BAREKENG: Journal of Mathematics and Its Applications
June 2025 Volume 19 Issue 2 Page 1409-1420
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss2pp1409-1420

A MACHINE LEARNING FRAMEWORK FOR SUICIDAL THOUGHTS PREDICTION USING LOGISTIC REGRESSION AND SMOTE ALGORITHM

Sarni Maniar Berliana[1*], Omas Bulan Samosir[2], Rafidah Abd Karim[3], Victoria Pena Valenzuela[4], Krismanti Tri Wahyuni[5], Andi Alfian[6]

[1,5]*Research Unit of Sustainable Development Goals, Politeknik Statistika STIS*
Jln. Otto Iskandardinata 63C, Jakarta, 13330, Indonesia

[2]*Demographic Institute, Faculty of Economics and Business, Universitas Indonesia*
Jln. Prof. Dr. Sumitro Djojohadikusumo UI, Depok, Jawa Barat 16424, Indonesia

[3]*Academy of Language Studies, University Teknologi MARA*
Perak Branch Tapah Campus, 35400, Malaysia

[4]*Department of Public Administration and Governance,*
College of Social Sciences and Philosophy, Bulacan State University
Diversion Road, Malolos, Bulacan, 3000, Philippines

[6]*BPS-Statistics of East Luwu Regency*
Jln. Ki Hajar Dewantara, Puncak Indah, Kabupaten Luwu Timur, Sulawesi Selatan, 92936 Indonesia

*Corresponding author's e-mail: * **sarni@stis.ac.id***

Article History:

Received: 19th November 2024
Revised: 2nd February 2025
Accepted: 4th March 2025
Published:1st April 2025

Keywords:

Balanced Accuracy;
Imbalanced Data;
Kappa;
Mental Health;
SDG 3;
Sensitivity;
Specificity.

ABSTRACT

Suicide, a global health challenge identified in Goal 3 of the global agenda for enhancing worldwide well-being, demands urgent attention. This study focused on predicting suicidal thoughts using machine learning, leveraging the 2021 National Women's Life Experience Survey (SPHPN) involving women aged 15 to 64. Analyzing 11,305 ever-married women, 504 (4.5%) reported experiencing suicidal thoughts. The outcome variable was binary (1 for suicidal thoughts, 0 for none). The study used seven predictors: age, education level, residence type, physical and sexual violence, smoking frequency, alcohol consumption, and depression. Ordinary logistic regression and SMOTE-based logistic regression were applied. The former identified physical violence, depression, and sexual violence as crucial factors, while the latter emphasized physical violence, sexual violence, and age. In cases of class imbalance, the SMOTE-enhanced model exhibited improved performance in terms of sensitivity, false positive rate, balanced accuracy, and Kappa statistic, with lower standard errors of parameter estimates. The findings highlight the importance of addressing violence and mental health in policies aimed at reducing suicidal thoughts among women. Policymakers can use these insights to develop targeted interventions, and healthcare providers can identify high-risk individuals for timely interventions. Community programs and public health campaigns should promote mental well-being and prevent suicidal behaviors using these findings. Future research should include more predictors, diverse populations, and longitudinal data to better understand causal relationships and timing. Interdisciplinary collaboration and advanced machine learning techniques can enhance predictive accuracy and model interpretability.

How to cite this article:

S. M. Berliana, O. B. Samosir, R. A. Karim, V. P. Valenzuela, K. T. Wahyuni and A. Alfian., "A MACHINE LEARNING FRAMEWORK FOR SUICIDAL THOUGHTS PREDICTION USING LOGISTIC REGRESSION AND SMOTE ALGORITHM," *BAREKENG: J. Math. & App.*, vol. 19, iss. 2, pp. 1409-1420, June, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

The intentional killing of oneself is known as suicide [1]. Globally, suicide is a persistent public health concern. Suicide takes the lives of 703,000 individuals worldwide each year, accounting for more than one out of every 100 deaths (1.3%) in 2019 [2]. Suicide is one of the leading causes of death worldwide, claiming more lives than homicide, HIV/AIDS, breast cancer, malaria, and war combined. Suicidal thought, sometimes referred to as suicidal ideation, is the contemplation of terminating one's life without necessarily acting on it, whereas suicide is the act of taking one's own life [3]. The United Nations Agency for Health highlights the significance of mental health promotion and suicide prevention as part of its global health agenda because this issue is a significant public health concern worldwide [4].

Mental health conditions like substance misuse, anxiety, and depression are frequently linked to suicidal thoughts. The global goals for sustainable development (SDGs), established by the United Nations in 2015, include specific mental health and well-being objectives. The third goal focuses on ensuring healthy lifestyles and promoting well-being for individuals of all ages. As part of this objective, Target 3.4 seeks to improve mental health and well-being and prevent and treat non-communicable diseases to reduce premature mortality by one-third [5]. The suicide mortality rate (Indicator 3.4.2), which emphasizes the worldwide commitment to lowering suicide rates, is one of the indicators for this goal.

Suicidal thoughts are critical to address because they are often indicative of underlying mental health issues that require immediate attention. These thoughts can vary in intensity and frequency, ranging from fleeting considerations to detailed planning [6]. These thoughts can be a precursor to more severe actions, such as suicide attempts or completed suicides, which have devastating effects on individuals, families, and communities. Moreover, suicidal ideation is often associated with feelings of hopelessness, despair, and intense emotional pain, which can significantly disrupt a person's ability to carry out daily activities [7]. By addressing these thoughts, we can help individuals regain a sense of hope and purpose, ultimately improving their overall well-being [8], [9]. Recognizing and addressing suicidal thoughts early can prevent these tragic outcomes and provide individuals with the support and treatment they need to improve their mental health.

