HANIF
Journal of Information Systems
https://journal.ilmubersama.com/index.php/hanif

Research Article

Comparative Evaluation of Remote Sensing, Socioeconomic, and Integrated Data for Predicting Regional Economic Growth in North Sumatra Using Machine Learning

Dian Septiana [1], Sisti Nadia Amalia [2], Fahmi Ashari S. Sihaloho[3]

[1] *Department of Statistics, Faculty of Mathematic and Natural Science, Universitas Negeri Medan, 20221, North Sumatera, Indonesia*
[2] *Department of Statistics, Faculty of Mathematic and Natural Science, Universitas Negeri Medan, 20221, North Sumatera, Indonesia*
[3] *Department of Economic Education, Faculty of Economics, Universitas Negeri Medan, 20221, North Sumatera, Indonesia*

ARTICLE INFORMATION

Received: February 00, 00
Revised: March 00, 00
Available Online: April 00, 00

KEYWORDS

Regional Economic Growth; Remote Sensing; Socioeconomic Indicators; Machine Learning; Support Vector Regression

CORRESPONDENCE

Phone: +62 822 93832028
E-mail: dianseptiana@unimed.ac..id

ABSTRACT

This study evaluates the predictive performance of remote sensing variables, socioeconomic indicators, and their integration for estimating regional economic growth across 33 districts and municipalities in North Sumatra, Indonesia. A quantitative cross-sectional design was employed using three predictor scenarios: remote sensing variables (Nighttime Light, NDVI, Built-up Area, and Land Surface Temperature), socioeconomic indicators (Human Development Index, Open Unemployment Rate, Fiscal Capacity Index, and Disaster Risk Index), and an integrated dataset. Four regression algorithms (Linear Regression, Support Vector Regression, Random Forest, and K-Nearest Neighbors) were optimized using RandomizedSearchCV and validated through Leave-One-Out Cross Validation. Model performance was evaluated using RMSE, MAE, and R^2, while permutation importance assessed predictor contributions. Support Vector Regression achieved the best predictive performance across all predictor scenarios. The socioeconomic dataset yielded the highest prediction accuracy (RMSE = 0.625, MAE = 0.480, R^2 = 0.198), outperforming the remote sensing-only dataset (RMSE = 0.685, MAE = 0.471, R^2 = 0.038) and the integrated dataset (RMSE = 0.647, MAE = 0.492, R^2 = 0.140). Although the R^2 values were relatively low, they reflect the complexity of regional economic growth and the influence of factors beyond those included in this study. Permutation importance identified the Human Development Index and Disaster Risk Index as the most influential predictors. These findings indicate that socioeconomic indicators are stronger predictors of regional economic growth, while remote sensing variables provide complementary spatial information. Although integrating remote sensing variables did not improve predictive accuracy, it offers valuable environmental context for more comprehensive data-driven regional development planning.

INTRODUCTION

Economic growth is one of the primary indicators used to evaluate regional development because it reflects the capacity of a region to generate production, create employment opportunities, and improve community welfare. In Indonesia, regional economic performance is measured using Gross Regional Domestic Product (GRDP), which serves as an important basis for development planning, policy evaluation, and resource allocation. According to Statistics Indonesia (BPS), GRDP at constant prices is used to measure regional economic growth, whereas GRDP at current prices describes the economic structure and production capacity of a region. Consequently, understanding the factors associated with regional economic growth is essential for supporting evidence based regional development policies [1].

https://doi.org/10.12345/xxxxx

Regional economic growth is influenced by interactions among physical, environmental, institutional, and socioeconomic factors. Human capital, labour market conditions, fiscal capacity, infrastructure, land development, and environmental characteristics collectively determine regional productivity and competitiveness. Consequently, regions with similar economic structures may exhibit different growth trajectories because they possess different development capacities and landscape characteristics. These variations indicate that analysing regional economic growth requires a comprehensive framework capable of representing both the physical environment and socioeconomic conditions.

