

Machine Learning-Based Prediction of Household Internet Access Across Indonesian Provinces Using Socioeconomic and Temporal Features

Rahmatika ^{1*}, Nasrulloh Isnain ², Ulfa Pauziah ³

^{1,2,3} Department of Informatics Engineering, Faculty of Engineering and Computer Science, Universitas Indraprasta PGRI, South Jakarta City, Special Capital Region of Jakarta, Indonesia.

ABSTRACT

Unequal household internet access remains an important dimension of the digital divide across Indonesian provinces, requiring predictive approaches that account for socioeconomic conditions and temporal dynamics. This study predicted provincial household internet access and evaluated the added predictive value of previous-year connectivity. Secondary data from Statistics Indonesia covering 34 provinces during 2020–2024 were analyzed using Linear Regression, Random Forest, Gradient Boosting, and XGBoost. Two feature configurations were compared: one using contemporaneous socioeconomic and information and communication technology (ICT) predictors and another incorporating previous-year internet access as a lagged feature. Model performance was assessed using an out-of-time holdout for 2024, GroupKFold cross-validation, and rolling-origin temporal validation. Adding the lagged feature reduced RMSE across all models. XGBoost achieved the best 2024 holdout performance, with an RMSE of 2.7260 and an (R^2) of 0.4003, representing a 54.49% reduction in RMSE compared with its No-Lag counterpart. Previous-year internet access was the most influential feature (55.57%), followed by adjusted expenditure per capita (21.45%) and computer ownership (18.16%). Rolling-origin validation yielded a mean (R^2) of 0.5866. The results also reveal differences in model performance across validation strategies, indicating that cross-validation alone may not fully reflect predictive performance on future observations. Incorporating historical connectivity provides additional information for capturing the temporal persistence of internet access across provinces. The variation in prediction errors across provinces further highlights the need to consider regional differences when evaluating predictive models. These findings indicate that temporally informed machine learning can support the analysis of regional internet-access disparities and provide a basis for monitoring changes in digital connectivity across Indonesia. The proposed evaluation approach also offers a practical framework for comparing predictive models under different temporal and regional validation settings.

ARTICLE HISTORY

Received: 29 August 2026.
Revised: 1 September 2026.
Accepted: 26 September 2026.
Available Online: 30 October 2026.
Published: 1 December 2026.

KEYWORDS

Digital Divide; Household Internet Access; Machine Learning; Temporal Validation; XGBoost; Indonesia.

