IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 524~535
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp524-535

Feature selection in supervised machine learning: a case study of poverty dataset in West Java, Indonesia

Sean Marshelle, Septian Rahardiantoro, Anang Kurnia
Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

Article Info

Article history:

Received May 13, 2025
Revised Dec 9, 2025
Accepted Dec 15, 2025

Keywords:

Chi-squared
ElasticNet
Genetic algorithm
Information gain
LASSO
Sequential forward selection
West Java

ABSTRACT

West Java, one of the largest provinces in Indonesia with a population exceeding 50 million, reported a poverty rate of 7.62% in 2023. Data from the national socio-economic survey or *survei sosial ekonomi nasional* (SUSENAS) show that poverty is multidimensional, encompassing aspects of employment, education, sanitation, housing, food security, technology, and government assistance. Addressing this complexity requires identifying the most influential factors that determine household welfare. This study applies and compares three feature selection approaches—filter, wrapper, and embedded—to the SUSENAS dataset to evaluate their effectiveness in identifying key poverty determinants. By prioritizing variables with the strongest predictive power, the study provides an evidence-based framework for more efficient and targeted poverty alleviation strategies. Results indicate that the information filter method combined with random forest (RF) and the least absolute shrinkage and selection operator (LASSO) embedded method combined with logistic regression (LR) deliver the best performance, improving model accuracy while reducing more than 65% of irrelevant features. The selected indicators highlight critical sectors such as food security, housing, and access to technology, which can serve as short-term policy priorities. In the long term, broader interventions in education, employment, sanitation, and government support are recommended. These findings demonstrate how data-driven feature selection can guide effective policy design for reducing poverty in West Java.

Corresponding Author:

Sean Marshelle
Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics
IPB University
Dramaga, Bogor-16680, West Java, Indonesia
Email: seanmarshelle@apps.ipb.ac.id

1. INTRODUCTION

Indonesia, a vast archipelago of approximately 17,380 islands—including the major islands of Sumatra, Java, Kalimantan, Sulawesi, and Papua—is the fourth most populous country globally, with a population exceeding 280 million. Despite notable economic progress and rising living standards, poverty remains a persistent challenge, primarily due to regional disparities. According to the World Bank, extreme poverty is living on less than $1.90 per day for low-income economies and $3.20 for lower-middle-income economies. As of March 2018, based on purchasing power parity (PPP), 5.7% of Indonesians lived under the $1.90 threshold, and 27.3% fell under the $3.20 line [1]. West Java, with over 50 million residents, is one of the most populous provinces and ranks fourth in the number of administrative regions. In 2023, the Central Bureau of Statistics or Badan Pusat Statistik (BPS) reported a provincial poverty rate of 7.62%, with Indramayu District recording the highest rate at 12.13% [2].

The national socio-economic survey or *survei sosial ekonomi nasional* (SUSENAS), conducted biannually by the BPS in March and September, provides comprehensive data on Indonesia's population's social and economic conditions. Covering key aspects such as employment, education, sanitation, housing, and food consumption [3]. SUSENAS serves as a vital source for formulating and evaluating development policies, including sectoral programs, the sustainable development goals (SDGs), the national medium-term development plan or *rencana pembangunan jangka menengah nasional* (RPJMN), and Astacita—the long-term national vision [4].

Diverse socio-economic factors influence poverty in West Java, necessitating a comprehensive analysis to identify the most significant indicators. Feature selection is a key strategy for managing high-dimensional data by isolating the most relevant variables, thereby improving model performance and reducing computational complexity [4]. This process enhances accuracy, minimizes overfitting, and supports efficient pattern recognition in machine learning applications. Feature selection can be applied before or integrated into the modeling process to mitigate prediction errors. It is generally categorized into three approaches: filter methods, which assess variables independently using statistical measures; wrapper methods, which evaluate subsets based on model performance; and embedded methods, which incorporate feature selection during model training, combining the strengths of both filter and wrapper techniques.

This study evaluates and compares feature selection methods to identify the most effective approach for analyzing large and complex datasets. Filter methods are computationally efficient and suitable for big data, but may overlook interactions among features. Wrapper methods integrate learning algorithms to select optimal subsets, offering strong performance for high-dimensional or imbalanced data, though at higher computational cost. Embedded methods combine the efficiency of filters with the predictive strength of wrappers, balancing accuracy and cost. These approaches are particularly relevant for multidimensional datasets such as SUSENAS, which captures numerous interrelated poverty indicators. Feature selection was applied as a preprocessing step to the West Java dataset, improving predictive performance and highlighting key poverty determinants.

2. METHOD

This study aims to identify key indicators influencing poverty in Indonesia using a dataset comprising 226 variables from household, individual, and neighborhood-level sources. Exploratory data analysis is conducted to assess data quality, followed by a 70:30 train-test split. Initial models using random forest (RF) and logistic regression (LR) are developed without feature selection and evaluated with confusion matrix metrics. Feature selection is then performed using filter methods–information gain (IG) and chi-squared (CS)–and wrapper methods–sequential forward selection (SFS) and genetic algorithm (GA)–to improve model accuracy. Embedded approaches, including least absolute shrinkage and selection operator (LASSO) and ElasticNet, are also applied for simultaneous feature selection and model training. All analyses are conducted using Python, as illustrated in Figure 1.

