge of contraceptives, and knowledge of ASRH, or whether there was a dependence between them. If there was a dependence, then the two variables would have a statistical relationship with each other.

Table **1** shows the results of the Chi-Square test between dependent variables and independent variables. Age, education, and domicile showed a p-value less than significant level, indicating a statistically significant relationship between these variables with unintended pregnancy and pre-marital sex.

3.3. Classification of Unintended Pregnancy and Pre-marital Sex

Random forest was used to classify unintended pregnancy and pre-marital sex to make a model. The model was then used to predict unintended pregnancy and pre-marital sex among adolescents. Using random forest classification, the factors that influence unintended pregnancy and pre-marital sex among adolescents were determined. Following are the results of the random forest classification.

Table 1. Chi-square test.

	Chi-Square	df	p-value	Sig.
Age	15.638	3	0.001	*
Education	38.149	3	0.000	*
Domicile	8.601	3	0.035	*
Knowledge of contraceptives	2.890	3	0.409	-
Knowledge of adolescent sexual and reproductive health (ASRH)	1.032	3	0.793	-

Note: *: Significant at the level of 5%; -: not significant

Table **2** shows the classification performance of unintended pregnancy with the original data and when SMOTE-NC was applied with varying trees. The accuracy obtained by the model using the original was reported to be 0.9816 for all trees, indicating that the overall unintended pregnancy variable was classified with 98.16% precision. The macro sensitivity for all trees was 0.9816, indicating that the model could classify women who did not have unintended pregnancies by 98.16%. However, the model could not classify women who had unintended pregnancies because the macro specificity was reported to be NA. NA value obtained by macro specificity and macro G-Means indicated that there is an imbalanced problem in the unintended pregnancy data. Thus, the imbalanced data needs to be handled using SMOTE-NC.

Table 2. Classification performance of unintended pregnancy.

	Accuracy	Macro Sensitivity	Macro Specificity	Macro G-Means	N tree	M try
Original	0.9816	0.9816	NA	NA	50	2
	0.9816	0.9816	NA	NA	100	2
	0.9816	0.9816	NA	NA	250	2
	0.9816	0.9816	NA	NA	500	2
SMOTE-NC	0.8177	0.9961	0.0770	0.2759	50	2
	0.8259	**0.9974**	**0.0878**	**0.2955**	100	2
	0.8248	0.9962	0.0820	0.2855	250	2
	0.8285	0.9962	0.0839	0.2884	500	2

Abbreviation*: NA: Not applicable

The classification performance of unintended pregnancy using SMOTE-NC showed a decrease in accuracy. Although it showed decreased accuracy, the macro specificity and macro G-means increased. Moreover, it was found to have the highest macro sensitivity. It means that the imbalanced problem in the data has been resolved. The model using SMOTE-NC showed better results than the model using original data. The highest accuracy was produced by model SMOTE-NC with 500 trees and the highest macro sensitivity, macro specificity, and macro G-means were reported with 100 trees.

Table **3** shows the classification performance of pre-

marital sex using the original data and SMOTE-NC. The overall pre-marital sex variable was classified with 94.48% precision using the model with original data. Moreover, the model with original data classified women who did not have pre-marital sex by 94.48%. Macro specificity and macro G-means of the model produced NA value, indicating the presence of an imbalanced problem in pre-marital sex data.

Table 3. Classification performance of pre-marital sex.

	Accuracy	Macro Sensitivity	Macro Specificity	Macro G-Means	N tree	M try
Original	0.9448	0.9448	NA	NA	50	2
	0.9448	0.9448	NA	NA	100	2
	0.9448	0.9448	NA	NA	250	2
	0.9448	0.9448	NA	NA	500	2
SMOTE-NC	0.6388	**0.9675**	**0.0952**	**0.3012**	50	2
	0.6371	0.9611	0.0849	0.2846	100	2
	0.6426	0.9646	0.0912	0.2949	250	2
	0.6431	0.9658	0.0919	0.2963	500	2

Abbreviation: NA: Not applicable

The classification performance of pre-marital sex using SMOTE-NC showed that the accuracy was slightly decreased, but the macro specificity and macro G-means increased, of which its macro sensitivity was found to be the highest, indicating that the model using SMOTE-NC was better than the model using original data. Model SMOTE-NC using 500 trees also has the highest accuracy and using 50 trees obtained the highest macro sensitivity, macro specificity, and macro G-means.