Predicting suicidal thoughts involves identifying various risk factors and features that can indicate an individual's likelihood of experiencing such thoughts. Extensive research has been conducted on the determinants of suicidal thoughts, both in Indonesia and in other countries in the world. A study by [10] examining suicidal thoughts and attempts among Indonesian adolescent students discovered a significant association between these behaviors and factors, such as being bullied, feeling lonely, and having no close friends. The prevalence of suicidal ideation was higher among girls, while boys were more likely to report suicide attempts. Another study by [11] predicted self-harm and suicidal ideation during the COVID-19 pandemic in Indonesia. The findings revealed that, during the pandemic, 39.3% of participants reported self-harm and suicidal thoughts. Age, sex, religion, employment, place of residence, gender, sexual orientation, loneliness, disability status, and HIV status were key predictors. The other study by [12] examines the prevalence and determinants of suicidal behavior among adolescents in Bangladesh and Indonesia. The study utilizes data from a survey of global health for students (GSHS) to pinpoint essential factors linked to suicidal thoughts and attempts. The findings reveal that a significant proportion of adolescents in both Bangladesh and Indonesia experience suicidal behavior, with various socio-demographic and psychosocial factors contributing to this issue. Factors such as bullying, lack of parental support, and mental health issues are highlighted as significant contributors.

Given the categorical nature of the outcome variable, i.e., whether or not one has ever thought about committing suicide, logistic regression is commonly used to identify determinants [10], [11], [12]. However, class imbalance may make ordinary logistic regression biased towards the majority class because the model predicts the majority class more frequently, resulting in deceptively high accuracy while performing poorly on the minority class [13]. Imbalance classes cases are commonly found in medical diagnosis [14], fraud detection [15], [16], [17], machine fault detection [18], and defect classification [19].

Several methods can be applied to handle imbalance classes, some of which are the undersampling method [20], oversampling method [21], cluster-based algorithm [22], [23], and synthetic minority oversampling technique (SMOTE) algorithm [24]. Undersampling involves reducing the number of samples in the majority category to achieve a balanced class distribution. Undersampling minimizes the number of samples in the majority class to create a more balanced class distribution. This process can be performed randomly or by selecting the most informative samples. In contrast, random oversampling increases the representation of the minority class by duplicating its samples until both classes are balanced. A cluster-based

algorithm addresses both within-class and between-class imbalances simultaneously. Meanwhile, the SMOTE algorithm produces artificial samples for the underrepresented class by interpolating between existing samples within that class.

Studies consistently show that women report higher rates of suicidal ideation compared to men [25] [26]. This trend is observed across various age groups and cultural contexts [27], [28]. Women face unique risk factors such as hormonal changes [29], [30], [31], postpartum depression [32], [33], and higher rates of sexual abuse [34], [35] and domestic violence [36], [37]. Addressing these gender-specific risk factors require targeted research and interventions that consider the unique experiences of women. As far as we know, no research has utilized national-level data to study suicidal thoughts among Indonesian women. Therefore, this study seeks to identify the factors that contribute to suicidal thoughts among women in Indonesia and develop a predictive model for suicidal thoughts using a machine learning algorithm. This study's data showed a class imbalance between the women who had suicidal thoughts and the women who had never thought suicide; the ordinary classification method, which generally assumes proportional classes, is not appropriate to apply. This study will utilize the SMOTE algorithm, a popular method for addressing class imbalance in datasets, particularly in machine learning.

2. RESEARCH METHODS

2.1 Binary Logistic Regression

Binary logistic regression is employed when the outcome variable data type is categorical with a value of 0 or 1. This logistic regression model shows which predictor variables have significant effects on suicidal thoughts among women in Indonesia. Since the logistic regression model is not linear, the odds ratio is interpreted in order to analyze the model. The binary logistic regression model with k independent variables (x_i) and n observations is stated as [38].

$$\pi(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})},\ i = 1,\ldots,n. \tag{1}$$

where $\pi(\mathbf{x}_i)$ represents the probability that an observation belongs to a particular category of the binary outcome variable, commonly referred to as the "success probability" $\left(P(Y|\mathbf{x}_i)=1\right)$, $\mathbf{x}_i = \begin{bmatrix}1 & x_{i1} & x_{i2} & \cdots & x_{iM}\end{bmatrix}^T_{(M+1)\times 1}$, and $\boldsymbol{\beta} = \begin{bmatrix}\beta_0 & \beta_1 & \beta_2 & \cdots & \beta_M\end{bmatrix}^T_{(M+1)\times 1}$. The following logistic regression model is produced by applying a logit transformation to **Equation (1)**.

$$g(\mathbf{x}_i) = \ln\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \mathbf{x}_i^T\boldsymbol{\beta}, \tag{2}$$

The odds ratio (*OR*) is a metric that quantifies the relationship between a predictor variable and an outcome variable in a logistic regression model. It represents the ratio of the odds of an event happening in one group compared to the odds of it happening in another group. The odds of an event are calculated as the probability of the event happening divided by the probability of it not happening. In binary logistic regression, the odds of success are [39].

$$\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} = \exp(\mathbf{x}_i^T\boldsymbol{\beta}) \tag{3}$$

The odds ratio compares the odds of an outcome happening when a specific predictor variable is present to the chances of it happening when the predictor is absent. In this case, the odds of success when there is only one predictor are:

$$\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} = \exp(\beta_0 + \beta_1 x_1)$$

If we increase x_1 by one unit, the odds ratio is:

$$OR = \frac{\exp(\beta_0 + \beta_1(x_1 + 1))}{\exp(\beta_0 + \beta_1 x_1)} = \exp(\beta_1).$$

2.2 SMOTE Algorithm

One approach to addressing class imbalance in datasets is the SMOTE algorithm, which creates synthetic samples for the minority class to balance the dataset. This helps improve model performance by allowing the model to better learn the characteristics of the minority class. The primary goal of using SMOTE is to shift the classification decision boundary to favor the minority class. To achieve this, the technique generates synthetic samples that closely resemble the original minority class samples, based on the idea that the densities of the $K-$nearest neighbors are similar. The studies in **[40] [41]** provide an in-depth exploration of the probability density function of the patterns generated by SMOTE. These papers offer detailed mathematical formulations of the probability distribution for those interested.