1. Introduction

Internet access has become a fundamental component of contemporary socioeconomic development, supporting education, employment, public service delivery, participation in the digital economy, and social inclusion. Reliable connectivity expands access to online learning and educational resources, facilitates remote work and digitally mediated employment opportunities, and enables individuals to interact with increasingly digitalized public services (Kartiasih *et al.*, 2022; Lai & Widmar, 2021). Internet connectivity also supports participation in e-commerce and digital transactions, creating opportunities for economic activity in an increasingly digital environment (Ramírez-Correa *et al.*, 2025). Beyond these economic and functional benefits, internet use can contribute to social participation and the formation of social capital, including in Indonesia (Kharisma, 2022). Despite the continued expansion of internet connectivity in Indonesia, access remains uneven across regions. Previous studies have documented geographical disparities in digital development associated with differences in infrastructure, digital capabilities, and socioeconomic conditions. More developed regions generally exhibit stronger digital connectivity than less-developed and geographically disadvantaged areas (Jaya *et al.*, 2024; Kartiasih *et al.*, 2022). These disparities indicate that increasing national internet penetration does not necessarily translate into uniform household connectivity across provinces, highlighting the need to understand and predict regional variations in household internet access. Household internet access is associated with socioeconomic and technological conditions that shape households' capacity to participate in the digital environment. In Indonesia, regional disparities in internet access have been linked to differences in economic development, educational conditions, and the availability of digital resources (Kartiasih *et al.*, 2022). Education is relevant because higher educational attainment can strengthen digital capabilities, information-seeking behavior, and the ability to benefit from internet-based services. Economic capacity is similarly important, as households with greater financial resources are generally better positioned to afford digital devices, connectivity services, and other expenditures required for sustained internet use (Du *et al.*, 2021; Singh & Mohanty, 2025). Computer ownership also reflects household readiness to engage with digital technologies, although its relationship with internet access may evolve as connectivity increasingly occurs through multiple types of devices. At the provincial level, computer ownership, mean years of schooling, expected years of schooling, and adjusted expenditure per capita represent complementary dimensions of technological readiness, educational capacity, and economic resources. Examining these indicators jointly provides a basis for developing predictive models of household internet access across provinces with different socioeconomic characteristics. Machine learning (ML) provides a flexible framework for capturing nonlinear relationships and complex interactions that may not be adequately represented by conventional linear models. This capability is relevant to digital-access research, where connectivity may reflect interactions among technological, educational, and socioeconomic conditions. Previous research has examined the use of ensemble methods, including Random Forest and XGBoost, in relation to factors associated with the digital divide (Kong *et al.*, 2025). Random Forest, Gradient Boosting, and XGBoost have also been applied to prediction problems involving heterogeneous data and nonlinear relationships (Boateng *et al.*, 2025; Cui *et al.*, 2024; Tahir *et al.*, 2025). Accordingly, this study compares Linear Regression, used as an interpretable baseline, with Random Forest, Gradient Boosting, and XGBoost to evaluate their ability to predict household internet access across Indonesian provinces.

Despite the growing application of machine learning to prediction tasks, the temporal generalizability of predictive models remains an important concern. Studies that rely on static or cross-sectional predictors and conventional train-test evaluation may not adequately represent situations in which data distributions evolve over time (Dong *et al.*, 2025; Li *et al.*, 2024). Research on distribution shifts and temporal dynamics has also identified these factors as challenges for reliable prediction and forecasting (Deng *et al.*, 2024; Liu *et al.*, 2025). In the context of provincial internet-access prediction in Indonesia, further examination is needed to determine how well models generalize to future periods and whether incorporating previous-year internet access provides additional predictive value beyond contemporaneous socioeconomic and technological indicators. To address these research needs, this study develops a machine learning framework for predicting household internet access across Indonesian provinces using secondary province-level data from 2020 to 2024. The framework compares models using contemporaneous socioeconomic and information and communication technology (ICT) features with models that additionally incorporate lagged household internet access to represent temporal persistence in regional connectivity. Linear Regression, Random

Forest, Gradient Boosting, and XGBoost are evaluated using three complementary strategies: an out-of-time holdout test using 2024 observations, province-grouped cross-validation using GroupKFold, and rolling-origin temporal validation. The out-of-time holdout assesses performance on a future period excluded from model fitting, GroupKFold evaluates performance across province-based groups, and rolling-origin validation examines predictive performance across successive future periods. Together, these strategies provide complementary evidence about model performance across regional and temporal evaluation settings. This study has three objectives: to compare the predictive performance of multiple machine learning models for household internet access across Indonesian provinces; to assess the additional predictive value of lagged internet access alongside contemporaneous socioeconomic and ICT-related features; and to identify the features that contribute most strongly to model predictions. By evaluating these objectives through complementary validation strategies, the study provides an empirical basis for assessing the role of historical connectivity in provincial internet-access prediction and for informing the monitoring of regional digital-access disparities in Indonesia.

2. Methodology

2.1 Research Design and Data Source

This study employed a quantitative predictive design using secondary province-level data from Statistics Indonesia (Badan Pusat Statistik, BPS) for the period 2020–2024. The analysis covered 34 provinces with complete observations, resulting in a balanced dataset of 170 province-year observations. Household internet access was designated as the target variable, while household computer ownership, mean years of schooling (RLS), expected years of schooling (HLS), and adjusted expenditure per capita were used as contemporaneous predictors.

