Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 4 October 2024, pp: 1320-1329
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Predictive Model Comparison for Predicting Condom Use: Comparison of Conventional Logistic Regression and Other Machine Learning

Fadhaa Aditya Kautsar Murti

Department of Public Health, Universitas Indonesia, Indonesia

E-Mail: fadhaa.aditya@ui.ac.id

Received Jun 25th 2024; Revised Jul 15th 2024; Accepted Jul 22th 2024
Corresponding Author: Fadhaa Aditya Kautsar Murti

Abstract

Condom use at first sex remains an important issue as it shapes future sexual behavior. This study aimed to deploy and predict condom use using five different machine learning classification models. Dataset used for this study was from Indonesian Demographic and Health Survey (IDHS) 2017 with a population of interest was male adolescents. We evaluated five different models, namely logistic regression, naïve bayes, K-Nearest Neighbors, support vector machines, and decision tree. Performances of each model were assessed using metrics such as accuracy, specificity, sensitivity, ROC Curve, and AUC Score. Study found that different models exhibit different accuracy, specificity, sensitivity, ROC Curve, and AUC Score. The decision tree and naïve bayes models remained the models with the highest specificity and sensitivity, however the KNN model expressed the highest AUC score. Result from the conventional logistic regression also explained that condom use was associated with education level, age at first sex, and attitude towards condom use. The government is advised to create equal education opportunities for every adolescent and shape better knowledge and condom attitudes. Future studies are advised to enhance the performance of machine learning models using hyperparameter tuning and other methods.

Keywords: Condom Use, Demographic Health Survey, First Sex, Indonesian Adolescents, Machine Learning

1. INTRODUCTION

The reproductive health domain in Indonesia remains a significant issue. One of the main problems in this domain is the high number of HIV/AIDS cases and other Sexually Transmitted Infections (STIs). The latest prevalence of HIV showed 540,000 people living with HIV/AIDS (PLWHA) in 2022 [1]. This problem is exacerbated by the age of PLWHA in Indonesia, with a report from the first quarter of 2023 stating that more than 20% of new HIV cases are among adolescents [2].

It is well known that the prevalence and incidence of such issues are influenced by the consistency of condom use among adolescents. This is evident from the low number of adolescents who use condoms during their first sexual encounter [3]. Data from 2017 showed that less than 30% of adolescents used a condom during their first sexual experience [4]. It is, therefore, important to understand why condom use remains very low, given that condoms can reduce the probability of HIV infection by 90% [5].

Previous studies have explored why adolescents do not use condoms, citing factors such as low education levels, early sexual debut, peer pressure, and the nature of the partner at first sex [6], [7], [8], [9]. However, these studies primarily used conventional logistic regression to investigate the reasons for not using condoms. As statistical modeling and prediction methods have advanced, the application of machine learning in public health has increased. Previous studies have addressed the problem of unmet needs in HIV prevention using machine learning, particularly among Indian and Chinese populations [10], [11] .

However, the application of machine learning in Indonesia is relatively new. To our knowledge, other prediction methods besides conventional logistic regression have not been used to predict condom use in Indonesia. Therefore, this research aims to fill this gap by predicting condom use using machine learning methods, in comparison to conventional logistic regression. The novelty of using alternative prediction models is undoubtedly important, as it provides new perspectives for solving public health problems. This research will explore possible prediction models to achieve better population health outcomes, serving as a pioneer in public health technology.

The introduction section of this article explains why addressing the condom use issue is important and highlights the research gap: the unexplored alternative prediction methods for condom use compared to conventional logistic regression. The article then describes the materials and methods, including the dataset used and the statistical methods employed to achieve the research objectives. The results section explains the performance of various prediction models, including their evaluation metrics from sensitivity to AUC score. Among all prediction models, the best model will extract its top features to identify the most influential factors in predicting condom use. Fourthly, the discussion section addresses the results of this research in the socioeconomic context of Indonesian adolescents according to previous studies. Finally, this article concludes with recommendations for public health practitioners and statisticians for further prediction modeling.

2. MATERIALS AND METHOD

2.1. Study Population and Dataset Used

The dataset used in this study was from the Indonesian Demographic and Health Survey (IDHS) 2017. Population for this study was unmarried male adolescents in 2017. Inclusion criteria included adolescents who had sex. A total of 13,079 adolescents were collected and interviewed by the DHS Program during the data collection process, however a total of 8% adolescents admitted to having sex before and a total of 771 adolescents were analyzed due to missing data. IDHS used a complex sample design during the survey period. The sample selection method is as figure 1.

