ACCSCIENCE PUBLISHING

International Journal of Population Studies

RESEARCH ARTICLE

Predicting poverty using essential services access in the poorest region of Luzon, Philippines: A machine learning-based approach

Supplementary Files

1. Econometric models

The probit regression model is presented in Equation SI,

$$E(P = 1 \mid C) = \Phi (C^{F} + \beta) \quad \text{(SI)}$$

where E is the probability of a household becoming poor or not, Φ is the cumulative distribution function of the standard normal distribution, and β is the parameter under maximum likelihood estimation. The overall model can be expressed as in Equation SII,

$$P (K|C) = P(P = 1|C) = \Phi (\beta_0 + \beta_1 C) \quad \text{(SII)}$$

Since the dependent variable is a non-linear function of the regressors, the coefficient on C has no linear interpretation. The model was translated into a vector format as in Equation SIII,

$$K = \Phi + C\beta + \mu \quad \text{(SIII)}$$

where K is the probit, (o) is the outcome of a household in poverty, Φ is the cumulative standard normal distribution function, C is the vector of independent variables as shown in Table 1, β is the vector of coefficients, intercepts, or effects of independent factors on the poverty outcomes, and μ is the error term.

2. Performance evaluation metrics

We have analyzed the performance of each regression and classification algorithm. Various evaluation metrics for regression were applied, such as the mean squared error in Equation SIV,

$$MSE = \frac{1}{2}\sum_{i=1}^{n}(y - \hat{y})^2 \quad \text{(SIV)}$$

where n is the number of data units, y is the actual value, and $\check{y}$ is the predicted value. The root mean square error is presented in Equation SV:

$$RMSE = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(y - \hat{y})^2} \quad \text{(SV)}$$

It is the square root of the average squared distance between the observed value and the predicted value (Choubey *et al.*, 2020; Min *et al.*, 2022), and R^2 represents the coefficient of determination. It tells how much of the actual outcomes are replicated by the model. For classification, we utilized the accuracy of classification, which is the ratio of the correctly predicted outputs to the total number of input variables in Equation SVI,

$$\alpha = \frac{\sigma}{\delta} \quad \text{(SVI)}$$

where α is the classification accuracy, σ is the total number of correct predictions, and σ is the total number of predictions made. The confusion matrix is the matrix of output and shows the complete performance of the model (Onsay & Rabajante, 2024). The matrix is presented in Equation SVII:

Predicted values	Actual values		
		Positive (1)	**Negative (0)**
	Positive (1)	True positive	False positive
	Negative (0)	False negative	True negative
Predicted values	**Actual values**		
		Poor (1)	**Non-poor (0)**
	Poor (1)	True positive values	False positive values
	Non-poor (0)	False negative values	True negative values

(SVII)

The matrix accuracy is presented in Equation SVIII,

$$\alpha = \frac{\tau + \partial}{n} \quad \text{(SVIII)}$$

Where α is the matrix accuracy, τ is the true positive (TP) (Positive positive prediction), ∂ is the false negative (FN) (Positive negative predictions), and n is the total number of households. Precision is presented in Equation SIX,

$$\rho = \frac{\tau}{\tau + \varphi} \quad \text{(SIX)}$$

where ρ is the precision, τ is the TP (Positive positive prediction), and φ is the false positive (Negative positive prediction). It shows the correct positive results divided by the number of positive results divided by the classifier. Recall is presented in Equation SX,

$$\mathsf{C} = \frac{\tau}{\tau + \partial} \quad \text{(SX)}$$

where C is the recall, τ is the TP (Positive positive prediction), and ∂ is the FN (Positive negative predictions). It is the ratio of total correct positive results to all the relevant households. Finally, the F1 score formula is presented in Equation SXI:

$$\mathrm{F1} = 2\left[\frac{1}{\left(\frac{1}{\frac{\tau}{\tau+\varphi}}\right) + \left(\frac{1}{\frac{\tau}{\tau+\partial}}\right)}\right] \quad \text{(SXI)}$$

It reveals the test accuracy and the harmonic mean between ρ and C shows how robust and precise the classifier is (Choubey *et al.*, 2020; Min *et al.*, 2022; Onsay & Rabajante, 2024).