Recent advances in Earth observation technology have expanded the application of remote sensing beyond environmental monitoring into socioeconomic and regional economic analysis. Satellite observations provide spatially continuous information describing urban expansion, vegetation dynamics, land cover transformation, and human activities across extensive geographic areas. Donaldson and Storeygard demonstrated that satellite observations have become an important source of information for empirical economic research because they reveal spatial patterns that are difficult to capture using conventional statistical indicators alone [2]. Similarly, Burke *et al.* emphasized that remotely sensed data contribute significantly to sustainable development assessment by integrating environmental information with socioeconomic analysis [3].

Among various remotely sensed variables, nighttime light has become one of the most widely adopted proxies for economic activity because artificial illumination reflects the concentration of settlements, transportation networks, commercial centres, and industrial activities. The harmonized nighttime light dataset developed by Li *et al.* has substantially improved the temporal consistency of nighttime observations, enabling more reliable analyses of long term regional economic dynamics [4]. Zhao *et al.* further demonstrated that nighttime light has been extensively applied to analyse urbanization, regional development, disaster recovery, and socioeconomic activities across different geographical contexts [5]. Proville et al. also reported consistent relationships between nighttime light intensity and a wide range of socioeconomic indicators at the global scale, reinforcing the usefulness of nighttime light as a proxy for regional economic activity and human development [6].

Although Nighttime Light (NTL) is widely used as a proxy for economic activity and urbanization, it captures only one dimension of regional development because agricultural production, vegetation productivity, and land-based economic activities may contribute substantially to regional economies despite generating limited artificial illumination [7]. This finding has encouraged researchers to integrate additional remotely sensed variables capable of describing complementary characteristics of regional landscapes.

Several studies have shown that combining multiple remotely sensed variables improves the representation of regional economic conditions. NDVI reflects vegetation productivity, built up area represents urban expansion and infrastructure development, while Land Surface Temperature (LST) captures thermal characteristics associated with land cover transformation and urbanization. Pagaduan demonstrated that integrating nighttime light with land cover and net primary productivity produced more accurate estimates of subnational Gross Domestic Product than models relying solely on nighttime light [8]. Similar findings were reported by Kurnia *et al.*, who confirmed the usefulness of VIIRS nighttime light for modelling regional economic activity in Central Java [9]. Kamal *et al.* also showed that combining remote sensing variables with alternative digital information improved the estimation of Indonesia's Gross Domestic Product, indicating that economic conditions are better represented through the integration of multiple information sources [10]. More recently, Wang et al. further demonstrated that integrating nighttime light with OpenStreetMap data improved the assessment of regional economic development across multiple spatial scales compared with models based on a single data source, highlighting the advantage of data fusion for representing regional economic characteristics [11].

Remote sensing variables describe the spatial manifestation of economic activities, whereas socioeconomic indicators explain the structural capacity of a region to transform available resources into economic output. Human Development Index reflects the quality of human capital, the open unemployment rate represents labour market conditions, fiscal capacity indicates the financial capability of local governments to support development programmes, and disaster risk influences regional resilience through its effects on production activities and infrastructure. Rather than representing competing sources of information, remotely sensed and socioeconomic variables provide complementary perspectives for understanding regional economic growth.

Machine learning has become an increasingly important approach for analysing regional economic systems because it can model complex relationships among heterogeneous variables without requiring restrictive statistical assumptions. The

integration of machine learning with satellite imagery has further expanded the application of remote sensing in socioeconomic studies. Jean et al. demonstrated that deep learning models trained on daytime satellite imagery were able to predict poverty with remarkable accuracy, providing early evidence that remotely sensed data can effectively capture latent socioeconomic characteristics beyond conventional environmental applications [12]. Random Forest has been widely applied in remote sensing owing to its ability to model nonlinear relationships and evaluate variable importance [13]. Maxwell *et al.* also reported that machine learning algorithms substantially improve predictive performance in remote sensing applications involving multidimensional datasets [14]. Support Vector Regression has demonstrated competitive performance in nonlinear prediction problems and maintains good generalization ability across different sample characteristics [15]. At the national level, Muchisha *et al.* showed that machine learning algorithms provide reliable predictions of Indonesia's economic growth and outperform several conventional approaches [16].
Previous studies generally follow three research directions. The first uses nighttime light as the primary proxy for economic activity and regional development [4], [5]. The second integrates multiple remotely sensed variables, including vegetation indices, land cover, and built-up information, to improve the estimation of regional economic performance [8], [17], [18]. The third applies machine learning algorithms to predict economic indicators using either remotely sensed variables or socioeconomic data [14], [15], [16]. These developments demonstrate that remote sensing and machine learning have become important components of regional economic analysis. However, studies integrating remotely sensed variables describing landscape characteristics with socioeconomic indicators representing regional development capacity within a unified prediction framework remain relatively limited. Since regional economic growth reflects interactions between physical environments and socioeconomic conditions, combining these complementary information sources is expected to provide a more comprehensive representation of regional economic dynamics.