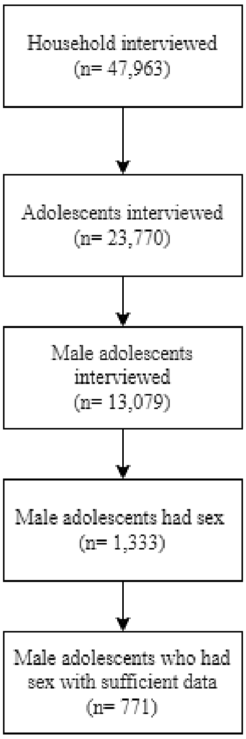

Figure 1. Sample Selection Method

2.2. Variables or Features

Dependent variable for this research was condom use at first sex, asked by a question, *"The first time you had sexual intercourse, did you or your partner use anything to prevent a pregnancy? What did you or your partner use?"*. Such a question included several options including condom, withdrawal, pill, emergency contraception, calendar method, and others. Adolescents were allowed to choose multiple answers. Those who answered with condom (along with other methods or not) were categorized as yes, elsewhere were classified as no.

Independent variables included intrapersonal, interpersonal, and environmental factors. Environmental factors included media exposure (television or TV, radio, and newspaper). Intrapersonal factors were age at first sex, partner at first sex, and peer influence. Partners at first sex were dichotomized as partner (synonymously as lover) and non-partner. Peer influence was whether the adolescent had a friend who already had sex or not.

Interpersonal factors were level of education, types of residence, HIV knowledge, and attitude towards condom use. HIV knowledge was a composite variable, derived from asking seven questions (including myths and truths) about HIV; higher scores corresponded to good knowledge of HIV. In contrast, attitude towards condom use was a composite variable from three questions (condom can be used to prevent pregnancy, condom can protect against HIV-AIDS and STIs, and condom can be reused). For each wrong answers, respondents received one score and therefore, higher scores represented poor attitude. Age at first sex, attitude towards condom use, and HIV knowledge were numerical variables. Variables chosen was similar to previous study [6], [12], [13].

2.3. Methodology

Bivariate and multivariate analysis was conducted using logistic regression. A p-value < 0.05 was considered statistically significant. Crude and adjusted odds ratio was provided to see the association between the independent and dependent variables. Since IDHS used two stages of sample selection and complex sample design, weight, strata, and cluster were taken into account during data analysis. Machine learning models in this study included logistic regression, decision tree, naive bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Reasons for why such models were chosen included that those models showed good performance in predicting risk behavior in the previous study [14]. Such models are also known for classification analysis.

Before analysis took place, data was pre-processed. Missing values were eliminated and so were adolescents who responded to some questions with a response of *don't know.* Pre-processed data was also split into 75% training dataset and 25% testing dataset. Dependent variable in the training dataset was also weighted using the SMOTE library to balance the disproportionate distribution of condom use so that the test dataset can perform very well. Before conducting bivariate analysis, weight was put into the coding and before building the model, numerical variables were standardized. Statistical analysis was done using a statistical software called STATA and machine learning models were built using Python 3.0. A comparison matrix of different machine learning models was assessed using accuracy, sensitivity, specificity, and AUC. Figure 2 explained the details of machine learning models. Each of machine learning classification models are explained in the 2.3 sub-section figure 2.

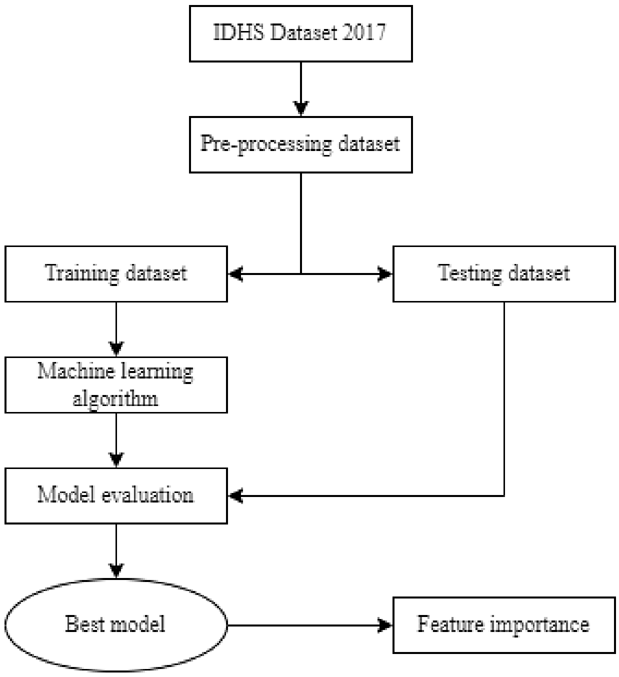

Figure 2. Machine learning flowchart

2.3.1. Logistic Regression

Logistic regression is a machine learning model to classify binary categories. The sigmoid classifier helps to understand the categorization of the outcome variables. The sigmoid function, σ(z), lies between 0—1. Mathematically, it can be expressed in the following equation (1) below.

$$\sigma(z) = \frac{1}{1 + e^{z}} \tag{1}$$

The Z notation in σ(z) is the multiplication of weight (wi) of each features (xi). Weight tells the model of how important the feature in the model is. The sigmoid function, also called logistic function, converts real values into probability. Since it predicts binary cases, often coded as 0 and 1, the probability for both categories equals to one [15].