Table **4** shows the classification performance of unintended pregnancy and pre-marital sex using the multilabel random forest with the original data and when SMOTE-NC was applied. The classification using original data produced an accuracy of 0.9292, which indicated that the model could classify the overall unintended pregnancy and pre-marital sex with 92.92% accuracy. However, macro sensitivity, macro specificity, and macro G-means showed NA value. It means that the model produced using original data could not predict the data in each class, indicating an imbalance problem in the data, which could be resolved using SMOTE-NC.

Table 4. Classification performance of unintended pregnancy and pre-marital sex.

	Accuracy	Macro Sensitivity	Macro Specificity	Macro G-Means	N tree	M try
Original	0.9292	NA	NA	NA	50	2
	0.9292	NA	NA	NA	100	2
	0.9292	NA	NA	NA	250	2
	0.9292	NA	NA	NA	500	2
SMOTE-NC	0.5165	**0.3092**	**0.7716**	**0.2934**	50	2
	0.5154	0.3024	0.7713	0.2822	100	2
	0.5095	0.3013	0.7705	0.2799	250	2
	0.5198	0.2996	0.7714	0.2766	500	2

Abbreviation: NA: Not Applicable

The performance of the multilabel random forest using SMOTE-NC showed decreased accuracy, but the macro sensitivity, macro specificity, and macro G-means increased. It means that the imbalanced problem in the data has been resolved. The random forest model using SMOTE-NC produced the highest accuracy with the number of trees at 500 while the model using 50 trees showed the highest macro sensitivity, macro specificity, and macro G-means.
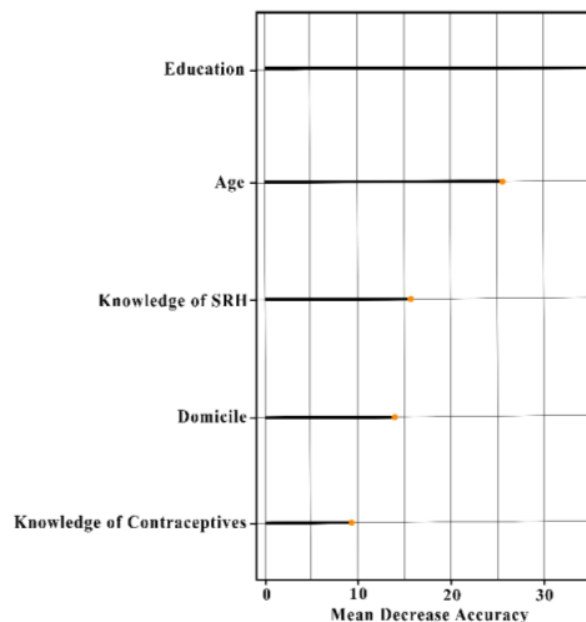
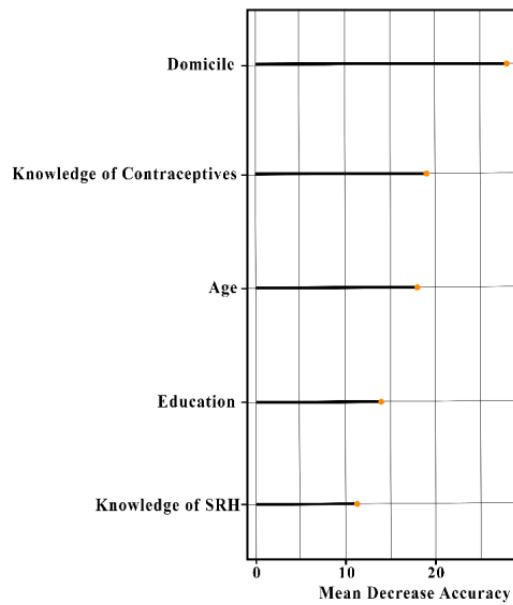**Fig. (2).** Variable importance of unintended pregnancy

Fig. (3). Variable importance of pre-marital sex

Fig. (**2**) shows the importance of each variable for unintended pregnancy using SMOTE-NC with 100 trees. The higher the MDA, the more important the variable is. In this case, it is education. The second most influential variable is age. This is followed by knowledge of SRH and domicile, while the least influential variable is knowledge of contraceptives. This means that knowledge of contraceptives is a variable that has the least influence over the other predictor variables.

