



Multidimensional livelihood-capital deprivation and income-defined household poverty in Sakon Nakhon, Thailand: Explanatory, predictive, and association-rule evidence

Pita Jarupunphol^a and Wichidtra Sudjarid^{b*}

^a Department of Digital Technology, Faculty of Science and Technology, Phuket Rajabhat University, Phuket, 83000, Thailand

^b Department of Environmental Science, Faculty of Science and Technology, Sakon Nakhon Rajabhat University, Sakon Nakhon, 47000, Thailand

Abstract

Article history:

Received: 19-05-2026

Revised : 29-06-2026

Accepted: 02-07-2026

Published: 18-07-2026

Keywords:

Multidimensional poverty;

Livelihood capital;

Machine learning;

Association rule mining;

Poverty screening

This study examines household poverty in Sakon Nakhon province, Thailand, using a livelihood-capital household survey containing 2,128 records and 73 variables. Household poverty in Sakon Nakhon, Thailand, remains a pressing policy concern, yet existing evidence rarely integrates explanatory, predictive, and co-occurrence evidence within a single locally grounded study. Poverty status was derived by comparing per-capita monthly income with the constant poverty-line field in the dataset; after preprocessing, 1,562 households were classified as poor and 497 as non-poor (poor-class share: 75.86%). Three evidence streams were separated: explanatory statistical analysis, leakage-controlled predictive modeling, and association rule mining (ARM). Group comparisons and robust logistic regression showed that Human Capital (OR = 0.318), Social Capital (OR = 0.568), Physical Capital (OR = 0.513), and agricultural green-book status (OR = 0.721) were associated with lower poverty odds, while household size (OR = 1.487) was associated with higher odds. Extra Trees achieved the strongest ROC-AUC (0.730); L2-regularized logistic regression achieved the strongest balanced accuracy (0.658). ARM identified recurring poverty-associated profiles in Kusuman district, characterized by low social participation and large households with low physical or social capital. This study demonstrates a reproducible, data-driven workflow for community-level poverty screening and policy prioritization in resource-constrained provincial settings. Its creative science contribution lies in integrating livelihood-capital theory, explainable statistical modeling, leakage-controlled machine learning, and association-rule mining into a single practical framework for identifying locally relevant poverty profiles and supporting sustainable community development.

* Corresponding author : wichidtra.s@snru.ac.th

DOI: <https://doi.org/10.55674/cs.v18i3.268007>

© [2026] The Author(s).

1. Introduction

Poverty in provincial Thailand is not only an income-based condition; it is also expressed through uneven access to education, skills, health, housing, infrastructure, social participation, and local problem-solving capacity, aligning with multidimensional poverty and livelihood-capital research [1 – 3]. Sakon Nakhon, a province in Northeastern Thailand, provides a relevant setting for examining how household livelihood conditions align with income-defined poverty in the available Sakon Nakhon dataset, because recent Thailand- and Sakon Nakhon-focused studies show

spatially uneven poverty factors and local livelihood-capital poverty applications [4, 5].

This article uses household survey data from Sakon Nakhon to study income-defined poverty through three deliberately separate analytical lenses. First, explanatory statistical analysis estimates associations between livelihood-capital domains and poverty status. Second, predictive machine learning evaluates whether non-income household conditions can screen for poverty risk without using direct income variables as predictors [6 – 8]. Third, association rule mining describes interpretable co-occurrence patterns among deprivation indicators and

poverty status. The three methods are used together because regression estimates adjusted associations, machine learning evaluates predictive discrimination, and ARM surfaces frequent co-occurrence profiles of deprivation, each answering a distinct empirical question. The study does not interpret any result as causal because the data are cross-sectional and observational.

Previous studies have examined poverty prediction, spatial patterns, and livelihood-capital factors in Thailand; however, few have simultaneously combined explanatory regression, machine-learning screening, and association-rule mining within a single locally grounded dataset, particularly at the provincial level in Northeastern Thailand. The present study addresses this gap. The study uses the available PPAOS-derived Sakon Nakhon household dataset; all findings should be interpreted within this dataset unless external representativeness can be independently verified.

In the context of creative science, the study is innovative because it combines social-science poverty theory with computational data analytics to generate interpretable and actionable knowledge. Rather than treating machine learning as a stand-alone prediction tool, the workflow integrates statistical explanation, predictive screening, and rule-based pattern discovery to support community development decisions in Sakon Nakhon. As such, the study has five objectives:

1. To audit the Sakon Nakhon household poverty dataset and document missingness, duplicates, class imbalance, and analytical limitations.
2. To estimate statistical associations between livelihood-capital domains, household characteristics, district context, and income-defined poverty.
3. To evaluate leakage-controlled machine-learning models for non-income poverty-risk screening.
4. To identify descriptive association rules that reveal recurring deprivation co-occurrence patterns related to poverty.
5. To integrate statistical, predictive, and ARM evidence without conflating association, prediction, co-occurrence, or causation.

2. Materials and Methods

Data Source and Unit of Analysis

The dataset was obtained from the Overview of Poor Households: 20 Provinces dashboard of Thailand's Provincial Poverty Alleviation Operating System (PPAOS), with a focus on Sakon Nakhon Province. The data were extracted for the period between January and May 2025. The analytical workbook contained 2,128 household records and 73 variables. The household was used as the unit of analysis. The data consist of administrative poverty-targeting records drawn from households already registered in the PPAOS system in Sakon Nakhon Province. They do not represent a random sample of all provincial households; accordingly, findings should not be generalized to the full provincial population. Variable interpretation was based primarily on the English field names in the workbook and was cross-checked against the accompanying data dictionary where

readable. Because some Thai-font encoding in the data dictionary limited direct text extraction, ambiguous fields were interpreted cautiously and were not used as primary explanatory anchors unless their meaning was clear from the workbook structure.

Outcome Definition

The primary outcome was an income-defined binary poverty variable. Per-capita monthly income was computed by dividing annual household income by 12 and the number of household members. A household was classified as poor when its per-capita monthly income was below the poverty line field available in the dataset. Records with missing income, missing poverty-line values, or non-positive household size were excluded from target-based analysis. The poverty line was constant across all records in this dataset extract, likely reflecting a single administrative threshold used in the PPAOS system at the time of extraction. This means all households are evaluated against the same threshold, which limits variation in the income line, simplifies target construction, and ensures transparent replicability. Because the poverty-line field was constant across this dataset extract, the resulting outcome should be interpreted as an income-defined analytical proxy for poverty. To prevent leakage, direct income variables, poverty-line fields, and financial-capital variables were excluded from the primary predictive feature set.