Based on this perspective, this study integrates nighttime light, NDVI, built up area, and Land Surface Temperature with the Human Development Index, open unemployment rate, fiscal capacity index, and disaster risk index to predict regional economic growth in North Sumatra. Linear Regression is employed as a baseline model and compared with Support Vector Regression, Random Forest, and K-Nearest Neighbours Regression using Leave One Out Cross Validation. Furthermore, permutation importance analysis is performed to identify the relative contribution of each predictor. The findings are expected to enrich the application of remote sensing and machine learning in regional economic analysis while providing empirical evidence to support data driven regional development planning.

METHOD

Research Design

This study employed a quantitative cross-sectional research design to compare the predictive performance of remote sensing variables, socioeconomic indicators, and their integration in estimating regional economic growth in North Sumatra Province, Indonesia. The unit of analysis consisted of 33 districts and municipalities, representing all administrative regions in North Sumatra.
Three predictor scenarios were evaluated:

1. Remote sensing dataset, consisting of Nighttime Light (NL), Normalized Difference Vegetation Index (NDVI), Built-up Area, and Land Surface Temperature (LST);
2. Socioeconomic dataset, consisting of Human Development Index (HDI), Open Unemployment Rate (OUR), Fiscal Capacity Index (FCI), and Disaster Risk Index (DRI);
3. Integrated dataset, combining all remote sensing and socioeconomic variables.

The response variable was the regional economic growth rate (%), measured at the district/city level.

To evaluate the predictive capability of each dataset, four regression algorithms were employed: Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), and K-Nearest Neighbors (KNN). Linear Regression was included as a baseline statistical model, whereas SVR, RF, and KNN were selected because they represent machine learning approaches capable of capturing complex relationships between predictor variables and economic growth. The predictive performance of all models was compared using identical validation and evaluation procedures to ensure a fair comparison across the three predictor scenarios.

The analytical workflow was implemented using the Python programming language. Data manipulation and numerical computation were performed using the Pandas and NumPy libraries, whereas machine learning model development, hyperparameter optimization, cross-validation, and performance evaluation were conducted using the Scikit-learn library. Remote sensing variables were extracted and processed using Google Earth Engine (GEE) prior to model development.

 https://doi.org/10.12345/xxxxx

 ISSN 3025-2342 (ONLINE)

To ensure reproducibility, all stochastic procedures were initialized using a fixed random seed (random_state = 123). The overall research workflow is illustrated in Figure 1.