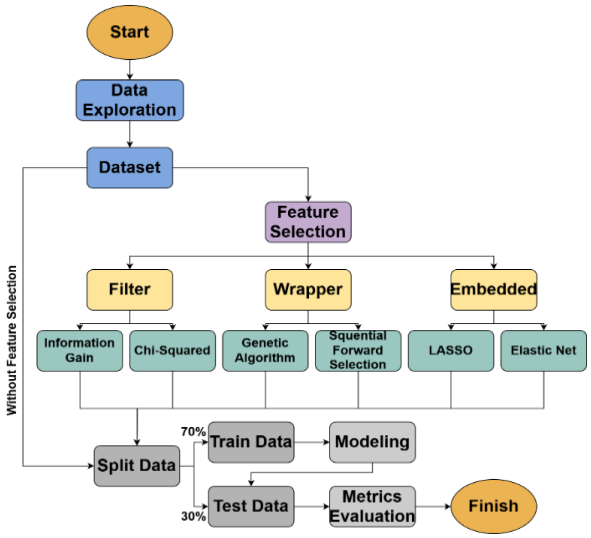

Figure 1. Flowchart on feature selection procedure

2.1. Data

This study utilizes primary data on poverty classification in West Java Province, Indonesia, sourced from the official BPS West Java website. Data were collected through household interviews across all districts and cities, resulting in 25,890 observations and 226 variables organized into thematic categories (Table 1). The binary response variable (Y) classifies households as 'poor' or 'not poor' based on the KAPITA variable, which measures per capita expenditure. Using the 2023 poverty threshold of IDR 495,229, 1,243 households are identified as 'poor'. Independent variables (X) originate from SUSENAS components, capturing multidimensional indicators across social, economic, and environmental sectors.

Table 1. Dataset independent feature categories

Feature categories	Number of variables	Feature categories	Number of variables
Food	22	Employment	5
Housing	13	Security	12
Sanitation	20	Technology	14
Health	11	Government assistance program	91
Education	9		
Economy	20	General information	9

2.2. Feature selection

Feature selection is a dimensionality reduction technique that identifies the most relevant variables while removing redundant or irrelevant ones, thereby improving interpretability, efficiency, and predictive performance. Although excluding irrelevant features may not always directly increase accuracy, it reduces noise and ambiguity during training, enabling more effective learning. Feature selection methods are generally categorized as supervised, unsupervised, or semi-supervised, depending on the availability of labeled data [5]. Furthermore, based on their search strategies and integration with learning algorithms, they are grouped into filter, wrapper, embedded, and hybrid approaches. Each approach demonstrates specific strengths and limitations [6]–[8], as summarized in Table 2.

Table 2. Comparatives feature selection methods

Method	Strength	Limitation
Filter	– Low computational cost	– Relies on features characteristics
	– Suitable for high-dimensional or large dataset	– Limited in handling bias
Wrapper	– Uses predictive algorithms	– High computational cost
	– Accounts for feature−model interactions	– Unsuitable for large datasets
		– Risk of overfitting
Embedded	– Considers feature–model interactions	– Dependent on parameter settings (e.g., alpha and threshold)
	– Computationally efficient	– Sensitive to feature correlation
Hybrid	– Combines strengths of multiple methods	– Dependent on chosen methods
	– Effective for complex data	– High computational cost

The filter method is a pre-processing approach that selects relevant variables based on statistical properties, independent of any learning algorithm. Its algorithm-agnostic nature makes it computationally efficient and broadly generalizable compared to wrapper methods [9]. However, the absence of model-based feedback can lead to suboptimal results, as important features may be excluded [5]. The filter approach generally follows two steps: ranking features by statistical scores and selecting those above a defined threshold. Common techniques include IG and CS. The functions for IG and CS are defined in (1) and (2), respectively. IG evaluates the dependency between predictors and the target variable using entropy, producing ranked scores from highest to lowest. To determine relevant features, IG values can be visualized, allowing users to eliminate variables with minimal contribution. In contrast, CS assesses association strength through chi values or p-values. Features with p-values between 0.01 and 0.05 are typically retained, as prior studies show they improve model performance with fewer variables [10].

$$IG(X,Y) = H(X) - H(X|Y) \tag{1}$$

$$X^2(t,c) = \frac{N(A_c D_c - C_c B_c)^2}{(A_c+C_c)(B_c+D_c)(A_c+B_c)(C_c+D_c)} \tag{2}$$

The wrapper method integrates classification algorithms into the training process, enabling direct evaluation of variable interactions. Unlike filter methods, wrappers generally yield more accurate and

problem-specific feature subsets because their evaluation is driven by algorithmic performance [11]. In this study, LR and RF were used as estimators, with SFS and GA representing wrapper approaches. SFS incrementally adds features that maximize predictive performance, while GA employs evolutionary principles to search for optimal feature combinations. To ensure reliability and reduce overfitting, both methods were applied with five-fold cross-validation, though this substantially increased computational cost. GA is particularly notable for its evolutionary optimization framework, which evaluates candidate subsets using a fitness function as in (3). The fitness function measures model performance—primarily classification accuracy—while also capturing stability and adaptability to data patterns. Features are encoded as binary strings, where "1" represents inclusion and "0" exclusion. High-fitness individuals are selected for reproduction, and crossover operations generate offspring by recombining subsets. To preserve diversity and avoid premature convergence, mutation introduces random alterations to the chromosome structure. In this study, GA was tuned with a population size of 10 chromosomes per generation, a crossover probability of 0.5, and a mutation rate of 10%. The maximum number of features was predefined to maintain computational feasibility. By balancing exploration and exploitation, GA identifies subsets that enhance classification performance, demonstrating its effectiveness in high-dimensional feature selection tasks [12].

$$fitness = \propto_{ov} \bar{S}_{ov} + \propto_{ss} \bar{S}_{ss} + \propto_{tr} \bar{S}_{tr} + \frac{\beta}{n_f} \quad (3)$$