2.3.2. Decision Tree

A decision tree is a way to make decisions based on questions and answers. It starts with a special node called the "root," which has no incoming connections. All other nodes have one incoming connection. There are two main types of nodes in a decision tree: internal nodes and leaf nodes. Internal nodes, or test nodes, have one or more outgoing connections leading to other nodes. Leaf nodes, also known as terminal or decision nodes, have no outgoing connections. Gini impurity is often used to address the method and the number of split in the trees, expressed by the equation (2) below.

$$\text{Gini(t)} = 1 - \sum_{j} [p(j|t)]^2 \tag{2}$$

The probability of $(p(j|t))^2$ explains how good a split is by evaluating how mixed the classes are in the resulting groups from the split. A lower Gini score indicates a purer split, meaning the classes are less mixed. The model selects the best split by choosing the one with the lowest Gini impurity compared to other potential splits. While decision tree can be used for both numeric and binary outcomes, in this research, decision tree focuses more on binary outcomes [16].

2.3.3. K-Nearest Neighbors (KNN)

KNN is one of the classification methods used in machine learning algorithms. It assigns the object according to another labeled object with the highest similarity. One priority in assigning an object is measuring the position of the object relative to the other labeled object. Various techniques are used, namely Manhattan distance and Euclidean distance. However, the most popular one is Euclidean distance. The equation of Euclidean distance is expressed by the equation (3) below.

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \tag{3}$$

In equation (3), the p and q represent the comparison of two objects, relatively, according to the n characteristics. Another priority in KNN is to determine the number of k, a concept to determine how many neighbors an object should be influenced by. Various techniques have been developed to address the appropriate number of k and one of the most popular calculations is that k equals to the square root of the number of observations found in training dataset [17], [18].

2.3.4. Naïve Bayes (NB)

NB is one of the classification methods where it lies on the probability of one event based on the probability of another event. Suppose that the probability of Y after X is determined by the function $f: X \rightarrow Y$ or $P(Y|X)$ where X can be a vector with n-number of attributes. Therefore, with the concept of NB, the calculation of $P(Y_i|X)$ can be described by the equation (4) below.

$$P(Y = y_i|X = x_k) = \frac{P(X = x_k|Y = y_i)\, P(Y = y_i)}{\sum_j P\left(X = x_k \middle| Y = y_j\right) P\left(Y = y_j\right)} \tag{4}$$

In each $P(Y|X)$, i and j denote the number of possibilities for Y and k refers to the number of possibilities of x. The NB assumes that each event in the $P(Y|X)$ is independently unrelated from each other [19].

2.3.5. Support Vector Classifier (SVM)

Support Vector Machines (SVM) are a type of discriminant technique. In classification tasks, discriminant machine learning techniques aim to find a discriminant function from an independent and identically distributed (iid) training dataset that accurately predicts labels for new instances. Geometrically, learning a classifier involves finding the best multidimensional surface that separates classes in the feature space. It separates the number by 2^n according to the theory of Vapnik-Chervonenkis (VC) and the VC capacity is equal to the number of training points N that the model can distinguish into 2^n unique labels. SVM can be described by the equation (5) below.

$$h_{(w,b)}(x) = g(w^T x + b) \tag{5}$$

In equation (5), w, b denotes the discriminant linear classifier and the $g(w^T x + b)$ is modelled from the probability that $P(y = 1|x;\, \theta)$. It translates into hyperplane that shapes and separates vectors into its classification result [20], [21].