Table 1. Variables and Operational Definitions

Code	Variable	Role	Unit
Y	Household Internet Access	Target	%
X1	Household Computer Ownership	Predictor	%
X2	Mean Years of Schooling (RLS)	Predictor	Years
X3	Expected Years of Schooling (HLS)	Predictor	Years
X4	Adjusted Expenditure per Capita	Predictor	Thousand IDR/person/year
Lag1	Previous-Year Household Internet Access	Temporal predictor	%

Source: Statistics Indonesia (BPS), 2020–2024; authors' compilation and processing.

To incorporate temporal information, a one-year lag of household internet access was constructed as follows:

$$\text{Lag1}_{i,t} = \text{InternetAccess}_{i,t-1} \quad (1)$$

where $\text{Lag1}_{i,t}$ denotes household internet access in province (i) during year (t-1). Because the 2020 observations did not have corresponding previous-year observations within the study period, the Lag-Augmented dataset covered 2021–2024.

2.2 Data Preprocessing and Exploratory Analysis

Data preprocessing involved checking missing values, duplicate observations, data types, province names, and the completeness of province-year records. Only provinces with complete observations were retained, and no imputation was required. Descriptive statistics and Pearson correlation analysis were then conducted to examine the distributions of the variables and their pairwise relationships. Multicollinearity among the contemporaneous predictors was assessed using the variance inflation factor (VIF). VIF values below 5 were considered acceptable for subsequent predictive modeling.

2.3 Feature Construction and Experimental Design

Two feature configurations were evaluated. The No-Lag configuration used the four contemporaneous predictors (X1–X4), whereas the Lag-Augmented configuration included the same predictors and the previous-year household internet access variable (Lag1). This design enabled a comparison of predictive performance with and without historical connectivity information.

Table 2. Experimental Feature Configurations and Temporal Data Split

Configuration	Predictors	Training Period	Training N	Test Period	Test N
No-Lag	X1, X2, X3, X4	2020–2023	136	2024	34
Lag-Augmented	X1, X2, X3, X4, Lag1	2021–2023	102	2024	34

Source: Authors' experimental design based on BPS data, 2020–2024.

The 2024 observations were designated as the out-of-time holdout set and excluded from model fitting and hyperparameter optimization. The No-Lag training set contained 136 observations from 2020–2023, while the Lag-Augmented training set contained 102 observations from 2021–2023 because the lagged feature was unavailable for 2020.

2.4 Machine Learning Models

Four regression algorithms were compared: Linear Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost). LR served as the linear baseline, whereas RF, GB, and XGBoost represented tree-based ensemble methods capable of modeling nonlinear relationships and interactions among predictors. The same four algorithms were evaluated under both feature configurations to support a consistent comparison of the predictive value of the lagged feature.

2.5 Model Validation and Hyperparameter Optimization

Model evaluation combined five-fold GroupKfold cross-validation with out-of-time holdout testing. Province was used as the grouping variable in GroupKfold so that observations from the same province were not distributed across the training and validation subsets within a fold. This procedure evaluated model performance across province-based groups, while the out-of-time holdout assessed predictions for the unseen year 2024. Hyperparameter optimization was conducted for Random Forest and XGBoost using RandomizedSearchCV on the Lag-Augmented training set. Root Mean Squared Error (RMSE) was used as the optimization criterion. Forty randomly sampled parameter combinations were evaluated using a fixed random seed of 42. The 2024 holdout set remained excluded from hyperparameter optimization to preserve its role as an independent temporal evaluation set.

2.6 Performance Evaluation

Model performance was evaluated using three regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). MAE measures the average magnitude of absolute prediction errors, whereas RMSE penalizes larger errors more strongly because the errors are squared before averaging. The (R^2) metric evaluates predictive performance relative to a benchmark based on the mean of the observed target values. Lower MAE and RMSE values indicate smaller prediction errors, while higher (R^2) values indicate better performance relative to this benchmark.