The embedded method provides an efficient approach to feature selection by integrating variable selection directly into the model training process. Positioned between filter and wrapper methods, it combines computational efficiency with predictive accuracy, eliminating the need for repeated model training [5]. Feature selection occurs during training, allowing the method to identify optimal subsets of variables based on the learning algorithm. Unlike wrapper methods, however, it does not iteratively evaluate all variable combinations [10]. Two widely applied embedded techniques are the LASSO and ElasticNet. LASSO applies L1 regularization, which reduces overfitting by shrinking less important coefficients to zero, making the model simpler and more interpretable, as formulated in (4) [13]. In this study, alpha values were tuned between 0.001 and 0.1, with five-fold cross-validation and a maximum of 10,000 iterations to identify the optimal threshold. ElasticNet combines L1 and L2 regularization to address multicollinearity while eliminating irrelevant features, as expressed in (5). Its performance relies on tuning the alpha parameter and the L1 ratio (0–1), where values near zero resemble ridge regression (L2) and higher values emphasize LASSO (L1). In this study, five-fold cross-validation was applied to identify the optimal balance, minimizing overfitting and enhancing accuracy. Additionally, a threshold based on the median coefficient value was used to refine feature selection and improve model interpretability.

$$\left(\hat{\beta}\right) = \arg min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2 \right\} \; subject \; to \; \sum_{j=1}^{k} \|\beta\|_1 < t \quad (4)$$

$$\min_{\beta} \left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \left(\lambda(\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2 \right) \right) \quad (5)$$

Hybrid methods combine the strengths of multiple techniques, such as filter–wrapper or filter–embedded combinations, to balance robustness and efficiency [14]. However, limitations remain. In hybrid approaches that rely on filters in the initial stage, potentially relevant features can be discarded because filters cannot model feature-to-feature interactions. This shortcoming may reduce the effectiveness of subsequent wrapper or embedded steps [6]. Moreover, wrapper–embedded combinations, while powerful, can be computationally expensive when applied to high-dimensional or large-scale datasets. Despite these challenges, selecting the appropriate feature selection strategy remains critical to optimizing model accuracy and interpretability, especially when analyzing complex, multidimensional data from large survey sources.

2.3. Classification approaches

To evaluate the effect of feature selection on model performance, it is essential to compare models developed with and without feature selection. LR and RF are utilized as classification techniques in this analysis. LR is a statistical method used for binary classification (e.g., "yes" or "no", "0" or "1") based on one or more independent variables. It establishes relationships between variables through regression analysis, similar to linear regression [15]. In contrast, RF, a machine learning algorithm, is particularly effective for modeling nonlinear relationships. It is an ensemble method, employing bagging to generate bootstrap samples and applying decision trees (DT). By training on multiple data subsets, RF mitigates overfitting and reduces the influence of outliers [16], [17].

3. RESULTS AND DISCUSSION

3.1. Data exploration

Exploratory data analysis identified three types of independent variables: continuous, coded, and categorical. Continuous variables measure values such as monthly per capita or food expenditures. Coded variables denote regional or demographic identifiers, including province, district, or city codes. Categorical variables capture survey-based classifications, such as employment sector, income group, or housing type. The predominance of categorical variables reflects the survey-driven nature of the dataset. Their distribution is illustrated in Figure 2, forming the basis for method selection.

Figure 3 illustrates the distribution of the response variable categories, which are imbalanced, with a disproportionate representation between classes 8 and 1. This imbalance may introduce bias and reduce the predictive accuracy of the model. To mitigate this issue, the distribution will be adjusted by undersampling class 0 to 80% of the instances in class 1. A missing value analysis was performed to ensure data integrity, and the results indicated that no missing values were present in any dataset's features.

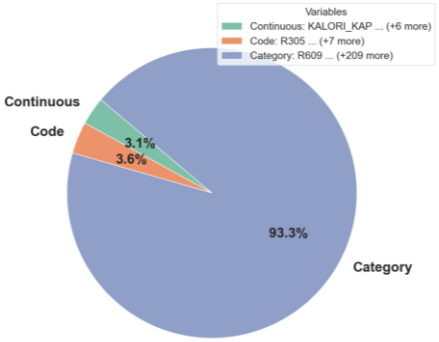

Figure 2. Data category distribution

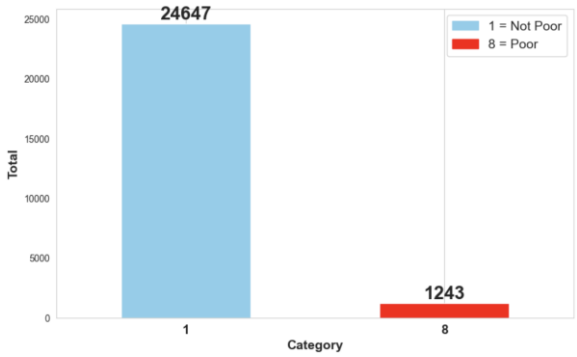

Figure 3. Distribution of response variable

3.2. Process features selected

IG method evaluates feature relevance by calculating entropy, where features that reduce uncertainty in the response variable are deemed more informative [18]. To determine the optimal threshold, entropy values were visualized through a line chart, with 0.01 selected as the cutoff. Using this criterion, 55 features were retained while 171 were eliminated, effectively reducing the dataset by more than 75%. The equilibrium point for feature selection is presented in Figure 4. In contrast, CS method measures dependency between predictors and the response variable through hypothesis testing [19]. While the overall CS statistic was 1.268044, additional refinement was required, prompting the use of p-values as a selection criterion. Therefore, p-values were employed as an additional selection criterion, enabling a more precise assessment of statistical significance [20]. Features with p-values ranging from 0.01 to 0.05 were retained, resulting in the selection of only 5 variables, as summarized in Table 3. Compared to IG, the CS method produced a much smaller feature subset, eliminating over 95% of variables, but offering higher statistical rigor in identifying the most significant predictors.