3. RESULTS AND DISCUSSION

3.1. Conventional Logistic Regression

Table 1. Conventional Logistic Regression to Predict Condom Use

Variables	Condom Use		cOR (95 CI)	AOR (95 CI)
	Yes	No		
Place of Residence				
Rural	35%	65%	0.53 (0.34–0.84)	0.66 (0.40–1.08)
Urban	23%	77%	Ref	Ref
Education				
Primary–junior high school	18%	82%	0.23 (0.12–0.46)	0.31 (0.15–0.63)**
High school	27%	73%	0.40 (0.24–0.67)	0.48 (0.27–0.83)**
Diploma–university degree	48%	52%	Ref	Ref
Age at first sex (17.47 ± 0.09)			1.16 (1.04–1.28)	1.12 (1.00–1.26)*
Peer Influence				
Don’t have/know friends who had sex	30%	70%	0.71 (0.14–3.67)	1.30 (0.21–8.05)
Have/know friends who had sex	23%	77%	Ref	Ref
Attitude towards Condom Use (3.22 ± 0.02)			0.43 (0.25–0.76)	0.52 (0.28–0.94)*
Partner at First Sex				
Partner	29%	71%	0.69 (0.36–1.30)	0.61 (0.30–1.22)
Non-partner	37%	63%	Ref	Ref
Newspaper Reading Frequency				
At least once a week	37%	63%	Ref	Ref
Less than once a week	29%	71%	0.68 (0.40–1.16)	0.94 (0.52–1.69)
Not at all	26%	74%	0.61 (0.34–1.09)	1.16 (0.59–2.27)
Radio Listening Frequency				
At least once a week	37%	63%	Ref	Ref
Less than once a week	33%	67%	0.80 (0.45–1.41)	0.81 (0.45–1.47)
Not at all	24%	76%	0.51 (0.28–0.92)	0.56 (0.30–1.04)
Television Watching Frequency				
At least once a week	29%	71%	Ref	Ref
Less than once a week	34%	66%	1.12 (0.72–2.20)	1.13 (0.45–1.47)
Not at all	10%	90%	0.26 (0.03–1.99)	0.22 (0.02–2.04)
HIV Knowledge (5.51 ± 0.05)			1.32 (1.08–1.62)	1.13 (0.928–1.39)

*significant for p<0.05
**significant for p<0.01
***significant for p <0.001

Logistic regression expressed a significant association between education, age at first sex, and attitude towards condom use to condom use, holding all variables constant. Lower odds of condom use were shown among primary to junior adolescents (AOR = 0.31; 95% CI = 0.15 – 0.63) and high school adolescents (AOR = 0.48; 95% CI = 0.27 – 0.83) compared to adolescents with a diploma or university degree. Similarly, lower odds of condom use decreased as the attitude towards condom use score increased. Since a high score on attitude towards condom use represented a poor attitude, adolescents with a poor attitude were less likely to use a condom at first sex. In addition, adolescents who had first sex at older age showed higher odds compared to those who had first sex at younger age (AOR = 1.12; 95% CI = 1.00–1.26)

Although other variables were not statistically significant, our model found some interesting risk measurements. Adolescents with a higher HIV knowledge score showed higher odds of using a condom at first sex. Since a higher HIV knowledge score translated to better knowledge, this implied that good HIV knowledge aligned with good sexual practices. Adolescents in rural areas showed lower odds of condom use (AOR = 0.66; 95% CI = 0.40–1.08), and so did adolescents who had their first sex with their partner (AOR = 0.61; 95% CI = 0.30–1.22). Regarding radio listening and television watching frequency, adolescents who did not listen to the radio (AOR = 0.56; 95% CI = 0.30–1.04) and did not watch television at all showed lower odds of using a condom (AOR = 0.22; 95% CI = 0.02–2.04). Inconsistency was found in the variable peer influence, where the crude and adjusted odds ratios showed a significant spike that crossed the 1.00 threshold.

3.2. Developing Machine Learning Model

Predictive model of condom use was developed using five models of machine learning, namely logistic regression (LR), decision tree (DT), naive bayes (NB), support vector machine (SVM), and K-Nearest Neighbors (KNN). Data was split into 75% training dataset and 25% testing dataset.

When identifying condom use, the NB model showed a recall score of 93%, suggesting a high specificity performance, although the model fails to correctly identify adolescents who did not use condom at first sex (recall score= 9%). In regards to identifying non-condom use, the DT model showed the highest recall score of 71%. Both LR and SVM showed similar performance in both identifying non-condom use (recall score= 60% for both models) and condom use (recall score= 59% for LR and 57% for SVM). In addition, the recall score for KNN was 46% and 77% in identifying non-condom use and condom use, respectively.

