BAREKENG: Journal of Mathematics and Its Applications
December 2026 Volume 20 Issue 4 Page 2693-2710
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss4pp2693-2710

CLASSIFICATION OF PRE-MARITAL SEX AMONG ADOLESCENTS USING COST-SENSITIVE WEIGHTED RANDOM FOREST ON IMBALANCED DATA

Nilam Novita Sari[1*], Bagus Sumargo[2], Dania Siregar[3], Endang Yuliani[4], Citra Komang Sari[5], Baihaqy Mahardika[6], Khansa Salsabil Fadya[7]

[1,2,3,4,5,6,7]*Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Negeri Jakarta*
Jln. Rawamangun Muka Raya No. 11, East Jakarta, 13220, Indonesia

*Corresponding author's e-mail: * **nilam.novita@unj.ac.id***

Article Info

Article History:

Received: 23rd July 2025
Revised: 30th April 2026
Accepted: 4th June 2026
Available Online: 1st July 2026

Keywords:

Adolescent Reproductive Health (ARH);
Cost Sensitive Weighted Random Forest;
imbalanced data;
pre-marital sex;
Random Forest.

ABSTRACT

Adolescents are a critical developmental phase characterized by significant physical, cognitive, and psychosocial changes, which increase vulnerability to risky behaviors, including premarital sexual activity. Such behavior may lead to serious consequences, such as unintended pregnancy, unsafe abortion, sexually transmitted infections, and HIV/AIDS. This study aims to classify premarital sexual behavior among adolescents using demographic, knowledge, and behavioral factors. The data were obtained from the 2019 Program Performance and Accountability Survey (SKAP), consisting of approximately 5,300 weighted adolescent respondents in East Java. The dataset exhibits extreme class imbalance, with only 0.3% of adolescents reporting premarital sexual behavior. To address this issue, three classification methods were applied: standard Random Forest (RF), Weighted Logistic Regression (WLR), and Cost-Sensitive Weighted Random Forest (CSWRF). Model performance was evaluated using 5-fold cross-validation and multiple metrics, including accuracy, sensitivity, specificity, precision, F1-score, G-Means, and Area Under the Curve (AUC). The results indicate that although the standard RF achieved very high accuracy (above 99%), it failed to identify the minority class, resulting in undefined specificity and poor discriminative ability (AUC ≈ 0.83). The WLR model improved classification balance, achieving reasonable specificity (0.8667) and AUC (0.9459), but produced more false positives. In contrast, the CSWRF model demonstrated the best overall performance, with high AUC values (above 0.95), improved specificity, and a better balance between sensitivity and specificity, indicating its effectiveness in handling highly imbalanced data. Variable importance analysis using Mean Decrease Accuracy (MDA) identified Knowledge of Adolescent Reproductive Health (ARH), residence type, and behavioral factors such as dating experience as the most influential predictors. Despite these findings, this study is limited by potential underreporting bias and exclusion of socio-cultural and psychological variables. Future research should incorporate additional predictors and explore advanced modeling approaches to improve classification performance.

How to cite this article:

N. N. Sari, B. Sumargo, D. Siregar, E. Yuliani, C. K. Sari, B. Mahardika and K. S. Fadya., "CLASSIFICATION OF PRE-MARITAL SEX AMONG ADOLESCENTS USING COST-SENSITIVE WEIGHTED RANDOM FOREST ON IMBALANCED DATA", *BAREKENG: J. Math. & App.*, vol. 20, no. 4, pp. 2693-2710, Dec, 2026.

1. INTRODUCTION

Adolescence is a critical transitional phase characterized by significant physical, cognitive, and psychosocial changes that influence decision-making and social behavior [1]. According to BKKBN, adolescents are individuals aged 10–24 years who are unmarried [2]. During this period, adolescents are particularly vulnerable to engaging in risky behaviors, including premarital sexual activity, which poses serious public health concerns [3].

Premarital sex refers to sexual activity among individuals who are not married [4]. This behavior is associated with various adverse outcomes such as unintended pregnancy, unsafe abortion, sexually transmitted infections (STIs), and HIV/AIDS, as well as broader social consequences including school dropout and unemployment [5], [6]. One of the main contributing factors is the lack of knowledge about Adolescent Reproductive Health (ARH), which leads to poor decision-making regarding sexual behavior [7].

Despite its importance, analyzing premarital sexual behavior remains challenging due to both methodological and data-related limitations. Traditional statistical approaches, such as logistic regression, often rely on assumptions of data balance and may perform poorly when applied to highly skewed social data. In addition, these methods may fail to capture complex nonlinear relationships and interactions among predictors, which are commonly present in behavioral and social phenomena. Consequently, machine learning approaches have gained increasing attention due to their flexibility, robustness, and superior predictive performance in handling high-dimensional and complex datasets [8]. To address class imbalance within traditional approaches, weighted logistic regression can be applied by assigning higher importance to the minority class, allowing it to serve as a baseline model for comparison with more advanced machine learning methods.

One widely used machine learning method is Random Forest (RF), an ensemble learning technique based on bootstrap aggregating and random feature selection [9]. It offers several advantages, such as the ability to handle large numbers of variables, fast computation, easy implementation, and high prediction accuracy [10], [11].

However, its performance significantly deteriorates when applied to imbalanced datasets, where one class heavily dominates the other [12]. In such cases, the model tends to be biased toward the majority class, resulting in poor detection of minority-class instances [13].

The prevalence of premarital sex among adolescents continues to rise. Globally, around 40% of adolescents aged 15–19 and 75% of those aged 20–24 engage in premarital sex [14]. In Indonesia, BKKBN reported that 60% of adolescents aged 16–17 and 20% of those aged 14–15 and 19–20 have engaged in sexual activity [15]. Furthermore, it is estimated that approximately one million adolescents experience premarital pregnancy, with 52% of these cases resulting in abortion [15]. These findings highlight that premarital sexual behavior among adolescents constitutes a critical public health issue with significant social and health consequences.

This issue is particularly relevant in the context of premarital sexual behavior data. Based on the SKAP survey, only 0.3% of adolescents in East Java reported engaging in premarital sex, while 99.7% did not. This indicates an extreme class imbalance problem. In addition, given the sensitive nature of sexual behavior, the reported prevalence is likely affected by underreporting bias, meaning that the true proportion may be higher than observed. Therefore, robust classification methods that can effectively handle both class imbalance and potential bias are essential for obtaining reliable insights.

Several strategies have been proposed to handle imbalanced data, including data-level (resampling methods) and algorithm-level approaches [16]. This study adopts an algorithm-level approach by adjusting the Random Forest model using class weighting to give more importance to the minority class [17].

Chen compared data-level methods (SMOTE, SMOTEBoost) with algorithm-level methods (Weighted Random Forest, Balanced Random Forest). They found that WRF and BRF outperformed SMOTE-based approaches in handling class imbalance [18].

Based on this background, this study aims to classify premarital sexual behavior among adolescents in East Java. The main contributions of this study are twofold: (1) the application of a cost-sensitive Weighted Random Forest to effectively address extreme class imbalance in sensitive social data, and (2) the identification of key risk factors associated with premarital sexual behavior among adolescents. The findings

are expected to provide both methodological contributions to imbalanced data classification and practical insights for adolescent reproductive health interventions.

2. RESEARCH METHODS

The research methods contain explanations in the form of paragraphs about the research design or descriptions of the experimental settings, data sources, data collection techniques, and data analysis conducted by the researcher. This guide will explain how to write the headings. If there are more than one heading, use the second level headings as below.

2.1 Data Source

The data used in this study are secondary data obtained from the 2019 *Program Performance and Accountability Survey* (SKAP) conducted by BKKBN Jakarta, specifically the Adolescent Module. The study population includes all adolescents aged 10–24 years in East Java, totaling approximately 9.123 million individuals. The sample consists of 5,300 weighted cases of unmarried adolescents aged 10–24 years residing in East Java.

The data used in this study are secondary data obtained from the 2019 Program Performance and Accountability Survey (Survei Kinerja dan Akuntabilitas Program (SKAP)) conducted by BKKBN, specifically from the Adolescent Module. The target population comprises all adolescents aged 10–24 years in East Java, totaling approximately 9.123 million individuals.