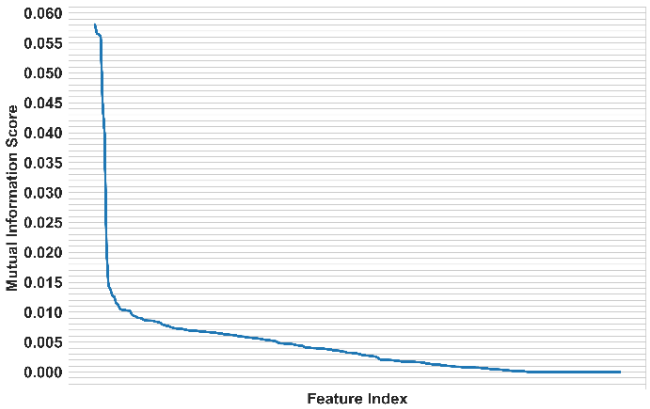

Figure 4. Chart of IG entropy value per variable

Table 3. Number of features selected

Method		Classifier		The ratio of the selected features to the original features (%)
		LR	RF	
Filter	IG	15	15	6.63
	CS	5	5	2.12
Wrapper	SFS	112	112	49.56
	GA	83	85	36.73-37.61
Embedded	LASSO	85	85	37.61
	ElasticNet	225	225	99.56

Wrapper-based approaches, including SFS and GA, apply classification models such as LR and RF to evaluate candidate feature subsets. SFS begins with an empty set and sequentially incorporates features that yield the greatest improvement in model performance, continuing until the stopping criteria are satisfied. Using LR and RF as estimators, SFS consistently selected 112 features, eliminating over 50% of the dataset. In contrast, GA adopts an evolutionary optimization process, encoding features as binary chromosomes and evaluating subsets based on a fitness function that integrates classification accuracy, stability, and adaptability [12]. Through genetic operations such as crossover and mutation, GA explores diverse feature combinations, enhancing robustness in feature selection. The results reveal that GA-LR selected 83 features, while GA-RF selected 85 features, both achieving more compact subsets compared to SFS. Overall, GA demonstrated superior efficiency in reducing dimensionality while maintaining relevant features, highlighting its effectiveness for high-dimensional poverty data.

LASSO is a regularization technique that simultaneously performs feature selection by shrinking insignificant coefficients to zero. This mechanism effectively removes irrelevant variables, reduces dimensionality, and improves interpretability while minimizing overfitting risks. The degree of shrinkage is determined by a tuning parameter that regulates the number of predictors retained [21]. In this study, LASSO demonstrated strong capability in dimensionality reduction by eliminating more than 60% of features, leaving only 85 variables for modeling. This result highlights its effectiveness in producing a more compact and interpretable model. On the other hand, ElasticNet integrates the L1 penalty of LASSO with the L2 penalty of ridge regression, making it particularly effective for handling correlated predictors and improving stability in high-dimensional data [22]. However, ElasticNet was more conservative in feature elimination, removing only one variable and retaining 225 for modeling. Overall, LASSO provided greater dimensionality reduction, whereas ElasticNet emphasized stability, demonstrating two complementary approaches to feature selection.

3.3. Feature selection performance

As presented in Table 4, the filter, wrapper, and embedded methods outperform the model without feature selection in the LR classification approach. Precision reflects the proportion of correctly identified positive instances among all predicted positives, while recall indicates the model's ability to capture all positive cases. Accuracy represents the overall proportion of correct predictions, and the F1-score offers a harmonic balance between precision and recall [23]. Among all methods, CS notably improved LR performance by increasing accuracy, recall, and F1-score, albeit with a higher computational cost. Similarly, LASSO, as an embedded method, achieved performance comparable to CS but required longer computation. Wrapper approaches such as SFS and GA also enhanced accuracy, precision, recall, and F1-score, though at substantial computational expense. Consistent with Aonpong *et al.* [24], LASSO has been shown to improve radiomics models while reducing feature dimensionality by up to 81%. CS, a univariate statistical method, selects features individually based on their significance to the response variable [25]. In this study, CS retained five features using p-value thresholds, although it does not account for feature interactions or redundancy.

Table 4. LR classification approach

Method		Metric					
		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Computation time	ANOVA test
No feature selection		83.29	96.49	85.57	90.70	6.78"	0.89
Filter	IG	79.82	96.76	81.54	88.50	0.09"	0.87
	CS	90.89	95.43	94.98	95.20	0.03"	0.94
Wrapper	SFS	88.10	99.37	88.08	93.39	42,479.45"	0.92
	GA	86.62	99.24	86.63	92.51	6,085.34"	0.91
Embedded	LASSO	88.68	99.33	88.70	93.72	93.52"	0.93
	ElasticNet	81.78	96.51	83.87	89.74	101.82"	0.88

As shown in Table 5, wrapper methods generally outperform models without feature selection, providing notable improvements in predictive performance. However, embedded methods sometimes yield less optimal outcomes, showing only a slight increase in accuracy but reductions in precision and F1-score compared to models using all features. Previous studies support these findings. For instance, Tchakoucht *et al.* [26] highlighted that the RF algorithm is highly effective for high-dimensional data due to its strong generalization ability. By constructing ensembles of DT on random subsets of features and aggregating predictions via majority voting, RF reduces sensitivity to irrelevant variables, thereby enhancing robustness. Similarly, Nadya *et al*. [27] reported that DT combined with IG improves accuracy, precision, and recall compared to using all features, while DT with LASSO yielded slightly lower results, differing by less than 1-2%. In line with these findings, this study also shows that IG-RF achieved higher accuracy than LASSO, ElasticNet, and the all-feature model by more than 2%. Interestingly, while most methods improved performance, CS produced lower accuracy than the no-feature-selection model, a contrasting outcome compared to its superior performance in LR.