The metrics were calculated as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

where y_i denotes the observed value, $\hat{y}_i$ the predicted value, $\bar{y}$ the mean of the observed values, and n the number of observations. Negative out-of-sample R^2 values were retained because they indicate predictive performance below the mean-based benchmark

2.7 Rolling-Origin Temporal Validation

Rolling-origin temporal validation was conducted to assess temporal generalizability beyond a single holdout evaluation. Three expanding-window scenarios were evaluated for the Lag-Augmented XGBoost model: training on 2021 and testing on 2022; training on 2021–2022 and testing on 2023; and training on 2021–2023 and testing on 2024. In each scenario, the model was trained using observations from years preceding the test year. This chronological separation prevented observations from future years from entering the training data for the corresponding evaluation.

2.8 Model Interpretation, Error Analysis, and Computational Environment

The final Lag-Augmented XGBoost model was examined using normalized feature-importance scores to assess the relative predictive contribution of Lag1 and the four contemporaneous predictors. These scores were interpreted as indicators of predictive contribution rather than causal effects. Prediction errors were assessed on the 2024 holdout set by comparing observed and predicted household internet access across provinces. Residuals and absolute errors were used to identify provinces with relatively large prediction deviations. All analyses were conducted in Python using Google Colab. The analysis used pandas and NumPy for data processing, scikit-learn for model development and validation, XGBoost for boosted-tree modeling, statsmodels for diagnostic analysis, and Matplotlib for visualization. A fixed random seed of 42 was applied to stochastic procedures to support computational reproducibility.

3. Results

This section presents the empirical results of the machine learning experiments for predicting household internet access across Indonesian provinces. The analysis covers descriptive statistics, correlation and multicollinearity diagnostics, model performance, cross-validation, feature importance, provincial prediction errors, and rolling-origin temporal validation.

3.1 Descriptive and Preliminary Analysis

Table 5. Descriptive Statistics of Study Variables, 2020–2024

Variable	Mean	Std. Dev.	Min	Max
Internet Access (Y), %	83.497	10.521	35.140	98.180
Computer Ownership (X1), %	17.103	8.051	4.190	40.780
Mean Years of Schooling (X2), years	8.739	1.060	6.690	11.490
Expected Years of Schooling (X3), years	13.118	0.743	10.550	15.660
Adjusted Expenditure (X4), thousand IDR	11,623.859	2,536.951	6,954	19,953

Source: Authors' calculations based on Statistics Indonesia (BPS) data, 2020–2024.

The mean household internet-access rate across the study observations was 83.497%. The annual average increased from 75.290% in 2020 to 90.471% in 2024, reflecting an overall improvement in household internet connectivity during the study period. Over the same period, the cross-provincial standard deviation decreased from 13.763 to 3.573 percentage points, indicating a narrower distribution of internet-access rates across provinces in 2024 than in 2020. This pattern suggests that the differences in internet-access rates among provinces became less pronounced over time. Nevertheless, the declining standard deviation does not imply that all provinces experienced the same rate of improvement or that regional disparities were completely eliminated. These descriptive findings provide an initial context for evaluating whether socioeconomic indicators and historical internet-access data can help explain and predict differences in connectivity across provinces.

Table 6. Correlation and Multicollinearity Diagnostics

Predictor	Correlation with Y	VIF
Computer Ownership (X1)	0.478	3.240
Mean Years of Schooling (X2)	0.538	2.639
Expected Years of Schooling (X3)	0.336	1.645
Adjusted Expenditure per Capita (X4)	0.633	2.791

Source: Authors' calculations.

Adjusted expenditure per capita had the strongest positive correlation with household internet access ($r = 0.633$), followed by mean years of schooling ($r = 0.538$), computer ownership ($r = 0.478$), and expected years of schooling ($r = 0.336$). All VIF values were below 5, indicating that the contemporaneous predictors met the study's specified threshold for acceptable multicollinearity.