Per-capita monthly income was calculated by dividing annual household income by 12 months and by the number of household members living in the household, as shown in equation (1), where PCI_m is per-capita monthly income, I_1 is annual household income, and N_h is the number of household members actually living in the household.

$$PCI_m = \frac{I_1}{12N_h} \quad (1)$$

Poverty status was transformed into a binary outcome variable, where households with per-capita monthly income below the poverty line were coded as poor, and all other households were coded as non-poor, as shown in equation (2), where Y_i represents the binary poverty status of the household i , $PCI_{m,i}$ represents the per-capita monthly income of the household i , and PL_1 represents the poverty line threshold.

$$Y_i = \begin{cases} 1, & \text{if } PCI_{m,i} < PL_1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Records with missing income, missing poverty line, or non-positive household size were excluded from target-based analysis. Direct income variables, poverty-line fields, and financial-capital fields were excluded from the primary predictive feature set to avoid leakage. This income-defined target is intentionally narrower than multidimensional poverty measurement, which treats poverty as a joint pattern of deprivation across domains such as education, health, and living standards [1, 3]. The dataset audit is summarized in Table 1.

Table 1 Dataset audit summary.

Item	Result
Raw records	2,128
Columns	73
Districts represented	7
Sub-districts represented	26
Exact duplicate rows	2
Target-valid records before exact-duplicate handling	2,059
Poor households in the target valid data	1,562
Non-poor households in the target valid data	497
Poor share	75.86%
Highest-missing variable	Agricultural Household Number, 62.69% missing

Note. Target-valid records had non-missing income, poverty-line, and household-size values. Poverty status was derived from per-capita monthly income. Missingness is reported for the original dataset.

Analytical Streams

The study uses four streams. The descriptive audit documents missingness, duplicates, distributions, and class imbalance. The explanatory stream uses group comparisons, chi-square tests, and robust logistic regression. The predictive stream compares a baseline classifier, L2-regularized logistic regression, random forest, Extra Trees, and gradient boosting under stratified cross-validation; these model families are widely used for interpretable or ensemble-based tabular prediction [9–12]. Extra Trees and Gradient Boosting were selected for their strong empirical performance on tabular datasets with moderate sample sizes; L2-regularized logistic regression was included as an interpretable baseline with regularization to prevent overfitting; Random Forest was included for ensemble comparison. The ARM stream constructs household transactions from district labels, poverty status, household-size indicators, green-book status, low/high capital items, and low component-score items; direct income and financial-capital items were excluded. Continuous capital scores were discretized into low/high categories using a median split applied to the full target-valid sample: a score at or below the sample median was coded as low; a score above the median was coded as high. This approach was chosen for interpretability in ARM itemset encoding.

Mathematical Formulation

Most of the mathematical formulations in this section follow standard presentations in the data mining and machine learning literature [13, 14].

Predictive Modeling Formulation

For predictive modeling, consistent with standard supervised classification notation, a feature vector

represents each household X_i , and the model estimates as in equation (3), where $\hat{y}_i$ is the predicted probability or class label for poverty, and $f(\cdot)$ is the fitted classifier. This formulation is applied across baseline models, including logistic regression, random forest, Extra Trees, and gradient boosting.

$$\hat{y}_i = f(X_i) \quad (3)$$

Logistic Regression

For logistic regression, the probability of poverty is defined as in equation (4), where the exponentiated coefficient e^{β_j} is interpreted as the odds ratio for the predictor x_{ij} , conditional on the other variables in the model. Because the data are observational, these odds ratios should be viewed as adjusted associations rather than causal effects.

$$P(y_i = 1 | X_i) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^p \beta_j x_{ij})}} \quad (4)$$

Loss Function (Binary Cross-Entropy)

For probabilistic binary classifiers, the classification loss is expressed as in equation (5), where y_i is the observed poverty label and $\hat{y}_i$ is the predicted probability. This represents the general probabilistic classification objective. Tree-based ensemble models optimize alternative criteria (e.g., Gini impurity, entropy, or gradient-based loss functions).

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (5)$$

Evaluation Metrics

Equations (6–9) define standard classification metrics derived from the confusion matrix: precision, recall (sensitivity), specificity, and F1-score. These metrics quantify, respectively, the correctness of positive predictions, the ability to identify positive cases, the ability to exclude negative cases, and the harmonic mean of precision and recall, which balances both types of error.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

Balanced accuracy in equation (10) is computed as the mean of recall for each class, which is equivalent to the average of sensitivity and specificity in the binary case. This measure counteracts class imbalance by weighting performance across minority and majority classes equally, making it more informative than overall accuracy for imbalanced datasets.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (10)$$

ROC-AUC summarizes how well the model separates the positive and negative classes over all possible classification thresholds by integrating the receiver operating characteristic curve, which plots the true positive rate against the false positive rate. Average precision, in contrast, summarizes the area under the precision–recall curve and is particularly informative in the presence of class imbalance, as it focuses on the precision–recall trade-off for the positive class rather than overall ranking performance [15 – 17].

Association Rule Mining (ARM)

In association rule mining, a rule of the form $X \rightarrow Y$ links an antecedent itemset X to a consequent itemset Y , where X and Y are typically disjoint itemsets drawn from the same transaction database. Equations (11–15) define the standard interestingness measures for such rules: support, confidence, lift, leverage, and conviction.

Support measures how frequently an itemset (or rule) occurs in the database, while confidence estimates the conditional probability of observing Y given X . Lift and leverage compare the observed co-occurrence of X and Y to what would be expected under independence, and conviction quantifies the degree to which the antecedent reduces the chance of observing the consequent's absence, with larger values indicating stronger dependence.

$$\text{Support}(A \rightarrow B) = P(A \cap B) \quad (11)$$

$$\text{Confidence}(A \rightarrow B) = P(B | A) = \frac{P(A \cap B)}{P(A)} \quad (12)$$

$$\text{Lift}(A \rightarrow B) = \frac{P(A \cap B)}{P(A)P(B)} \quad (13)$$

$$\text{Leverage}(A \rightarrow B) = P(A \cap B) - P(A)P(B) \quad (14)$$

$$\text{Conviction}(A \rightarrow B) = \frac{1 - P(B)}{1 - P(B|A)} \quad (15)$$

Rules were retained for substantive interpretation only when they met the planned thresholds: minimum support of 0.05, minimum confidence of 0.60, and minimum lift greater than 1.10. This use of support, confidence, lift, leverage, and conviction follows the ARM literature's emphasis on rule interestingness and post-mining filtering [18 – 21]. These rules describe co-occurrence patterns, not causal mechanisms.