Table 5. RF classification approach

Method		Metric					
		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Computation time	DMRT test
No feature selection		95.69	99.63	95.82	97.69	3.12"	A
Filter	IG	98.44	99.59	98.77	99.17	3.24"	B
	CS	91.64	91.37	91.64	91.51	0.74"	A
Wrapper	SFS	96.32	99.62	96.50	98.04	30,374.36"	A
	GA	96.99	99.69	97.14	98.40	254.25"	A
Embedded	LASSO	95.99	95.25	95.99	95.28	95.83"	A
	ElasticNet	95.76	97.16	95.76	96.23	97.94"	A

Annotation: methods sharing the same group notation indicate no significant differences.

CS method produced the lowest performance compared to other feature selection techniques and even to models without feature selection, despite its lower computational cost. CS selected only five features from 226 variables, which may be unsuitable for RF. As an ensemble of DT trained on bootstrap samples, RF relies on diverse feature subsets to maximize performance [17]. Consequently, excessive feature reduction can limit its ability to capture data variability. In contrast, a study by Agustina *et al*. [28] demonstrated that RF combined with wrapper-based feature selection achieved superior outcomes, improving accuracy, precision, recall, F1-score, and receiver operating characteristic (ROC) by approximately 2% compared to RF with filter methods. These findings highlight the importance of aligning feature selection strategies with model characteristics.

Feature selection using LR approaches, evaluated with a single-factor analysis of variance (ANOVA) test, showed no significant differences among methods (p-value=0.83). Consequently, further statistical testing with Duncan's multiple range test (DMRT) was not applied. In contrast, RF-based approaches yielded an ANOVA p-value <0.05, meeting the criteria for DMRT analysis. The results indicate that methods such as CS, SFS, GA, LASSO, and ElasticNet, as well as the baseline model without feature selection, did not differ significantly in performance. However, IG demonstrated a statistically significant improvement compared to other methods, highlighting its effectiveness in enhancing model performance within the RF framework.

3.4. The best approach

The LASSO method was identified as the most effective feature selection technique for the poverty dataset of West Java, Indonesia, particularly when applied with LR. By selecting 85 features, LASSO improved model performance by 3-5%, outperforming the CS method, which only increased accuracy, recall, and F1-score, albeit with lower computational cost. LASSO also achieved higher area under the curve (AUC) and ROC scores than CS, indicating stronger discriminative ability. According to Chicco and Jurman [29], CS may be more prone to threshold sensitivity, leading to weaker performance in predicting positive cases. As shown in Figure 5, LASSO consistently outperformed other methods with LR, while Figure 6 highlights IG-RF and LASSO-LR combinations as having excellent AUC values (>0.9), consistent with Çorbacıoğlu and Aksel [30].

Figure 7 shows the confusion matrices of the classification models used in this study, with Figure 7(a) representing the RF-IG model and Figure 7(b) representing the LR-LASSO model. Both approaches effectively classified class "1" (not poor) and class "8" (poor), with the majority of predictions being correct. However, LR+LASSO showed a higher number of false negatives, with 835 samples misclassified as "not poor." This indicates that while the method performs well overall, it struggles in accurately identifying the "poor" class. As noted by Zhang *et al.* [31], LASSO often exhibits a higher false-negative rate in test sets, exceeding 30% compared to training. This suggests class imbalance, which is also evident in this study's dataset.