The research by Diasanty and Sutiawan showed the same result. Education and age were reported to have a significant effect on the incidence of unintended pregnancy in adolescents [20]. Adolescents aged less than 20 years were found to be more at risk of experiencing unintended pregnancy compared to adolescents greater than 20 years [21]. Meanwhile, Mardiyah *et al*. showed that the average age for unintended pregnancy in adolescents was between 15-19 years [6].

Fig. (**3**) shows the importance of each variable of pre-marital sex using SMOTE-NC with 50 trees. The highest MDA is domicile, which means that domicile has the greatest influence over the other predictor variables. This is followed by knowledge of contraceptives, age, education, and knowledge of SRH. This means knowledge of SRH is a variable that has the least influence over the other predictor variables.

Fig. (4). Variable importance of unintended pregnancy and pre-marital sex.

The results obtained are in line with the results of research conducted by Sari *et al*., which shows that education, knowledge of contraceptives, knowledge of adolescent sexual RH, and domicile are the factors that have significant relationships to pre-marital sex behavior in adolescents [22]. Research in Malang City found that adolescents in urban areas are more likely to be at risk of pre-marital sex behavior than in rural areas [23]. People in urban areas tend to be more permissive regarding dating behavior. However, this finding contradicts research conducted by Rusmiati and Hastono, which shows that the area where the adolescents live does not significantly affect pre-marital sexual behavior in them [24, 25]. Adolescents who have better knowledge of contraceptives have a lower risk of pre-marital sex than adolescents who do not [22].

Fig. (**4**) shows the importance of each variable for unintended pregnancy and pre-marital sex using SMOTE-NC with 50 trees. The age variable has the highest MDA compared to other variables. It means that age is the most influential variable in the classification of unintended pregnancy and pre-marital sex. The second most influential variable that affects the classification of unintended pregnancy and pre-marital sex is education. This is followed by domicile and knowledge of contraceptives. Lastly, the variable that has the least effect on classification of unintended pregnancy and pre-marital sex is knowledge of SRH.

CONCLUSION

Based on the analysis above, it can be concluded that there is a significant relationship between unintended pregnancy and pre-marital sex with age, education, and domicile variable. Besides that, the classification using original data produced high accuracy, but macro specificity and macro G-mean produced NA value, while the model using SMOTE-NC produced lower accuracy but higher macro specificity, macro sensitivity, and macro G-means, which resolved the problem of imbalanced data. In terms of importance, education and domicile were found to be the most important variables of unintended pregnancy and pre-marital sex, respectively. Meanwhile, the most important variable in the classification of unintended pregnancy and pre-marital sex was reported to be age and education.

LIST OF ABBREVIATIONS

ASRH = Adolescent Sexual and Reproductive Health

SRH = Sexual and Reproductive Health

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used in the studies that are the basis of this research.

CONSENT FOR PUBLICATION

Informed consent was obtained.

STANDARDS OF REPORTING

STROBE guidelines were followed.

AVAILABILITY OF DATA AND MATERIALS

Not applicable.

FUNDING

CONFLICT OF INTEREST

The authors declare no conflicts of interest, financial or otherwise.

ACKNOWLEDGEMENTS

REFERENCES

[1]   UNFPA. Population and Development, I: Program of Action. Cairo: Int Conf Popul Dev 1994; pp. 1689-99.

[2]   Morris JL, Rushwan H. Adolescent sexual and reproductive health: The global challenges. Int J Gynaecol Obstet 2015; 131(Suppl. 1): S40-2.
[http://dx.doi.org/10.1016/j.ijgo.2015.02.006] [PMID: 26433504]

[3]   Tebeu PM, Fomulu JN, Khaddaj S, de Bernis L, Delvaux T, Rochat CH. Risk factors for obstetric fistula: a clinical review. Int Urogynecol J Pelvic Floor Dysfunct 2012; 23(4): 387-94.
[http://dx.doi.org/10.1007/s00192-011-1622-x] [PMID: 22143450]

[4]   Blanc AK, Tsui AO, Croft TN, Trevitt JL. Patterns and trends in adolescents' contraceptive use and discontinuation in developing countries and comparisons with adult women. Int Perspect Sex Reprod Health 2009; 35(2): 063-71.
[http://dx.doi.org/10.1363/3506309] [PMID: 19620090]

[5]   Singh Susheela, Darroch Jacqueline E, Darroch SS. Adding it up: costs and benefits of contraceptive services estimates for 2012. Guttmacher Inst United Nations Popul Fund (UNFPA) 2012.