The SKAP 2019 survey employed a multistage sampling design. In the first stage, villages/urban wards were selected using Probability Proportionate to Size (PPS) sampling based on the number of households. In the second stage, one cluster within each selected village/ward was chosen using PPS. In the third stage, 35 households were selected from each cluster using systematic random sampling based on household listing conducted by enumerators.

All adolescents aged 10–24 years who were unmarried and had resided in the selected household for at least six months were included as respondents. These included biological children, stepchildren, adopted children, and other dependents recorded as household members. The number of adolescent respondents varied across households and clusters depending on household composition.

To obtain population-representative estimates, sampling weights provided in the SKAP dataset were applied. The weighting process began with the calculation of design weights, defined as the inverse of the sampling fraction at each stage of the sampling design. Subsequently, normalization weighting was performed to produce normalized weights for households, families, women of reproductive age, and adolescents. This normalization step aims to avoid inflated values and to align the weighted data with the actual sample size. The final weight assigned to each observation represents the number of individuals in the population that the sample unit represents. Therefore, weighting was applied prior to analysis to ensure unbiased estimation.

The initial dataset consisted of 1,700 adolescent respondents in East Java. After data cleaning, including the removal of incomplete and inconsistent observations, the final dataset comprised 5,300 weighted samples used for analysis.

2.2 Methods

2.2.1 Random Forest

Random Forest (RF) is an ensemble learning method that combines the predictions of multiple decision trees to improve regression and classification performance [19]. For a given input vector, RF constructs K regression trees using the values of evidential features and outputs the average prediction. To reduce correlation between trees and increase diversity, RF applies bootstrap aggregation (bagging), where each tree is trained on a random subset of the original data generated by resampling with replacement. This ensures some samples may appear multiple times while others may be excluded from training, enhancing robustness to data variability and improving prediction accuracy [19].

When growing each tree, RF selects the optimal split from a randomly chosen subset of features, reducing inter-tree correlation and lowering generalisation error, albeit at the cost of slightly weaker individual trees. Trees are grown without pruning, making the model computationally efficient. Samples excluded from the bootstrap set form the out-of-bag (OOB) dataset, which is used to compute unbiased estimates of generalisation error without requiring a separate validation set [20]. As the number of trees increases, the error stabilises, preventing overfitting.

Additionally, RF provides a measure of feature importance. This is achieved by permuting a feature's values while keeping other variables constant and observing the decrease in OOB prediction accuracy. Such capability allows RF to identify the most influential evidential features for model building, making it an effective and interpretable approach for complex datasets [21].

2.2.2 Weighted Random Forest

The Weighted Random Forest (WRF) method adopts the standard RF algorithm for highly imbalanced datasets using a cost-sensitive learning approach. The class weights are computed using a cost-sensitive scheme defined as [22]:

$$W_i = \frac{N}{C \cdot c_i} \cdot \eta, \tag{1}$$

where W_i is the weight for class i, $N$is the total number of samples, C is the number of classes, and c_i is the number of samples in class i. The adjustment factor $\eta$is derived from the imbalance ratio, defined as:

$$\eta = \frac{\max(c_i)}{\min(c_i)}. \tag{2}$$

A cost matrix $C_{ij}$is defined to assign higher penalties to misclassification of minority-class instances:

$$C_{ij} = \begin{cases} 0, & \text{if } i = j \\ W_i, & \text{if } i \neq j \end{cases}. \tag{3}$$

These weights are incorporated during tree construction by modifying the Gini criterion and during prediction using weighted majority voting. The final prediction is obtained by aggregating weighted votes across all trees:

$$WP = \sum_{i=1}^{ntree} W_i \cdot Pt_i, \tag{4}$$

where WP is the final weighted prediction, $ntree$ is the number of trees in the forest, and Pt_i is the prediction produced by the i-th tree.

To reduce bias toward the majority class, higher misclassification costs are assigned to the minority class through class weights. These weights influence the model in two stages: (1) during tree construction, they adjust the Gini criterion for split selection, and (2) in terminal nodes, predictions are made using weighted majority votes, where each class's vote is proportional to its weight and case count. The final RF prediction aggregates these weighted votes across trees. Class weights, a key tuning parameter, are optimized using the out-of-bag accuracy estimate [18].

2.2.3 Weighted Logistic Regression

Weighted Logistic Regression is an extension of the standard logistic regression model designed to handle imbalanced data by assigning different weights to each observation. In this approach, higher weights are given to the minority class, while lower weights are assigned to the majority class, allowing the model to reduce bias toward the dominant class [23].

Let $\pi(x) = P(y = 1 \mid x)$denote the probability of the positive class. The logistic regression model is defined as:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}. \tag{5}$$

In the weighted setting, the likelihood function is modified by incorporating weights $w_i$for each observation:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{w_i y_i} [1 - \pi(x_i)]^{w_i (1-y_i)}. \quad (6)$$

The corresponding log-likelihood is given by:

$$L\ell(\beta) = \sum_{i=1}^{n} w_i \, [y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))]. \quad (7)$$

The parameter estimates are obtained using Maximum Likelihood Estimation (MLE), typically solved through iterative numerical methods such as the Newton-Raphson algorithm.

Weighted logistic regression is widely used due to its interpretability and simplicity. However, it assumes a linear relationship between predictors and the log-odds of the response variable, which may limit its ability to capture complex nonlinear patterns compared to machine learning approaches such as Random Forest.

2.2.4 Imbalance Data

Imbalanced data distribution arises when the sample sizes among different classes in a dataset are markedly disproportionate, with one or more majority classes containing substantially more instances than the minority classes. This uneven distribution poses a significant challenge in machine learning, as most conventional algorithms operate under the assumption of balanced class proportions and therefore tend to exhibit bias in favor of the majority class [24].

Conventional learning algorithms are typically developed under the assumption of balanced datasets, which can lead to biased decision boundaries and poor generalization for rare instances. Such limitations stem from overlapping feature spaces, limited sample sizes, and the substantial misclassification costs associated with minority classes. Addressing these challenges necessitates a thorough understanding of the dataset's characteristics and the real-world implications of failing to detect minority cases [25].

2.2.5 SMOTE-N

The Synthetic Minority Oversampling Technique (SMOTE) is an oversampling method that generates synthetic minority class samples to address class imbalance and reduce overfitting [26]. Using the k-nearest neighbors (kNN) approach, SMOTE selects a minority instance, identifies its k nearest neighbors, and creates new samples by interpolating between the instance and randomly chosen neighbors. The process depends on the number of minority samples (T), the oversampling ratio (N), and the number of neighbors (k). Synthetic samples are generated until the desired oversampling ratio is achieved, either by random selection when N is small or iteratively when N exceeds the original dataset size [27]. SMOTE is typically applied to continuous data. Therefore, the SMOTE-N method was developed to handle categorical data [28].

SMOTE-N determines the nearest neighbors using a modified Value Difference Metric (VDM). The VDM measures the similarity between feature values across all feature vectors by constructing a matrix that records the distances between corresponding feature values. The distance (δ) between two feature values is calculated based on this definition [26].

$$\delta(V_1, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k, \quad (8)$$

where:

C_1 : the total number of occurrences of feature value V_1;

C_{1i} : is the number of occurrences of feature value V_1 for class i;

C_2 : the total number of occurrences of feature value V_2;

C_{2i} : is the number of occurrences of feature value V_2 for class i.

2.2.6 Classification Performance

A confusion matrix is a table used to evaluate the performance of a classification algorithm. It provides a clear visualization and summary of how well the model performs in categorizing data [29]. Table 1 presents an example of a confusion matrix.

Table 1. Confusion Matrix

Actual	Predicted	
	Positive	**Negative**
Positive	TP	FN
Negative	FP	TN

Based on the table above, the four main components used to calculate performance metrics are [29]:

1. True Positive (TP) – The number of positive cases correctly predicted as positive.
2. True Negative (TN) – The number of negative cases correctly predicted as negative.
3. False Positive (FP) – The number of negative cases incorrectly predicted as positive (Type I error).
4. False Negative (FN) – The number of positive cases incorrectly predicted as negative (Type II error).