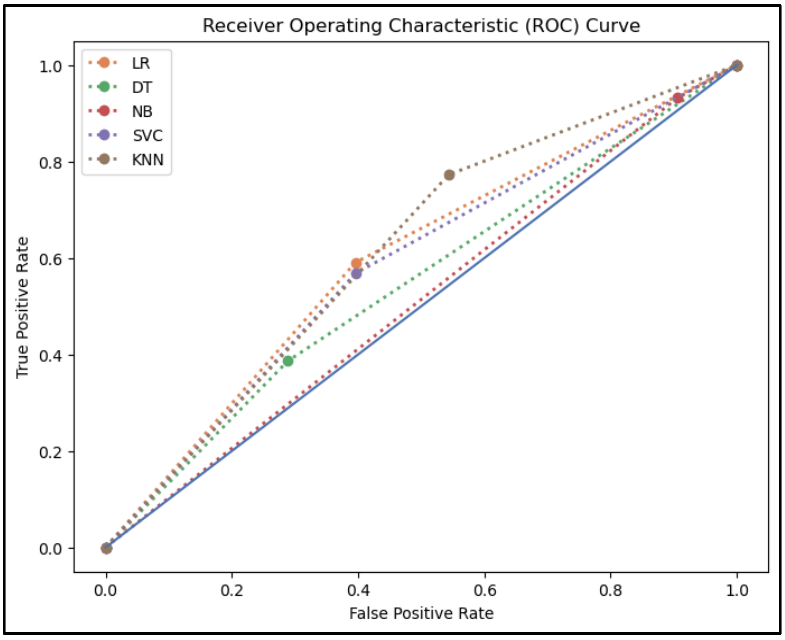

Figure 3. ROC Comparison for each machine learning model

Table 2. Baseline comparison for each machine learning model

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.601	0.590	0.604	0.597
Decision Tree	0.637	0.386	0.711	0.549
Naive Bayes	0.284	0.931	0.094	0.513
SVC	0.595	0.568	0.604	0.586
K-Nearest Neighbors	0.528	0.772	0.456	0.615

In the base model of our machine learning algorithm, naive bayes remained as the most sensitive model with low overall performance, contrasting the Decision Tree model which showed the highest specificity with lowest sensitivity. KNN showed the highest AUC score with great performance on sensitivity, average performance on accuracy, and bad performance on specificity. Both logistic regression and SVC showed similar performance. Highest AUC score was shown by KNN followed by logistic regression. All models yielded an AUC score above 0.50.

Given that KNN performed the best, according to its high AUC score compared to other models, KNN would be the baseline to extract essential features that were deemed as important in predicting condom use. Using the F-Regression method to extract best features, we found similarity between the KNN model and the conventional logistic regression model. In KNN model, the top features included education (annotated as *kat_pendidikan*), place of residence (annotated as *qtype*), and attitude towards condom use (annotated as *sikap*), followed by radio listening frequency and HIV knowledge. Significant variables in the conventional logistic regression were at the top five of important variables in KNN, showing a compatibility of KNN in predicting condom use. Details were given in figure 4.

3.3. Discussion

Our study highlighted the performance of machine learning in the public health field. To our knowledge, this was among the first studies to apply machine learning for predicting reproductive health problems in Indonesia. It also shed light on how useful machine learning could be compared to other conventional methods, given its capability. In our findings, different models performed differently, with the highest accuracy score obtained in the LR model. The highest sensitivity score was seen in the NB model, and the highest specificity was observed in the DT model. The AUC score was the highest in the KNN model. After extracting the best features in the KNN model for predicting condom use, we found that education, type of residence, attitude towards condom use, radio listening frequency, and HIV knowledge were among the top features to predict condom use. With an AUC score of 61,5%, KNN showed similar results of predicting condom use to the conventional logistic regression in this research.