3.2 Baseline and Lag-Augmented Model Performance

The No-Lag models used the four contemporaneous predictors (X1–X4), whereas the Lag-Augmented models additionally incorporated previous-year household internet access (Lag1). All models were evaluated on the 2024 out-of-time holdout set. The mean internet-access rate was 81.758% in the training data and 90.471% in the holdout set. Their corresponding standard deviations were 10.854 and 3.573 percentage points, respectively.

Table 7. Comparison of No-Lag and Lag-Augmented Model Performance

Model	No-Lag RMSE	No-Lag (R^2)	Lag-Augmented RMSE	Lag-Augmented (R^2)	RMSE Improvement (%)
Linear Regression	7.3422	-3.3502	6.2810	-2.1836	14.45
Random Forest	5.3702	-1.3272	3.4589	0.0345	35.59
Gradient Boosting	6.1201	-2.0226	3.7952	-0.1623	37.99
XGBoost	5.9904	-1.8958	2.7260	0.4003	54.49

Source: Authors' calculations.

All No-Lag models produced negative (R^2) values on the 2024 holdout set, indicating that their predictions performed worse than the mean-based benchmark for that evaluation set. Incorporating Lag1 reduced RMSE across all four algorithms. The largest reduction was observed for XGBoost (54.49%), followed by Gradient Boosting (37.99%), Random Forest (35.59%), and Linear Regression (14.45%). The Lag-Augmented XGBoost model achieved the lowest holdout RMSE (2.7260) and the highest holdout (R^2) (0.4003) among the evaluated configurations. These results indicate that adding previous-year household internet access improved holdout prediction accuracy for all four algorithms, with the largest RMSE improvement observed for XGBoost.

3.3 Cross-Validation and Hyperparameter Optimization

Five-fold GroupKFold cross-validation was applied to the Lag-Augmented models, using province as the grouping variable. The results are summarized in Table 8.

Table 8. GroupKFold Cross-Validation Performance

Model	MAE	RMSE	(R^2)
Linear Regression	2.3970	3.4061	0.7883
Random Forest	2.6748	4.3545	0.7690
Gradient Boosting	2.8893	4.6913	0.7271
XGBoost	2.9292	4.8356	0.6850

Source: Authors' calculations.

Linear Regression achieved the lowest MAE and RMSE and the highest (R^2) among the four models under GroupKFold cross-validation. These results differ from the 2024 holdout results, where Lag-Augmented XGBoost achieved the best performance. The difference reflects the results obtained under distinct evaluation procedures and should be considered when interpreting model generalizability. Hyperparameter tuning improved the 2024 holdout performance of Random Forest but did not improve that of XGBoost. The untuned Lag-Augmented XGBoost model retained the best holdout performance, with an RMSE of 2.7260 and an (R^2) of 0.4003. It was therefore selected for the subsequent feature-importance and provincial error analyses, based on its performance on the designated future-period holdout set.

3.4 Feature Importance and Provincial Error Analysis

Feature-importance analysis was conducted using the retained Lag-Augmented XGBoost model to examine the relative predictive contribution of its input variables. The normalized importance scores are presented in Table 9.

Table 9. XGBoost Feature Importance

Feature	Importance (%)
Previous-Year Internet Access (Lag1)	55.57
Adjusted Expenditure per Capita (X4)	21.45
Computer Ownership (X1)	18.16
Mean Years of Schooling (X2)	3.88
Expected Years of Schooling (X3)	0.94

Source: Authors' calculations.

Previous-year household internet access (Lag1) had the highest normalized feature-importance score (55.57%). Adjusted expenditure per capita (21.45%) and computer ownership (18.16%) followed, while mean years of schooling (3.88%) and expected years of schooling (0.94%) had lower scores. These values describe the relative importance assigned by the fitted model and should not be interpreted as causal effects. The provincial error analysis identified Papua as the province with the largest absolute prediction error, at 9.671 percentage points. It was followed by Bali (5.084), West Kalimantan (3.867), Maluku (3.512), and Central Kalimantan (2.952). The results indicate that prediction errors varied across provinces, with Papua showing the largest deviation between observed and predicted household internet access in the 2024 holdout set.