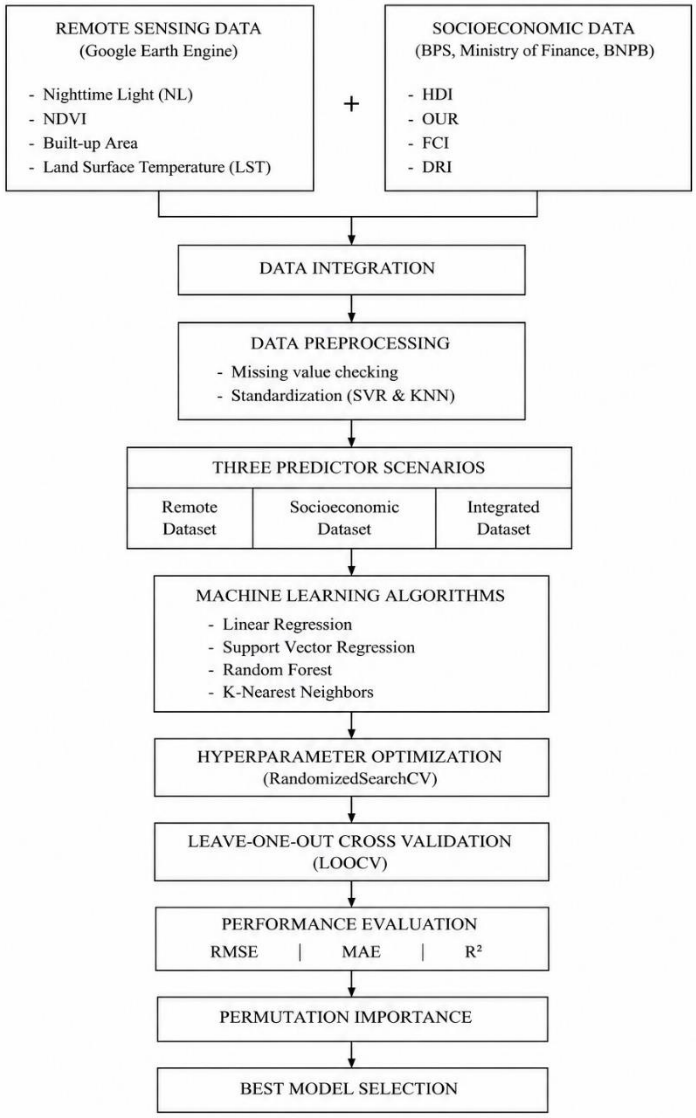

Figure 1. Research Workflow

Data Collection

This study employed a cross-sectional dataset comprising 33 districts and municipalities in North Sumatra Province, Indonesia. The response variable was the annual regional economic growth rate (%) at the district/city level. Two categories of predictor variables were used: remote sensing variables and socioeconomic variables.

Remote sensing variables

Remote sensing variables were extracted using Google Earth Engine (GEE), a cloud-based geospatial analysis platform that enables large-scale processing of satellite imagery. GEE provides direct access to numerous satellite image collections and allows efficient preprocessing, including cloud masking, temporal compositing, and spatial aggregation [19].

Four remote sensing indicators were generated:
1. Nighttime Light (NL) derived from the VIIRS Day/Night Band annual composite;
2. Normalized Difference Vegetation Index (NDVI) derived from MODIS vegetation products;
3. Built-up Area, derived from ESA WorldCover land-cover products;
4. Land Surface Temperature (LST) derived from MODIS Terra Land Surface Temperature products.

For each district/city, annual composite images representing the study period were generated in GEE. Subsequently, zonal statistics were performed using the administrative boundary of each district/city to calculate the mean value of each remote sensing variable. The resulting values were exported from GEE in tabular format for subsequent statistical analysis. The selected remote sensing variables were chosen to represent complementary dimensions of regional characteristics. NL has been widely used as a proxy for economic activity and urbanization, whereas NDVI reflects vegetation conditions and land productivity. Built-up Area represents urban expansion and infrastructure development, while Land Surface Temperature (LST) captures environmental conditions associated with land-use change and urbanization [2], [3], [4], [12].

Socioeconomic variables

Socioeconomic variables were obtained from official government institutions. The Human Development Index (HDI) and Open Unemployment Rate (OUR) were collected from Statistics Indonesia (BPS). The Fiscal Capacity Index (FCI) was obtained from the Ministry of Finance, while the Disaster Risk Index (DRI) was collected from the National Disaster Management Agency (BNPB). All variables correspond to the 2025 observation year to ensure temporal consistency with the remote sensing variables.

The selected socioeconomic variables represent key dimensions of regional development that are theoretically associated with economic growth. HDI reflects the quality of human capital through education, health, and living standards, whereas OUR describes labor market conditions and workforce utilization. FCI represents the financial capacity of local governments to support regional development through public investment and service provision, while DRI captures regional vulnerability to natural hazards that may disrupt economic activities and development. Collectively, these variables provide complementary information on the socioeconomic conditions influencing regional productivity and economic performance.