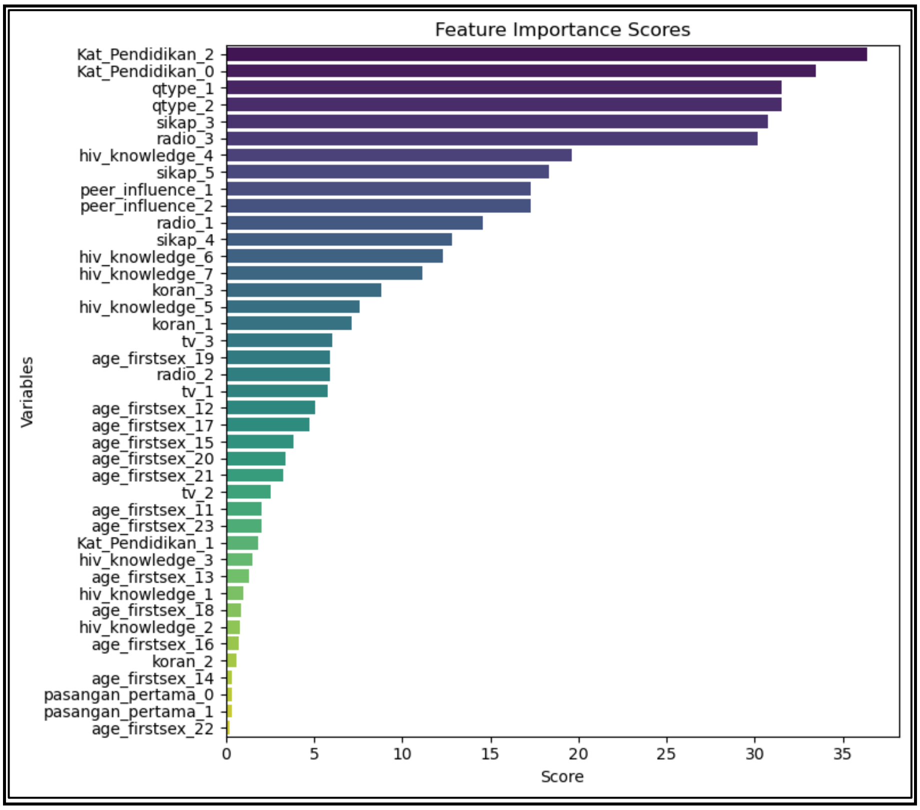

Figure 4. Feature/ Variables important in the KNN model

One thing to note is that it is important to realize that the performance of models addressed different problems and also could vary according to the dataset used [22]. The importance of performance evaluation for different models, therefore, remains undoubtedly important, as different machine learning models serve different purposes in public health schemes. Previous studies of ML usage include Govindan and Maduravasagam (2018) in India and Zhang et al. (2023) in China. A previous study in India deployed a machine learning model to see reasons for not using condoms among students [10]. The model used in the study was SVM, with the accuracy of the model at 73%, with no other models for comparison. In China, several models of machine learning were deployed [11]. The NB model showed the highest sensitivity of 47%. The AUC score for KNN, LR, SVM, and NB was 56%, 77%, 51%, and 72%, respectively, in that study. Differences in our study compared to previous studies are due to, again, the nature of machine learning with different populations and problem questions.

In our conventional logistic regression model, lower odds defined a lower probability to use a condom at first sex. When a p-value approach was taken into consideration, among statistically significant variables, education acted as the most powerful variable to decrease the probability of condom use. This finding aligned with the result from feature importance revealed that education level (*kat_education*) showed the highest score among all features. Result showed that lower education levels were associated with lower odds of condom use for both adolescents in primary–junior school and high school compared to adolescents with a diploma or university degree. This finding was similar to previous studies in other countries [7]. A possible explanation is that those in higher education had better knowledge about condoms and HIV [23]. In addition, higher education was also shown to be a protective factor, meaning that those at a lower education level were at risk of engaging in risky sexual behavior [24].

In regard to attitudes towards condom use, a higher score was associated with lower odds of using a condom. Since higher scores constituted poor attitudes, poor attitudes corresponded to condomless sex. A previous study highlighted how important attitude towards condom use was [25]. This was because people with positive attitudes exhibited a strong intent to adopt a behavior [26], which was condom use in this study. In our study, we found that age at first sex was also associated with condom use, where a lower age corresponded with lower condom use. This finding was similar to the previous study [27]. The association was due to the fact that younger adolescents and early sexual debut showed a higher probability of engaging in risky behavior [28].

Although some variables were not significantly associated, looking at their adjusted odds ratio provided some perspectives. Adolescents who had higher scores of HIV knowledge (corresponding to better HIV knowledge) were more likely to use condoms. Suboptimal HIV knowledge showed to be a risk factor for risky behavior [29]. Health knowledge increased the likelihood of practicing protective behavior for self-protection purposes. A narrative review showed that knowledge affected the adoption of health behavior, especially when it came to a high score of health knowledge (knowledgeable) [30]. In addition, adolescents who had first sex with their partner showed lower odds of using condoms. This finding was similar to a previous study with a highlight on how different types of partners influenced the risk of HIV/AIDS and other Sexually Transmitted Infections (STIs) [31]. Non-steady partners or casual partners (relatives, strangers, prostitutes) were seen as a risk; therefore, a risk preparation behavior was performed, which was condom use. In addition, adolescents in rural areas showed lower odds of using condoms. A previous study showed that disparity among rural and urban adolescents indeed existed [32]. Healthcare services were lacking in rural areas and stigmas prevailed, therefore preventing rural adolescents from practicing safe sexual behavior [33]. In our study, we also found inconsistencies between media exposure and peer influence, where the adjusted and crude odds differ significantly. There might be potential confounding factors [34], exhibited by different odds ratios before and after adjustment in the model.

Our study added various contributions to public health. First, machine learning in public health is relatively new, and our study could serve as a benchmark for future studies where similarities between important features from machine learning and significant variables from the conventional logistic regression were stated clearly in this study. Second, we added a more diverse classification model than the previous study to enrich the application of machine learning. In addition, we also used a dataset from IDHS 2017, the latest data to look at sexual health on a population level as it was deemed representative. However, our study also had its flaws. First, we used secondary data where many explanatory variables (such as religiosity) were not found in the dataset. Second, we only deployed the base mode of machine learning without tuning the hyperparameters. Future studies are suggested to do hyperparameter tuning for the best model provided.

4. CONCLUSION

Logistic regression multivariate analysis confirmed the association between attitude towards condom use, education level, and age at first sex with condom use at first sex. Adolescents with poor attitudes, lower education levels, and lower ages at first sex were less likely to use condoms at first sex. Machine learning models were deployed to predict condom use at first sex and evaluated through various metrics. The DT and NB models remained the models with the highest specificity and sensitivity. The government is advised to create equal education opportunities for every adolescent and shape better knowledge and condom attitudes. Future studies are advised to enhance the performance of machine learning models using hyperparameter tuning and other methods.