3. Machine learning regression

The following regression methods were utilized for predicting poverty levels based on socioeconomic indicators.

3.1. Linear regressions

This approach is suitable for modeling continuous outputs through the utilization of the least squares methodology. The application of the lasso shrinkage method effectively reduced the prediction error (Hastie *et al.*, 2009; Muñetón-Santa & Manrique-Ruiz, 2023) (Equation SXII),

$$\hat{\beta}^{lasso} = \underbrace{argmin}_{\beta} = \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \quad \text{(SXII)}$$

subject to Equation SXIII:

$$\sum_{j=1}^{p}\left|\beta_j\right| \le t \quad \text{(SXIII)}$$

The lasso penalty is showcased through the equivalent Lagrange form (Equation SXIV):

$$\sum_{i}^{p}|\beta_j| \quad \text{(SXIV)}$$

That limitation makes the solutions non-linear in the y_i (Hastie *et al.*, 2017; Muñetón-Santa & Manrique-Ruiz, 2023). Due to the possibility of certain coefficients being precisely zero, the lasso acts akin to continuous subset selection, especially when t is small (Equation SXV):

$$\hat{\beta}^{lasso} = \underbrace{\arg\min}_{\beta} \left\{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \le t\right\} \quad \text{(SXV)}$$

We utilized it because it offers a straightforward approach to understanding the linear relationship between poverty and predictors such as cell phone ownership, electricity, and Internet access. While limited by its assumption of linearity, it provides a foundational understanding of how these factors impact poverty levels.

3.2. Extreme gradient boosting

Among the top supervised learning algorithms, this particular one stands out. Its objective function and regularization encompass the loss function, enabling the model to assess actual and predicted values. Extreme Gradient Boosting (XGBoost), an efficient and scalable gradient boosting framework, employs various regularization techniques, such as column subsampling during tree construction, tree depth

regularization, and L1 and L2 regularization on leaf values. These methods help in managing model complexity and preventing overfitting (Chen *et al*., 2015; Chen & Guestrin, 2016) (Equation SXVI):

$$\hat{y}^{peci} = \sum_{h=1}^{S} m_s\left(x_g\right) = \sum_{h=1}^{S} v_z\left(x_g\right) \quad \text{(SXVI)}$$

Where $\check{y}^{peci}$ is the predicted value, $m_s\ (x_g)$ is the prediction of the s-th tree in the ensemble for the g-th household, $m_s\ (x_g)$is the leaf weight, z $(x_g)$is the assigned value to the leaf node corresponding to the g-th household, s is the total number of trees, z $(x_g)$is the index of the leaf node for the i-th sample obtained by following the learned decision rules in s, and x_g is the feature vector of the feature vector of the g-th household (Onsay & Rabajante, 2024). We employed it because it possesses the ability to handle complex interactions and non-linear relationships. XGBoost can capture intricate patterns between poverty and predictors such as cell phone usage, electricity availability, and internet access. Despite its complexity, XGBoost excels in predictive accuracy, potentially providing nuanced insights into poverty prediction.

3.3. Polynomial regressions

Polynomial regression was utilized to ascertain the optimal coefficients that effectively conform to the data, aiming to minimize the disparity between predicted and observed values of the dependent variable (Heiberger *et al*., 2009; Ostertagová, 2012) (Equation SXVII):

$$\hat{y}^{peci} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \ldots \beta_n X^n \quad \text{(SXVII)}$$

where $\check{y}^{peci}$ is the predicted value of the dependent variable, X is the independent variable, β is the coefficients or parameters of the polynomial regression model, and n is the degree of the polynomial, determining the highest power of x in the equation. We applied it because it can explore more intricate patterns that linear models might overlook, as it can analyze non-linear relationships.

