PARAMETER
JURNAL MATEMATIKA, STATISTIKA DAN TERAPANNYA

e-ISSN. 2829-6303
Vol. 05, No. 2, September 2026
Pages 267-286

Check for updates

Machine Learning Approaches for Household Smoking Classification in West Java: A Comparison between Encoded Logistic Regression and XGBoost

Boy Riansyah [1*], Felasofa Rahmatanti [2], Jasmin Nur Hanifa [3], Wildatul Maulidiy [4], Bagus Sartono [5], Aulia Rizki Firdawanti [6]

School of Data Science, Mathematics, and Informatics IPB University
Raya Dramaga St, Kabupaten Bogor, 16680, West Java, Indonesia

E-mail Correspondence Author: **hanifajasmin@apps.ipb.ac.id*

Abstract
Smoking behavior within households is a significant public health issue, particularly in densely populated regions such as West Java. This study aims to compare the classification performance of Logistic Regression and XGBoost in identifying household smoking status under several feature representation schemes, including baseline models, One-Hot Encoding, and Weight of Evidence (WoE). The analysis used 23.854 household observations from 2024 National Socioeconomic Survey (SUSENAS) in West Java after preprocessing. Model performance was evaluated using 5-fold stratified cross-validation based on accuracy, sensitivity, specificity, and Area Under the Curve (AUC) metrics. Among the evaluated configurations, XGBoost with WoE Encoding achived the highest observed accuracy (0.739), sensitivity (0.782), and AUC (0.797), although the differences in AUC across models were relatively small. Gain-based feature importance identified the Food–Non-Food Expenditure Ratio (X1) as the most influential predictor. The dominance of this expenditure ratio confirms that household expenditure structure, as an indicator of economic condition and welfare, plays a crucial role in distinguishing smoking households. These findings support the hypothesis that poorer households tend to have a higher probability of smoking.

Keywords: *Classification, encoded logistic regression, machine learning, weight of evidence, GBoost.*

: https://doi.org/10.30598/parameterv5i2pp267-286

Research Article ***Open Acces***
Email: jurnalparameter@gmail.com
Homepage: https://ojs3.unpatti.ac.id/index.php/parameter

Submitted: Desember 2025
Accepted: September 2026

1. INTRODUCTION

Smoking is a public health issue with significant global impacts. The World Health Organization (WHO) reports that tobacco consumption causes more than 8 million deaths annually. Of these, more than 7 million deaths are directly attributable to tobacco use, while approximately 1.2 million deaths occur among non-smokers exposed to secondhand smoke. These findings emphasize that smoking harms not only active smokers but also passive smokers, particularly within household settings [1]. In Indonesia, the prevalence of active smokers continues to increase, including among children and adolescents. Based on the 2023 Indonesian Health Survey (SKI), the number of active smokers is estimated to reach approximately 70 million people, with a substantial proportion coming from younger age groups. Data from the Global Youth Tobacco Survey (GYTS) 2019 indicate that the proportion of smokers among students aged 13–15 years increased from 18.3% in 2016 to 19.2% in 2019. Consistent with these findings, SKI 2023 reports that individuals aged 15–19 years constitute one of the age groups with notable smoking prevalence. Consequently, the increasing prevalence of smoking among children and adolescents strengthens the urgency of analyzing smoking behavior at the household level [2].

The growing number of active smokers in Indonesia also increases health risks for non-smokers due to exposure to secondhand smoke (SHS), particularly within households. A smoking household is defined as a household with at least one active smoker, thereby increasing SHS exposure for other members [3]. Globally, SHS causes approximately 600,000 deaths and affects more than one-third of the world's population. In Indonesia, this risk is more severe due to high exposure within homes; SKI 2023 records that 78% of households are exposed to cigarette smoke indoors, a proportion significantly higher than in many other countries [4].

At the regional level, West Java exhibits a higher smoking prevalence than the national average. SKI 2023 reports that smoking prevalence in West Java reaches 32.98%, exceeding the national average of 28.99% [5]. In addition, Statistics Indonesia (BPS) reports that West Java ranks second in terms of the highest number of poor residents in Indonesia, totaling 3.67 million people in 2024. The combination of high smoking prevalence and high poverty levels places smoking households in a high-risk group, both in terms of health outcomes and economic burden due to tobacco expenditure. This indicates that smoking behavior has implications not only for health but also for household socioeconomic welfare, consistent with prior studies highlighting the health and economic consequences of household smoke exposure [6]. Numerous studies have documented the health and economic impacts of household smoke exposure. Previous research finds that tobacco consumption significantly reduces household food expenditure allocation, increasing the risk of malnutrition and food insecurity, particularly among low-income families [7]. Moreover, literature reviews show that exposure to cigarette smoke during pregnancy consistently increases the risk of low birth weight, with most studies reporting significant reductions in birth weight. These findings confirm that cigarette smoke harms not only household members but also fetal development and increases the risk of neonatal complications [8].

Most studies on tobacco consumption have primarily focused on individual characteristics, while analyses using household as the unit of analysis remain relatively limited. Therefore, this study examines smoking households in West Java Province based on demographic, economic, and welfare, and health indicators. From a statistical perspective, the predominance of categorical predictors and potentially complex intervariable relationships motivates the use of models with different structural characteristics. Logistic Regression provides a relatively transparent linear classification framework, whereas XGBoost can capture nonlinear relationships and interactions among predictors. In addition, categorical feature representation may affect model performance. One-Hot Encoding and Weight of Evidence provide different approaches for representing categorical information in classification tasks [9]. Accordingly, the contribution of this study lies in systematically comparing Logistic Regression and XGBoost under alternative feature representation schemes in the context of household smoking classification. The primary objective is to compare their classification performance using accuracy, sensitivity, specifity, and Area Under the Curve (AUC) and to identify the configuration that provides the most effective predictive performance.