The SMOTE oversampling algorithm developed by **[24]** follows these steps

a. Select a minority class sample

   Choose a sample at random, denoted as x_i, from the minority group.

b. Find $K-$nearest neighbors

   i. Determine the $K-$nearest neighbors of x_i within the minority class using a distance measure, typically Euclidean distance.

   ii. The formula for Euclidean distance between x_i and x_j is $d(x_i, x_j) = \left(\sum_{m=1}^{M}(x_{im} - x_{jm})^2\right)^{1/2}$ where x_{im} and x_{jm} are the $m-$th features of samples x_i and x_j $(m = 1,\ldots,M)$, respectively.

c. Randomly choose one of the $K-$nearest neighbors

   Choose one of the $K-$nearest neighbors at random, denoted as x_{zi}.

d. Generate a synthetic sample

   i. Create a new synthetic sample x_{new} by interpolating between x_i and x_{zi}.

   ii. The interpolation formula is $x_{\text{new}} = x_i + \delta(x_{zi} - x_i)$ where δ is a random number between 0 and 1.

e. Repeat Steps 1-4

   Continue repeating the steps outlined above until the required number of synthetic samples is created.

2.3 Model Evaluation

A confusion matrix is employed to assess the effectiveness of a model in a classification problem. It shows the counts of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*).

From this matrix, several key evaluation metrics—such as accuracy, sensitivity, specificity, and Kappa—are derived to assess the effectiveness of a logistic regression model. Accuracy represents the ratio of correct predictions (including both true positives and true negatives) to the total number of predictions. Sensitivity measures the percentage of true positives that the model accurately identifies. Specificity, on the other hand, quantifies the proportion of true negatives that the model correctly identifies. Kappa measures the level of agreement between the observed and expected accuracy (based on random chance), adjusting for the likelihood of agreement happening by chance. The following are the formulas for each metric **[42]**.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$Kappa = \frac{2(TP \cdot TN - FN \cdot FP)}{(TP + FP)(FP + TN) + (TP + FN)(FN + TN)} \tag{7}$$

where the confusion matrix is presented in **Table 1**.

An accuracy value of 1 (100%) means all predictions are correct. However, accuracy can be misleading in imbalanced datasets where one class is much more prevalent than the other. When the sensitivity value is 1, the model correctly identifies all positive cases. Similarly, when the specificity value is 1, the model accurately identifies all negative cases. A Kappa value 1 indicates perfect agreement between the model's predictions and the actual outcomes. A Kappa value of 0 suggests no agreement beyond what would be expected by chance, while negative values indicate a deal worse than chance.

Table 1. Confusion Matrix

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

To evaluate the model's prediction performance, both "in-sample" data (the data used to train the model) and "out-of-sample" data (new, unseen data not involved in the model fitting process) are utilized. Eighty percent of the data is used to train the model and make in-sample predictions. The remaining portion, the validation (or test) data set, tests the model's ability to predict out-of-sample values on data it has not encountered before. Four metrics—sensitivity or true positive rate (*TPR*), false positive rate (*FPR*) calculated from $1 - Specificity$, balanced accuracy obtained from $(Sensitivity + Specificity)/2$, and Kappa—are used to compare the performance of each model.

2.4 Data and Research Variables

We used the 2021 National Women's Life Experience Survey (SPHPN), which included female respondents ages 15 to 64. The SPHPN is a survey conducted by the Ministry of Women's Empowerment and Child Protection (Kemen PPPA) in collaboration with BPS-Statistics Indonesia and the Demographic Institute of the University of Indonesia (LDUI). This survey aims to obtain information related to the prevalence of violence toward women in Indonesia at the national estimate level. The number of samples is 12,800 households spread across 34 provinces in 160 regencies/municipalities spread across 946 sub-districts with a response rate of 94.85% **[43]**.

The outcome variable is binary with values 1 (having suicidal thoughts) and 0 (not having suicidal thoughts). Seven predictors are used in this study, which are age (continue), an education level (0=no education, 1=incomplete primary, 2=complete primary, 3=complete secondary, 4=higher), residence type (0=rural, 1=urban), the experience of physical violence (0=no, 1=yes), the experience of sexual violence (0=no, 1=yes), smoking frequency (0=never, 1=sometimes, 2=everyday), alcohol consumption (0=no, 1=yes), and depression experience (0=no, 1=yes).

3. RESULTS AND DISCUSSION

The 2021 SPHPN data was used to obtain a unit of analysis consisting of 11,305 ever-married women in this study, and 504 of them (4.5%) reported having suicidal thoughts at least once in their lifetime, with the mean age of women was 41.6 years. There is an imbalance in the dataset, as those who have suicidal thoughts are less likely to occur than those who do not. Imbalanced data is common in an outcome variable and can affect model performance. For this reason, using logistic regression with the SMOTE algorithm instead of the standard logistic regression model is appropriate.

Table 2. Distribution Percentage of Women Who Have Had Suicidal Thoughts by Background Characteristics, Indonesia, 2021

Background Characteristics	Suicide Thoughts				Total	
	No		Yes			
	n	%	n	%	n	%
Urban/rural classification						
Rural	4,823	96.2%	191	3.8%	5,014	100%
Urban	5,978	95.0%	313	5.0%	6,291	100%
Education level						
No education	378	95.0%	20	5.0%	398	100%
Incomplete primary	1,170	94.1%	74	5.9%	1,244	100%
Complete primary	3,293	96.8%	110	3.2%	3,403	100%
Complete secondary	5,002	95.3%	249	4.7%	5,251	100%
Higher	958	94.9%	51	5.1%	1,009	100%
Smoking frequency						
Never	10,447	95.7%	467	4.3%	10,914	100%
Sometimes	239	92.3%	20	7.7%	259	100%
Everyday	115	87.1%	17	12.9%	132	100%
Alcohol consumption						
No	10,737	95.6%	490	4.4%	11,227	100%
Yes	64	82.1%	14	17.9%	78	100%
Experience of physical						
No	10,011	96.5%	358	3.5%	10,369	100%
Yes	790	84.4%	146	15.6%	936	100%
Experience of sexual						
No	10,293	96.3%	397	3.7%	10,690	100%
Yes	508	82.6%	107	17.4%	615	100%
Depression						
No	10,782	95.7%	482	4.3%	11,264	100%
Yes	19	46.3%	22	53.7%	41	100%
Total	**10,801**	**95.5%**	**504**	**4.5%**	**11,305**	**100%**