REFERENCES

[1] The Joint United Nations Programme on HIV/AIDS, “HIV/AIDS Indonesia.” Accessed: Jul. 15, 2024. [Online]. Available: https://www.unaids.org/en/regionscountries/countries/indonesia

[2] Kementerian Kesehatan Republik Indonesia, “Laporan Eksekutif Perkembangan HIV AIDS dan Penyakit Infeksi Menular Seksual (PIMS) Triwulan I tahun 2023,” DKI Jakarta, 2023.

[3] T. Shafii, K. Stovel, and K. Holmes, “Association between condom use at sexual debut and subsequent sexual trajectories: A longitudinal study using biomarkers,” *Am J Public Health*, vol. 97, no. 6, pp. 1090–1095, Jan. 2007, doi: 10.2105/AJPH.2005.068437.

[4] National Population and Family Planning Board (BKKBN), Statistics Indonesia (BPS), Ministry of Health (Kemenkes), and ICF, “Indonesia Demographic and Health Survey 2017: Adolescent Reproductive Health.,” DKI Jakarta, 2018.

[5] K. F. Norr *et al.*, “Condom use increased after a peer group intervention implemented by community volunteers in Malawi,” *BMC Public Health*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12889-024-18991-z.

[6] T. H. T. Do, L. C. Le, J. A. Burgess, and D. S. Bui, “Determinants of condom use at sexual debut among young Vietnamese,” *International Journal of Adolescent Medicine and Health*, vol. 26, no. 3. Walter de Gruyter GmbH, pp. 403–410, Aug. 01, 2014. doi: 10.1515/ijamh-2013-0322.

[7] Y. K. Cheah, A. Suleiman, M. Ramly, C. C. Kee, and K. K. Lim, “The Role of Education and Demographic Factors in Condom Use among Male Drug Users in Malaysia,” *Malaysian Journal of Medicine and Health Sciences*, vol. 20, no. 1, pp. 52–59, Jan. 2024, doi: 10.47836/mjmhs.20.1.8.

[8] J. Lammers, S. J. G. van Wijnbergen, and D. Willebrands, “Condom use, risk perception, and HIV knowledge: A comparison across sexes in Nigeria,” *HIV/AIDS - Research and Palliative Care*, vol. 5, pp. 283–293, Oct. 2013, doi: 10.2147/HIV.S31687.

[9] I. G. N. E. Putra, D. Astuti, D. A. Widyastari, and D. Phuengsamran, “Prevalence and determinants of condom use among male adolescents in Indonesia,” *Int J Adolesc Med Health*, vol. 33, no. 2, Apr. 2018, doi: 10.1515/ijamh-2018-0141.

[10] B. Govindan and K. Maduravasagam, “Deploying machine learning to find out the reasons for not using condom in a questionnaire-based study of 120 patients,” *Indian J Sex Transm Dis AIDS*, vol. 39, no. 1, pp. 50–54, Jan. 2018, doi: 10.4103/ijstd.IJSTD_64_17.

[11] F. Zhang *et al.*, “Application of machine learning for risky sexual behavior interventions among factory workers in China,” *Front Public Health*, vol. 11, 2023, doi: 10.3389/fpubh.2023.1092018.

[12] B. Parajuli, C. Adhikari, and N. Tripathi, “Determinants of condom use during last sexual intercourse among male college youth of Kaski, Nepal: A cross-sectional survey,” *PLoS One*, vol. 16, no. 12 December, Dec. 2021, doi: 10.1371/journal.pone.0261501.

[13] K. Y. L. Dewi, P. P. Januraga, and N. P. Widarini, “Determinants of condom use by male adolescents in Indonesia during their first sexual intercourse: 2017 national survey data analysis,” *HIV and AIDS Review*, vol. 22, no. 3, pp. 231–236, 2023, doi: 10.5114/hivar.2023.131592.

[14] J. Fieggen, E. Smith, L. Arora, and B. Segal, “The role of machine learning in HIV risk prediction,” *Frontiers in Reproductive Health*, vol. 4, 2022, doi: 10.3389/frph.2022.1062387.

[15] D. Jurafsky and J. Martin, *Speech and Language Processing*. Prentince Hall, 2023.

[16] V. Lee, Lin Liu, and R. Jin, “Decision Trees: Theory and Algorithms,” in *Data Classification: Algorithms and Applications*, 2014.

[17] Z. Zhang, “Introduction to machine learning: K-nearest neighbors,” *Ann Transl Med*, vol. 4, no. 11, Jun. 2016, doi: 10.21037/atm.2016.03.37.