[6]   Lopez LM, Grey TW, Hiller JE, Chen M. Education for contraceptive use by women after childbirth. Cochrane Database Syst Rev 2015; 2015: CD001863.
[http://dx.doi.org/10.1002/14651858.CD001863.pub4]

[7]   Mardiyah S, Pamungkas CE, Juliana R. Prevention of unwanted pregnancy (Ktd) in adolescent women who are sexually active in the work area of kerembong poskesdes, lombok central. J Kebidanan 2019; 4: 40-2.

[8]   Ardiansyah M. East Java has the highest cases of pregnant students outside of marriage. 2016. Available from: https://www.merdeka.com/peristiwa/jawa-timur-paling-tinggi-kasus-hamil-luar-nikah.html

[9]   Afkhamzadeh A, Rahmani K, Felehgary M, Farhadifar F, Faraji O. Risk factors for unintended pregnancy in women: A nested case-control study. Shiraz E Med J 2020; 2020: 21.

[10]   Mai V, Kittisuksathit S. Factors influencing pre-marital sexual intercourse among unmarried young individuals in Cambodia. Makara J Health Res 2019; 23: 143-9.
[http://dx.doi.org/10.7454/msk.v23i3.1157]

[11]   Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Deep synthetic minority over-sampling technique. J Artif Intell Res 2002; 16: 321-57.
[http://dx.doi.org/10.1613/jair.953]

[12]   Zain I, Fithiasari K, Permatasari E, *et al*. Imbalanced data analysis of adolescent risk behavior of drug abuse using random forest. In: AECon 2020, December 19-20. Purwokerto, Indonesia 2021.
[http://dx.doi.org/10.4108/eai.19-12-2020.2309145]

[13]   Fithriasari K, Hariastuti I, Wening KS. Handling imbalance data in classification model with nominal predictors. Int J Comput Sci Appl Mathe 2020; 6(1): 33.
[http://dx.doi.org/10.12962/j24775401.v6i1.6643]

[14]   Sain H, Purnami SW. Combine sampling support vector machine for

imbalanced data classification. Proc Comput Sci 2015; 72: 59-66.
[http://dx.doi.org/10.1016/j.procs.2015.12.105]

[15] Sari NN, Zain I, Fithriasari K, Muhaimin A. ABR. BR+ for addressing imbalanced multilabel data classification combined with resampling technique. In: Proceedings of The 6th Asia-Pacific Education And Science Conference, AECon,19-20 December 2020. Purwokerto, Indonesia 2021.

[16] Breiman L. Random forests. Mach Learn 2001; 45(1): 5-32.
[http://dx.doi.org/10.1023/A:1010933404324]

[17] Ali J, Khan R, Ahmad N, Maqsood I. Random Forests and Decision Trees. Int J Comput Sci Issues 2012; 9: 272-8.

[18] Biau G. Analysis of a Random Forests Model. J Mach Learn 2012; 13: 1063-95.

[19] Han H, Guo X, Yu H. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. 7th IEEE International Conference on Software Engineering and Service Science (ICSESS). 26-28 August 2016; Beijing. 2016; pp. 219-4.

[20] Heumann C, Schomaker M. Introduction to Statistics and Data Analysis. Switzerland: Springer 2016.
[http://dx.doi.org/10.1007/978-3-319-46162-5]

[21] Diasanty N, Sutiawan R. Contraceptive failure with unwanted pregnancies (KTD) in high-risk aged women in Indonesia (advanced analysis of 2012 IDHS data) = Contraceptive failure with unwanted pregnancy on high-risk aged women in Indonesia analysis of IDHS 2012. In: Fkm Ui. 2014; 2014: p. 18.

[22] Febriana F, Sari LK. Factors affecting unwanted pregnancy in Indonesia, 2017. Semin Nat Official Stat 2021; 2020(1): 1041-51.
[http://dx.doi.org/10.34123/semnasoffstat.v2020i1.592]

[23] Sari IP, Nasution SL, Lailatul Alfiah L. Factors affecting premarital sexual behavior in adolescents in south sumatra. Jurnal Ilmu Kesehatan Masyarakat 2022; 13(1): 50-61.
[http://dx.doi.org/10.26553/jikm.2022.13.1.50-61]

[24] Anas RH. The surveillance of pre-marital sexual behavior in malang. Heal Nations 2018; 2

[25] Rusmiati D, Hastono SP. Adolescent attitudes toward virginity and sexual behavior in dating. Nat Public Health J 2015; 10(1): 29.
[http://dx.doi.org/10.21109/kesmas.v10i1.815]