As shown in Table 2, over half of the study sample—55.6% and 55.4%, respectively—lived in urban areas and completed secondary school. Ninety-six percent of women had never smoked, and 99.3% had never drank alcohol. The percentage of women who experienced physical and sexual violence was 5.4% and 8.3%, respectively. On the other hand, 99.6% of women reported never having experienced depression. Women who live in cities, smoke, consume alcohol, have experienced physical or sexual violence, or suffer from depression are more likely to have suicidal thoughts. Meanwhile, women with higher or lower levels of education are more likely to have suicidal thoughts.

3.1 Factors Associated with Suicidal Thoughts Among Women in Indonesia

Table 3 displays the parameter estimates for the traditional logistic regression model using the training data. Except for education level and place of residence, all predictors have significant effects on suicidal thoughts. Women who have higher education levels, live in urban areas, have experienced physical or sexual

violence, smoke, consume alcohol, or suffer from depression are more likely to have suicidal thoughts. Meanwhile, a one-year increase in women's age will decrease the likelihood of suicidal thoughts.

Table 3. Ordinary Logistic Regression for Suicidal Thoughts, Indonesia, 2021

	Coefficient	Std. Error	*z* value	*p*-value
Intercept	-3.47	0.28	-12.25	0.00
Age	-0.16	0.06	-2.88	0.00
Incomplete primary	0.24	0.31	0.78	0.44
Complete primary	-0.39	0.30	-1.33	0.18
Complete secondary	-0.06	0.29	-0.22	0.83
Higher	0.08	0.33	0.23	0.82
Urban	0.18	0.11	1.59	0.11
Having physical violence experience	1.25	0.13	9.30	0.00
Having sexual violence experience	0.95	0.16	6.13	0.00
Sometimes smoking	0.21	0.28	0.74	0.46
Everyday smoking	1.05	0.32	3.33	0.00
Consuming alcohol	0.86	0.37	2.30	0.02
Having depression experience	2.80	0.38	7.42	0.00

For the second model, we ran a logistic regression model combined with the SMOTE algorithm to increase the representation of the minority class. After oversampling the outcome variable, we achieved an equal representation of each class, totaling 8,641 observations for women with suicidal thoughts and those without. The number of observations for each class of variables used in this study is shown in **Table 4**.

After oversampling the outcome variable, we trained a new model using the updated dataset. The parameter estimates based on the training data are presented in **Table 5**. All predictors have significant effects on suicidal thoughts. Women who live in urban areas have experienced physical or sexual violence, smoke, consume alcohol, or suffer from depression are more likely to have suicidal thoughts. Similar to the finding in the traditional logistic regression model, a one-year increase in women's age will decrease the likelihood of suicidal thoughts.

Table 4. Number of Observations After Applying SMOTE Algorithm

Class	Education Level	Area	Physical Violence	Sexual Violence	Smoking Frequency	Alcohol Consumption	Depression	Suicidal Thoughts
1	622	7,234	14,161	15,147	16,418	17,035	17,031	8,641
2	2,078	10,048	3,121	2,135	490	247	251	8,641
3	4,429				374			
4	8,505							
5	1,648							
Total	**17,282**	**17,282**	**17,282**	**17,282**	**17,282**	**17,282**	**17,282**	**17,282**

Table 5. Logistic Regression using SMOTE Algorithm for Suicidal Thoughts, Indonesia, 2021

	Coefficient	Std. Error	*z* value	*p*-value
Intercept	-0.14	0.09	-1.59	0.11
Age	-0.17	0.02	-9.52	0.00
Incomplete primary	-0.01	0.10	-0.15	0.88
Complete primary	-0.69	0.09	-7.66	0.00
Complete secondary	-0.32	0.09	-3.65	0.00
Higher	-0.30	0.10	-2.99	0.00
Urban	0.22	0.03	6.41	0.00
Having physical violence experience	1.28	0.05	25.18	0.00
Having sexual violence experience	1.07	0.06	17.29	0.00
Sometimes smoking	-0.06	0.10	-0.60	0.55
Everyday smoking	0.98	0.13	7.31	0.00

Consuming alcohol	0.54	0.18	3.05	0.00
Having depression experience	2.34	0.26	8.90	0.00

We produced a feature importance visualization, which is shown in **Figure 1**, as an additional means of comprehending the obtained model. The visualization shows which variables have high feature importance scores. It is evident that experiencing physical violence, depression, and sexual violence are all important features. This visualization also suggests that higher education does not appear to be useful.