2. METHOD

2.1 Data

This study uses 2024 SUSENAS data for West Java Province obtained from Statistics Indonesia (BPS). The unit of analysis is households, with a total of 26.012 observations. The response variable is binary, identifying household smoking status: 1 for Smoking Households and 0 for Non-Smoking Households. The response variable was constructed by combining household-level data with household member level data from the 2024 National Socioeconomic Survey (SUSENAS). Household smoking status was determined using information collected for each household member on current tobacco smoking and electronic cigarette use. A household was classified as a smoking household if at least one household member reported currently smoking tobacco products or electronic cigarettes. Otherwise, the household was classified as a non-smoking household. A total of 21 predictor variables representing demographic characteristics, health conditions, and economic–welfare aspects of households are included. These variables represent multidimensional factors potentially influencing household smoking behavior.

Table 1. List of Predictor Variables

No	Predictor Variable Name	Representing Factor	No	Predictor Variable Name	Representing Factor
X1	Food and Non-Food Expenditure Ratio	Economic & Welfare	X11	Presence of School-Age Household Members	Demographic
X2	Per Capita Protein Expenditure	Health	X12	Proportion of Working Household Members	Economic & Welfare
X3	Per Capita Fat Expenditure	Health	X13	Internet Usage	Economic & Welfare
X4	Residential Location	Demographic	X14	Mobile Phone Usage	Economic & Welfare
X5	Occupation Sector of Household Head	Economic & Welfare	X15	Car Ownership	Economic & Welfare
X6	Housing Ownership Status	Economic & Welfare	X16	Access to Sanitation	Health
X7	Number of Household Members	Demographic	X17	Status of Household Receiving Credit in the Past Year	Economic & Welfare
X8	Main Source of	Economic &	X18	Household Status as	Economic &

No	Predictor Variable Name	Representing Factor	No	Predictor Variable Name	Representing Factor
	Household Financing	Welfare		JKN Contribution Assistance Recipient	Welfare
X9	Presence of Toddlers	Demographic	X19	Household Status as Cash Social Assistance Recipient	Economic & Welfare
X10	Tingkat Highest Education Level in the Household	Demographic	X20	Household Status as Job-Loss Insurance Recipient	Economic & Welfare
			X21	Inadequate Consumption of Nutritious Food	Health

All predictor variables are categorical, except for several numerical variables: Food–Non-Food Expenditure Ratio (X1), Protein Consumption per Capita (X2), Fat Consumption per Capita (X3), Number of Adult Household Members (X7), and Proportion of Working Household Members (X12).

2.2 Preprocessing and Cross Validation

Preprocessing includes handling special characters, removing numerical outliers, and addressing class imbalance in the response variable. Numerical outliers were identified using the interquartile range (IQR) method. Observation with value below $Q_1 - 1.5(IQR)$ or above $Q_3 + 1.5(IQR)$ in at least one numerical predictor were removed. Observations containing an outlying value in at least one numerical predictor were excluded from the dataset. As a result, the initial sample of 26.012 household was reduced to 23.854 households, which were subsequently used in the classification analysis.

The remaining data were then evaluated using 5-fold stratified cross-validation to preserve the class proportions in each fold., a technique widely recommended in the literature for providing relatively unbiased and stable model performance estimates compared to single data splits [10], while preventing overfitting. Response-dependent preprocessing procedures, including WoE estimation and random oversampling, were performed exclusively using the training data within each fold. The WoE values estimated from the training data were then applied to the corresponding validation data. Random oversampling is conducted by duplicating minority-class observations only in the training folds, while the validation folds were retained in their original class distribution to prevent data leakage.

XGBoost hyperparameter tuning is conducted using grid search evaluated through 5-fold stratified cross-validation. In this scheme, tuning and evaluation are performed using the same data folds; thus, nested cross-validation is not applied. Although evaluation is conducted on validation data within each fold, using the same folds for tuning and evaluation may slightly overestimate performance compared to independent data. Therefore, this study focuses on relative performance comparisons across model configurations.

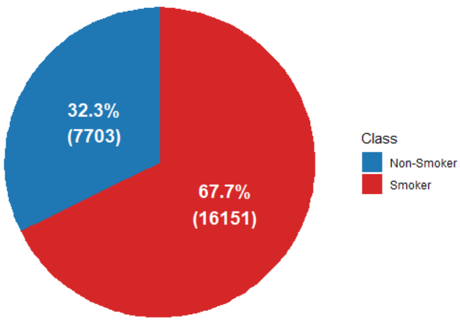


Figure 1. Class Distribution

Based on **Figure 1**, smoking households constitute the majority class, whereas non-smoking households constitute the minority class. Random oversampling was selected to balance the response classes by duplicating randomly selected observations from the minority class **[11]**. This approach was prefered over synthetic oversampling methods such as SMOTE because the dataset contains a large proportion of categorical predictors, and the study aimed to preserve the observed combinations of houshold characteristics rather than generate synthetic household profiles. Random oversampling also provides a consistent resampling procedure for both Logistic Regression and XGBoost. To prevent data leakage, oversampling was performed exclusively within each training fold, while the corresponding validation fold retained its original class distribution.

2.3 Logistic Regression Based on *Weight of Evidence* (WoE)

This study applies logistic regression based on Weight of Evidence (WoE) as one of the main analytical methods. Logistic regression is used to model the probability of a binary outcome through the logit function and has been widely applied in socio-economic studies **[12]**. The WoE technique transforms each category of predictor variables into numerical values based on log-odds, calculated from the ratio of the distribution between events (smoking households) and non-events (non-smoking households), thereby producing a more informative and structured predictor representation **[13]**. Formally, for the *i*-th predictor variable with categories or bins $B_{i1}, B_{i2}, \dots, B_{iK}$ where j =1,2,…,K. The Weight of Evidence (WoE) value for the *j*-th bin is defined as follow.

$$WoE_{ij} = \ln\left(\frac{P(X_i \in B_{ij}|Y = 0|)}{P(X_i \in B_{ij}|Y = 1|)}\right) \tag{1}$$

where $Y = 1$ denotes smoking households (events) and $Y = 0$ denotes non-smoking households (non-events). The above formulation measures the relative concentration of non-smoking and smoking households within each category or bin. A positive WoE value indicates that the proportion of non-smoking households in bin B_{ij} is larger than that of smoking households, whereas a negative WoE value indicates a higher concentration of smoking households. The magnitude of WoE_{ij} reflects the strength of association between the predictor category and the likelihood of being classified as a smoking household. By transforming categorical and binned numerical variables into this log-odds scale, the WoE encoding produces predictors that are directly aligned with the logistic regression model structure and improves estimation stability in the presence of multi-category variables.