4. Machine learning classification

The following classification methods were utilized to predict poverty based on socioeconomic indicators.

4.1. Gaussian Naïve Bayes

It is a straightforward yet efficient approach to probabilistic classification. This method, tailored for binary outcomes, relies on the assumption of feature independence and Gaussian distribution for continuous features. The algorithm involves computing the variance and means of each feature for every class in the training set. When classifying a new instance, it utilizes the Bayes theorem to calculate the posterior probability of each class based on the observed feature values. Ultimately, the method assigns the new instance to the class with the highest posterior probability. Gaussian Naïve Bayes performs well in binary classification tasks where features follow a Gaussian distribution and is known for its simplicity and effectiveness. The equation is presented in Equation SXVIII:

$$\hat{y}^{\text{peci}}\left(C_i|x\right) = \frac{y(c_i)\ y(x|\ c_i)}{y(x)} \quad \text{(SXVIII)}$$

where $\hat{y}^{\text{peci}}\left(C_i|x\right)$ uses the Bayes theorem to determine the posterior probability of each class given the observed features (Aji *et al*., 2023; Hastie *et al*., 2009). We utilized it due to its simplicity and efficiency in handling high-dimensional data, making it suitable for scenarios with multiple predictors such as cell phone ownership, electricity, and Internet access. Despite its assumption of feature independence, it can provide quick insights into the likelihood of poverty based on these factors.

4.2. Logistic regressions

It is a powerful method for binary classification tasks. It employs a logistic or sigmoid function to model the relationship between input features and the binary target variable. In logistic regression, predictions are made by calculating the odds of the target variable belonging to each class based on a predefined threshold.

Table S1. Household characteristics and access to essential services distribution

Parameters	Frequency (*n*)	Distribution (%)
Household income conditions		
Poor	8,930	63.69
Non-poor	5,091	36.31
Total	14,021	100.00
Electricity connection		
With electricity connection	13,402	95.59
Without electricity connection	619	4.41
Total	14,021	100.00
Cell phone ownership		
With cell phone	9,432	67.27
Without cell phone	4,589	32.73
Total	14,021	100.00
Internet connection		
With Internet connection	5,489	39.15
Without Internet connection	8,532	60.85
Total	14,021	100.00

Table S2. Average mean square error of machine learning regressors for poverty prediction by electricity connection, cell phone ownership, and internet access

Locals	Linear	Extreme gradient boost	Polynomial
Abucayan	0.14876	0.24438	0.20202
Balaynan	0.67194	0.09461	0.25940
Digdigon	0.72641	0.26171	0.25855
Hiwacloy	0.95092	0.24117	0.26509
Lamon	0.79998	0.26523	0.92734
Maysalay	0.90877	0.13824	0.10900
Payatan	0.73840	0.20488	0.45252
Pinaglabanan	0.83128	0.24595	0.19348
San Isidro West	0.60314	0.24773	0.76032
Scout Fuentebella	0.30080	0.19865	0.73912
Tabgon	0.56353	0.19232	0.51038
Tamban	0.90649	0.19719	0.64333
Isarog	0.67920	0.21100	0.42887
Bagumbayan Grande	0.73257	0.25290	0.20061
Bagumbayan Pequeño	0.53298	0.24464	0.11320
Belen	0.41891	0.22682	0.08345
La Purisima	0.83403	0.15813	0.17443
Panday	0.70679	0.10450	0.25292
San Benito	0.17682	0.26075	0.22577
San Isidro	0.41113	0.19592	0.68586
San Jose	0.79122	0.26292	0.19391
San Juan Evangelista	0.17414	0.22497	0.02704
San Juan Bautista	0.96403	0.20571	0.83993
Poblacion	0.58380	0.21348	0.29327
Buyo	0.20732	0.20805	0.25234
Catagbacan	0.82282	0.25534	0.10340
Matacla	0.67494	0.24285	0.81332
San Pedro	0.16025	0.23350	0.61937
Tagongtong	0.12326	0.19337	0.13543
Ranggas	0.39772	0.22662	0.35871
Cagaycay	0.62066	0.24745	0.24364
Gimaga	0.44155	0.17798	0.19991
Halawi Gogon	0.16699	0.25805	0.22754
Maymatan	0.71081	0.25073	0.22436
Napawon	0.06585	0.23695	0.13755
Salog	0.87709	0.14441	0.18578
Taytay	0.95171	0.13297	0.19423
Salog	0.54781	0.20693	0.20186
Goa	0.55213	0.21451	0.32068