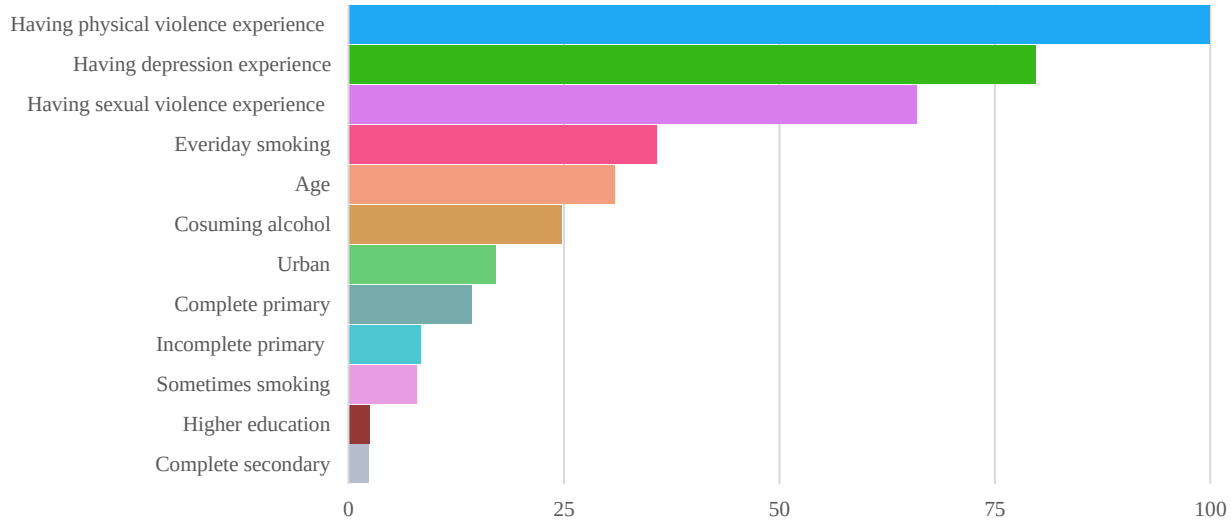

Figure 1. Feature Importance of Ordinary Logistic Regression

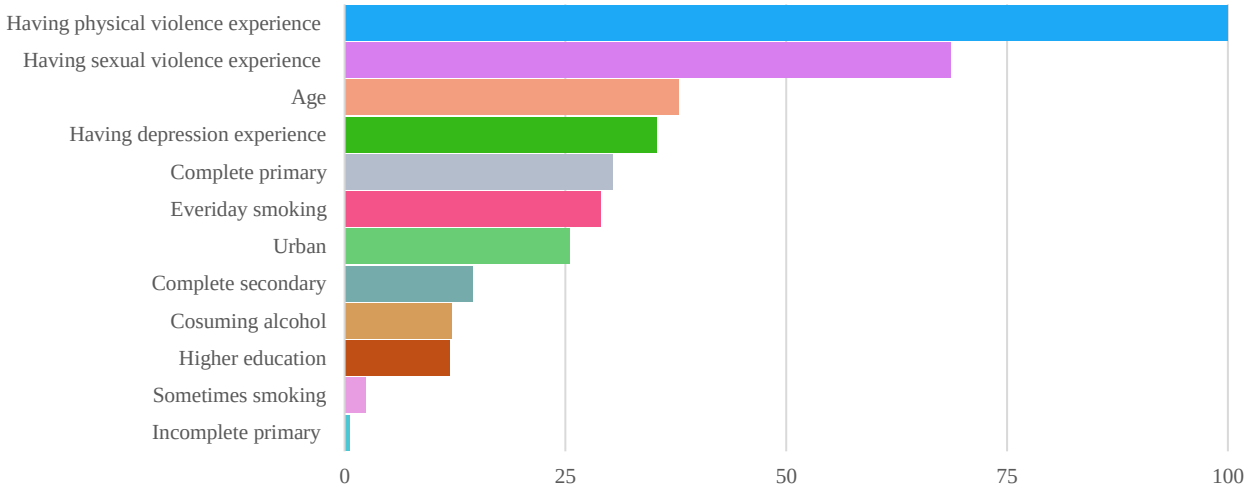

Figure 2. Feature Importance of Logistic Regression using SMOTE Algorithm

The visualization of feature importance for the logistic regression model that fitted using SMOTE-based data is presented in **Figure 2**. As before, the visualization shows that experiencing physical violence and sexual violence are important features. The third important feature is age. This visualization also shows that incomplete primary education is not an important feature.

3.2 Model Performance Evaluation

The estimates shown in **Table 2** are calculated using the default probability of success of $p = 0.5$. The estimates produced are based on the assumption of a proportional number of cases between women who have had suicidal thoughts and those who have not. However, as shown in **Table 1**, the distribution for each class of the outcome variable suicidal thoughts is not 50/50. In this study, we calculate a more suitable threshold to optimize the predictions for each class. We obtain an optimum threshold of 0.043, which should give more reliable predictions instead of the default threshold 0.5. We then compare the model performance using the new threshold and the model using the default 0.5 threshold. The confusion matrices for the default threshold

of 0.5 and the threshold of 0.043 are shown in **Table 6**. The confusion matrices are calculated for both in-sample and out-of-sample observations.

Table 6. Evaluation Metrics for Ordinary Logistic Regression and SMOTE Algorithm

Metric	Ordinary Logistic Regression				SMOTE Algorithm	
	In-Sample		Out-of-Sample		In-Sample	Out-of-Sample
	$p = 0.5$	$p = 0.043$	$p = 0.5$	$p = 0.043$		
Accuracy	0.9559	0.8166	0.9566	0.8274	0.6535	0.8226
Sensitivity	0.9990	0.8318	0.9986	0.8444	0.8346	0.8407
Specificity	0.0347	0.4901	0.0500	0.4600	0.4724	0.4300
Kappa	0.0611	0.1303	0.0866	0.1298	0.3070	0.1143

With a 0.5 threshold, the accuracy is high for both in-sample and out-of-sample data, but the Kappa statistic and specificity are significantly low. This condition means our sample has a tendency to predict women who have had suicidal thoughts more often than they have. Changing the threshold, we obtained lower accuracy for both in-sample and out-of-sample, but the specificity and Kappa statistic increased. The sensitivity for both in-sample and out-of-sample cases decreased when the 0.043 threshold was used. This means the true cases of women who have had suicidal thoughts are less likely to be found. These findings highlight the importance of threshold choice when handling imbalanced data sets.