The application of WoE allows logistic regression models to handle categorical variables more efficiently without requiring the construction of a large number of dummy variables, which may increase model complexity and hinder interpretability. In this study, the dataset includes 16 categorical predictor variables, which, if represented using conventional dummy encoding, would result in a substantially larger number of derived variables depending on the number of categories in each predictor. In general, dummy encoding generates $\Sigma(k_i - 1)$ new variables, where k_i denotes the number of categories of the i-th variable, leading to dimensionality expansion. This condition increases the risk of estimation instability and complicates result interpretation. Therefore, WoE is selected as a transformation strategy that reduces dimensional complexity while maintaining a balance between model performance, estimation stability, and interpretability. In this study, WoE values are computed using only the training data within each fold to prevent information leakage. The resulting WoE values are then used to transform both training and testing data within the same fold, ensuring that the constructed model remains simple and interpretable. In general, the logistic regression model used in this study can be expressed as follows:

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (2)$$

In **Equation (1)**, p represents the probability that a household belongs to the smoking household category, while β_0 denotes the model intercept. The regression coefficients $\beta_i$indicate the magnitude of contribution of the i-th predictor to changes in the log-odds, and X_i represents the predictor variables included in the model, including categorical predictors transformed using WoE. Through the logistic function, the resulting logit values are converted back into probabilities using the following equation:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} \quad (3)$$

This formulation allows the model to directly estimate event probabilities, enabling intuitive interpretation of each variable's effect while remaining consistent even when involving transformed categorical predictors. Recent studies also show that combining logistic regression with WoE transformation improves model stability and performance when dealing with categorical variables or imbalanced data distributions **[14], [15]**. These findings support the use of WoE in this study, particularly given the multi-categorical nature of the data and the relatively high degree of class imbalance.

2.4 Extreme Gradient Boosting (XGBoost)

XGBoost is a tree-based gradient boosting algorithm that builds models additively using decision trees, where each additional tree is designed to reduce the prediction errors of the preceding ensemble **[16]**. In this study, XGBoost was applied using two categorical variable transformation approaches: One-Hot Encoding (OHE) and WoE Encoding, to evaluate whether different representations of categorical information affect household smoking classification performance. OHE converts each category into a binary indicator, enabling categorical variables to be processed directly by decision trees. Meanwhile, WoE transforms each category into a log-odds scale based on the response distribution within each category, producing a numerical representation with statistical interpretability that is compatible with boosting algorithms. These two encoding schemes are used to evaluate the effect of categorical variable representation on XGBoost performance in classifying smoking households in West Java. The XGBoost algorithm generally constructs additive predictive models as follows:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathcal{F} \quad (4)$$

where $\widehat{y_1}$ represents the prediction for the i-th observation, K the total number of trees in the ensemble and each f_k is a decision tree belonging to the function space $\mathcal{F}$. In the context of XGBoost, the function space is defined as the set of all binary decision trees with leaves T and leaf weights w.

In this study, the XGBoost modeling process begins with hyperparameter tuning on key parameters such as max_depth, n_estimators, learning_rate, alpha (L1), lambda (L2), gamma, and min_child_weight. The selection of parameters for tuning follows prior studies [17]. By adhering to best practices from previous research, this study conducts targeted tuning to obtain a stable, efficient, and well-generalizing XGBoost configuration before cross-validation. After the tuning process is completed and optimal parameter configurations are obtained, all XGBoost models are evaluated using 5-fold cross-validation. This approach provides a consistent and replicable methodological framework for comparing XGBoost performance under OHE and WoE encoding schemes in predicting household smoking status in West Java.

3. RESULTS AND DISCUSSION

3.1. Characteristics and Distribution of Smoking Households

Prior to regression modeling, all predictor variables were evaluated for potential multicollinearity. Based on Table 2, all variables met the criteria with Variance Inflation Factor (VIF) values below the thresholds of 5 and 10, indicating that multicollinearity was not a concern in this study.

Table 2. Multicollinearity Assessment

No	VIF	Status	No	VIF	Status
X1	1.376	Safe	X11	1.902	Safe
X2	2.296	Safe	X12	1.699	Safe
X3	2.008	Safe	X13	3.585	Safe
X4	1.305	Safe	X14	3.122	Safe
X5	1.127	Safe	X15	1.007	Safe
X6	1.110	Safe	X16	1.056	Safe
X7	2.944	Safe	X17	1.040	Safe
X8	1.474	Safe	X18	1.137	Safe
X9	1.344	Safe	X19	1.022	Safe
X10	2.635	Safe	X20	1.035	Safe
			X21	1.048	Safe

Based on the distribution of numerical variables in **Figure 2** by household smoking status, differences in characteristics between smoking and non-smoking households are observed. In general, smoking households exhibit higher median values for the number of adult household members, the food–non-food expenditure ratio, the proportion of working household members, and fat consumption. A higher food–non-food expenditure ratio among smoking households indicates a potential crowding-out effect, whereby households purchasing cigarettes allocate a larger portion of their budget to tobacco rather than non-food necessities [18], [19]. Higher fat consumption among smoking households reflects less healthy dietary patterns, with greater total energy and fat intake compared to non-smoking households. This finding is supported by observational studies and reviews reporting a consistent correlation between smoking and higher fat and calorie intake [20]. In contrast, protein consumption shows no substantial difference between smoking and non-smoking households.