Several performance metrics can be derived from these values, as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{9}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{10}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 - Score = \frac{2 * recall * precision}{recall + precision} \tag{14}$$

$$G - Means = specificity * sensitivity \tag{15}$$

2.3 Research Variables

The variables used in this study consist of dependent and independent variables, as shown in Table 2.

Table 2. Research Variables

No	Variable Name	Definition	Description
Dependent Variable			
1	Ever had sexual intercourse (Y_1)	Indicates whether the respondent has ever engaged in sexual intercourse	Categories: - No - Yes
		Independent Variables	
			Demographics
2	Age (X_1)	Age of the respondent at the time of the survey	Continuous
3	Gender (X_2)	Biological sex of the respondent	Categories: - Male - Female
4	Education Level (X_3)	Highest level of formal education attained by the respondent	Categories: - Never attended school - Elementary School, Junior High School, Senior High School

No	Variable Name	Definition	Description
			- Diploma I, Diploma II, Diploma III, Bachelor's Degree, Master's Degree, Doctorate (PhD)
5	Residence Type (X_4)	Type of area where the respondent resides	Categories: Urban / Rural
Knowledge			
6	Knowledge of Contraception (X_5)	Respondent's awareness of contraceptive methods	Categories: - No - Yes
7	Knowledge of Female Fertility Period (X_6)	Respondent's knowledge of the fertile period in the menstrual cycle	Categories: - No - Yes
8	Knowledge of Pregnancy (X_7)	Respondent's understanding of pregnancy-related concepts	Categories: - No - Yes
9	Knowledge of HIV-AIDS (X_8)	Respondent's knowledge of HIV/AIDS transmission and prevention	Categories: - No - Yes
10	Knowledge of Adolescent Reproductive Health (ARH) (X_9)	Respondent's knowledge of general adolescent reproductive health (ARH)	Categories: - No - Yes
Behavioral			
11	Ever been in a romantic relationship (X_{10})	Indicates whether the respondent has ever been in a romantic relationship	Categories: - No - Yes
12	Dating Style (X_{11})	Type of physical interaction in romantic relationships	Categories: - Never been in relationship - No physical contact - Holding hands - Hugging - Kissing - Fondling

2.4 Research Design

The stages of analysis in this study are as follows:

1. Data Pre-processing: Data cleaning was performed to ensure data quality. Missing values were handled using imputation techniques. For continuous variables (age), missing values were imputed using the means to reduce the influence of outliers. For categorical variables, missing values were imputed using the mode (most frequent category). All categorical variables were transformed into numerical forms using label encoding (binary variables coded as 0/1 and ordinal variables encoded based on their natural order) to ensure compatibility with the machine learning model.
2. Data Splitting: The dataset was partitioned using 5-fold cross-validation to ensure robust model evaluation and to reduce variability due to random sampling.
3. Model Building: A Cost-Sensitive Weighted Random Forest model was trained using the training folds. The model incorporates class weights to address class imbalance.
4. Model Parameters and Tuning: The number of trees ($ntree$) was treated as a tuning parameter and evaluated at several values (50, 100, 250, and 500 trees). The optimal number of trees was selected based on out-of-bag (OOB) error. The number of variables randomly selected at each split ($mtry$) was set to the default value, i.e., $\sqrt{p}$, where $p$is the number of predictors.
5. Model Evaluation: The trained model was evaluated on the validation folds using multiple performance metrics suitable for imbalanced classification, including accuracy, sensitivity (recall), specificity, precision, F1-score, and G-Mean.
6. Performance Assessment: In addition to confusion matrix-based metrics, Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) were used to evaluate the model's

discriminative ability. Precision–Recall (PR) curves were also employed to provide a more informative evaluation under class imbalance conditions.

7. Variable Importance: Variable importance measures were computed based on the mean decrease in the Gini index to identify the most influential predictors of premarital sexual behavior among adolescents.

3. RESULTS AND DISCUSSION

Before conducting the analysis, a descriptive analysis is first carried out to provide an overview of the data.

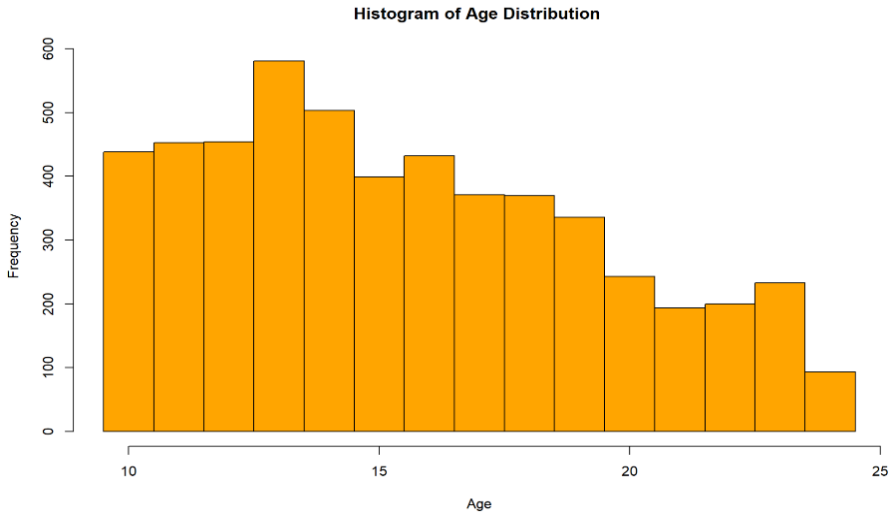

Figure 1. Respondents' Age Distribution
*(**Source**: Microsoft Excel)*

Based on Fig. 1, the respondents' ages fall within the adolescent range of 10–24 years. The largest group of respondents is those aged 13 years, totalling 581 individuals, while the smallest group is those aged 24 years.

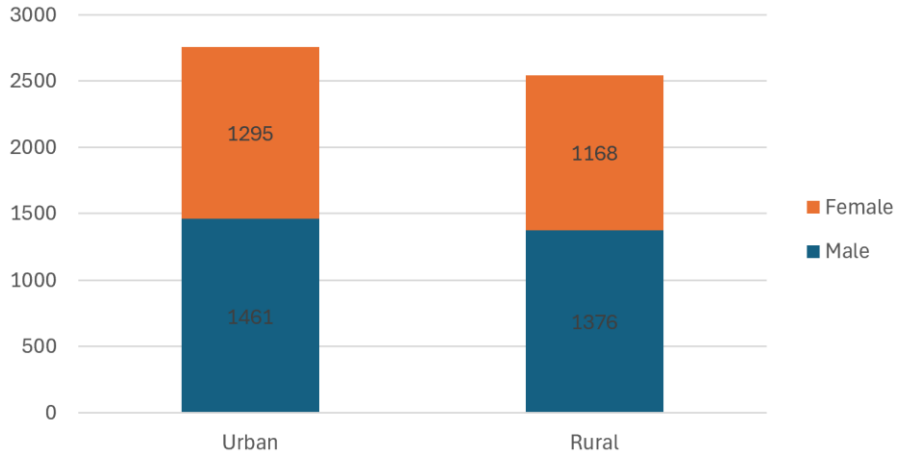

Figure 2. Graph of Adolescents' Gender and Place of Residence
*(**Source**: Microsoft Excel)*

Fig. 2 presents the distribution of respondents based on place of residence and gender. Of the 5,300 respondents, 2,837 are male, and 2,463 are female. The number of respondents living in urban areas is greater than that of those residing in rural areas.