3. Results and Discussion

Descriptive Overview

The analytical sample covers seven districts, but district representation is uneven, consistent with spatial heterogeneity in poverty observed in regional analyses [4]. The seven district records include KutBak 610, Mueang Sakon Nakhon 526, Akat Amnuai 423, Kusuman 309, Wanon Niwat 141, Phanna Nikhom 109, and Phang Khon only 10. Poverty rates derived from the transparent income-per-capita rule ranged from 67.0% in Akat Amnuai to 86.4% in Kusuman, with Phang Khon showing 100% poverty

based on only 10 records. However, district-level estimates should be interpreted with caution due to uneven sample sizes; the Phang Khon rate should not be interpreted substantively.

Fig. 1 shows that missingness was concentrated in a small number of administrative variables, particularly the agricultural household number and survey date fields. Most analytical variables had comparatively low missingness. A17 is missing for 7.24% of households, while the derived poverty target is missing for 3.24% because poverty status cannot be calculated when required income or household-size information is incomplete. Income per capita, the overall capital index (ch 0), and several natural-capital and income components have approximately 2% missingness.

Fig. 2 presents the distribution of monthly per-capita household income used to construct the binary poverty target, consistent with empirical patterns of income inequality in poverty studies [1, 3]. Per-capita income was calculated by dividing annual household income by 12 and by the number of household members residing in the household. The dashed vertical line represents the poverty threshold of THB 2,629.053 per person per month; eligible households below this value were classified as poor.

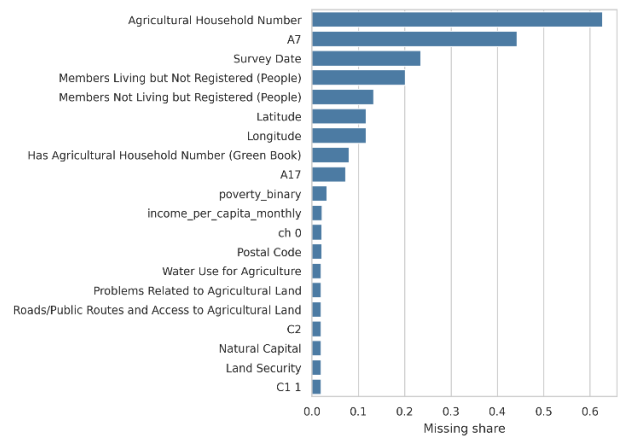


Fig. 1 Missingness structure among the top 20 variables.

Note. Missing share is the proportion of 2,128 records without an observed value. A7 includes government and social-welfare transfers; A17 includes daily-wage income, remittances, and other reported income. In the processed dataset, ch 0 is the mean of the five livelihood-capital domain scores; C1.1 denotes agricultural income and C2 denotes remittances from children working elsewhere. Some missing values are structurally expected and should not be interpreted as data-entry errors.

The distribution is strongly right-skewed. Most households are concentrated at relatively low-income levels, while a much smaller number extend into the upper-income tail. The median monthly per-capita income is approximately THB 1,486, and the 75th percentile is approximately THB 2,625, almost identical to the poverty threshold. Consequently, 1,562 of the 2,059 eligible records (75.9%) were classified as poor, producing a moderately imbalanced prediction target. Income values were clipped at the 99th percentile in the histogram for visual clarity; this display adjustment did not alter the underlying poverty

classification. Nevertheless, the target represents income poverty only and should not be interpreted as a comprehensive measure of multidimensional deprivation.

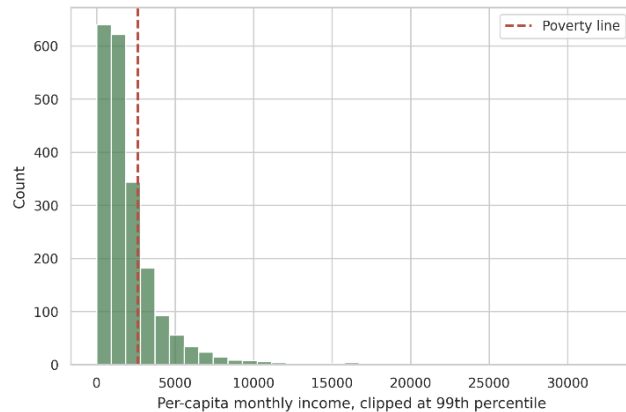


Fig. 2 Income distribution used in poverty-target construction.

Fig. 3 shows lower Human, Physical, and Social Capital scores among poor households, whereas Natural Capital showed weaker separation. This pattern is consistent with livelihood-capital research indicating that non-income deprivation domains may remain informative for poverty assessment [1 – 3, 22, 23].

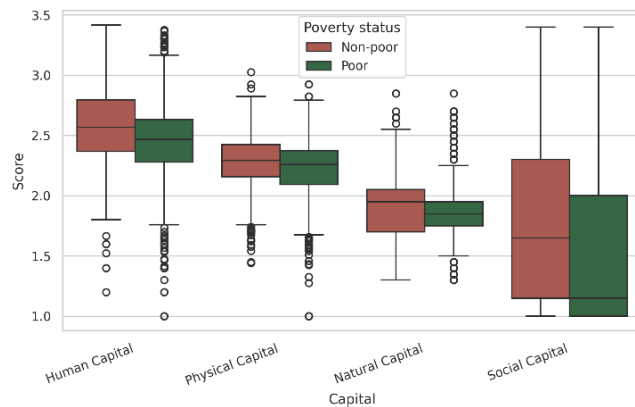


Fig. 3 Capital-domain distributions by poverty status.

Fig. 4 indicates that livelihood-capital indicators are related but not interchangeable, as the observed correlations suggest overlap among deprivation domains while still preserving meaningful conceptual distinctions between human, physical, natural, and social capital. This pattern is consistent with multidimensional poverty measurement approaches that treat deprivation domains as distinct but related constructs [1]. Accordingly, modeling these indicators as separate domains is preferable to collapsing all non-income deprivation into a single, undifferentiated score, because separate-domain treatment allows the analysis to identify which forms of livelihood capital are most closely associated with income-defined poverty. This approach also

supports clearer interpretation for policy users by distinguishing whether poverty risk is more strongly linked to skills and education, material living conditions, natural resource constraints, or social participation.

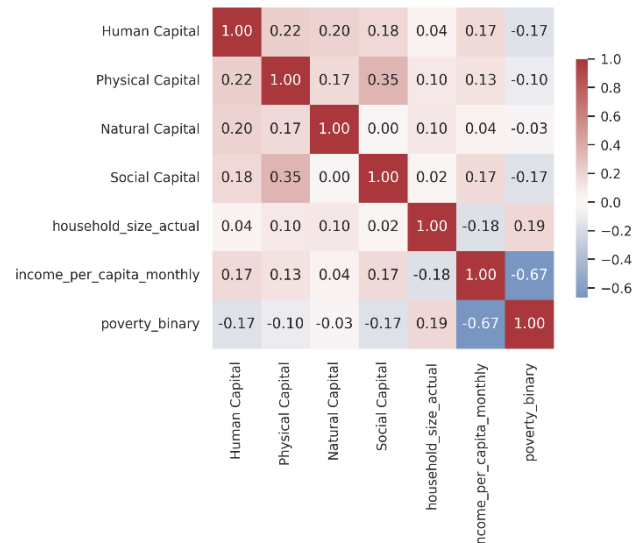


Fig. 4 Correlation heatmap for key variables.