3.5 Rolling-Origin Temporal Validation

Rolling-origin temporal validation was conducted to assess the temporal generalizability of the Lag-Augmented XGBoost model using expanding training windows. Each scenario used observations from years preceding the test year, with the results presented in Table 10.

Table 10. Rolling-Origin Temporal Validation

Training Period	Test Year	MAE	RMSE	(R^2)
2021	2022	3.8544	6.6872	0.5473
2021–2022	2023	2.5080	4.0293	0.8121
2021–2023	2024	2.1108	2.7260	0.4003
Mean	—	2.8244	4.4808	0.5866

Source: Authors' calculations.

Across the three temporal evaluations, the model achieved mean MAE and RMSE values of 2.8244 and 4.4808, respectively, and a mean (R^2) of 0.5866. MAE and RMSE decreased across the successive evaluation scenarios as the training window expanded. However, (R^2) varied across test years, reaching 0.8121 in 2023 before decreasing to 0.4003 in 2024. Thus, the results show that prediction errors and relative model performance varied across temporal evaluation periods.

4. Discussion

The results demonstrate that incorporating temporal information improved the prediction of household internet access across Indonesian provinces. The No-Lag models produced negative (R^2) values on the 2024 holdout set, whereas adding previous-year internet access reduced RMSE across all four algorithms. XGBoost achieved the largest RMSE reduction (54.49%), with an RMSE of 2.7260 and an (R^2) of 0.4003. These findings suggest that household internet access exhibits temporal persistence at the provincial level and that contemporaneous socioeconomic indicators alone may be insufficient to predict connectivity accurately in a future period. The improvement observed across all algorithms also indicates that the predictive value of historical connectivity was not limited to a single modeling approach. The importance of temporal information is further supported by the feature-importance results. Lag1 accounted for the largest normalized importance score (55.57%), followed by adjusted expenditure per capita (21.45%) and computer ownership (18.16%). This pattern indicates that historical connectivity and socioeconomic capacity provide substantial predictive information for the model. The findings are consistent with previous studies reporting associations between socioeconomic conditions, education, digital resources, and disparities in internet adoption (Kartiasih *et al.*, 2022; Singh & Mohanty, 2025). However, feature-importance scores represent relative predictive contributions within the fitted model and should not be interpreted as evidence of causal influence. In practical terms, the results suggest that historical connectivity may be a useful input when developing models to monitor changes in regional internet access. A further finding is the difference between cross-validation and temporal holdout performance. Linear Regression achieved the highest (R^2) under GroupKFold cross-validation (0.7883), whereas XGBoost performed best on the 2024 out-of-time holdout. This difference indicates that strong performance across province-based validation groups does not necessarily translate into superior prediction of a future period. The mean household internet-access rate increased from 81.758% in the training data to 90.471% in the 2024 holdout set. This difference in the target distribution highlights the importance of temporally separated evaluation when assessing models intended to predict evolving digital-access indicators. The contrasting results also suggest that model selection should consider the intended prediction setting rather than relying exclusively on cross-validation metrics. Therefore, reporting multiple evaluation strategies can provide a more comprehensive assessment of model performance and reduce reliance on a single validation result.