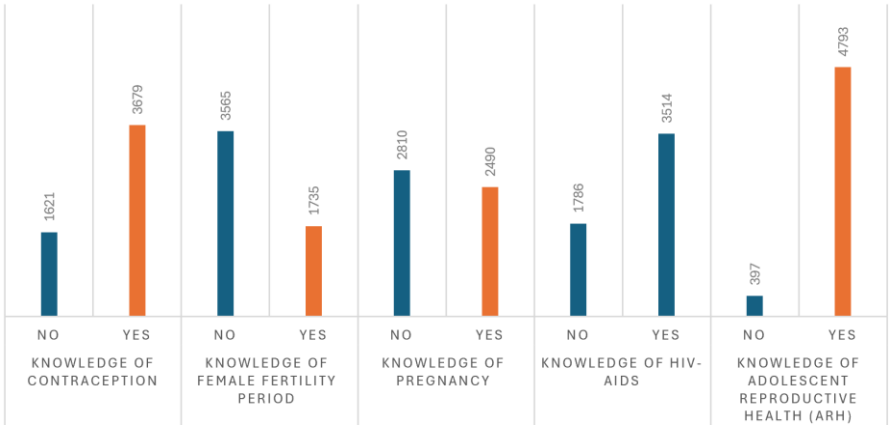

Figure 3. Graph of Adolescents' Knowledge
*(**Source**: Microsoft Excel)*

Based on Fig. 3, it is evident that respondents with knowledge about family planning (KB) far outnumber those without such knowledge, indicating the success of family planning outreach in the community. However, knowledge regarding the fertile period is predominantly lacking, with only 1,735 out of 5,300 respondents being aware of it. This low level of awareness highlights a significant knowledge gap about the reproductive cycle, an important aspect for family planning and the prevention of unplanned pregnancies.

In terms of pregnancy knowledge, the number of respondents who are aware and those who are unaware is relatively balanced, though the latter group is slightly larger. This suggests the need for enhanced education on pregnancy and its associated risks.

Regarding HIV/AIDS, the majority of respondents possess knowledge of the topic, reflecting the effectiveness of public health campaigns, although around 1,700 respondents remain uninformed. Lastly, in terms of knowledge about Adolescent Reproductive Health (ARH), only 397 respondents lack awareness, demonstrating that KRR education has successfully reached a wide audience.

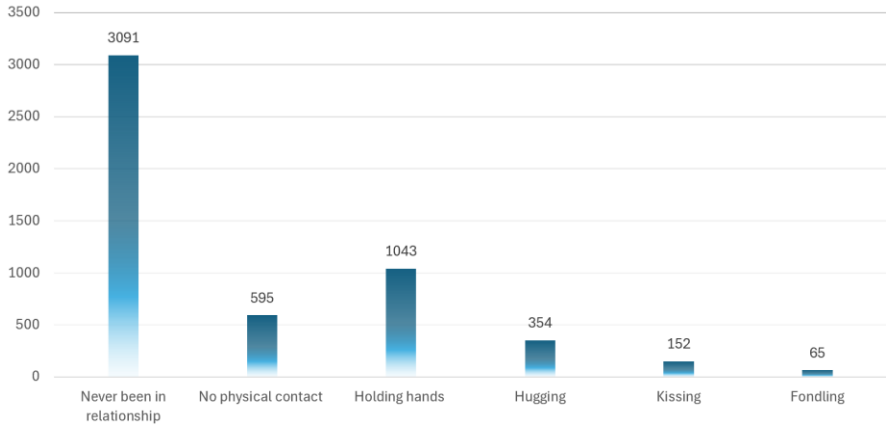

Figure 4. Adolescent Dating Styles
*(**Source**: Microsoft Excel)*

Fig. 4 shows that of the 5,300 adolescent respondents, 3,091 reported never having dated, while 2,209 stated that they had dated. Among those who had dated, the most common dating behavior was holding hands, reported by 1,043 respondents. This was followed by dating without engaging in any particular activity (354 respondents), dating with kissing as the primary behavior (152 respondents), and the least common being dating involving touching/stimulating, reported by 65 respondents.

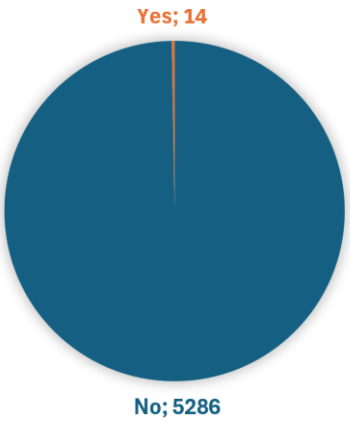

Figure 5. Graph of Adolescent Premarital Sexual Behavior
*(**Source**: Microsoft Excel)*

Fig. 5 above shows data on premarital sexual behavior. The number of adolescents who engaged in premarital sex is much smaller compared to those who did not, with only 14 adolescents (0.26%) reporting such behavior.

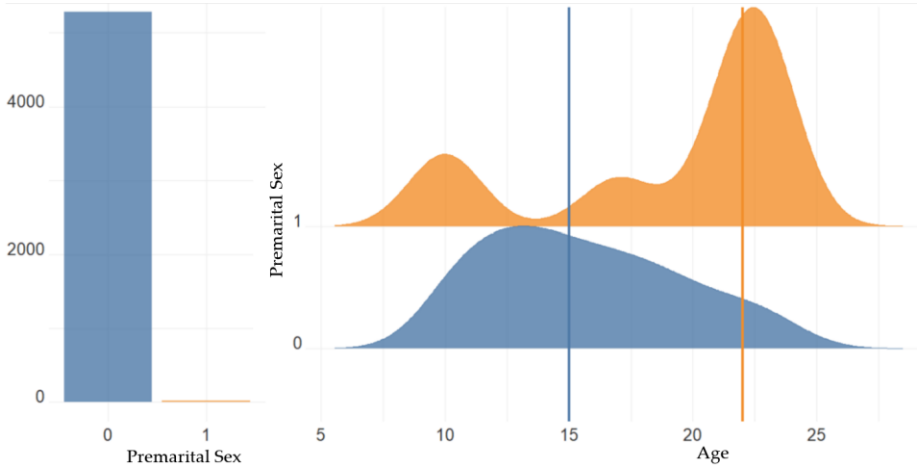

Figure 6. Age and Premarital Sex Density Plot
*(**Source**: Python)*

Fig. 6 presents the density plot between the variables of age and premarital sex. The figure indicates that the average age of adolescents who did not engage in premarital sex was 15 years, while the average age of those who did was 22 years. Adolescents who did not engage in premarital sex were most numerous at age 13, whereas those who engaged in premarital sex were most numerous at age 23. The plot also reveals distinct data patterns between adolescents who engaged in premarital sex and those who did not, suggesting a significant difference in the ages of these two groups. The next step will involve performing classification on the premarital sex data.

Classification was performed using three methods: standard RF (without handling class imbalance), weighted logistic regression, and CSWRF, a modified version of RF designed to address class imbalance by assigning different weights to each class to balance them. The confusion matrix of the methods is presented in the following tables.

Table 3. Confusion Matrix of Random Forest

n = 50			**n = 100**		
	Actual 0	**Actual 1**		**Actual 0**	**Actual 1**
Prediction 0	5234(TN)	3 (FN)	Prediction 0	5234 (TN)	3 (FN)
Prediction 1	52 (FP)	11 (TP)	Prediction 1	52 (FP)	11 (TP)

n = 250			**n = 500**		
	Actual 0	**Actual 1**		**Actual 0**	**Actual 1**
Prediction 0	5233 (TN)	3 (FN)	Prediction 0	5234 (TN)	3 (FN)
Prediction 1	53 (FP)	11 (TP)	Prediction 1	52 (FP)	11 (TP)

Table 4. Confusion Matrix of Weighted Random Forest

n = 50			**n = 100**		
	Actual 0	**Actual 1**		**Actual 0**	**Actual 1**
Prediction 0	5237 (TN)	4 (FN)	Prediction 0	5239 (TN)	3 (FN)
Prediction 1	49 (FP)	11 (TP)	Prediction 1	47 (FP)	11 (TP)
n = 250			**n = 500**		
	Actual 0	**Actual 1**		**Actual 0**	**Actual 1**
Prediction 0	5239 (TN)	3 (FN)	Prediction 0	5239 (TN)	3 (FN)
Prediction 1	47 (FP)	11 (TP)	Prediction 1	47 (FP)	11 (TP)

Table 5. Confusion Matrix of Weighted Logistic Regression

	Actual 0	**Actual 1**
Prediction 0	4973 (TN)	2 (FN)
Prediction 1	313 (FP)	12 (TP)

The confusion matrices presented in Table 3 are obtained by aggregating prediction results from all 5 cross-validation folds, providing an overall representation of model performance across the dataset. For the standard Random Forest (RF) model, the aggregated confusion matrix shows that the model is able to correctly classify the majority class, as indicated by the high number of true negatives (TN ≈ 5233–5234). However, although some true positives (TP = 11) are observed, the number of false negatives (FN = 3) and false positives (FP ≈ 52–53) indicates that the model still struggles to consistently identify the minority class. This finding contrasts with the fold-wise average evaluation, where specificity and G-Mean were reported as zero or undefined. The discrepancy arises because the aggregated confusion matrix combines predictions from all folds, allowing minority class detections from different folds to accumulate, resulting in non-zero values.