[18] S. Zhang and J. Li, “KNN Classification With One-Step Computation,” *IEEE Trans Knowl Data Eng*, vol. 35, no. 3, pp. 2711–2723, Mar. 2023, doi: 10.1109/TKDE.2021.3119140.

[19] I. Wickramasinghe and H. Kalutarage, “Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation,” *Soft comput*, vol. 25, no. 3, pp. 2277–2293, Feb. 2021, doi: 10.1007/s00500-020-05297-6.

[20] Andrew Ng, “Support Vector Machines.” Accessed: Jul. 15, 2024. [Online]. Available: https://see.stanford.edu/materials/aimlcs229/cs229-notes3.pdf

[21] M. Awad and R. Khanna, “Support Vector Machines for Classification,” in *Efficient Learning Machines*, Apress, 2015, pp. 39–66. doi: 10.1007/978-1-4302-5990-9_3.

[22] O. Samuel, T. Zewotir, and D. North, “Application of machine learning methods for predicting under-five mortality: analysis of Nigerian demographic health survey 2018 dataset,” *BMC Med Inform Decis Mak*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12911-024-02476-5.

[23] J. Etowa *et al.*, “Factors Associated with Condom Knowledge, Attitude, and Use among Black Heterosexual Men in Ontario, Canada,” *Scientific World Journal*, vol. 2021, 2021, doi: 10.1155/2021/8862534.

[24] F. Jackson and Z. T. Haile, “Association between educational attainment and risky sexual behaviour among Ghanaian female youth,” *Afr Health Sci*, vol. 23, no. 1, pp. 301–308, 2023, doi: 10.4314/ahs.v23i1.32.

[25] T. E. Senn, L. A. J. Scott-Sheldon, and M. P. Carey, “Relationship-Specific Condom Attitudes Predict Condom Use Among STD Clinic Patients with both Primary and Non-primary Partners,” *AIDS Behav*, vol. 18, no. 8, pp. 1420–1427, Aug. 2014, doi: 10.1007/s10461-014-0726-y.

[26] S. M. Md Nor, H. Abdullah, Z. Zaremohzzabieh, W. M. Wan Jaafar, and S. Osman, “The Influence of Healthy Lifestyle Attitude, Subjective Norm, and Perceived Behavioural Control on Healthy Lifestyle Behaviour among Married Individuals in Malaysia: Exploring the Mediating Role of Healthy Lifestyle Intention,” *International Journal of Academic Research in Business and Social Sciences*, vol. 13, no. 17, Dec. 2023, doi: 10.6007/ijarbss/v13-i17/19834.

[27] K. Omona and J. K. Ssuka, “Early sexual debut and associated factors among adolescents in Kasawo Sub-county, Mukono district, Uganda,” *Cogent Public Health*, vol. 10, no. 1, Dec. 2023, doi: 10.1080/27707571.2023.2183561.

[28] T. Yosef, T. Nigussie, D. Getachew, and M. Tesfaye, “Prevalence and factors associated with early sexual initiation among college students in southwest Ethiopia,” *Biomed Res Int*, vol. 2020, 2020, doi: 10.1155/2020/8855276.

[29] Q. Zhou *et al.*, “HIV knowledge, sexual practices, condom use and its associated factors among international students in one province of China: a cross-sectional study,” *BMJ Open*, vol. 12, no. 8, Aug. 2022, doi: 10.1136/bmjopen-2021-058108.

[30] F. A. Rincón Uribe *et al.*, “Health knowledge, health behaviors and attitudes during pandemic emergencies: A systematic review,” *PLoS One*, vol. 16, no. 9, p. e0256731, 2021, doi: 10.1371/journal.pone.0256731.

[31] P. Van Empelen and G. Kok, “Condom use in steady and casual sexual relationships: Planning, preparation and willingness to take risks among adolescents,” *Psychol Health*, vol. 21, no. 2, pp. 165–181, Apr. 2006, doi: 10.1080/14768320500229898.
[32] F. Aditya Kautsar Murti and K. Sari, “Rural-Urban Disparity of Condom Use at the Last Sex among Indonesian Male Adolescents,” *Jurnal Kesehatan Reproduksi*, vol. 14, no. 2, pp. 105–114, 2023, [Online]. Available: http://journaliakmitangsel2.iakmi.or.id
[33] L. Hu *et al.*, “Condom Use and Related Factors among Rural and Urban Men Who Have Sex With Men in Western China: Based on Information-Motivation-Behavioral Skills Model,” *Am J Mens Health*, vol. 14, no. 1, Jan. 2020, doi: 10.1177/1557988319899799.
[34] International Agency for Research on Cancer, “Dealing with confounding in the analysis,” in *Cancer Epidemiology*, World Health Organization, 1999.