Fig. 5 visually represents descriptive variation in poverty rates across districts. Because district sample sizes were uneven, particularly for Phang Khon, these comparisons should be interpreted as dataset-specific signals rather than stable population estimates.

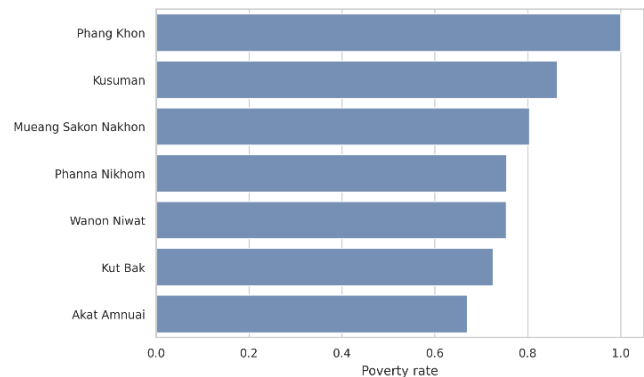


Fig. 5 District poverty-rate variation.

Statistical Findings

Table 2 presents group differences between poor and non-poor households, showing that poor households had lower Human, Physical, and Social Capital scores and larger household size. The most consistent non-income correlates of income-defined poverty were Human Capital (OR = 0.318), Social Capital (OR = 0.568), and household size (OR = 1.487). The strongest standardized differences were household size, Human Capital, and Social Capital, all in the small-effect range. Natural Capital showed a negligible difference, with a confidence interval that crossed zero.

Table 3 shows that higher Human Capital ($OR = 0.318, p < 0.001$), Physical Capital ($OR = 0.513, p = 0.023$), and Social Capital ($OR = 0.568, p < 0.001$) were associated with lower adjusted odds of poverty, whereas larger household size was associated with higher odds ($OR = 1.487, p < 0.001$). Natural Capital was not clearly associated with poverty after adjustment ($OR = 0.713, p = 0.103$). District effects should be interpreted as dataset-specific, given the uneven, unweighted sample.

Chi-square tests indicated modest associations between poverty and district ($\chi^2 = 48.089, p < 0.001$, Cramér's $V = 0.153$) and agricultural green-book status ($\chi^2 = 20.580, p < 0.001$, Cramér's $V = 0.104$). Household member mismatch was not clearly associated with poverty ($p = 0.170$). Overall, these findings are best interpreted as screening signals rather than strong determinants of poverty.

Table 2 Group differences in livelihood-capital indicators and household size by poverty status.

Variable	Poor, <i>M</i>	Non-poor, <i>M</i>	Diff.	95% CI	<i>p</i>	<i>g</i>
HC	2.442	2.576	-.134	[-.168, -.100]	<.001	-.410
PC	2.215	2.274	-.059	[-.081, -.036]	<.001	-.248
NC	1.892	1.916	-.023	[-.056, .010]	.166	-.076
SC	1.510	1.750	-.240	[-.304, -.176]	<.001	-.403
HHS	3.182	2.403	.779	[.623, .934]	<.001	.441

Note. *M* = mean; *Diff.* = mean difference (*Poor* – *Non-poor*); *CI* = 95% confidence interval; *p* = two-sided Welch's *t*-test probability; *g* = Hedges' standardized mean difference. HC = Human Capital; PC = Physical Capital; NC = Natural Capital; SC = Social Capital; HHS = household size.

Table 3 Robust logistic regression associations with poverty.

Predictor	Odds ratio	95% CI	<i>p</i> -value
Human Capital	0.318	[0.221, 0.458]	<0.001
Physical Capital	0.513	[0.288, 0.914]	0.023
Natural Capital	0.713	[0.475, 1.071]	0.103
Social Capital	0.568	[0.422, 0.764]	<0.001
Household size	1.487	[1.368, 1.618]	<0.001
Agricultural green-book status	0.721	[0.549, 0.946]	0.018
Kusuman district	2.042	[1.234, 3.376]	0.005
Kut Bak district	2.275	[1.624, 3.186]	<0.001
Mueang Sakon Nakhon district	1.303	[0.841, 2.019]	0.236
Phanna Nikhom district	1.228	[0.697, 2.163]	0.478
Wanon Niwat district	1.137	[0.652, 1.980]	0.651

Note. Odds ratios were estimated using multivariable logistic regression with HC3 robust standard errors. Values below 1 indicate lower adjusted odds of poverty. Akat Amnuai was the reference district. Results indicate associations, not causal effects.

Table 4 Cross-validated model performance.

Model	Accuracy	Balanced accuracy	Precision	Recall	F1	ROC-AUC	Average precision
Extra Trees	0.735	0.641	0.827	0.822	0.825	0.730	0.887
Logistic L2	0.668	0.658	0.855	0.678	0.756	0.724	0.881
Random Forest	0.764	0.636	0.820	0.883	0.850	0.718	0.880
Gradient Boosting	0.776	0.563	0.784	0.974	0.868	0.704	0.867
eline Prior	0.759	0.500	0.759	1.000	0.863	0.500	0.759

Note. Values are means from repeated stratified five-fold cross-validation. Poor households were treated as the positive class. Because the target was imbalanced, balanced accuracy, ROC-AUC, and average precision were prioritized over accuracy and F1-score alone.

Table 4 compares cross-validated predictive performance across the evaluated model families and demonstrates why model evaluation must be separated from simple class-prevalence behavior. The Baseline Prior model achieved high

recall and F1 because it predicted the poor class for every household. This is a known problem in imbalanced classification, where majority-class prediction can inflate conventional performance metrics when class prevalence is

high [15 – 17]. However, the same baseline produced balanced accuracy and ROC-AUC values of 0.500, indicating no discriminative ability beyond class prevalence. Extra Trees achieved the strongest ROC-AUC and average precision, while L2-regularized logistic regression achieved the highest balanced accuracy. Gradient Boosting produced the highest F1 but weak balanced accuracy, suggesting overemphasis on the majority poor class.

Predictive Model Comparison

Fig. 6 visually confirms that Extra Trees and Logistic L2 provide the strongest discrimination among the evaluated models, with ROC-AUC values that clearly exceed the baseline classifier and indicate better separation between poor and non-poor households across classification thresholds. The gap above the baseline indicates that non-income livelihood features carry meaningful screening signals. However, the magnitude remains moderate, thereby supporting cautious use for prioritization rather than definitive poverty classification.