Data Preprocessing

Before model development, all predictor variables were examined for missing values and consistency across districts/cities. Since Support Vector Regression (SVR) and K-Nearest Neighbors (KNN) are sensitive to differences in variable scales, predictor variables were standardized using the StandardScaler, which transforms each variable into zero mean and unit variance. In contrast, Linear Regression and Random Forest were trained using the original predictor values because Random Forest is inherently scale-invariant, while Linear Regression does not require standardization for prediction. The preprocessing procedure was implemented within a machine learning pipeline to ensure that scaling was performed exclusively on the training data during cross-validation, thereby preventing data leakage.

Machine Learning Models

Four regression algorithms were employed in this study, namely Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), and K-Nearest Neighbors (KNN). These algorithms were selected because they represent different regression paradigms, ranging from conventional statistical modeling to nonlinear machine learning approaches. Linear Regression was used as the baseline model, whereas SVR, RF, and KNN were adopted to investigate their capability in modelling both linear and nonlinear relationships between remote sensing variables, socioeconomic indicators, and regional economic growth.

Linear Regression

Linear Regression (LR) was employed as the baseline model because of its simplicity, computational efficiency, and ease of interpretation. The model assumes a linear relationship between predictor variables and the response variable by estimating regression coefficients that minimize the residual sum of squares. Owing to these characteristics, Linear Regression is commonly used as a benchmark for comparing the predictive performance of more advanced machine learning models [20]. The regression model is expressed as

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon \tag{1}$$

where y denotes the regional economic growth rate, x_i represents the predictor variables, β_i denotes the regression coefficients, β_0 is the intercept, and ε is the random error term.

Support Vector Regression

 https://doi.org/10.12345/xxxxx

Support Vector Regression (SVR) was selected because of its strong generalization capability for regression problems involving nonlinear relationships and limited sample sizes. Instead of minimizing only the prediction error, SVR applies the Structural Risk Minimization principle to balance model complexity and prediction accuracy, thereby reducing the risk of overfitting [15]. This characteristic makes SVR particularly suitable for cross-sectional datasets with relatively few observations. The regression function is defined as

$$f\left(x\right)=w^{T}\phi\left(x\right)+b \tag{2}$$

where $\phi(x)$ denotes the nonlinear mapping function, w is the weight vector, and b is the bias term.

Both linear and Radial Basis Function (RBF) kernels were considered during hyperparameter optimization. The RBF kernel is expressed as

$$K\left(x_i,x_j\right)=\exp\left(-\gamma\left\|x_i-x_j\right\|^2\right) \tag{3}$$

where γ controls the influence of neighbouring observations. The optimal kernel and corresponding hyperparameters were determined through cross-validation.

Random Forest

Random Forest (RF) is an ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and random feature selection. The final prediction is obtained by averaging the predictions generated by individual trees, resulting in improved predictive accuracy and reduced variance. Moreover, RF effectively captures nonlinear relationships and interactions among predictor variables without requiring assumptions regarding data distribution [13]. The prediction is computed as

$$y=\frac{1}{B}\sum_{b=1}^{B}T_b\left(x\right) \tag{4}$$

where B is the total number of decision trees and $T_b(x)$ represents the prediction produced by the b-th tree.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) regression is a non-parametric learning algorithm that estimates the response variable using the average value of the nearest neighboring observations in the predictor space. Because KNN makes no assumptions regarding the functional relationship between predictors and the response, it is capable of modelling local nonlinear patterns [21], [22]. The prediction is calculated as

$$y=\frac{1}{k}\sum_{i=1}^{k}y_i \tag{5}$$

where k denotes the number of nearest neighbors and y_i represents the observed response values of neighboring observations.

Hyperparameter Optimization

Hyperparameter optimization was performed using Randomized Search Cross Validation (RandomizedSearchCV) implemented in the Scikit-learn library. Compared with exhaustive Grid Search, RandomizedSearchCV explores randomly selected combinations of hyperparameters within predefined search spaces, thereby substantially reducing computational cost while maintaining competitive optimization performance. This approach is particularly suitable for studies involving multiple machine learning algorithms and relatively limited computational resources [23].