Rolling-origin validation provides additional evidence of performance across successive future-year evaluations. The Lag-Augmented XGBoost model achieved positive (R^2) values for 2022 (0.5473), 2023 (0.8121), and 2024 (0.4003), with a mean (R^2) of 0.5866 across the three scenarios. However, performance varied across test years, and the 2024 (R^2) was lower than the values obtained for 2022 and 2023. Thus, although the model maintained positive predictive performance relative to the mean-based benchmark in all three evaluations, its generalization performance was not uniform over time. This variation emphasizes the need to evaluate predictive models across multiple temporal windows, particularly when regional indicators may change at different rates. Provincial error analysis further revealed variation in prediction accuracy across regions. Papua recorded the largest absolute prediction error (9.671 percentage points), followed by Bali (5.084), West Kalimantan (3.867), Maluku (3.512), and Central Kalimantan (2.952). These differences indicate that the model did not predict household internet access equally accurately across all provinces. A single national model may not fully capture regional variations in connectivity trajectories, although the present results alone do not establish the specific geographic, infrastructural, or socioeconomic factors responsible for the observed errors. Further analysis incorporating additional regional indicators would be needed to investigate these differences. Identifying such error patterns is important because aggregate performance metrics may conceal substantial prediction deviations in individual provinces. Overall, the findings highlight the predictive value of combining contemporaneous socioeconomic indicators with historical connectivity information and evaluating models through temporally appropriate validation strategies. Several limitations should be acknowledged. First, the observation period covers only 2020–2024, restricting the assessment of longer-term temporal patterns. Second, the analysis includes 34 provinces with complete observations, which limits the scope of regional inference to the included units. Third, the predictor set is relatively limited and does not directly represent infrastructure availability, network quality, internet affordability, urbanization, or digital literacy. Future research could extend the observation period, incorporate these additional indicators where reliable data are available, and apply explainability techniques such as

SHAP to investigate the contribution of individual features to model predictions. Future studies could also assess whether province-specific models or regional groupings improve prediction accuracy in areas with larger errors. Such extensions could provide a more detailed understanding of prediction behavior and support more comprehensive monitoring of regional digital-access disparities.

5. Conclusion

This study investigated the prediction of household internet access across Indonesian provinces using socioeconomic and information and communication technology (ICT) indicators. It compared Linear Regression, Random Forest, Gradient Boosting, and XGBoost under No-Lag and Lag-Augmented configurations using province-level data for 2020–2024. An out-of-time holdout, GroupKFold cross-validation, and rolling-origin temporal validation were employed to assess predictive performance under different evaluation settings. The results showed that models relying exclusively on contemporaneous predictors performed poorly on the 2024 holdout, whereas incorporating previous-year household internet access reduced RMSE across all four algorithms. The Lag-Augmented XGBoost model achieved the best holdout performance, with an RMSE of 2.7260 and an (R^2) of 0.4003, representing a 54.49% reduction in RMSE compared with its No-Lag counterpart. Lag1 also had the highest normalized feature-importance score (55.57%), followed by adjusted expenditure per capita and computer ownership. The findings demonstrate the predictive value of historical connectivity and highlight the importance of selecting validation strategies that reflect the intended prediction task. Although Linear Regression achieved the strongest performance under GroupKFold cross-validation, XGBoost performed best on the 2024 out-of-time holdout, illustrating that cross-regional validation performance does not necessarily translate into superior future-period prediction. Rolling-origin validation yielded positive (R^2) values across all three evaluation periods, with a mean (R^2) of 0.5866, although performance varied across years. These results provide an empirical basis for using temporally validated machine learning to support the monitoring of regional digital-access disparities in Indonesia. In particular, incorporating historical connectivity may help improve the assessment of provincial internet-access patterns when future observations are not yet available. Several limitations should be acknowledged. The study covers a relatively short observation period, includes 34 provinces with complete observations, and uses a restricted set of socioeconomic and ICT predictors. Consequently, the findings should be interpreted within the scope of the available data and the evaluated prediction settings. Future research could extend the time series and incorporate additional indicators, including broadband infrastructure, network quality, internet affordability, urbanization, and digital literacy. Explainability techniques such as SHAP could also be applied to examine individual feature contributions more comprehensively. Further investigation of province-specific prediction errors may help determine whether regional models or additional contextual variables can improve predictive accuracy in provinces where the current model performs less accurately.

Acknowledgments

The authors would like to thank Statistics Indonesia (Badan Pusat Statistik, BPS) for providing the publicly available secondary data used in this study. The authors also acknowledge the open-source Python libraries and Google Colab environment used for data processing, model development, and machine learning analysis. These resources supported the implementation of the experimental procedures and the evaluation of the predictive models. The authors recognize the importance of accessible public data and open-source computational tools in facilitating reproducible research and further investigation into regional digital-access disparities in Indonesia.