The combined Fig. 7 presents the classification performance of the best model. The confusion matrix in Fig. 7(a) summarizes pooled out-of-fold predictions, while the ROC curve in Fig. 7(b) shows moderate discrimination above chance. Together, these results support use of the model for screening or prioritization rather than automated poverty eligibility decisions [7, 8, 15 – 17]. This supports the interpretation that the model may help prioritize screening by identifying households that warrant closer review or follow-up by local official. **Explainability**

Table 6 presents normalized impurity-based importance values from the Extra Trees model. The ten leading predictors accounted for 32.9% of total importance, indicating that predictive information was distributed across multiple household characteristics. Household composition was the most prominent feature group, led by household size and membership-related variables. Human-capital indicators, housing conditions, welfare benefits, and Social Capital also contributed meaningfully, supporting the ablation finding that component-level characteristics add information beyond broad capital-domain composites.

Nevertheless, these values indicate only the relative predictive contribution. Closely ranked features should not be interpreted as reliably different, and impurity-based importance may be affected by correlated or highly variable predictors. Accordingly, the results do not indicate statistical significance, direction of association, or causality [6, 12].

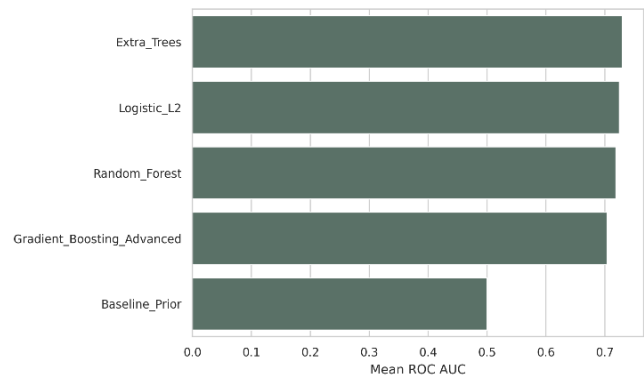
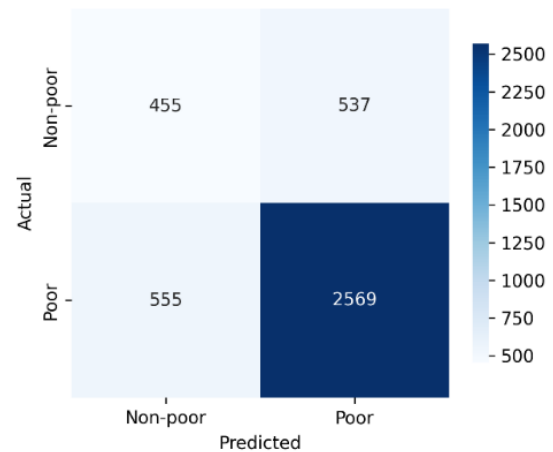
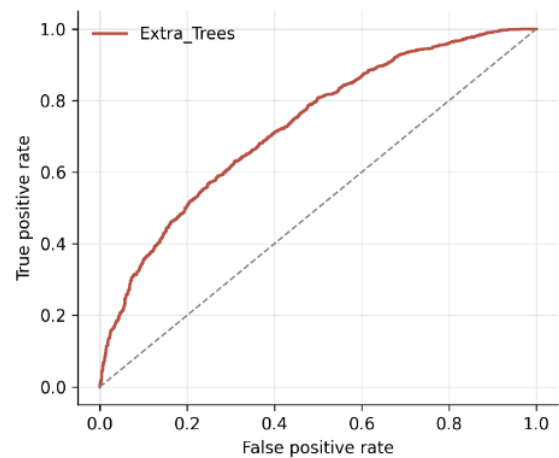


Fig. 6 ROC-AUC comparison across predictive models.



(a) Confusion matrix for the best model.



(b) ROC curve for the best model.

Fig. 7 Confusion matrix and ROC curve for the best model.

Table 5 Ablation results.

Specification	Balanced accuracy	F1	ROC-AUC	Precision (Avg)
Full non-income with geography	0.566	0.869	0.704	0.870
No geography	0.569	0.868	0.694	0.862
Capital domains only	0.558	0.864	0.686	0.864

Note. Values are five-fold cross-validation means. Average precision is reported in the final column. Differences are descriptive and do not establish statistical superiority.

Table 6 Top feature-importance values.

Feature	Importance
Household size	0.054
Occupation and skills for income generation	0.044
Members living and registered	0.032
Knowledge from knowledgeable people for problem solving	0.032
Number of members in the house registration	0.031
Human Capital	0.028
Housing condition	0.028
Housing sanitation	0.027
Government welfare benefits	0.026
Social Capital	0.026

Note. Values are normalized Extra Trees impurity-based importance scores. They indicate relative predictive contribution only and do not represent direction, statistical significance, or causality.

Fig. 8 shows that household composition, human-capital-related indicators, housing conditions, welfare support, and social capital were among the most influential predictors. These importance values indicate relative predictive contribution only and should not be interpreted as causal effects [6, 12].

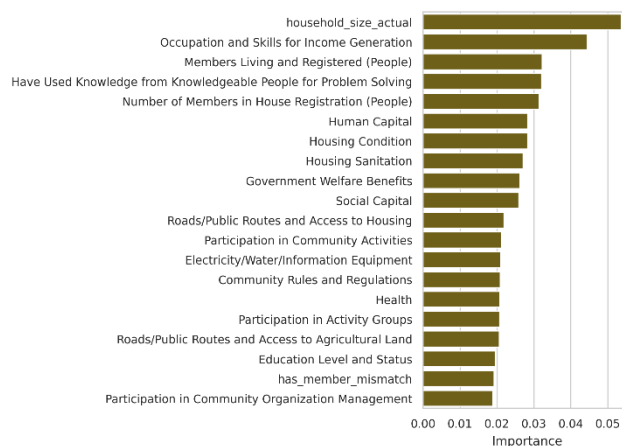


Fig. 8 Feature-importance ranking for the best model.

Association Rule Mining Findings

ARM was used because the survey contains many ordinal and binary-like household-condition variables that can be encoded as itemsets. Each household transaction included indicators of poverty status, district, household size, green book status, low- and high-capital items, and low-component score items. Direct income and financial-capital items were excluded.

Table 7 reports selected poverty-associated rules, indicating that no single item dominates them. The strongest poverty rules combine location, household size, infrastructure, physical capital, and social-participation indicators. The recurring association between large households and low social participation may reflect greater per-capita resource pressure and reduced adult availability

for community activities, as well as constraints on access to local institutions that limit opportunities for participation. Nevertheless, district-level ARM findings should be interpreted with caution given the uneven sample sizes across districts. Lift values are above 1.10 but not extreme, so these rules are meaningful descriptive signals rather than deterministic poverty profiles, as emphasized in association-rule mining literature [18 – 20]. Similarly, lift values only slightly above the reporting threshold should be interpreted as useful screening signals rather than strong determinants of poverty.