The search space was defined individually for each algorithm according to their respective hyperparameters. For SVR, optimization included the kernel function, regularization parameter (C), kernel coefficient (γ), and epsilon (ε). RF optimization considered the number of decision trees, maximum tree depth, minimum samples required for node splitting, minimum samples required at leaf nodes, and the maximum number of features considered at each split. For KNN regression, the optimization process included the number of neighbors, distance metric, and weighting scheme.

A total of 30 randomly sampled hyperparameter combinations were evaluated for each algorithm. The optimal hyperparameter configuration was selected based on the lowest cross-validation prediction error obtained during the optimization process. To ensure reproducibility, all stochastic procedures were initialized using a fixed random seed (random_state = 123).

Table 1. Hyperparameter search space for machine learning models

Model	Hyperparameter	Search Space
SVR	Kernel	Linear, RBF, Polynomial
	C	0.01, 0.1, 1, 5, 10, 50, 100, 500
	Gamma	scale, auto, 0.001, 0.01, 0.1, 1
	Epsilon	0.001, 0.01, 0.05, 0.10, 0.20, 0.50
Random Forest	Number of Trees	100–700
	Maximum Depth	None, 2, 3, 5, 7, 10
	Minimum Samples Split	2, 3, 5, 10
	Minimum Samples Leaf	1, 2, 3, 5
	Maximum Features	sqrt, log2, None
KNN	Number of Neighbors	2–9
	Distance Metric	Manhattan, Euclidean
	Weight Function	Uniform, Distance

Model Validation

Model validation was performed using Leave-One-Out Cross Validation (LOOCV) because the dataset consisted of only 33 observations, making conventional train-test splitting or k-fold cross-validation less suitable. LOOCV is a special case of k-fold cross-validation in which the number of folds equals the number of observations ($k = n$). Consequently, one observation is used as the testing sample, while the remaining $n - 1$ observations are used to train the model. This procedure is repeated until every observation has served exactly once as the testing sample, ensuring that all available data contribute to both model training and evaluation [20], [23]. The prediction error across all validation iterations is estimated as

$$CV_{LOO} = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, y_i\right) \tag{6}$$

where CV_{LOO} denotes the Leave-One-Out Cross Validation error, n is the total number of observations, $L(\cdot)$ represents the loss function, y_i is the observed value of the i-th observation, and $\hat{y}_i$ is the predicted value obtained from a model trained using all observations except the i-th observation [23].

The prediction error was computed for each iteration and subsequently averaged across all observations to obtain an overall estimate of model performance. Compared with conventional train-test splitting, LOOCV maximizes the amount of training data available in each iteration, which is particularly advantageous when the sample size is limited. Furthermore, LOOCV provides a nearly unbiased estimate of prediction performance while reducing the dependence of model evaluation on a single random partition of the dataset [23].

In this study, LOOCV was implemented using the LeaveOneOut and cross_val_predict() functions available in the Scikit-learn library. The predicted values generated from each validation iteration were aggregated to compute the overall prediction accuracy using the evaluation metrics described in next section. The validation procedure was applied consistently to all machine learning models and predictor scenarios to ensure a fair comparison of predictive performance [20].

Performance Evaluation

The predictive performance of each machine learning model was evaluated using three widely adopted regression metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). These metrics were selected because they provide complementary information regarding prediction accuracy. RMSE measures the magnitude of prediction errors while assigning greater penalties to large deviations, MAE quantifies the average absolute prediction error regardless of its direction, and R^2evaluates the proportion of variance in the response variable explained by the prediction model. The combined use of these metrics provides a comprehensive assessment of model performance and facilitates an objective comparison among the evaluated algorithms [20], [23].

The Root Mean Square Error (RMSE) is calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i\right)^2} \tag{7}$$

where n is the number of observations, y_i denotes the observed value, and $\hat{y}_i$ represents the predicted value.

The Mean Absolute Error (MAE) is defined as

$$MAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i\right|} \tag{8}$$

and the Coefficient of Determination is computed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{9}$$

where $\bar{y}$ is the mean of the observed response values.