References

- Boateng, S. A., Xi, J., Fumey, M. P., & Kumi, J. K. (2025). Exploring the nonlinear relationship of environmental sustainability factors and economic growth in West Africa: Novel machine learning evidence. *Sustainable Development*. <https://doi.org/10.1002/sd.70216>
- Cui, C., Zhuang, D., Bao, Y., & Li, Z. (2024). Population density prediction at township scale supported by machine learning method: A case study in Inner Mongolia. *ISPRS International Journal of Geo-Information*, 13(12), Article 426. <https://doi.org/10.3390/ijgi13120426>
- Deng, S., Sprangers, O., Li, M., Schelter, S., & de Rijke, M. (2024). Domain generalization in time series forecasting. *ACM Transactions on Knowledge Discovery from Data*. <https://doi.org/10.1145/3643035>
- Dong, J.-D., Crichton, K., Yamada, A., Sawaya, Y., Cranor, L. F., & Christin, N. (2026). Accurate, generalizable, and practical behavioral models to identify impending user exposure to malicious websites. *ACM Transactions on the Web*, 20(1), Article 1, 1–31. <https://doi.org/10.1145/3768587>
- Du, H., Du, H., Zhou, N., Cao, H., Zhang, J.-T., Zhang, J.-T., Chen, A., & King, R. B. (2021). Economic inequality is associated with lower internet use: A nationally representative study. *Social Indicators Research*, 155(3), 789–803. <https://doi.org/10.1007/s11205-021-02632-8>
- Jaya, I. G. N. M., Pahlevi, S. M., Susenna, A., Agustina, L., Kusumasari, D., Sukma, Y. A. A., Hernikawati, D., Rahmi, A., Pravitasari, A. A., & Kristiani, F. (2024). Framework for monitoring the spatiotemporal distribution and clustering of the digital society index of Indonesia. *Sustainability*, 16(24), Article 11258. <https://doi.org/10.3390/su162411258>
- Kartiasih, F., Nachrowi, N. D., Wisana, I. D. G. K., & Handayani, D. (2022). Inequalities of Indonesia's regional digital development and its association with socioeconomic characteristics: A spatial and multivariate analysis. *Information Technology for Development*. <https://doi.org/10.1080/02681102.2022.2110556>
- Kharisma, B. (2022). Surfing alone? The internet and social capital: Evidence from Indonesia. *Journal of Economic Structures*, 11(1). <https://doi.org/10.1186/s40008-022-00267-7>
- Kong, H., Fang, H., & Zhang, G. (2025). Predicting and explaining life satisfaction among older adults using tree-based ensemble models and SHAP: Evidence from the digital divide survey. *PLOS ONE*, 20, Article e0337938. <https://doi.org/10.1371/journal.pone.0337938>
- Lai, J., & Widmar, N. O. (2021). Revisiting the digital divide in the COVID-19 era. *Applied Economic Perspectives and Policy*, 43(1), 458–464. <https://doi.org/10.1002/aep.13104>
- Li, Z., Zhou, F., Wang, Z., Xu, X., Liu, L., & Yin, G. (2024). Measuring and classifying IP usage scenarios: A continuous neural trees approach. *Scientific Reports*, 14, Article 5144. <https://doi.org/10.1038/s41598-024-55750-x>
- Liu, W., Ning, K., Zhang, L., Marier, S. M., Liu, Y., Chen, Y., & Xia, F. (2025). Deep learning for time series forecasting: A survey. *International Journal of Machine Learning and Cybernetics*. <https://doi.org/10.1007/s13042-025-02560-w>
- Ramírez-Correa, P., Mariano, A. M., & Santos, M. R. (2025). Digital and sustainable education and social inclusion: A bibliometric review with the consolidated meta-analytical approach. *Sustainability*, 17(13), Article 5677. <https://doi.org/10.3390/su17135677>