The evaluation metrics for the oversampled data using SMOTE are also displayed in **Table 6**. The accuracy became lower when applying the SMOTE algorithm relative to a regular data set. However, oversampling the minority class, as evidenced by the specificity and Kappa statistic values, improves the outcome variable's predictions. In the meantime, when the SMOTE algorithm was applied to the regular data, the sensitivity decreased. It is necessary to consider the costs associated with reducing the sensitivity. The model with the SMOTE algorithm and the model with the 0.043 threshold has better prediction performances than the model with the standard 0.5 threshold. It is also worth mentioning that, as **Table 5** and **Table 3** show, the standard errors for the estimated parameters in the logistic regression model using the SMOTE algorithm are smaller than those of the standard errors of the traditional logistic regression model. This means that logistic regression using the SMOTE algorithm is better than the ordinary logistic regression model.

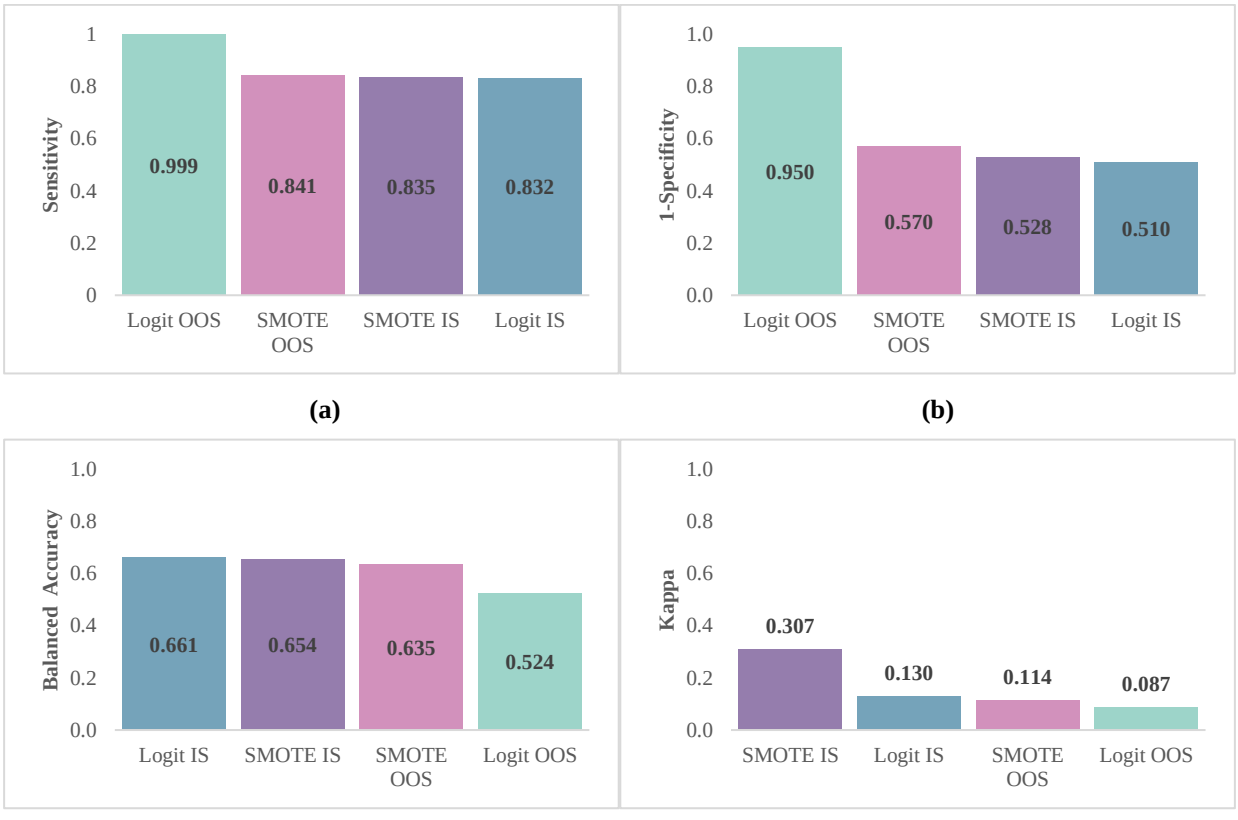

Figure 3. Model Performance by (a) Sensitivity (True Positive Rate), (b) False Positive Rate (1-Specificity), (c) Balanced Accuracy, and (d) Kappa

The differences in model performances for each model are easier to observe using visualizations. We evaluate the performances of the ordinary logistic regression model and the logistic regression model using the SMOTE algorithm, both in-sample (IS) and out-of-sample (OOS). We visualize four metrics, which are the sensitivity (true positive rate), the false positive rate, the balanced accuracy, and the Kappa, as shown in **Figure 3(a)** – **Figure 3(d)**, respectively.

The sensitivity of the standard logistic regression model out-of-sample performs very well, as shown in **Figure 3(a)**, indicating that it may predict which women have had suicidal thoughts took place. However, further analysis must be conducted to determine whether these results are reliable. The SMOTE algorithm-based logistic regression performs similarly in-sample and out-of-sample, which indicates that it generalizes well.

Figure 3(b), shows that the ordinary logistic regression model has a higher out-of-sample false positive rate than its in-sample counterpart and the logistic regression model based on the SMOTE algorithm. This indicates that the high sensitivity in **Figure 3(a)** was unreliable. Given its high false positive rate, it seems it achieved a high sensitivity merely by correctly predicting the majority class. Although the false positive rate is higher than we would prefer, we find a good balance in and out-of-sample for the logistic regression model based on the SMOTE method.

Figure 3(c) shows the balanced accuracy, which accounts for any possible class imbalance by considering both the specificity and the sensitivity. The logistic regression model based on the SMOTE algorithm is slightly lower than the in-sample logistic regression model for balanced accuracy. However, using the ordinary logistic regression model, we observe a discrepancy between in-sample and out-of-sample balanced accuracy. Meanwhile, the balanced accuracy of the logistic regression model based on the SMOTE algorithm for both in-sample and out-of-sample are relatively similar, indicating that this model generalizes better.