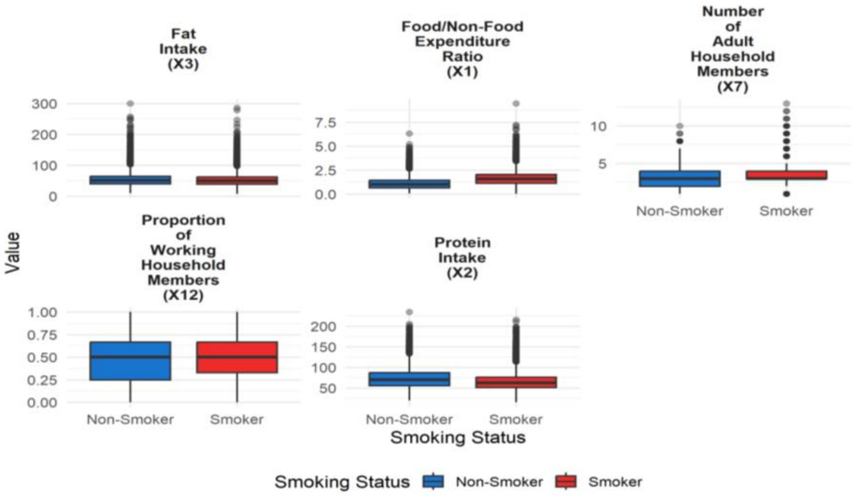


Figure 2. Distribution of Numerical Variables by Smoking Status

The visualization (Figure 2) also reveals the presence of extreme values (outliers) in several numerical variables, which may significantly affect logistic regression performance, particularly when WoE transformation is applied, as it may generate low-frequency bins and unstable WoE values. Therefore, outlier handling during preprocessing is necessary to maintain model stability and improve reliability.

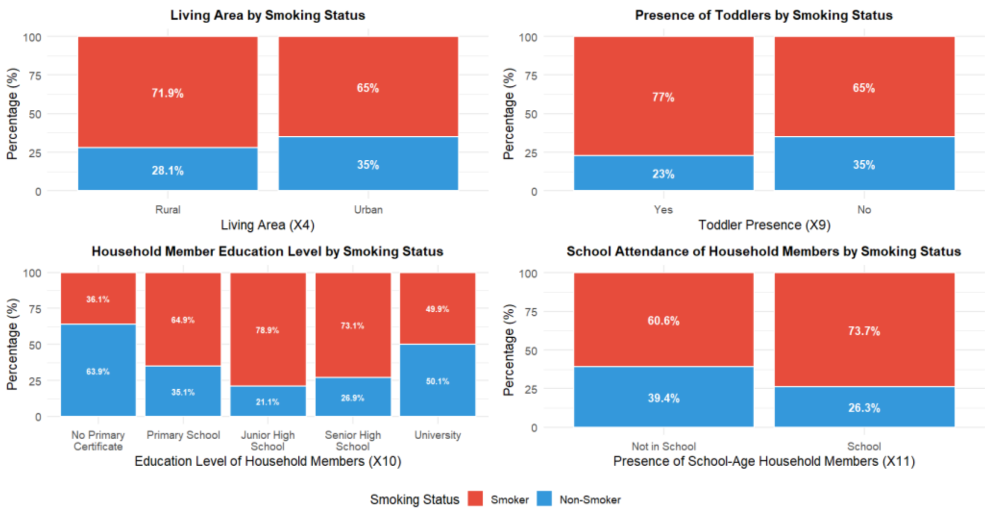
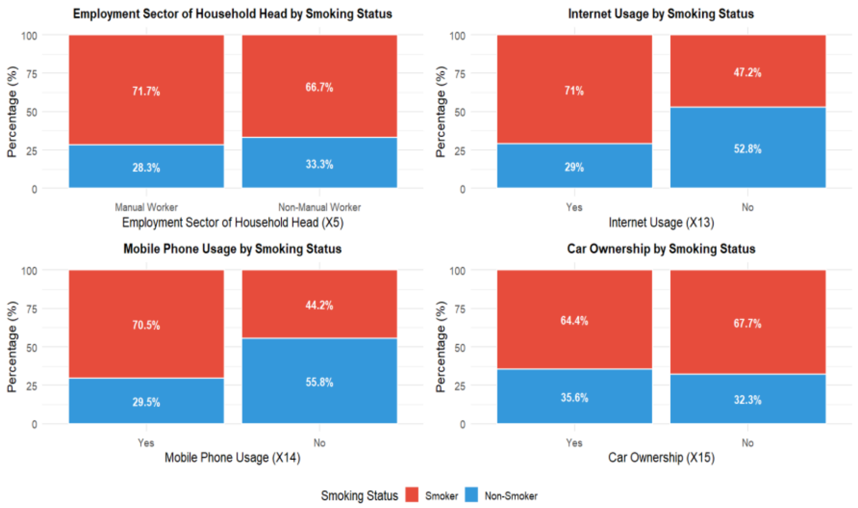


Figure 3. Distribution of Demographic Categorical Variables

Based on the distribution of demographic variables in Figure 3 by household smoking status, a higher proportion of smoking households is observed in rural areas compared to urban areas. This pattern is consistent with prior studies indicating that smoking prevalence tends to be higher in rural regions than in urban areas, particularly in developing countries including Indonesia. Several studies highlight that rural environments often have more permissive social norms toward smoking, easier access to tobacco products, and relatively weaker health supervision mechanisms compared to urban settings [21].

Educational attainment within smoking households is dominated by junior and senior high school graduates, indicating a concentration at the secondary education level. This finding aligns with literature suggesting that individuals with low to medium education levels face higher smoking risks due to limited access to health information, lower health literacy, and stronger social environmental influences. The presence of toddlers is higher in smoking households compared to households without toddlers. This finding supports claims that households with toddlers or larger household structures are more likely to include smokers [22]. The proportion of household members still attending school is also higher in smoking households. Consistent with previous studies, adolescents and household members within educational environments are more likely to be exposed to peer influence and smoking behavior at home. Prior research emphasizes that household exposure to smoking is a dominant risk factor for adolescent smoking initiation, in addition to school and community influences [23].



(a)

(b)

Figure 4. Distribution of Economic and Welfare Categorical Variables

Based on the distribution of variables within the economic and welfare category in Figure 4 (a) by household smoking status, a higher proportion of household heads working in manual labor sectors is observed among smoking households. This indicates an association between physically demanding occupations and smoking prevalence [24]. Additionally, smoking households show higher percentages of internet and mobile phone usage compared to non-smoking households. Meanwhile, car ownership proportions are nearly identical between smoking and non-smoking households. Figure 4 (b) shows that smoking households have higher proportions across almost all categories of housing tenure status compared to non-smoking households, except for official housing. The dominant source of household financing originates from working household members.