In contrast, the Weighted Random Forest (WRF) model demonstrates a more balanced classification performance. The number of false positives decreases (FP ≈ 47–49), while true positives remain stable (TP = 11). At the same time, false negatives are relatively low (FN = 3–4), indicating that the model is able to better capture minority class instances compared to the standard RF. This improvement confirms that the use of class weighting effectively reduces bias toward the majority class and enhances the model's ability to detect minority cases.

Meanwhile, the Weighted Logistic Regression (WLR) model shows a different pattern. Although it achieves a slightly higher number of true positives (TP = 12) and lower false negatives (FN = 2), it produces a substantially larger number of false positives (FP = 313). This indicates that the model tends to overcompensate for class imbalance by predicting more observations as belonging to the minority class, which reduces its precision and may lead to overestimation of premarital sexual behavior.

It is important to emphasize that performance metrics derived from the aggregated confusion matrix may differ from the average metrics obtained from 5-fold cross-validation. In fold-wise evaluation, each fold is assessed independently, and extreme cases—such as folds where the minority class is not detected at all—can result in zero or undefined specificity and G-Mean. However, when predictions are aggregated, even a small number of correct minority classifications across folds will produce non-zero values.

Therefore, the aggregated confusion matrix provides a global performance overview, while the average cross-validation results better reflect the stability and consistency of the model. In this study, the results indicate that the standard RF model is unstable in handling imbalanced data, whereas WRF provides a more balanced trade-off between detecting minority and majority classes. Although WLR improves sensitivity, its tendency to generate excessive false positives limits its practical reliability. The performance of both methods using various values of ntree is presented in the following tables.

The performance of the three methods, Random Forest (RF), Cost-Sensitive Weighted Random Forest (CSWRF), and Weighted Logistic Regression (WLR), across various values of *ntree* is presented in Table 6, Table 7, and Table 8. In addition, the classification results are further evaluated using aggregated confusion matrices derived from the 5-fold cross-validation to provide a more comprehensive assessment of model performance.

Table 6. Performance of Random Forest (Without Imbalance Handling)

	Mean Accuracy	Mean Sensitivity	Mean Specificity	Mean Precision	Mean F1-Score	Mean G-Means	Mean AUC	ntree
1	0.9977360	0.9977353	NA	1.0000000	0.9988663	NA	0.8307	50
2	0.9981130	0.9981120	NA	1.0000000	0.9990550	NA	0.8297	100
3	0.9977360	0.9977353	NA	1.0000000	0.9987663	NA	0.8291	250
4	0.9977358	0.9977351	NA	1.0000000	0.9988662	NA	0.8282	500

Table 6 presents the performance of the RF model without addressing the class imbalance. The table shows that the average accuracy obtained exceeds 99% for each ntree value of 50, 100, 250, and 500. This indicates very high accuracy, and similarly, the average sensitivity, average precision, and average F1-Score also reflect excellent performance. Although the average accuracy, sensitivity, precision, and F1-Score demonstrate very good results, the values for specificity and average G-Means are reported as NA.

This indicates a problem related to class imbalance, as the prediction results show that instances of category 1 (the minority class) tend to be misclassified as category 0 (the majority class). As a result, the predicted values for category 1 are all classified as 0, making it impossible to compute specificity, which leads to an NA value. The random forest modeling results above indicate the need for handling the class imbalance issue in the dataset. One approach to address this problem is by using the Weighted Random Forest method.

In addition, the AUC values for the Random Forest model range from 0.8282 to 0.8307 across different *ntree* values. Although these values indicate a moderate level of discriminative ability, they are relatively low compared to models that properly handle class imbalance. This suggests that, despite achieving high accuracy, the model is not effective in distinguishing between the minority and majority classes.

The relatively stable AUC values across different *ntree* settings also indicate that increasing the number of trees does not significantly improve the model's discriminative performance. This further confirms that the main limitation of the standard Random Forest lies not in model complexity, but in its inability to address the imbalance in class distribution. Therefore, relying solely on accuracy and AUC in this context may lead to misleading conclusions, as the model still fails to correctly identify minority class instances.

Table 7. Performance of Weighted Random Forest

	Mean Accuracy	Mean Sensitivity	Mean Specificity	Mean Precision	Mean F1-Score	Mean G-Means	Mean AUC	ntree
1	0.9902	0.9907	0.7667	0.9909	0.9953	0.8639	0.9836	50
2	0.9906	0.9911	0.7667	0.9997	0.9953	0.8641	0.9509	100
3	0.9906	0.9911	0.7667	0.9905	0.9953	0.8641	0.9822	250
4	0.9906	0.9911	0.7667	0.9905	0.9953	0.8631	0.9845	500

Table 7 presents the performance of the CSWRF, which is an RF method that applies different weights to each class. The weight assignment is based on a formula, giving higher weights to the minority class (those who engage in premarital sex) and lower weights to the majority class (those who do not engage in premarital sex). This approach ensures that the model does not ignore the minority class.

As shown in Table 7, the average accuracy, average sensitivity, average precision, and average F1-score all remain very high, above 98%, although these metrics are slightly lower than those obtained from the standard random forest. However, with the use of the weighted random forest, the average specificity and average G-Means show significant improvement. Specifically, the average specificity exceeds 16%, and the average G-Means exceeds 39% across all ntree values.

The increase in average specificity and G-Means indicates that the weighted random forest is capable of handling imbalanced data, as the resulting model can accurately classify the minority class. In other words, the model is no longer biased toward the majority class, unlike the previous random forest model. The improvement in specificity suggests that the model is starting to correctly identify individuals in category 1 (the minority group), i.e., those who engage in premarital sex. Moreover, the significant increase in G-Means highlights a better balance between sensitivity and specificity.

The AUC values of the CSWRF model range from 0.9509 to 0.9845 across different *ntree* values, indicating excellent discriminative performance. Compared to the standard Random Forest, the CSWRF

achieves substantially higher AUC values, reflecting its improved ability to distinguish between adolescents who engage in premarital sexual behavior and those who do not.

The consistently high AUC values across different *ntree* settings also suggest that the model is stable and robust, and that increasing the number of trees does not significantly affect its discriminative capability. This further confirms that the improvement in performance is primarily driven by the incorporation of class weights rather than model complexity. Overall, the combination of high AUC, improved specificity, and higher G-Means demonstrates that the CSWRF model provides a more balanced and reliable classification performance for highly imbalanced data.

Table 8. Performance of Weighted Logistic Regression

Mean Accuracy	Mean Sensitivity	Mean Specificity	Mean Precision	Mean F1-Score	Mean G-Means	Means AUC
0.9406	0.9408	0.8667	0.9996	0.9692	0.8866	0.9459

Table 8 presents the performance of the weighted logistic regression model. The results indicate that the model achieves relatively good performance, with a specificity of 0.8667, a G-Means value of 0.8866, and an AUC of 0.9459, suggesting that it is capable of distinguishing between classes with a reasonable level of accuracy while maintaining a balance between sensitivity and specificity. As a parametric model, logistic regression provides interpretable coefficients and a straightforward probabilistic framework, making it useful for understanding the direction and magnitude of the relationships between predictors and the outcome variable.