The last figure visualizes the Kappa statistic, as shown in **Figure 3(d)**. When we apply the SMOTE algorithm, the in-sample Kappa statistic is the highest, and the lowest Kappa statistic is for the out-of-sample ordinary logistic regression model. These findings confirm that applying the SMOTE algorithm increases the predictive performance of the logistic regression model that exhibits imbalance classes.

3.3 Limitations

This study focuses on ever-married women aged 15 to 64 from the 2021 National Women's Life Experience Survey (SPHPN). As a result, it may not fully represent suicidal thoughts among never-married women or women outside this age range, which limits the generalizability of the results. The data on suicidal thoughts and other predictors rely on self-reports, which can be affected by response biases, inaccuracies, and underreporting, particularly on sensitive subjects like mental health and violence. Additionally, the cross-sectional design of the study captures data at a single point in time, making it difficult to infer causality or understand the temporal relationships between predictors and suicidal thoughts.

4. CONCLUSIONS

In summary, our study presents several key findings and recommendations:

a. The ordinary logistic regression model identifies physical violence, depression, and sexual violence as the three most influential factors associated with suicidal thoughts. Meanwhile, physical violence, sexual violence, and age are the three most important features based on the SMOTE algorithm. In imbalance class cases, the logistic regression model using the SMOTE algorithm gives better performance in predictions based on the sensitivity, false positive rate, balanced accuracy, and Kappa statistic. The SMOTE algorithm-based logistic regression generalizes better based on the relatively balanced values for each evaluation metric used in this study. Moreover, the standard errors of the SMOTE algorithm-based logistic regression parameter estimates are smaller than those of the ordinary logistic regression model.

b. The findings highlight the importance of addressing physical and sexual violence, as well as mental health conditions like depression, in policies aimed at reducing suicidal thoughts among women. Policymakers can use these insights to develop targeted interventions and support programs. Healthcare providers can utilize the identified predictors to screen and identify

individuals at higher risk of suicidal thoughts, thereby providing timely and appropriate interventions. This research underscores the need for increased awareness and education on the impacts of violence and mental health issues. Community programs and public health campaigns can leverage these findings to promote mental well-being and prevent suicidal behaviors.

c. Future research can expand upon these findings by incorporating a wider array of predictors, examining diverse population groups, and employing longitudinal data to unravel the causal relationships and timing of suicidal thoughts. Encouraging interdisciplinary collaboration among experts in psychology, sociology, public health, and data science is essential to developing a more comprehensive understanding of the factors contributing to suicidal thoughts. Additionally, applying advanced machine learning techniques such as random forests, support vector machines, or neural networks can enhance predictive accuracy and model interpretability.

ACKNOWLEDGMENT

We are thankful to the Center for Research and Community Service (PPPM) Politeknik Statistika STIS for the support to complete this study.

REFERENCES

[1] World Health Organization, "Suicide," World Health Organization, 2024. [Online]. Available: https://www.emro.who.int/health-topics/suicide/feed/atom.html. [Accessed 15 October 2024].

[2] World Health Organization, "Suicide worldwide in 2019: Global Health Estimates," 2021.

[3] E. D. Klonsky, A. M. May and B. Y. Saffer, "Suicide, Suicide Attempts, and Suicidal Ideation," *Annual Review of Clinical Psychology*, vol. 12, pp. 307-330, 2016.

[4] World Health Organization, "Mental Health, Brain Health and Substance Use," World Health Organization, June 2021. [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data. [Accessed 15 October 2024].

[5] World Health Organization, "SDG Target 3.4 Non-communicable diseases and mental health," 2024. [Online]. Available: https://www.who.int/data/gho/data/themes/topics/sdg-target-3_4-noncommunicable-diseases-and-mental-health. [Accessed 2024 October 2024].

[6] T. Joiner and M. D. Rudd, Suicide Science: Expanding the Boundaries, Kluwer Academic Publishers, 2000.

[7] J. D. Ribeiro, X. Huang, K. R. Fox and J. C. Franklin, "Depression and hopelessness as risk factors for suicide ideation, attempts and death: meta-analysis of longitudinal studies," *The British Journal of Psychiatry*, vol. 212, no. 5, pp. 279-286, 2018.

[8] World Health Organization, World mental health report: transforming mental health for all, 2022.

[9] M. K. Nock, G. Borges and Y. Ono, Suicide: Global Perspectives from the WHO World Mental Health Surveys, Cambridge University Press, 2012.

[10] I. G. N. E. Putra, P. A. E. S. Karin and N. L. P. Ariastuti, "Suicidal ideation and suicide attempt among Indonesian adolescent students," *International Journal of Adolescent Medicine and Health*, vol. 33, no. 5, pp. 1-12, 2021.

[11] A. Liem, B. Prawira, S. Magdalena, M. J. Siandita and J. Hudiyana, "Predicting self-harm and suicide ideation during the COVID-19 pandemic in Indonesia: a nationwide survey report," *BMC Psychiatry*, vol. 22, pp. 1-10, 2022.

[12] M. Marthoenis and S. M. Y. Arafat, "Rate and Associated Factors of Suicidal Behavior among Adolescents in Bangladesh and Indonesia: Global School-Based Student Health Survey Data Analysis," *Scientifica*, 2022.

[13] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, New York, Springer Science, 2005, pp. 853-867.

[14] R. B. Rao, S. Krishnan and R. S. Niculescu, "Data mining for improved cardiac care," *ACM SIGKDD Explorations Newsletter*, vol. 8, no. 1, pp. 3-10, 2006.

[15] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 27, pp. 1-54, 2018.

[16] W. Wei, J. Li, L. Cao, Y. Ou and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol. 16, p. 449–475, 2013.

[17] M. Herland, T. M. Khoshgoftaar and R. A. Bauder, "Big Data fraud detection using multiple medicare data sources," *Journal of Big Data*, vol. 5, no. 1, pp. 1-21, 2018.