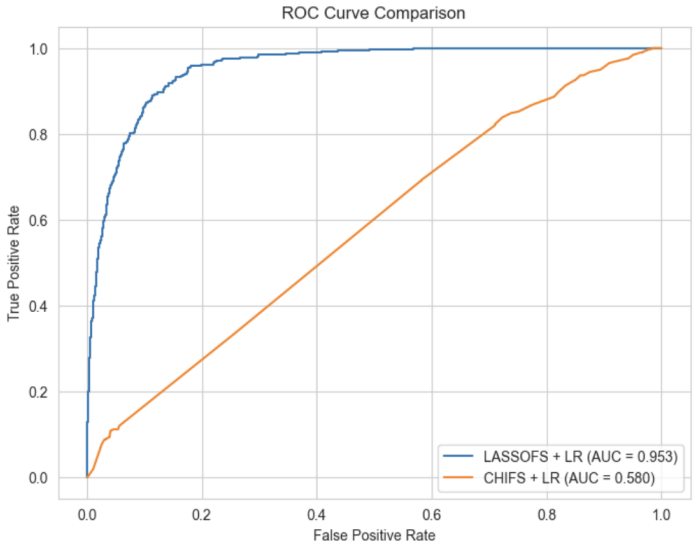

Figure 5. ROC curve and AUC score comparison LASSO-RL and CS-RL

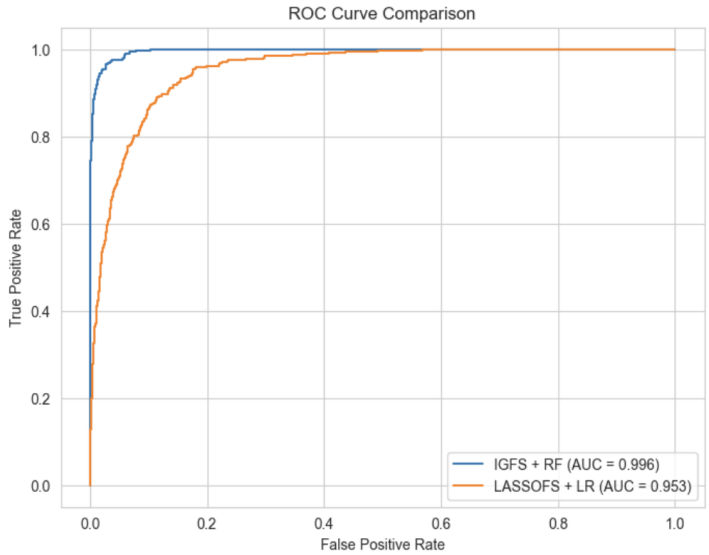

Figure 6. ROC curve and AUC score comparison LASSO-LR and IG-RF

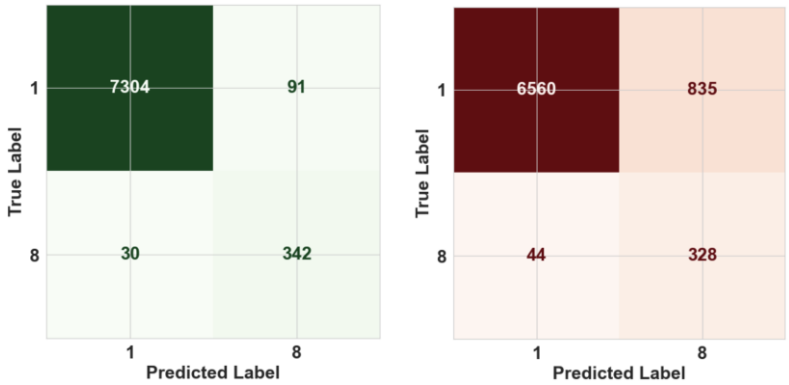

Figure 7. Confusion matrix of (a) RF+IG and (b) RL+LASSO models

3.5. Application impact

Table 6 presents the feature categories used in this study. General information includes the respondent's location (city/district) and household size. Education covers literacy level, schooling, and reading habits. Economy relates to non-food expenditures, such as clothing or vacations. Employment reflects income sources and job level. Technology captures access through electronic ownership. Security records crime experiences. Health includes smoking habits and healthcare visits. Food measures hunger experiences and food expenditure. Housing addresses home size, electricity, and cooking fuel. Sanitation assesses water sources, hygiene practices, and septic facilities. Government assistance records aid frequency and value received. LASSO-LR and IG-RF were identified as the most effective approaches for predicting poverty in the West Java dataset. Key determinants of poverty include economic conditions, access to technology, food security, housing, and participation in government programs. LASSO-LR additionally highlighted sanitation as a significant factor. According to Mora and Rivera [32], households with internet access experience lower poverty levels compared to those without. Government initiatives such as family hope program or *program keluarga harapan* (PKH) and non-cash food assistance or *bantuan pangan non-tunai* (BPNT) play crucial roles in reducing poverty by improving access to education, healthcare, and food security [33]. As stated in Headey and Martin [34] rising food prices negatively affect household income, deepening poverty gaps and vulnerability.

Table 6. Selected feature categories based on optimal approaches

Features categories	LASSO-LR	IG-RF	Features categories	LASSO-LR	IG-RF
General information	7	2	Health	9	-
Education	2	-	Food	12	7
Economy	9	1	Housing	11	2
Employment	3	-	Sanitation	8	-
Technology	11	1	Government assistance program	13	2
Security	-	-			

Figure 8 presents the features commonly selected by both models, which include KALORI_KAP, KARBO_KAP, LEMAK_KAP, and R1702, representing food and nutritional consumption per capita. Features such as R2001B, R2001K, and R2001L reflect household access to technology, while R1817 relates to housing conditions, specifically cooking fuel. In Indonesia, cooking fuel serves as a key poverty indicator. Many rural households in West Java, particularly those below the poverty line, still rely on firewood, whereas urban and financially stable households predominantly use gas. This aligns with Pangaribowo and Iskandar [35], which reported that nearly 90% of the poorest rural households in Indonesia continue to use firewood, underscoring the strong relationship between economic conditions and reliance on traditional fuels.

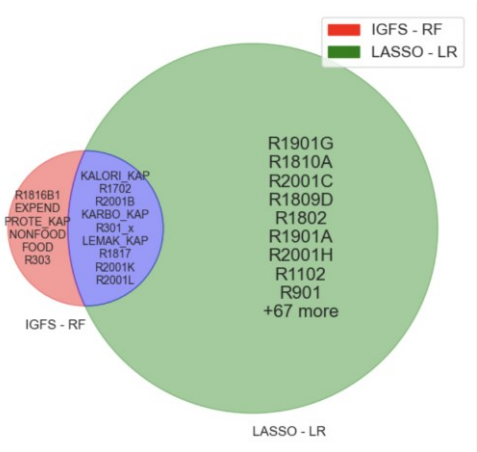

Figure 8. Features selected from best approaches

The findings from the West Java study highlight priority sectors for poverty reduction. Ensuring access to adequate nutrition and affordable food ingredients is essential, as food adequacy strongly influences household welfare. Access to technology also emerges as a key factor, enabling the government to better

target households eligible for assistance programs, ensuring alignment with SDGs 1 ("no poverty") and 2 ("zero hunger"). The feature R301_x, representing household size, indicates that larger households face higher poverty risks. This aligns with Soseco [36], which reported that household size significantly affects education and net wealth. These results suggest that population control and improved access to education can enhance household income, strengthen food security, and ultimately reduce poverty and hunger in West Java.

This study demonstrates that feature selection effectively identifies key indicators of poverty in West Java, Indonesia. While the approach can be adapted to other provinces, varying socio-economic and environmental conditions may produce different indicators. For short-term strategies, the combination of IG and RF provides optimal predictive performance, whereas LASSO with LR offers robust long-term solutions. The results highlight household size, food consumption, technology access, housing conditions, and participation in government assistance programs as primary poverty indicators. These findings provide valuable benchmarks for policymakers in West Java to design targeted poverty reduction strategies.