Table S3. Average R^2 of machine learning regressors for poverty prediction by electricity connection, cell phone ownership, and Internet access

Locals	Linear	Extreme Gradient Boost	Polynomial
Abucayan	0.77774	0.78712	0.72017
Balaynan	0.25456	0.93689	0.66279
Digdigon	0.20009	0.76979	0.66364
Hiwacloy	0.30450	0.79033	0.65710
Lamon	0.12652	0.76627	0.43210
Maysalay	0.01773	0.89326	0.81319
Payatan	0.18810	0.82662	0.46967
Pinaglabanan	0.09522	0.78555	0.90284
San Isidro West	0.32336	0.78377	0.16187
Scout Fuentebella	0.62570	0.83285	0.18307
Tabgon	0.36297	0.83918	0.41181
Tamban	0.02001	0.83431	0.27886
Isarog	0.24730	0.82050	0.49332
Bagumbayan Grande	0.19393	0.77860	0.72158
Bagumbayan Pequeño	0.39352	0.78686	0.80899
Belen	0.50759	0.80468	0.83874
La Purisima	0.09247	0.87337	0.74776
Panday	0.21971	0.92700	0.66927
San Benito	0.74968	0.77075	0.69642
San Isidro	0.51537	0.83558	0.23633
San Jose	0.13528	0.76858	0.72828
San Juan Evangelista	0.75236	0.80653	0.89515
San Juan Bautista	0.34500	0.82579	0.08226
Poblacion	0.34270	0.81802	0.62892
Buyo	0.71918	0.82345	0.66985
Catagbacan	0.10368	0.77616	0.94910
Matacla	0.25156	0.78865	0.10887
San Pedro	0.76625	0.79800	0.30282
Tagongtong	0.80324	0.83813	0.78676
Ranggas	0.52878	0.80488	0.56348
Cagaycay	0.30584	0.78405	0.67855
Gimaga	0.48495	0.85352	0.72228
Halawi Gogon	0.75951	0.77345	0.69465
Maymatan	0.21569	0.78077	0.69783
Napawon	0.86065	0.79455	0.78464
Salog	0.04941	0.88709	0.73641
Taytay	0.10230	0.89853	0.72796
Salog	0.37869	0.82457	0.72033
Goa	0.37437	0.81699	0.60151

Table S4. Classification accuracies of machine learning algorithms for poverty prediction by electricity, cell phone ownership, and Internet