However, its performance remains inferior to that of the CSWRF model, particularly in terms of AUC and overall balance as reflected by G-Means. This is likely due to the inherent limitations of logistic regression, which assumes a linear relationship between the predictors and the log-odds of the response variable and may not adequately capture complex nonlinear patterns or interactions among variables. In contrast, the CSWRF model, as an ensemble-based machine learning method, is more flexible and better suited for modeling complex data structures. Therefore, CSWRF provides a more robust and reliable approach for classifying premarital sexual behavior in the presence of imbalanced data.

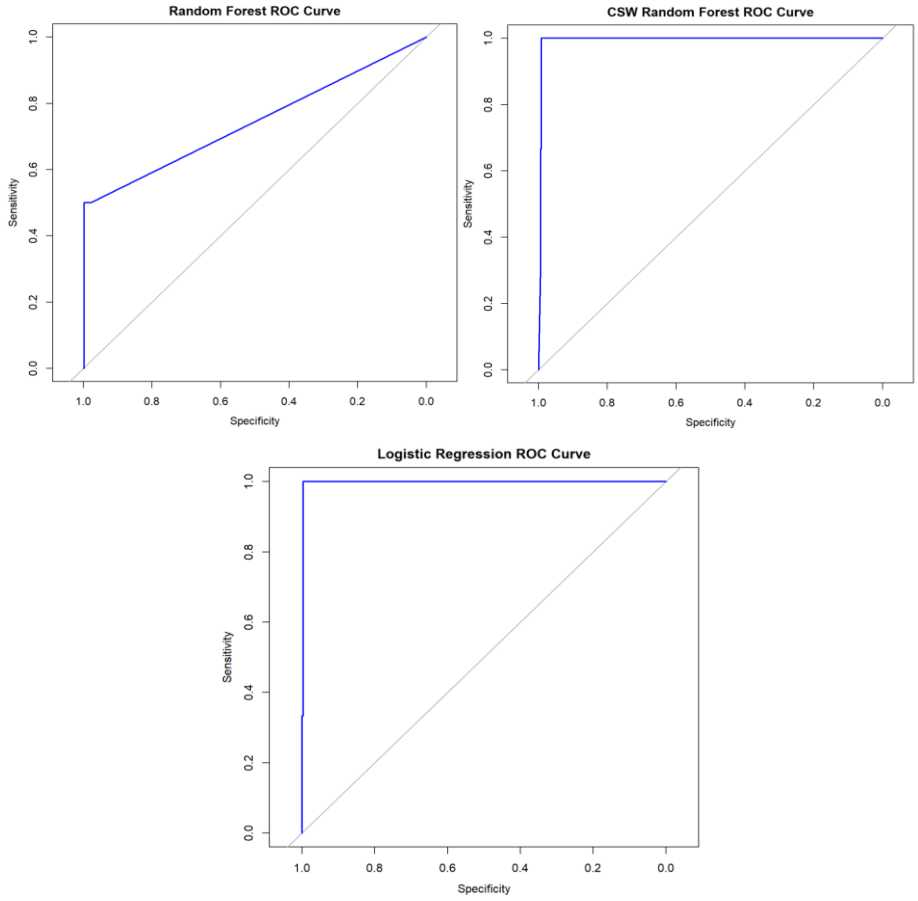

Figure 7. ROC Curve Plot
*(**Source**: R Programming)*

To further evaluate the discriminative performance of the models, Receiver Operating Characteristic (ROC) curves were constructed for the Random Forest (RF), Cost-Sensitive Weighted Random Forest (CSWRF), and weighted logistic regression models, as shown in Fig. 7.

The ROC curve illustrates the trade-off between sensitivity (true positive rate) and specificity (true negative rate) across different classification thresholds. A model with better classification performance will have a curve that is closer to the top-left corner and a higher Area Under the Curve (AUC). Based on the results, the standard Random Forest model exhibits poor discriminative ability, as its ROC curve lies close to the diagonal line. This indicates that the model performs only slightly better than random guessing, which is consistent with previous findings that the RF model fails to properly classify the minority class due to severe class imbalance.

In contrast, the CSWRF model demonstrates a significantly improved ROC curve, which is positioned much closer to the top-left corner. This indicates a strong ability to distinguish between adolescents who engage in premarital sexual behavior and those who do not. The high AUC value (above 0.97) further confirms the excellent discriminative performance of the CSWRF model. This improvement is attributed to the incorporation of class weights, which allows the model to better capture patterns in the minority class.

Similarly, the weighted logistic regression model also shows good discriminative performance, with its ROC curve approaching the top-left region. However, its AUC remains lower than that of CSWRF, indicating slightly inferior performance. This suggests that although weighting improves logistic regression, it is still limited in capturing complex nonlinear relationships compared to ensemble-based methods such as CSWRF.

Overall, the ROC analysis confirms that the CSWRF model outperforms both the standard Random Forest and weighted logistic regression in terms of discriminative ability. This is consistent with the performance results, where CSWRF achieves higher AUC values (above 0.95), along with improved specificity and G-Means compared to the standard Random Forest model, which shows lower AUC (around 0.83) and undefined specificity due to its inability to detect the minority class.

Compared to weighted logistic regression, although the latter demonstrates relatively good balance in terms of specificity and G-Means, its AUC remains lower than that of CSWRF. In addition, the confusion matrix indicates that weighted logistic regression tends to produce a higher number of false positives, suggesting less stable classification performance. These findings highlight the importance of incorporating cost-sensitive learning to improve model performance when dealing with highly imbalanced datasets, as it enhances both discriminative ability and the balance between sensitivity and specificity.

Mean Decrease Accuracy (MDA) is one of the metrics used in the random forest algorithm to assess the importance of each predictor variable used in the model [30]. MDA is calculated by measuring the drop in model accuracy when the values of a particular variable are randomly permuted [31]. The greater the decrease in accuracy, the more important that variable is [32]. In other words, the higher the MDA value, the more important or influential the variable is in contributing to the classification model of premarital sexual behavior. The MDA of CSWRF can be seen in Figure 8.

The results show that Knowledge of Adolescent Reproductive Health (ARH) is the most influential variable. This finding is consistent with Selvia's [33] study, which reported a significant relationship between reproductive health knowledge and premarital sexual attitudes (p-value = 0.037). The results indicate that higher levels of knowledge are associated with more appropriate attitudes toward premarital sexual behavior. This suggests that reproductive health knowledge functions as a protective factor, enabling adolescents to better understand risks and make informed decisions.

This finding is further supported by Fauziah's [34] study, which highlights that limited knowledge of reproductive health increases the likelihood of risky sexual behavior among adolescents. Adolescents who lack adequate knowledge are more vulnerable to engaging in unsafe sexual practices, while those with better knowledge tend to demonstrate more responsible behavior.