4. CONCLUSION

This study identifies the combination of LASSO with LR and IG with RF as the most effective approaches for reducing features aligned with poverty indicators in West Java, Indonesia. Using SUSENAS data, both approaches successfully improved model performance while substantially reducing dataset dimensionality. LASSO-LR increased accuracy by more than 5% compared to models without feature selection, though with higher computational cost and no statistically significant differences. Meanwhile, IG-RF improved accuracy, recall, and F1-score by 2-3%, with statistically significant differences across methods, and reduced over 65% of the dataset's features. These reductions enable clearer identification of priority sectors that government policies should target. In the short term, interventions should emphasize food sufficiency and affordability, improved housing conditions, and birth control programs. In the longer term, policies can extend to strengthening economic opportunities, employment, technology access, health, and sanitation. By focusing on these identified areas, the West Java government can implement more targeted and efficient poverty alleviation strategies, ultimately reduce poverty levels and address hunger simultaneously. This study thus provides a practical framework for evidence-based policymaking.

FUNDING INFORMATION

This research was supported by funding from the School of Data Science, Mathematics, and Informatics, IPB University.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sean Marshelle	✓	✓	✓		✓		✓	✓	✓	✓	✓		✓	✓
Septian Rahardiantoro	✓	✓		✓	✓	✓	✓			✓		✓		✓
Anang Kurnia	✓	✓		✓	✓	✓	✓			✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

The data used in this study are available upon reasonable request from the corresponding author, [SM], or from the Central Bureau of Statistics of West Java Province, Indonesia.

REFERENCES

[1] H. Hill, "What's happened to poverty and inequality in Indonesia over half a century?," *Asian Development Review*, vol. 38, no. 1, pp. 68–97, 2021, doi: 10.1162/adev_a_00158.

[2] BPS West Java, "Poverty in regencies/cities in West Java Province 2018-2023 *(in Indonesian*: *Kemiskinan kabupaten/kota di Provinsi Jawa Barat 2018-2023*)," *Badan Pusat Statistik*. 2024. Accessed: Feb. 20, 2025. [Online]. Available: https://jabar.bps.go.id/id/publication/2024/08/19/14c9831d5dd0fda3387a5054/kemiskinan-kabupaten-kota-di-provinsi-jawa-barat-2018-2023.html

[3] BPS, "March 2023 national socio-economic survey (KOR) (in Indonesian: *Survei sosial ekonomi nasional 2023 Maret (KOR)*)," *Badan Pusat Statistik*. 2023. Accessed: Feb. 20, 2025. [Online]. Available: https://silastik.bps.go.id/v3/index.php/mikrodata/detail/ZnZSZms4aStzN2JUSVY1QklqZ08rdz09

[4] B. Venkatesh and J. Anuradha, "A review of feature selection and its methods," *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3–26, 2019, doi: 10.2478/cait-2019-0001.

[5] S. Wang, J. Tang, and H. Liu, "Feature selection," in *Encyclopedia of Machine Learning and Data Mining*, New York, United States: Springer, 2017, pp. 503–511, doi: 10.1007/978-1-4899-7687-1_101.

[6] N. Pudjihartono, T. Fadason, A. W. K. -Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, pp. 1–17, 2022, doi: 10.3389/fbinf.2022.927312.

[7] Y. Syahidin and A. I. Suryani, "Hybrid for feature selection algorithms in the field of medical records," *International Journal of Psychology and Health Science*, vol. 1, no. 3, pp. 111–119, 2023, doi: 10.38035/ijphs.v1i3.339.

[8] G. Wei, J. Zhao, Y. Feng, A. He, and J. Yu, "A novel hybrid feature selection method based on dynamic feature importance," *Applied Soft Computing*, vol. 93, 2020, doi: 10.1016/j.asoc.2020.106337.

[9] N. S. -Maroño, A. A. -Betanzos, and M. T. -Sanromán, "Filter methods for feature selection–a comparative study," in *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, Berlin, Heidelberg: Springer, 2007, pp. 178–187, doi: 10.1007/978-3-540-77226-2_19.

[10] D. C. Sami'un, A. Sugiharto, and F. Jie, "Chi square feature selection for improving sentiment analysis of news data privacy treats," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 18, pp. 6601–6610, 2024.

[11] F. Z. Janane, T. Ouaderhman, and H. Chamlal, "A filter feature selection for high-dimensional data," *Journal of Algorithms & Computational Technology*, vol. 17, 2023, doi: 10.1177/17483026231184171.

[12] F. Amini and G. Hu, "A two-layer feature selection method using genetic algorithm and Elastic Net," *Expert Systems with Applications*, vol. 166, 2021, doi: 10.1016/j.eswa.2020.114072.

[13] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[14] A. Kaur, K. Guleria, and N. K. Trivedi, "Feature selection in machine learning: methods and comparison," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 789–795, doi: 10.1109/icacite51222.2021.9404623.

[15] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic regression model optimization and case analysis," in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, 2019, pp. 135–139, doi: 10.1109/iccsnt47585.2019.8962457.

[16] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867x20909688.

[17] G.-W. Cha, H.-J. Moon, and Y.-C. Kim, "Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables," *International Journal of Environmental Research and Public Health*, vol. 18, no. 16, 2021, doi: 10.3390/ijerph18168530.

[18] T. Z. Win and N. S. M. Kham, "Information gain measured feature selection to reduce high dimensional data," in *ICCA 2019 Proceedings*, 2019, pp. 68–73.

[19] M. Nassar, H. Safa, A. Al Mutawa, A. Helal, and I. Gaba, "Chi squared feature selection over Apache Spark," in *Proceedings of the 23rd International Database Applications & Engineering Symposium on - IDEAS '19*, 2019, pp. 1–5, doi: 10.1145/3331076.3331110.