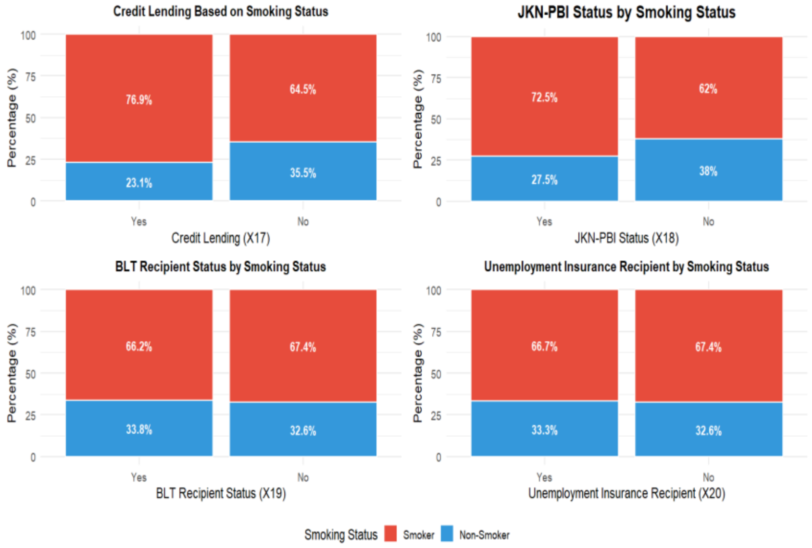


Figure 5. **Distribution of Economic and Welfare Categorical Variables**

Other economic and welfare variables in Figure 5 show that smoking households have higher proportions across all social assistance categories. All variables within the social assistance category indicate a higher prevalence of smoking households receiving health insurance assistance and government credit programs. These findings suggest that smoking households tend to experience greater socioeconomic vulnerability, making them more likely recipients of social protection programs. This pattern aligns with prior research indicating that lower socioeconomic status correlates with higher exposure to household cigarette smoke [22].

Figure 6. Distribution of Health-Related Categorical Variables

Figure 6 illustrates differences in sanitation access and food insecurity between smoking and non-smoking households. For sanitation access, smoking households exhibit higher proportions within the inadequate sanitation category, indicating an overlap between smoking behavior and environmental vulnerability. A similar pattern appears for access to healthy and nutritious food, where smoking prevalence is consistently higher across all categories of limited food access. This finding is consistent with previous studies showing that smokers tend to have less healthy dietary patterns and lower consumption of nutritious foods compared to non-smokers **[20]**. The exploratory data analysis (EDA) results in this study serve as substantive validation for the classification modeling outcomes. The alignment between empirical patterns identified during EDA and variables deemed important by the models indicates that the developed models not only achieve strong predictive performance but also capture statistically relevant and substantively meaningful relationships. This provides supporting evidence that the resulting models do not suffer from overfitting.

3.2. Weight of Evidence (WoE) Transformation

Based on **Table 3**, the WoE values indicate the direction and strength of the relationship between each bin and the log-odds of smoking household status. Positive WoE values indicate that the proportion of non-smoking households in the corresponding bin is greater than that of smoking households, while negative WoE values indicate the opposite tendency.

Table 3. Binning Results and WoE Transformation for Smoking Household Variables

No	Bin	Frequency	WoE	No	Bin	Frequency	WoE
X1	<0.835	4771	1.172	X11	Attending School	12538	-0.289
	(0.835, 1.208]	4771	0.403		Not Attending School	11316	0.289
	(1.208, 1.553]	4770	-0.232	X12	≤0.33	9253	0.108
	(1.553, 1.983]	4771	-0.705		>0.33	14601	-0.070
	>1.983	4771	-1.189	X13	Yes	20359	-0.168
X2	<49.66	4771	-0.354		No	3495	0.860
	(49.66, 59.21]	4771	-0.261	X14	Yes	21114	-0.145
	(59.21, 68.303]	4770	-0.053		No	2740	0.982
	(68.303, 80.903]	4771	0.082	X15	Yes	2618	0.138
	>80.903	4771	0.509		No	21236	-0.017
X3	<36.347	4771	0.011	X16	Adequate	16660	0.098
	(36.347, 44.492]	4771	-0.028		Inadequate	7194	-0.243
	(44.492, 52.726]	4770	-0.017	X17	Yes	5530	-0.471
	(52.726, 63.430]	4771	0.017		No	18324	0.126
	>63.430	4771	0.016	X18	Yes	12430	-0.233
X4	Rural	8176	-0.217		No	11424	0.233

No	Bin	Frequency	WoE	No	Bin	Frequency	WoE
X5	Urban	15678	0.107	X19	Yes	671	0.043
	Non Manual Worker	20479	0.029		No	23183	-0.001
	Manual Worker	3375	-0.181	X20	Yes	430	0.075
X6	Owner-Occupied	1255	0.011		No	23424	-0.001
	Rent Free	2489	-0.061	X21	Yes	1931	-0.132
	Official Housing	79	0.766		No	21873	0.001
	Rented	1255	-0.109		Don't Know	41	-0.263
	Others	3	0.047		Refused	9	-1.339
X7	1	2149	1.912				
	2	4887	0.404				
	3	6336	-0.274				
	4	6629	-0.446				
	>5	3853	-0.616				
X8	Working Household Member	21283	-0.199				
	Investment	48	1.077				
	Remittances (Cash/Goods)	1968	1.443				
	Pension	555	1.52				
X9	Yes	4765	-0.458				
	No	19089	0.102				
X10	Primary School	4877	0.128				
	Junior	5398	-0.575				
	Senior	8957	-0.271				
	Higher Education	3462	0.727				
	No Primary School Certificate	1160	1.299				

Based on **Table 3**, numerical variables resulting from binning exhibit changes in WoE values across bins that reflect patterns in their relationship with the logit. Meanwhile, for categorical variables, differences in the likelihood of being classified as a smoking household across categories can be identified through WoE transformation. Thus, the application of WoE enables the conversion of both categorical and numerical information into a stable and interpretable log-odds scale for logistic regression modeling.