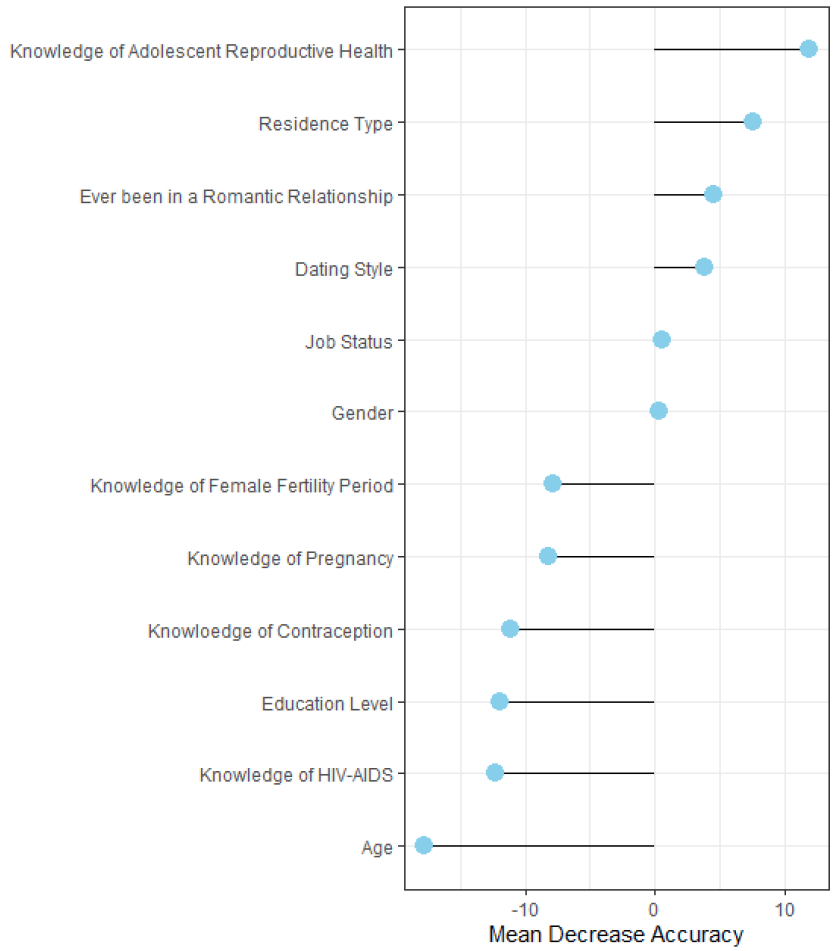

Figure 8. Mean Decrease Accuracy (MDA)
*(**Source**: R Programming)*

The second most important variable is residence type (urban/rural), indicating the role of environmental and socio-cultural factors. This aligns with findings from [34], which emphasize that external factors such as environment, media exposure, and social context significantly influence adolescent behavior. Adolescents in urban areas may experience greater exposure to information and social interaction, which can increase both opportunities and risks related to premarital sexual behavior.

Behavioral variables, including dating style and having been in a romantic relationship, also show substantial importance in the model. This is consistent with [34] findings that peer influence and social interactions play a crucial role in shaping adolescent sexual behavior. Adolescents who are involved in romantic relationships are more likely to experience emotional closeness and peer pressure, which may increase the likelihood of engaging in premarital sexual activity. Similarly, more permissive dating styles reflect behavioral patterns that are closely associated with higher sexual risk.

Interestingly, variables such as age, education level, and HIV/AIDS knowledge show relatively low importance in the model. This finding is in line with Fauziah's [34] study, which reported that the relationship between age and premarital sexual behavior is not always consistent across studies. In some cases, age and education do not directly influence behavior, as their effects may be mediated by other factors such as parental supervision, peer influence, and socio-cultural environment.

Overall, the MDA results indicate that adolescent premarital sexual behavior is influenced by a combination of knowledge, behavioral, and environmental factors. The dominance of ARH knowledge highlights the importance of cognitive factors, while the significance of behavioral and environmental variables reflects the complex and multidimensional nature of adolescent sexual behavior.

The findings of this study have important implications for adolescent reproductive health interventions. From a policy perspective, these findings suggest that interventions should prioritize improving reproductive health knowledge while also addressing behavioral and environmental influences. Educational institutions should strengthen comprehensive reproductive health education, while programs targeting peer influence and

healthy relationship behavior should be developed. In addition, community and family involvement is essential to create a supportive environment that reduces risky sexual behavior among adolescents.

4. CONCLUSION

This study shows that the standard Random Forest model fails to handle extreme class imbalance, as it cannot properly identify the minority class despite high accuracy. The weighted logistic regression model improves classification balance; however, its performance remains inferior to the Cost-Sensitive Weighted Random Forest (CSWRF), particularly in terms of discriminative ability. The CSWRF model demonstrates the best overall performance, as evidenced by improvements in specificity and strong ROC results. The MDA analysis further indicates that Knowledge of Adolescent Reproductive Health (ARH), residence type, and behavioral factors such as dating experience are the most influential predictors of premarital sexual behavior. These findings highlight the importance of knowledge, environmental, and behavioral factors in shaping adolescent behavior. This study contributes by demonstrating the effectiveness of cost-sensitive machine learning for highly imbalanced social data and identifying key determinants of adolescent premarital sexual behavior. Future research should incorporate additional socio-cultural variables and explore more advanced modeling approaches to further improve predictive performance. needed.

Author Contributions

Nilam Novita Sari: Conceptualization, Methodology, Supervision, Validation, Writing – Review and Editing. Bagus Sumargo: Data Curation, Software. Dania Siregar: Formal Analysis, Software, Validation. Endang Yuliani: Formal Analysis, Investigation. Citra Komang Sari: Investigation, Writing – Original Draft, Writing – Review and Editing. Baihaqy Mahardika: Data Curation, Resources, Writing – Original Draft. Khansa Salsabil Fadya: Writing – Original Draft, Writing – Review and Editing. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research was supported by the Jakarta State University under the Faculty Basic Research Scheme in 2025.

Acknowledgement

The authors would like to thank the Faculty of Mathematics and Natural Science, Jakarta State University, for supporting this research and all individuals associated with this research work.

Declarations

The authors declare that there are no conflicts of interest regarding the publication of this study.

Declaration of Generative AI and AI-assisted technologies

Generative AI tools (e.g., ChatGPT) were used solely for language refinement (grammar, spelling, and clarity). The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. The authors reviewed and approved all final text.

REFERENCES

[1] WHO, *Handout for Module A Introduction*. 2018.

[2] BKKBN, "SURVEI KINERJA DAN AKUNTABILITAS PROGRAM KKBPK (SKAP) KELUARGA," *Puslitbang KB dan KS Badan Kependud. dan Kel. Berencana Nas.*, pp. 5–24, 2019, [Online]. Available: https://id.scribd.com/document/492839675/SKAP-KELUARGA-2019

[3] F. Bobyanti, "KeNAKALAN REMAJA," *JERUMI J. Educ. Relig. Humanit. Multidiciplinary*, vol. 1, no. 2, pp. 476–481, 2023, doi: https://doi.org/10.57235/jerumi.v1i2.1402.