Locals	Logistic	Random forest	Gaussian Naïve Bayes
Abucayan	0.86446	0.90771	0.86536
Balaynan	0.88938	0.93263	0.89028
Digdigon	0.83682	0.88007	0.83772
Hiwacloy	0.88542	0.92867	0.88632
Lamon	0.86184	0.90509	0.86274
Maysalay	0.81759	0.86084	0.81849
Payatan	0.82105	0.86430	0.82195
Pinaglabanan	0.81382	0.85707	0.81472
San Isidro West	0.89015	0.93340	0.89105
Scout Fuentebella	0.87521	0.91846	0.87611
Tabgon	0.89054	0.93379	0.89144
Tamban	0.81868	0.86193	0.81958
Isarog	0.85541	0.89866	0.85631
Bagumbayan Grande	0.66239	0.70564	0.66329
Bagumbayan Pequeño	0.75769	0.80094	0.75859
Belen	0.73094	0.77419	0.73184
La Purisima	0.67468	0.71793	0.67558
Panday	0.82601	0.86926	0.82691
San Benito	0.80734	0.85059	0.80824
San Isidro	0.75120	0.79445	0.75210
San Jose	0.81725	0.86050	0.81815
San Juan Evangelista	0.66077	0.70402	0.66167
San Juan Bautista	0.85525	0.89850	0.85615
Poblacion	0.76354	0.80679	0.76444
Buyo	0.83852	0.88177	0.83942
Catagbacan	0.80994	0.85319	0.81084
Matacla	0.81843	0.86168	0.81933
San Pedro	0.81838	0.86163	0.81928
Tagongtong	0.86045	0.90370	0.86135
Ranggas	0.82914	0.87239	0.83004
Cagaycay	0.85888	0.90213	0.85978
Gimaga	0.88785	0.93110	0.88875
Halawi Gogon	0.83834	0.88159	0.83924
Maymatan	0.83433	0.87758	0.83523
Napawon	0.81286	0.85611	0.81376
Salog	0.83987	0.88312	0.84077
Taytay	0.86479	0.90804	0.86569
Salog	0.84813	0.89138	0.84903
Goa	0.82406	0.86731	0.82496

Table S5. Classification accuracies of machine learning algorithms for poverty prediction by electricity connection, cell phone ownership, and internet access at Pipeline (Δ)

Locals	Logistic	Random forest	Gaussian Naïve Bayes
Abucayan	0.90936	0.92376	0.91026
Balaynan	0.93428	0.94868	0.93518
Digdigon	0.88172	0.89612	0.88262
Hiwacloy	0.93032	0.94472	0.93122
Lamon	0.90674	0.92114	0.90764
Maysalay	0.86249	0.87689	0.86339
Payatan	0.86595	0.88035	0.86685
Pinaglabanan	0.85872	0.87312	0.85962
San Isidro West	0.93505	0.94945	0.93595
Scout Fuentebella	0.92011	0.93451	0.92101
Tabgon	0.93544	0.94984	0.93634
Tamban	0.86358	0.87798	0.86448
Isarog	0.90031	0.91471	0.90121
Bagumbayan Grande	0.70729	0.72169	0.70819
Bagumbayan Pequeño	0.80259	0.81699	0.80349
Belen	0.77584	0.79024	0.77674
La Purisima	0.71958	0.73398	0.72048
Panday	0.87091	0.88531	0.87181
San Benito	0.85224	0.86664	0.85314
San Isidro	0.79610	0.81050	0.79700
San Jose	0.86215	0.87655	0.86305
San Juan Evangelista	0.70567	0.72007	0.70657
San Juan Bautista	0.90015	0.91455	0.90105
Poblacion	0.80844	0.82284	0.80934
Buyo	0.88342	0.89782	0.88432
Catagbacan	0.85484	0.86924	0.85574
Matacla	0.86333	0.87773	0.86423
San Pedro	0.86328	0.87768	0.86418
Tagongtong	0.90535	0.91975	0.90625
Ranggas	0.87404	0.88844	0.87494
Cagaycay	0.90378	0.91818	0.90468
Gimaga	0.93275	0.94715	0.93365
Halawi Gogon	0.88324	0.89764	0.88414
Maymatan	0.87923	0.89363	0.88013
Napawon	0.85776	0.87216	0.85866
Salog	0.88477	0.89917	0.88567
Taytay	0.90969	0.92409	0.91059
Salog	0.89303	0.90743	0.89393
Goa	0.86896	0.88336	0.86986

By learning coefficients for each feature through model fitting on training data using an optimization process, logistic regression effectively classifies new instances (Hastie *et al*., 2009). We have modified the model into Equation SXIX:

$$E = j + C\beta + k \quad \text{(SXIX)}$$

where E is the logit, (p) is the log [p/(1−p)], E is the probability of being poor, j is the intercept or individual effects, C is the vector of independent variables, β is the vector of coefficients, intercepts, or effects, and μ is the error term (Onsay & Rabajante, 2024). We applied it because it is ideal for modeling poverty prediction. Logistic regression's linear nature allows for the examination of the relationship between predictors (cell phone ownership, electricity, and internet access) and the probability of poverty. While assuming linearity, it remains effective in capturing associations and is interpretable for understanding how each predictor influences poverty.