3.3. Model Performance Comparison

The evaluation results for Logistic Regression and XGBoost under several data-handling scenarios are presented as follows:

Table 4. Mean Classification Performance Comparison

Methodology	Accuracy (Mean ± SD)	Sensitivity (Mean ± SD)	Specificity (Mean ± SD)	AUC (Mean ± SD)
Regresi Logistik Base	0.725 ±0.006	0.706 ± 0.009	0.733 ± 0.006	0.793 ± 0.009
Regresi Logistik WoE	0.725 ± 0.009	0.745 ± 0.009	0.686 ± 0.013	0.786 ± 0.009
XGBoost One Hot Encoding	0.739 ± 0.008	0.781 ± 0.011	0.655 ± 0.015	0.791 ± 0.009
XGBoost WoE Encoding	0.739 ±0.004	0.782 ± 0.006	0.650 ± 0.005	0.797 ± 0.007

Table 4 presents the mean classification performance, among the evaluated models, XGBoost with WoE Encoding achieved the highest average AUC (0.797), accuracy (0.739), and sensitivity (0.782) among the evaluated model configurations. However, the differens in AUC compared with Baseline Logistic Regression (0.793) and XGBoost with OHE (0.791) were relatively small. The objective of this study was not to test the statistical significance of performance differences between models, but to compare the magnitude of their classification metrics and identify the configuration with the most favorable overall performance. Therefore, the results are interpreted descriptively, with XGBoost with WoE Encoding considered the best-performing configuration based on its observed metric values, rather than as statistically significant superiority over the other models. This model also showed relatively high sensitivity, indicating its ability to identify smoking households, although this was accompanied by a lower specifity of 0.650. Overall, XGBoost with WoE Encoding provided the most favorable classification performance among the evaluated configurations based on the observed values of the performance metrics.

3.4. Feature Importance of the Best Model

To further examine the predictive structure of the selected model, this section presents feature importance derived from the XGBoost with WoE Encoding. Feature importance in this study represents the relative contribution of each predictor to the model's prediction and should not be interpreted as evidence of a casual effect on household smoking behavior. The analysis is used as a complementary descriptive tool to identify which variables contribute most strongly to classification and to compare these patterns with the exploratory data analysis. serves as a complementary evaluation tool to link model-driven findings with prior exploratory insights. **Figure 7** presents the feature importance results obtained from the XGBoost model with WoE encoding, highlighting the relative contribution of demographic, economic–welfare, and health-related variables in distinguishing smoking and non-smoking households.

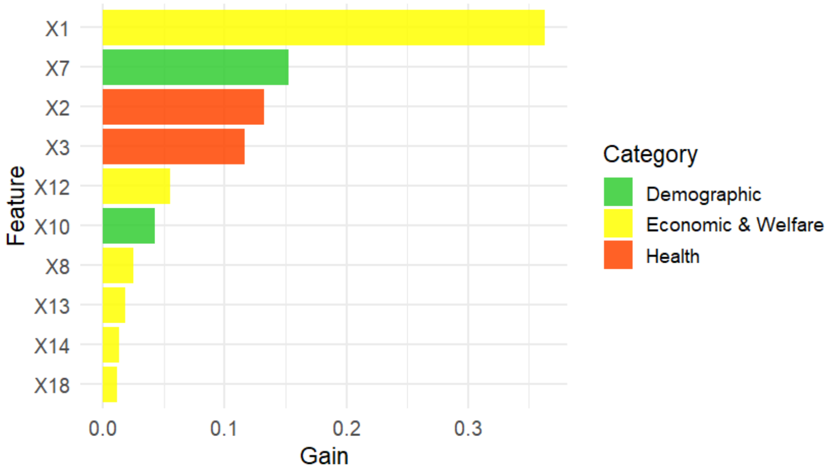


Figure 7. Feature Importance Ranking

The consistency between the feature importance results and the findings from exploratory data analysis (EDA) further supports the substantive validity of the developed model. Based on the evaluation, XGBoost with WoE Encoding emerged as the best-performing model. The feature importance results presented in **Figure 7** identified three variables with the highest importance: the Food–Non-Food Expenditure Ratio (X1), Number of Adult Household Members (X7), and Per Capita Protein Consumption (X2). In this study, feature importance was measured using the Gain metric from XGBoost, which represents the relative contribution of a feature to improving the model's predictive performance. Therefore, the ranking of X1, X7, and X2 as the most important features is based on their respective Gain values. The numerical distribution visualization shows that smoking households tend to have higher median values for the Food–Non-Food Expenditure Ratio and the Number of Adult Household Members, while exhibiting relatively lower protein consumption compared with non-smoking households. The consistency between these descriptive patterns and the variables identified as the most informative by the model indicates that XGBoost with WoE transformation captures key empirical characteristics identified during the exploratory data analysis stage. This provides evidence that the model structure is consistent and substantively meaningful, rather than being driven solely by statistical optimization.

It should be emphasized that feature importance analysis in this study is not intended as a tool for structural or causal interpretation. Instead, it is used as an empirical validation mechanism to assess the consistency between model outputs and prior empirical findings. Feature importance is employed to evaluate whether the variables identified as key predictors align with the conceptual framework and existing literature, thereby reinforcing the substantive validity of the modeling results. Overall, feature importance results from the XGBoost model with WoE Encoding are consistent with empirical studies showing that smoking behavior in Indonesia is strongly influenced by household socioeconomic conditions. National analyses based on SUSENAS and Riskesdas data indicate that low-income households have a higher probability of smoking compared to more affluent groups. This phenomenon is associated with the role of cigarettes as a coping mechanism under economic stress and the tendency of poorer households to allocate a greater share of expenditure to addictive commodities. These findings are consistent with literature identifying socioeconomic vulnerability as a primary determinant of smoking behavior in Indonesia **[19], [25]**.