[4] Y. Y. Yedemie, "PSYCHOSOCIAL PREDICTORS OF ATTITUDE TOWARD PREMARITAL SEXUAL PRACTICE

AMONG UNIVERSITY STUDENTS, ETHIOPIA," *Front. Psychol.*, vol. 15, no. November, pp. 1–7, 2024, doi: https://doi.org/10.3389/fpsyg.2024.1369964.
[5] R. B. Shrestha, "PREMARITAL SEXUAL BEHAVIOUR AND ITS IMPACT ON HEALTH AMONG ADOLESCENTS," *J. Heal. Promot.*, vol. 7, no. June, pp. 43–52, 2019, doi: https://doi.org/10.3126/jhp.v7i0.25496.
[6] D. L. Sufyan and Y. Nurdiantami, "PEER INFLUENCE AND DATING AS PREDICTORS OF PRE- MARITAL SEXUAL BEHAVIOR AMONG INDONESIA UNMARRIED YOUTH," vol. 30, no. Ichd, pp. 235–241, 2020, doi: https://doi.org/10.2991/ahsr.k.201125.041.
[7] V. Nuraldila and D. S. Yuhandini, "KETERKAITAN PENGETAHUAN TENTANG KESEHATAN REPRODUKSI REMAJA DENGAN PERILAKU SEKS PRA NIKAH PADA SISWA-SISWI KELAS XI DI SMA PGRI 1 KABUPATEN MAJALENGKA TAHUN 2017," *Care J. Ilm. Ilmu Kesehat.*, vol. 5, no. 3, p. 431, 2017, doi: https://doi.org/10.33366/cr.v5i3.710.
[8] T. Kee and W. K. O. Ho, "OPTIMIZING MACHINE LEARNING MODELS FOR URBAN SCIENCES: A COMPARATIVE ANALYSIS OF HYPERPARAMETER TUNING METHODS," *Urban Sci.*, vol. 9, no. 9, 2025, doi: https://doi.org/10.3390/urbansci9090348.
[9] A. Ridwan, U. Muzakir, and S. Nurhidayati, "OPTIMIZING E-COMMERCE INVENTORY TO PREVENT STOCK OUTS USING THE RANDOM FOREST ALGORITHM APPROACH," *Int. J. Softw. Eng. Comput. Sci.*, vol. 4, no. 1, pp. 107–120, 2024, doi: 10.35870/ijsecs.v4i1.2326.
[10] J. L. Speiser, M. E. Miller, J. Tooze, and L. Edward, "A COMPARISON OF RANDOM FOREST VARIABLE SELECTION METHODS FOR CLASSIFICATION PREDICTION MODELING," *Expert Syst. Appl.*, vol. 134, pp. 93–101, 2019, doi: https://doi.org/10.1016/j.eswa.2019.05.028.
[11] V. Ignatenko, A. Surkov, and S. Koltcov, "RANDOM FORESTS WITH PARAMETRIC ENTROPYBASED INFORMATION GAINS FOR CLASSIFICATION AND REGRESSION PROBLEMS," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: https://doi.org/10.7717/peerj-cs.1775.
[12] N. N. Sari, I. Zain, K. Fithriasari, and A. Muhaimin, "BR+ FOR ADDRESSING IMBALANCED MULTILABEL DATA CLASSIFICATION COMBINED WITH RESAMPLING TECHNIQUE," 2021, doi: https://doi.org/10.4108/eai.19-12-2020.2309179.
[13] I. Zain *et al.*, "IMBALANCED DATA ANALYSIS OF ADOLESCENT RISK BEHAVIOR OF DRUG ABUSE USING RANDOM FOREST," 2021, doi: https://doi.org/10.4108/eai.19-12-2020.2309145.
[14] W. A. Tegegne, "SELF-ESTEEM, PEER PRESSURE, AND DEMOGRAPHIC PREDICTORS OF ATTITUDE TOWARD PREMARITAL SEXUAL PRACTICE AMONG FIRST-YEAR STUDENTS OF WOLDIA UNIVERSITY: IMPLICATIONS FOR PSYCHOSOCIAL INTERVENTION," *Front. Psychol.*, vol. 13, no. August, pp. 1–7, 2022, doi: https://doi.org/10.3389/fpsyg.2022.923639.
[15] J. D. Kasingku, A. Hubert, and F. Sanger, "PERAN PENDIDIKAN AGAMA DALAM MEMBENTENGI REMAJA DARI PERGAULAN BEBAS," *Educatio*, vol. 9, no. 4, pp. 2114–2122, 2023, doi: https://doi.org/10.31949/educatio.v9i4.6061.
[16] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, "CROSS-VALIDATION FOR IMBALANCED DATASETS: AVOIDING OVEROPTIMISTIC AND OVERFITTING APPROACHES [RESEARCH FRONTIER]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, pp. 59–76, 2018, doi: https://doi.org/10.1109/MCI.2018.2866730.
[17] H. Akbar and W. K. Sanjaya, "KAJIAN PERFORMA METODE CLASS WEIGHT RANDOM FOREST PADA KLASIFIKASI IMBALANCE DATA KELAS CURAH HUJAN," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 3, no. 1, 2023, doi: https://doi.org/10.20885/snati.v3i1.30.
[18] C. Chen, A. Liaw, and L. Breiman, "USING RANDOM FOREST TO LEARN IMBALANCED DATA," *Univ. California, Berkeley*, no. 1999, p. 24, 2008.
[19] L. Breiman, "RANDOM FORESTS," *Mach. Learining*, vol. 45, pp. 5–32, 2001, doi: https://doi.org/10.1201/9780429469275-8
[20] S. Janitza and R. Hornung, *ON THE OVERESTIMATION OF RANDOM FOREST'S OUT-OF-BAG ERROR*, vol. 13, no. 8. 2018. doi: https://doi.org/10.1371/journal.pone.0201904
[21] F. Cappelli, G. Castronuovo, S. Grimaldi, and V. Telesca, "RANDOM FOREST AND FEATURE IMPORTANCE MEASURES FOR DISCRIMINATING THE MOST INFLUENTIAL ENVIRONMENTAL FACTORS IN PREDICTING CARDIOVASCULAR AND RESPIRATORY DISEASES," *Int. J. Environ. Res. Public Health*, vol. 21, no. 7, 2024, doi: https://doi.org/10.3390/ijerph21070867.
[22] D. Krishnan and S. Singh, "COST-SENSITIVE BOOTSTRAPPED WEIGHTED RANDOM FOREST FOR DOS ATTACK DETECTION IN WIRELESS SENSOR NETWORKS," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2021-Decem, no. December 2021, pp. 375–380, 2021, doi: https://doi.org/10.1109/TENCON54134.2021.9707254.
[23] G. Zeng, "A COMPREHENSIVE STUDY OF COEFFICIENT SIGNS IN WEIGHTED LOGISTIC REGRESSION," *Heliyon*, vol. 10, no. 15, p. e35040, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e35040.
[24] R. A. Danquah, "HANDLING IMBALANCED DATA: A CASE STUDY FOR BINARY CLASS PROBLEMS," 2020, [Online]. Available: http://arxiv.org/abs/2010.04326
[25] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "IMBALANCED DATA PROBLEM IN MACHINE LEARNING: A REVIEW," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3531662.
[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "DEEP SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.doi: https://doi.org/10.1613/jair.953
[27] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF For Noise Identification In Imbalanced Data Classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3413–3423, 2022, doi: https://doi.org/10.1016/j.jksuci.2021.01.014.
[28] I. Zain, A. T. Rumiati, E. O. Permatasari, and N. N. Sari, "IMBALANCED DATA ANALYSIS OF ADOLESCENT UNINTENDED PREGNANCY AND PRE-MARITAL SEX USING UNIVARIATE AND BIVARIATE RANDOM FOREST," *Open Public Health J.*, vol. 16, no. 1, pp. 1–7, 2023, doi: https://doi.org/10.2174/18749445-v16-e230111-2022-65.
[29] A. A. Salih and A. M. Abdulazeez, "EVALUATION OF CLASSIFICATION ALGORITHMS FOR INTRUSION DETECTION SYSTEM: A REVIEW," *J. Soft Comput. Data Min.*, vol. 2, no. 1, pp. 31–40, 2021, doi: https://doi.org/10.30880/jscdm.2021.02.01.004.

[30] Z. Hweju, F. Kopi, and K. Abou-El-Hossein, "PARAMETER IMPORTANCE ANALYSIS: RANDOM FOREST APPROACH," *J. Phys. Conf. Ser.*, vol. 2256, no. 1, 2022, doi: https://doi.org/10.1088/1742-6596/2256/1/012019.

[31] H. Han, X. Guo, and H. Yu, "VARIABLE SELECTION USING MEAN DECREASE ACCURACY AND MEAN DECREASE GINI BASED ON RANDOM FOREST," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 0, pp. 219–224, 2016, doi: https://doi.org/10.1109/ICSESS.2016.7883053.

[32] F. Martinez-Taboada and J. I. Redondo, "THE SIESTA (SEAAV INTEGRATED EVALUATION SEDATION TOOL FOR ANAESTHESIA) PROJECT: INITIAL DEVELOPMENT OF A MULTIFACTORIAL SEDATION ASSESSMENT TOOL FOR DOGS," *PLoS One*, vol. 15, no. 4, pp. 1–10, 2020, doi: https://doi.org/10.1371/journal.pone.0230799.

[33] S. Nur'aini and A. Muhlisin, "THE RELATIONSHIP BETWEEN KNOWLEDGE OF REPRODUCTIVE HEALTH AND ADOLESCENTS ATTITUDES TO PREMARITAL SEXUALITY," *Indones. J. Glob. Heal. Res.*, vol. 2, no. 4, pp. 799–806, 2019, doi: https://doi.org/10.37287/ijghr.v2i4.250.

[34] A. N. Fauziah and S. Maesaroh, "PENGARUH UMUR DAN TINGKAT PENDIDIKAN TERHADAP PERILAKU SEKS PRANIKAH PADA REMAJA DI RW 03 KALURAHAN MOJOSONGO SURAKARTA INFLUENCE THE AGE AND LEVEL EDUCATION TOWARD PREMARITAL SEX BEHAVIOR OF ADOLESCENT OF RW 3 , MOJOSONGO DISTRICT OF SURAKARTA," *Indones. J. Med. Sci.*, vol. 4, no. 2, pp. 202–207, 2017.