Table 8 reports selected high-lift deprivation-structure rules, showing that capital-domain scores are internally coherent with their component indicators. For example, low health and low education strongly co-occur with low Human Capital. These rules do not all make poverty the consequence, but they show how components of deprivation cluster.

Fig. 9 ranks the strongest retained ARM rules. The rules describe recurring deprivation profiles involving household size, social participation, district context, and capital-related conditions; they represent descriptive co-occurrence patterns rather than causal mechanisms [18– 21]. Most rules have very high confidence, often approaching 1.00, with support ranging approximately from 0.05 to 0.32. The dense upper-right cluster therefore represents rules that are both common and highly reliable within the sample.

The strongest-lift rule, low health combined with low education predicting low Human Capital, has lower support (0.074) but comparatively high confidence (0.894) and the highest lift (3.573). Other prominent rules connect district and participation indicators with Social Capital. The figure demonstrates a trade-off between prevalence and association strength. High-support rules describe broadly recurring deprivation structures, whereas lower-support, higher-lift rules identify narrower but more distinctive patterns.

Figure 10 ranks the 12 strongest retained association rules by lift. All displayed rules had lift values above 3.0, indicating that the antecedent–consequent combinations occurred at least three times more often than expected under independence. The strictest rule linked low health and low educational attainment with low Human Capital (support = 0.074, confidence = 0.894, lift = 3.573). Kusuman households with low activity-group participation showed low Social Capital with perfect confidence (support = 0.086, confidence = 1.000, lift = 3.298).

Agricultural-land problems in Akat Amnuai were also associated with low Natural Capital (support = 0.104, confidence = 0.918, lift = 3.293). Other rules connected to infrastructure, environmental constraints, and local resource conditions, with some reaching confidence = 1.000 and support up to 0.321. Because many rules link component indicators to their derived capital-domain scores, they primarily demonstrate descriptive construct coherence rather than causal effects or independent predictors of poverty.

Table 7 Selected poverty-associated association rules.

Rule	Antecedent	Sup.	Conf.	Lift	Lev.	Conv.
R1	KUS + LAGP	.079	.916	1.207	.014	2.860
R2	KUS + LCAP	.079	.916	1.207	.014	2.860
R3	KUS + LSC	.079	.916	1.207	.014	2.860
R4	LEWI + LHHS	.075	.912	1.201	.013	2.731
R5	LHHS + LSC	.089	.906	1.194	.015	2.575
R6	LHHS + LPC	.069	.904	1.191	.011	2.507
R7	LCRC + LHHS	.197	.896	1.181	.030	2.318
R8	LHHS + LMA	.182	.891	1.174	.027	2.206

Note. All rules have Poor household as the consequent. Sup. = support; Conf. = confidence; Lev. = leverage; Conv. = conviction. KUS = Kusuman District; LAGP = low activity-group participation; LCAP = low community-activity participation; LSC = low Social Capital; LEWI = low electricity, water, and information equipment; LHHS = large household size; LPC = low Physical Capital; LCRC = low community-rule compliance; LMA = low mutual assistance.

Table 8 Selected high-lift deprivation co-occurrence rules.

Rule	Antecedent	Consequent	Sup.	Conf.	Lift	Lev.	Conv.
R1	LH + LHE	LHC	.074	.894	3.573	.053	7.081
R2	KUS + LAGP	LSC	.086	1.000	3.298	.060	Inf.
3	AAK + LALP	LNC	.104	.918	3.293	.072	8.843
R4	LH + LOS	LHC	.147	.815	3.255	.102	4.042
R5	MSN + LAGP	LSC	.149	.962	3.174	.102	18.523

Note. Sup. = support; Conf. = confidence; Lev. = leverage; Conv. = conviction; Inf. = infinite conviction. LH = low health; LHE = low highest educational attainment; LHC = low Human Capital; KUS = Kusuman District; LAGP = low activity-group participation; LSC = low Social Capital; AAK = Akat Amnuai District; LALP = low agricultural-land problems; LNC = low Natural Capital; LOS = low occupational skills; MSN = Mueang Sakon Nakhon District.

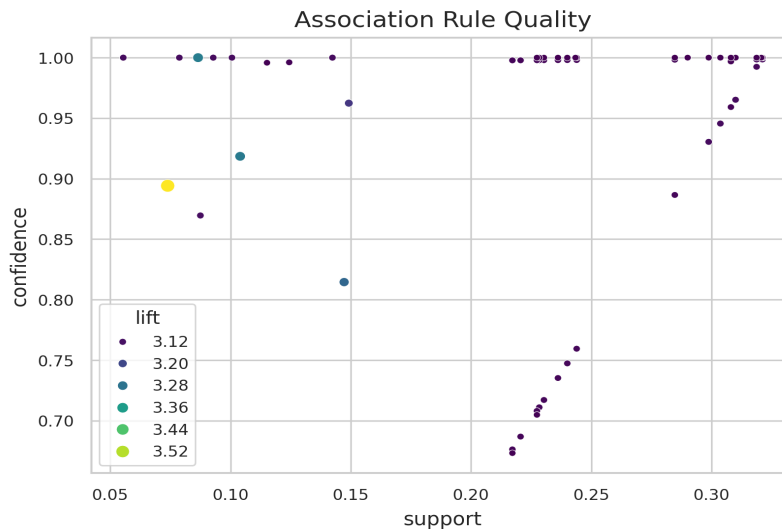


Fig. 9 ARM rule scatter plot.

Figure 11 shows that the number of retained association rules declined as minimum support or confidence thresholds became stricter, while minimum lift was fixed at 1.10. At support = 0.03, increasing confidence from 0.50 to 0.70 reduced the rule count from 12,457 to 8,174. At confidence

= 0.50, increasing support from 0.03 to 0.10 reduced the count to 9,321. Applying both stricter thresholds retained 6,063 rules, approximately 51% fewer than under the most permissive setting. Nevertheless, the large number of retained rules indicates substantial redundancy and co-

occurrence within the survey indicators. Consequently, substantive interpretation should focus on stable, nontrivial, and theoretically meaningful rules rather than rule volume alone. Threshold sensitivity assesses the robustness of rule

generation, not causality or predictive validity, consistent with recent work on rule validation and statistical testing in data mining [21].