The number of household members (X7) emerges as an important predictor. Empirically, households with more adult members have a higher likelihood of being classified as smoking households due to the increased probability that at least one member smokes. This reflects a probabilistic mechanism whereby a larger number of adults naturally increases behavioral variation within the household, including smoking behavior. Thus, X7 serves as an indicator of the likelihood of smoker presence rather than a direct behavioral determinant. Protein consumption per capita (X2) is also identified as a significant variable, indicating that dietary consumption patterns are relevant to household smoking status. Empirical evidence suggests that smoking households often exhibit suboptimal dietary patterns, particularly for nutritionally valuable commodities such as protein. This result is consistent with the exploratory analysis showing that cigarette expenditure may crowd out household budgets allocated for nutritious food, resulting in lower protein consumption among smoking households. Fat consumption per capita (X3) also contributes significantly to the model. Unlike protein, fat intake is relatively more affordable and thus remains accessible even when households allocate resources to cigarettes. In addition, the proportion of working household members (X12) and the highest educational attainment within the household (X10) exert meaningful influence on smoking behavior. The presence of working household members may increase smoking frequency due to available income for cigarette purchases while simultaneously shaping household consumption habits. Educational attainment exhibits a varied pattern, with relatively higher smoking prevalence at the secondary education level and declining prevalence at higher education levels, reflecting greater health awareness and healthier lifestyle choices among more educated groups. Other variables such as the primary source of household financing (X8), internet usage (X13), mobile phone usage (X14), and the presence of school-attending household members (X11) highlight the role of information access and socioeconomic dynamics in shaping smoking behavior. Overall, feature importance results indicate that economic and welfare factors dominate as key predictors, followed by demographic and health-related conditions. These findings provide empirical evidence that policy interventions targeting socioeconomically vulnerable households have greater potential to reduce smoking prevalence.

3.4. Study Limitations and Implications

Although XGBoost demonstrated superior performance compared with the other methods, several limitations should be acknowledged. First, the model was developed using cross-sectional SUSENAS data, which do not capture the temporal dynamics of household smoking behavior. Second, the use of households as the unit of analysis may limit the ability of the study to capture individual-level determinants, such as age, gender, personal income, smoking intensity, local cigarette prices, peer influence, and social norms. Third, the study evaluated only two main approaches, namely Logistic Regression and XGBoost, and did not include alternative models that may achieve better performance, such as CatBoost, LightGBM, or neural network-based models. Furthermore, the findings of this study are predictive in nature and should not be interpreted as causal relationships. Accordingly, feature importance in XGBoost reflects the relative contribution of features to model predictions and should not be interpreted as a causal effect on smoking behavior. Nevertheless, this study offers important implications. The performance of XGBoost with WoE transformation demonstrates the effectiveness of tree-based boosting methods in capturing nonlinear relationships and complex interactions among variables, providing opportunities for the application of similar algorithms in behavioral health studies. Furthermore, the Food–Non-Food Expenditure Ratio emerged as the most important feature for model prediction, indicating that household expenditure allocation patterns are relevant to the classification of smoking households. This finding may provide considerations for household-oriented health and social policy planning, although it should not be interpreted as evidence of a causal relationship.

Furthermore, the findings of this study are consistent with previous research examining factors influencing household cigarette expenditure in East Java in 2018. The study found that the education level of the household head, age of the household head, employment sector, and average per capita expenditure significantly influenced cigarette expenditure, with the effect of per capita expenditure increasing as cigarette consumption increased. These findings are related to the present study through the Food–Non-Food Expenditure Ratio, which emerged as the most important feature in classifying smoking households in West Java. This ratio indirectly represents household economic conditions and expenditure allocation patterns, thereby providing additional insight into economic characteristics associated with the classification of smoking households **[26]**.

4. CONCLUSION

This study demonstrates that Logistic Regression and XGBoost can provide comparable classification performance for identifying smoking household in West Java when combined with appropriate categorical feature representations. Among the evaluated configurations, XGBoost with WoE Encoding showed a slight numerical advantage in overall predictive performance, particularly in identifying smoking households, while Logistic Regression with WoE remains useful when greater model transparency is required. Gain-based feature importance identified the Food–Non-Food Expenditure Ratio (X1) as the leading predictive feature, indicating that household expenditure structure contains relevant information for household smoking classification. Overall, the study contributes a comparative frramework for evaluating linear and tree-based classification models with alternative categorical encoding strategies using household-level socioeconomic, demographic, and health information. These findings are predictive and should not be interpreted as casual relationships.

Acknowledgments

The authors would like to express their gratitude to **Dr. Ir. Budi Susetyo, M.S.** and **Gerry Alfa Dito, S.Si., M.Si.** for their valuable suggestions, criticisms, and feedback provided during the manuscript preparation process. These contributions played an important role in improving the quality of this paper, making it clearer and more accessible to readers from both statistical and non-statistical backgrounds

Funding Information

This research and the associated publication costs were fully supported by the School of Data Science, Mathematics and Informatics, IPB University. The authors gratefully acknowledge this institutional support in enabling the completion and dissemination of this work.

Author Contributions Statement

Boy Riansyah conceived the main research concept and study design and contributed to data collection, data processing, and data analysis using the XGBoost method with OHE for categorical variable transformation. **Felasofa Rahmatanti** contributed to data collection, data processing, and data analysis, including multicollinearity assumption checking, and performed baseline logistic regression analysis. **Jasmin Nur Hanifa** contributed to data collection and data processing, performed WoE transformation, and conducted data analysis using logistic regression with WoE. **Wildatul Maulidiyah** contributed to data collection, data processing, and data analysis using the XGBoost method with WoE. **Bagus Sartono** and **Aulia Rizki Firdawanti** provided methodological review and substantive input on the manuscript. All authors were involved in the research process, including data collection, data cleaning, and exploratory data analysis, and participated in drafting and revising the manuscript and approved the final version for publication.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests, personal relationships, or professional affiliations that could have influenced the work reported in this paper. Accordingly, the authors state that there is no conflict of interest in connection with the submission and publication of this manuscript.

Data Availability

The data that support the findings of this study were obtained from the 2024 National Socioeconomic Survey (SUSENAS) conducted by Statistics Indonesia (Badan Pusat Statistik, BPS) of West Java through a licensed purchase by the authors. Restrictions apply to the availability of these data, as they are not publicly accessible and are subject to BPS data usage regulations. The data were used under permission for the purpose of this research and are therefore not openly available. Derived data supporting the findings of this study may be available from the corresponding author upon reasonable request, subject to approval from the data provider.