[20] E. Beh, "Exploring how to simply approximate the p-value of a chi-squared statistic," *Austrian Journal of Statistics*, vol. 47, no. 3, pp. 63–75, 2018, doi: 10.17713/ajs.v47i3.757.

[21] V. Fonti, *Feature selection using LASSO*. Amsterdam, Netherlands: Vrije Universiteit Amsterdam, 2017.

[22] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[23] M. Azer, H. H. Zayed, M. E. A. Gadallah, and M. Taha, "Multi platforms fake accounts detection based on federated learning," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 4, pp. 3837–3848, 2024, doi: 10.11591/ijai.v13.i4.pp3837-3848.

[24] P. Aonpong *et al.*, "Comparison of machine learning-based radiomics models for early recurrence prediction of hepatocellular carcinoma," *Journal of Image and Graphics*, vol. 7, no. 4, pp. 117–125, 2019, doi: 10.18178/joig.7.4.117-125.

[25] E. Hokijuliandy, H. Napitupulu, and Firdaniza, "Application of SVM and chi-square feature selection for sentiment analysis of Indonesia's National Health Insurance mobile application," *Mathematics*, vol. 11, no. 17, 2023, doi: 10.3390/math11173765.

[26] T. A. Tchakoucht, B. Elkari, Y. Chaibi, and T. Kousksou, "Random forest with feature selection and k-fold cross validation for predicting the electrical and thermal efficiencies of air based photovoltaic-thermal systems," *Energy Reports*, vol. 12, pp. 988–999, 2024, doi: 10.1016/j.egyr.2024.07.002.

[27] F. E. P. Nadya, M. F. I. Ferdiansyah, V. R. S. Nastiti, and C. S. K. Aditya, "Implementation of feature selection strategies to enhance classification using XGBoost and decision tree," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 18–194, 2024, doi: 10.15294/sji.v11i1.48145.

[28] T. Agustina, M. Masrizal, and I. Irmayanti, "Performance analysis of random forest algorithm for network anomaly detection using feature selection," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 2, pp. 1116–1124, 2024, doi: 10.33395/sinkron.v8i2.13625.

[29] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Mining*, vol. 16, no. 1, 2023, doi: 10.1186/s13040-023-00322-4.

[30] Ş. K. Çorbacıoğlu and G. Aksel, "Receiver operating characteristic curve analysis in diagnostic accuracy studies: a guide to interpreting the area under the curve value," *Turkish Journal of Emergency Medicine*, vol. 23, no. 4, pp. 195–198, 2023, doi: 10.4103/tjem.tjem_182_23.

[31] J. Zhang *et al.*, "Development and validation of a LASSO prediction model for cisplatin induced nephrotoxicity: a case-control study in China," *BMC Nephrology*, vol. 25, no. 1, 2024, doi: 10.1186/s12882-024-03623-w.
[32] F. G. -Mora and J. M. -Rivera, "Exploring the impacts of internet access on poverty: a regional analysis of rural Mexico," *New Media & Society*, vol. 25, no. 1, pp. 26–49, 2021, doi: 10.1177/14614448211000650.
[33] H. Simangunsong and D. Sihotang, "The impact of economic conditions on social assistance programs and poverty alleviation," *Law and Economics*, vol. 17, no. 2, pp. 73–91, 2023, doi: 10.35335/laweco.v17i2.2.
[34] D. D. Headey and W. J. Martin, "The impact of food prices on poverty and food security," *Annual Review of Resource Economics*, vol. 8, no. 1, pp. 329–351, 2016, doi: 10.1146/annurev-resource-100815-095303.
[35] E. H. Pangaribowo and D. D. Iskandar, "Exploring socio-economic determinants of energy choices for cooking: the case of eastern Indonesian households," *Environment, Development and Sustainability*, vol. 25, no. 7, pp. 7135–7148, 2022, doi: 10.1007/s10668-022-02362-y.
[36] T. Soseco, "Household size, education, and household wealth in Indonesia: evidence from quantile regression," *Jurnal Ekonomi Indonesia*, vol. 10, no. 3, pp. 281–297, Feb. 2022, doi: 10.52813/jei.v10i3.72.

BIOGRAPHIES OF AUTHORS

Sean Marshelle obtained a Bachelor of Animal Science from Universitas Padjadjaran, Indonesia. He has served as a research and development supervisor in an agricultural company in Indonesia. In the research and development division, he is involved in data analysis. To further his interest in this field, he continues his studies in a master's program in statistics and data science at IPB University, Indonesia. He has researched various topics, including data visualization, machine learning, and field trial projects. He can be contacted at email: seanmarshelle@apps.ipb.ac.id.

Septian Rahardiantoro earned a Bachelor of Statistics and a Master of Science from IPB University in Indonesia. In 2023, he earned a doctor of philosophy (Ph.D.) from Okayama University. He is a lecturer at the Department of Statistics, IPB University, Bogor, Indonesia. His main teaching and research interests include data science, statistical machine learning, and statistical modeling in environmental science. He can be contacted at email: septianrahardiantoro@apps.ipb.ac.id.

Anang Kurnia is a Statistics and Data Science Professor and Vice Dean of Academic, Students, and alumni affairs at the School of Data Science, Mathematics, and Informatics at IPB University, Indonesia. He is also chairman of the Indonesian Statistical Association (ISI-Indonesia). He is the former head of the Department of Statistics at IPB University (2014-2023) and the former head of the Indonesian Statistics Higher Education Association (2014-2018). His main teaching and research interests include statistical machine learning, statistical inference, generalized linear mixed models, data science, and small area estimation. He has published several research articles in international journals in the area of statistics and data science. He can be contacted at email: anangk@apps.ipb.ac.id.