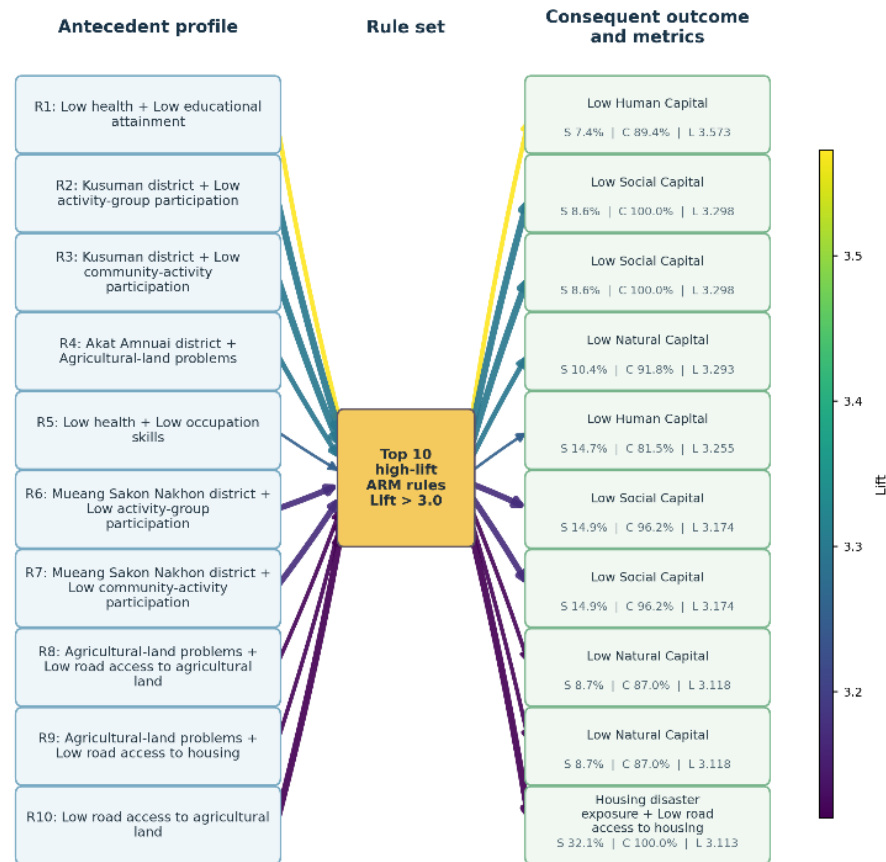


Fig. 10 ARM rule ranking.

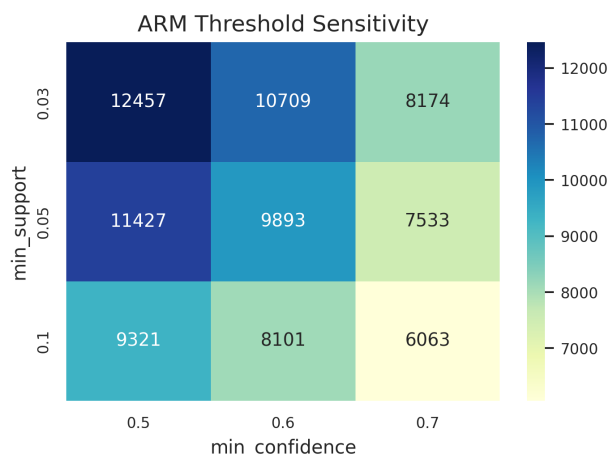


Fig. 11 ARM threshold sensitivity.

The results support a multidimensional interpretation of income-defined poverty in the Sakon Nakhon dataset, consistent with established multidimensional poverty frameworks [1, 3]. Poor households had lower levels of

Human Capital and Social Capital, and both domains remained associated with poverty in a robust logistic regression. These results matter because they point beyond income transfers alone: skills, education, health, local participation, mutual assistance, and community problem-solving capacity appear linked to poverty status in the survey, which is consistent with recent evidence that livelihood capital and social capital are associated with poverty perception, livelihood resilience, and household vulnerability [2, 24 – 28].

Implications

The findings support a multidimensional interpretation of income-defined poverty in the Sakon Nakhon dataset. Human and Social Capital, household size, and selected Physical Capital indicators were the most consistent non-income correlates across group comparisons, logistic regression, predictive modeling, and association-rule mining. This pattern is consistent with livelihood-capital research that links skills, education, health, social participation, material conditions, and household resources to poverty vulnerability [1 – 3, 22 – 29].

Natural Capital showed weaker adjusted statistical associations with income-defined poverty, although ARM identified coherent co-occurrence patterns among natural-resource and agricultural-land indicators. This difference reflects the distinct purposes of the analytical methods: regression estimates adjusted associations, whereas ARM identifies recurring patterns within the observed data. Predictive performance was moderate. Extra Trees provided useful discrimination, but the results do not support automated decisions on poverty eligibility. Instead, the model may assist local officials in prioritizing households for further assessment, together with administrative verification and contextual judgment [7, 8, 15 – 17, 30].

ARM added interpretable local profiles, particularly involving larger households, lower social participation, low social or physical capital, and selected district conditions. These profiles may support follow-up surveys, social support coordination, and locally tailored poverty-reduction planning, but they should not be interpreted as causal rules [18–21]. The findings suggest that local poverty responses should combine income support with human capital development, social participation mechanisms, household composition awareness, and infrastructure-sensitive targeting. These implications remain specific to the available PPAOS-derived dataset and should not be generalized as province-wide causal estimates.

4. Conclusion

This study integrated explanatory statistics, leakage-controlled machine learning, and association-rule mining to examine income-defined household poverty in Sakon Nakhon. Household size, Human Capital, Social Capital, and selected Physical Capital indicators were the most consistent non-income correlates of poverty. Extra Trees provided moderate predictive discrimination, supporting screening and prioritization rather than automated eligibility decisions. ARM identified interpretable deprivation profiles involving large households, low social participation, low capital conditions, and selected district contexts. The study also demonstrates that routinely collected household data can support locally relevant poverty assessment when statistical association, predictive performance, and descriptive co-occurrence evidence are interpreted separately. Practical use should combine income-based identification with livelihood-capital assessment and administrative review.

5. Suggestions

The cross-sectional administrative dataset supports association, prediction, and descriptive co-occurrence analysis, but not causal inference. The data were not a random provincial sample. Sampling weights were unavailable, and district sizes were uneven. In addition, the poverty target was an income-based proxy using a constant poverty threshold. Some variables also had substantial or structurally expected missingness. In this case, future research should validate the workflow using data from additional provinces, incorporate sampling weights where

available, compare income-defined poverty with validated multidimensional poverty measures, and use longitudinal data to examine poverty transitions over time.

6. Acknowledgement

This research was financially supported by the Program Management Unit on Area Based Development (PMU A), Contract No. A11F680121.

7. Declaration of Generative AI in Scientific Writing

During the preparation of this manuscript, the authors used generative AI tools to assist with language editing and clarity improvement. All scientific content, data analysis, interpretation, and conclusions were performed and verified by the authors. The authors take full responsibility for the integrity and accuracy of the published work.