REFERENCES

[1] WHO Indonesia, "Pernyataan WHO: Indonesia sehat dan sejahtera melalui cukai dan harga produk tembakau yang lebih tinggi," WHO. Accessed: Jan. 09, 2026. [Online]. Available: https://www.who.int/indonesia/news/detail/09-10-2020-pernyataan-who-indonesia-sehat-dan-sejahtera-melalui-cukai-dan-harga-produk-tembakau-yang-lebih-tinggi

[2] S. N. Tarmizi, "Perokok Aktif di Indonesia Tembus 70 Juta Orang, Mayoritas Anak Muda," Kementerian Kesehatan. Accessed: Jan. 09, 2026. [Online]. Available: https://kemkes.go.id/id/perokok-aktif-di-indonesia-tembus-70-juta-orang-mayoritas-anak-muda

[3] A. Asyary and M. Veruswati, "Smoking Behavior and Cigarette Expenditure in a Household: Evidence for Smoke-Free Houses Initiation in Indonesia," *Int. J. Prev. Med.*, vol. 14, no. 1, Jan. 2023, doi: 10.4103/ijpvm.ijpvm_113_21.

[4] N. D. Rahmawati, A. Ahsan, and H. Andriani, "78 persen rumah tangga Indonesia teracuni asap rokok dari perokok aktif, berdampak pada kesehatan bayi," Universitas Indonesia. Accessed: Jan. 09, 2026. [Online]. Available: https://scholar.ui.ac.id/en/clippings/78-persen-rumah-tangga-indonesia-teracuni-asap-rokok-dari-perokok-2/

[5] "Persentase Penduduk Berumur 15 Tahun ke Atas yang Merokok Tembakau selama Sebulan Terakhir Menurut Provinsi (Persen), 2025," 2025. Accessed: Jan. 09, 2026. [Online]. Available: https://www.bps.go.id/assets/statistics-table/2/MTQzNSMy/persentase-penduduk-berumur-15-tahun-ke-atas-yang-merokok-tembakau-selama-sebulan-terakhir-menurut-provinsi.html

[6] "Rilis Kemiskinan dan Ketimpangan," Badan Pusat Statistik Jawa Barat. Accessed: Jan. 09, 2026. [Online]. Available: https://jabar.bps.go.id/id/news/2025/01/15/729/rilis-kemiskinan-dan-ketimpangan.html

[7] E. Rastiyanto Amrullah *et al.*, "Konsumsi Tembakau Dan Implikasinya Terhadap Pengeluaran Pangan Rumah Tangga: Pendekatan Fraksional Logit."

[8] V. A. B. Aly Syahbana, Y. H. Darundiati, and N. A. Yunita Dewanti, "Paparan Asap Rokok Sebagai Faktor Risiko Kejadian Berat Bayi Lahir Rendah (Bblr) : Kajian Literatur Sistematik," *Jurnal Kesehatan Masyarakat (Undip)*, vol. 10, no. 3, pp. 312–318, May 2022, doi: 10.14710/jkm.v10i3.32761.
[9] W. Zhu, R. Qiu, and Y. Fu, "Comparative Study on the Performance of Categorical Variable Encoders in Classification and Regression Tasks," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.09682
[10] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 14, pp. 1137–1145, 1995.
[11] B. Santoso, H. Wijayanto, K. A. Notodiputro, and B. Sartono, "Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 58, p. 012031, Mar. 2017, doi: 10.1088/1755-1315/58/1/012031.
[12] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 2013.
[13] N. Siddiqi, *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecard*, 2nd ed. New Jersey: John Wiley & Sons, Inc, 2017.
[14] G. Zeng, "On the analytical properties of category encodings in logistic regression," *Commun. Stat. Theory Methods*, vol. 52, no. 6, pp. 1870–1887, Mar. 2023, doi: 10.1080/03610926.2021.1939382.
[15] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, Nov. 2015, doi: 10.1016/j.ejor.2015.05.030.
[16] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Jan. 2019.
[18] "Survei Sosial Ekonomi Nasional (SUSENAS) 2024: Keterangan Konsumsi/Pengeluaran Makanan & Bukan Makanan, dan Pendapatan/Penerimaan Rumah Tangga," Jakarta, 2024.
[19] A. Swarnata, F. Z. Kamilah, G. Melinda, and V. Adrison, "The Impoverishing Effect of Tobacco Use in Indonesia," *Nicotine and Tobacco Research*, vol. 26, no. 10, pp. 1331–1338, Sep. 2024, doi: 10.1093/ntr/ntae088.
[20] G. Heydari, F. Heidari, M. Yousefifard, and M. Hosseini, "Smoking and Diet in Healthy Adults: A Cross-Sectional Study in Tehran, Iran, 2010," 2014. [Online]. Available: http://ijph.tums.ac.ir
[21] R. I. Mardhiyyah, "Pemodelan Faktor Sosial Ekonomi dan Demografi terhadap Status Merokok Penduduk Jawa Barat Menggunakan Regresi Logistik Biner," Institut Teknologi Sepuluh Nopember, 2023.
[22] S. Nadhiroh, K. Djokosujono, and D. M. Utari, "Socioeconomic characteristics, paternal smoking and secondhand tobacco smoke exposure among infants in Jakarta, Indonesia," *Tob. Induc. Dis.*, vol. 18, no. May, May 2020, doi: 10.18332/tid/120077.
[23] R. Amelia and A. Mu'nim, "Regresi Logistik Biner dengan Proses Resampling dalam Menduga Faktor Determinan Merokok Remaja," *Seminar Nasional Official Statistics*, vol. 2022, no. 1, pp. 704–716, Nov. 2022, doi: 10.34123/semnasoffstat.v2022i1.1123.
[24] A. Katainen, "Smoking and workers' autonomy: A qualitative study on smoking practices in manual work," *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine*, vol. 16, no. 2, pp. 134–150, Mar. 2012, doi: 10.1177/1363459311403944.
[25] N. Kusumawardani, I. Tarigan, Suparmi, and A. Schlotheuber, "Socio-economic, demographic and geographic correlates of cigarette smoking among Indonesian adolescents: results from the 2013 Indonesian Basic Health Research (RISKESDAS) survey,"

Glob. Health Action, vol. 11, no. sup1, pp. 54–62, Dec. 2018, doi: 10.1080/16549716.2018.1467605.

[26] Rizka Ayu Oktaviani, N. Nurcholis, and F. Suhaeri, "Pengeluaran Rokok Rumah Pada Rumah Tangga Yang Mengkonsumsinya," *Jurnal Ekonomi dan Bisnis*, vol. 26, no. 1, pp. 1–10, Jun. 2022, doi: 10.24123/jeb.v26i1.4841.