4.3. Random forests

Random forests excel in identifying important variables and uncovering relationships. Renowned for its effectiveness in assessing variable importance, this algorithm adeptly captures non-linear relationships between explanatory and dependent variables (Schonlau & Zou, 2020). The equation is presented in Equation SXX:

$$\hat{y}^{drp} = \frac{1}{2}\sum_{v=1}^{t} \gamma_v\ (c) \quad \text{(SXX)}$$

Where $\check{y}^{drp}$ is the predicted value of the target variable, t is the total number of decision trees in the random forest, and γ_v (c) represents the prediction of the v-th decision tree for the input features. Random forest is a popular ensemble learning method for classification tasks. By combining the outcomes of multiple decision trees, it delivers accurate predictions. Known for its robustness in handling noisy input, high-dimensional data, and feature interactions, random forest effectively reduces the risk of overfitting while generating strong and precise predictions. In addition, it can assess feature importance, offering valuable insights into the key features influencing the classification process (Onsay & Rabajante, 2024). We have chosen it for its robustness and ability to handle complex interactions among predictors, as it can capture non-linear relationships between poverty and cell phone ownership, electricity, and Internet access.

References

Aji, P., Wijaya, D.R., Hernawati, E., Yualinda, S., Yualinda, S., Frasanta, M.A.H., *et al.* (2023). Poverty level prediction based on e-commerce data using naïve bayes algorithm and similarity-based feature selection. *International Journal of Applied Information Technology*, 7(2):114-126.

https://doi.org/10.25124/ijait.v7i02.5374

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. p. 785-794.

https://doi.org/10.1145/2939672.2939785

Chen, T., & He, T. (2025). xgboost: eXtreme Gradient Boosting (R package version 1.7.11.1) [Computer software]. Available from: https://cran.ms.unimelb.edu.au/web/packages/xgboost/vignettes/xgboost.pdf

Choubey, D.K., Kumar, P., Tripathi, S., & Kumar, S. (2020). Performance evaluation of classification methods with PCA and PSO for diabetes. *Network Modeling Analysis in Health Informatics Bioinformatics*, 9:5.

https://doi.org/10.1007/s13721-019-0210-8

Hastie, T., Tibshirani, R., Friedman, J.H., & Friedman, J.H. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Vol. 2. New York: Springer; p. 1-758.

Heiberger, R.M., Neuwirth, E., Heiberger, R.M., & Neuwirth, E. (2009). Polynomial regression. In: R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics. Germany: Springer Science; p. 269-284.

Min, P.P., Gan, Y.W., Hamzah, S.N.B., Ong, T.S., & Sayeed, M.S. (2022). Poverty prediction using machine learning approach. *Journal of Southwest Jiaotong University*, 57(1):137-146.

https://doi.org/10.35741/issn.0258-2724.57.1.12

Muñetón-Santa, G., & Manrique-Ruiz, L.C. (2023). Predicting multidimensional poverty with machine learning algorithms: An open data source approach using spatial data. *Social Sciences*, 12(5):296.

https://doi.org/10.3390/socsci12050296

Onsay, E.A., & Rabajante, J.F. (2024). Combining machine learning (ML) and participatory rural appraisal (PRA) for disaster risk preparedness (DRP): Evidence from the poorest region of luzon, philippines. *International Journal of Disaster Risk Reduction*, 112(1):104809.

https://doi.org/10.1016/j.ijdrr.2024.104809

Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Engineering*, 48:500-506.

https://doi.org/10.1016/j.proeng.2012.09.545

Schonlau, M., & Zou, R.Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3-29.

https://doi.org/10.1177/1536867X20909688