8. CRediT Author Statement

Pita Jarupunphol: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing – Original Draft, Visualization. **Wichidtra Sudjarid:** Resources, Investigation, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

9. Research Involving Human and Animal Rights

This study used secondary household survey data. The authors did not conduct new experiments involving humans or animals.

10. Ethics Approval and Consent to Participate

The analysis was conducted using de-identified secondary data. No personally identifiable information was used in the analysis. Because the authors did not directly recruit participants or conduct new data collection for this manuscript, additional participant consent was not required for the present secondary analysis.

11. Declaration of Competing Interest

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.

12. References

- [1] Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics*, 95(7–8), 476–487. <https://doi.org/10.1016/j.jpubeco.2010.11.006>
- [2] Lu, Y., Chen, L., & Meng, Y. (2023). How do public services supply, livelihood capital, and livelihood strategies affect subjective poverty? *PLOS ONE*, 18(10), e0292651. <https://doi.org/10.1371/journal.pone.0292651>
- [3] Suppa, N., & Kanagaratnam, U. (2025). The global Multidimensional Poverty Index: Harmonised level estimates and their changes over time. *Scientific Data*, 12, 153. <https://doi.org/10.1038/s41597-024-04269-x>
- [4] Gomez-Mendez, I., & Amornbunchornvej, C. (2026). Regional and spatial dependence of poverty factors in

- Thailand, and its use in Bayesian hierarchical regression analysis. *Statistical Journal of the IAOS*, 42(1), 322–337. <https://doi.org/10.1177/18747655261416694>
- [5] Songleknok, S., & Kuptabut, S. (2023). Combining unsupervised and supervised learning to predict poverty households in Sakon Nakhon, Thailand. *International Journal of Advanced Computer Science and Applications*, 14(12), 818–828. <https://doi.org/10.14569/IJACSA.2023.0141283>
- [6] Atzmueller, M., Furnkranz, J., Kliegr, T., & Schmid, U. (2024). Explainable and interpretable machine learning and data mining. *Data Mining and Knowledge Discovery*, 38, 2571–2595. <https://doi.org/10.1007/s10618-024-01041-y>
- [7] Corral, P., Henderson, H., & Segovia, S. (2025). Poverty mapping in the age of machine learning. *Journal of Development Economics*, 172, 103377. <https://doi.org/10.1016/j.jdeveco.2024.103377>
- [8] Verme, P. (2025). Predicting poverty. *The World Bank Economic Review*, 39(4), 1039–1064. <https://doi.org/10.1093/wber/lhae044>
- [9] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- [10] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- [11] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- [12] Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, 55, 3503–3568. <https://doi.org/10.1007/s10462-021-10088-y>
- [13] Joshi, A. V. (2023). *Machine Learning and Artificial Intelligence*. Springer Cham. <https://doi.org/10.1007/978-3-031-12282-8>
- [14] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Foulds, J. R. (2025). *Data mining: Practical machine learning tools and techniques* (5th ed.). Morgan Kaufmann.
- [15] Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., Magwood, O., Sheikh, Y., McInnes, M., & Holzinger, A. (2023). Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 329–341. <https://doi.org/10.1109/TPAMI.2022.3145392>
- [16] Kim, M., & Hwang, K.-B. (2022). An empirical evaluation of sampling methods for the classification of imbalanced data. *PLOS ONE*, 17(7), e0271260. <https://doi.org/10.1371/journal.pone.0271260>
- [17] Liu, Y., Li, Y., & Xie, D. (2024). Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks. *Journal of Statistical Computation and Simulation*, 94(1), 183–203. <https://doi.org/10.1080/00949655.2023.2238235>
- [18] Chen, H., Yang, M., & Tang, X. (2024). Association rule mining of aircraft event causes based on the Apriori algorithm. *Scientific Reports*, 14, 13440. <https://doi.org/10.1038/s41598-024-64360-6>
- [19] Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2023). A survey on the use of association rules mining techniques in textual social media. *Artificial Intelligence Review*, 56, 1175–1200. <https://doi.org/10.1007/s10462-022-10196-3>
- [20] Hahsler, M., Grun, B., & Hornik, K. (2005). arules: A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15), 1–25. <https://doi.org/10.18637/jss.v014.i15>
- [21] Mouakher, A., Hajje, F., & Ayouni, S. (2022). Efficient mining support-confidence based framework generalized association rules. *Mathematics*, 10(7), 1163. <https://doi.org/10.3390/math10071163>
- [22] Bebbington, A. (1999). Capitals and capabilities: A framework for analyzing peasant viability, rural livelihoods and poverty. *World Development*, 27(12), 2021–2044. [https://doi.org/10.1016/S0305-750X\(99\)00104-7](https://doi.org/10.1016/S0305-750X(99)00104-7)
- [23] Scoones, I. (2009). Livelihoods perspectives and rural development. *The Journal of Peasant Studies*, 36(1), 171–196. <https://doi.org/10.1080/03066150902820503>
- [24] Tang, L., Xu, Y., Wang, W., & Wang, Y. (2023). Impact of livelihood capital and rural site conditions on livelihood resilience of farm households: Evidence from contiguous poverty-stricken areas in China. *Environmental Science and Pollution Research*, 30, 123808–123826. <https://doi.org/10.1007/s11356-023-30426-7>
- [25] Zhao, X., & Lan, F. (2023). The impact of livelihood capital endowment on household poverty alleviation: The mediating effect of land transfer. *Land*, 12(7), 1346. <https://doi.org/10.3390/land12071346>
- [26] Chuong, H. N. (2023). The role of social capital in shaping livelihood for rural Vietnamese households. *PLOS ONE*, 18(12), e0295292. <https://doi.org/10.1371/journal.pone.0295292>
- [27] Fan, Y., Ning, J., & Qin, H. (2023). Investigating the effectiveness of livelihood capital in reducing re-poverty risk: An empirical analysis of policy withdrawal and income structures in rural China. *Frontiers in Environmental Science*, 11, 1175315. <https://doi.org/10.3389/fenvs.2023.1175315>
- [28] Xiao, H., Liang, X., Chen, C., & Xie, F. (2022). The impact of multidimensional poverty on rural households' health: From a perspective of social capital and family care. *International Journal of Environmental Research and Public Health*, 19(21), 14590. <https://doi.org/10.3390/ijerph192114590>
- [29] Habib, N., Ariyawardana, A., & Aziz, A. A. (2023). The influence and impact of livelihood capitals on livelihood diversification strategies in developing countries: A systematic literature review. *Environmental Science and Pollution Research*, 30, 69882–69898. <https://doi.org/10.1007/s11356-023-27638-2>
- [30] Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603, 864–870. <https://doi.org/10.1038/s41586-022-04484-9>