[18] N. Kerdprasop and K. Kerdprasop, "Data preparation techniques for improving rare class prediction," in *Recent Researches in Computational Techniques, Non-Linear Systems and Control*, 2011.

[19] P. B. Polak, J. D. Prusa and T. M. Khoshgoftaar, "Low-shot learning and class imbalance: a survey," *Journal of Big Data*, vol. 11, pp. 1-37, 2024.

[20] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection," in *Proceedings of the 14th International Conference on Machine Learning, 179-186*, San Francisco, 1997.

[21] C. X. Ling and C. Li, "Data mining for direct marketing: problems and solutions," in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD98)*, New York, 1998.

[22] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 40-49, 2004.

[23] W.-C. Lin, C.-F. Tsai, Y.-H. Hu and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, Vols. 409-410, pp. 17-26, 2017.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321-357, 2002.

[25] S. S. Canetto and I. Sakinofsky, "The Gender Paradox in Suicide," *Suicide and Life-Threatening Behavior*, vol. 28, no. 1, pp. 1-23, 1998.

[26] D. L. Schrijvers, J. Bollen and B. G. Sabbe, "The gender paradox in suicidal behavior and its impact on the suicidal process," *Journal of Affective Disorders*, vol. 138, no. 1-2, pp. 19-26, 2012.

[27] L. M. Range and M. M. Leach, "Gender, Culture, and Suicidal Behavior: A Feminist Critique of Theories and Research," *Suicide and Life-Threatening Behavior*, vol. 28, no. 1, pp. 24-36, 1998.

[28] K. R. Andriolo, "Gender and the Cultural Construction of Good and Bad Suicides," *Suicide and Life-Threatening Behavior*, vol. 28, no. 1, pp. 37-49, 1998.

[29] A. Wikman, J. Sacher, M. Bixo, A. L. Hirschberg, H. K. Kallner, C. N. Epperson, E. Comasco and I. S. Poromaa, "Prevalence and correlates of current suicidal ideation in women with premenstrual dysphoric disorder," *BMC Women's Health*, vol. 22, pp. 1-7, 2022.

[30] T. Eisenlohr-Moul, M. Divine, K. Schmalenberger, L. Murphy, B. Buchert, M. Wagner-Schuman, A. Kania, S. Raja, A. B. Miller, J. Barone and J. Ross, "Prevalence of lifetime self-injurious thoughts and behaviors in a global sample of 599 patients reporting prospectively confirmed diagnosis with premenstrual dysphoric disorder," *BMC Psychiatry*, vol. 22, pp. 1-15, 2022.

[31] E. Osborn, J. Brooks, P. M. S. O'Brien and A. Wittkowski, "Suicidality in women with Premenstrual Dysphoric Disorder: a systematic literature review," *Archives of Women's Mental Health*, vol. 24, p. 173–184, 2021.

[32] S. Doucet and N. Letourneau, "Coping and Suicidal Ideations in Women with Symptoms of Postpartum Depression," *Clinical medicine. Reproductive health*, vol. 2, pp. 9-19, 2009.

[33] L. d. A. Quevedo, C. C. Scholl, M. B. d. Matos, R. A. d. Silva, F. M. d. C. Coelho, K. A. T. Pinheiro and R. T. Pinheiro, "Suicide Risk and Mood Disorders in Women in the Postpartum Period: a Longitudinal Study," *Psychiatric Quarterly*, vol. 92, p. 513–522, 2021.

[34] S. Stepakoff, "Effects of Sexual Victimization on Suicidal Ideation and Behavior in U.S. College Women," *Suicide and Life-Threatening Behavior*, vol. 28, no. 1, pp. 107-126, 1998.

[35] S. S. Brokke, T. B. Bertelsen, N. I. Landrø and V. Ø. H. 2022, "The effect of sexual abuse and dissociation on suicide attempt," *BMC Psychiatry*, vol. 22, pp. 1-8, 2022.

[36] S. J. White, J. Sin, A. Sweeney, T. Salisbury, C. Wahlich, C. M. M. Guevara, S. Gillard, E. Brett, L. Allwright, N. Iqbal, A. Khan, C. Perot, J. Marks and N. Mantovani, "Global Prevalence and Mental Health Outcomes of Intimate Partner Violence Among Women: A Systematic Review and Meta-Analysis," *Trauma Violence Abuse*, vol. 25, no. 1, pp. 494-511, 2024.

[37] K. M. Devries, J. Y. Mak, L. J. Bacchus, J. C. Child, G. Falder, M. Petzold, J. Astbury and C. H. Watts, "Intimate Partner Violence and Incident Depressive Symptoms and Suicide Attempts: A Systematic Review of Longitudinal Studies," *PLOS Medicine*, vol. 10, no. 5, pp. 1-11, 2013.

[38] D. W. Hosmer Jr., S. Lemeshow and R. X. Sturdivant, Applied Logistic Regression, Hoboken, New Jersey: John Wiley & Sons, Inc., 2013.

[39] A. Agresti, An Introduction to Categorical Data Analysis, 3 ed., Hoboken, NJ: John Wiley & Sons, Inc., 2019.

[40] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Information Sciences 505 (2019) 32–64*, vol. 505, pp. 32-64, 2019.

[41] D. Elreedy, A. F. Atiya and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning*, vol. 113, p. 4903–4923, 2024.

[42] H. Dalianis, Clinical Text Mining: Secondary Use of Electronic Patient Records, Switzerland: Springer, 2018.

[43] Ministry of Women Empowerment and Child Protection, "The 2021 National Women's Life Experience Survey [Survei Pengalaman Hidup Perempuan Nasional (SPHPN) 2021]," MoWECP [Kemen PPPA], Jakarta, 2022.

[44] Lifeline, "Suicide," 2024. [Online]. Available: https://toolkit.lifeline.org.au/topics/suicide/feelings-and-effects-of-suicide. [Accessed 15 October 2024].