Lower RMSE and MAE values indicate better predictive accuracy because they correspond to smaller prediction errors. Conversely, a higher R^2value indicates that a larger proportion of the variability in regional economic growth is explained by the model. In this study, model performance was primarily compared using RMSE as the principal evaluation metric because it penalizes large prediction errors more heavily than MAE, making it particularly suitable for regression-based prediction problems. MAE and R^2were employed as complementary measures to provide a more comprehensive evaluation of predictive performance [20], [23].

Variable Importance

To interpret the contribution of each predictor, Permutation Importance was applied to the best-performing model. For each variable, predictor values were randomly permuted while the remaining variables were kept unchanged. The decrease in model performance after permutation represents the importance of the corresponding variable [24]. The permutation importance score is defined as

$$PI_j = E - E_j^{perm} \tag{9}$$

where Edenotes the original prediction error, and E_j^{perm}denotes the prediction error after permuting variable j.

RESULTS AND DISCUSSION

Comparative Performance of Machine Learning Models

The predictive performance of the four machine learning algorithms was compared under three predictor scenarios: remote sensing variables, socioeconomic variables, and their integration. Model performance was assessed using Leave-One-Out Cross Validation (LOOCV) together with three evaluation metrics, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). Lower values of RMSE and MAE indicate higher prediction accuracy, whereas larger R^2values indicate better explanatory capability. The comparison of predictive performance for all evaluated models is presented in Table 2.

Table 2. Predictive performance of machine learning models under different predictor scenarios

Predictor Scenario	Model	RMSE	MAE	R^2
Remote Sensing	LR	0.752	0.567	-0.161
	SVM	**0.685**	**0.471**	**0.038**
	RF	0.790	0.610	-0.283
	KNN	0.792	0.614	-0.289
Socioeconomic	LR	0.702	0.547	-0.012
	SVM	**0.625**	**0.480**	**0.198**
	RF	0.653	0.482	0.124
	KNN	0.657	0.494	0.112
Integrated	LR	0.654	0.504	0.121
	SVM	**0.647**	**0.492**	**0.140**
	RF	0.684	0.496	0.039
	KNN	0.682	0.525	0.043

Table 2 shows that SVR consistently achieved the best predictive performance across all predictor scenarios. Compared with LR, RF and KNN, SVR produced the lowest RMSE and MAE values while simultaneously yielding the highest coefficient of determination. This consistent performance indicates that SVR was more effective in modelling the relationship between predictor variables and regional economic growth, particularly for the relatively small cross-sectional dataset used in this study.

Among the three predictor scenarios, the socioeconomic dataset produced the highest prediction accuracy. Using only four socioeconomic variables, SVR achieved an RMSE of 0.625, MAE of 0.480, and an $R^2$of 0.198, outperforming both the remote sensing and integrated datasets. These findings suggest that socioeconomic characteristics provide more direct information regarding regional economic performance than remotely sensed environmental indicators. Variables describing human development, labour market conditions, fiscal capacity, and disaster risk appear to explain regional economic growth more effectively than satellite-derived variables alone.

The remote sensing dataset yielded the lowest predictive performance, with the best model (SVR) obtaining an RMSE of 0.685 and an $R^2$of 0.038. Although remotely sensed variables such as Nighttime Light, NDVI, Built-up Area, and Land Surface Temperature contain valuable spatial information, these variables alone were insufficient to explain the complexity of regional economic growth. Economic performance is influenced by multiple socioeconomic dimensions that cannot be directly observed from satellite imagery, thereby limiting the predictive capability of remote sensing variables when used independently.

Interestingly, integrating remote sensing and socioeconomic variables did not produce the highest predictive accuracy. Although the integrated dataset improved upon the remote sensing-only scenario, its performance remained slightly inferior to the socioeconomic dataset, with SVR achieving an RMSE of 0.647 and an $R^2$of 0.140. This result indicates that incorporating additional remote sensing variables did not substantially enhance prediction accuracy under the present dataset. One possible explanation is the limited sample size (33 districts and municipalities), which may reduce the ability of machine learning models to fully exploit the complementary information contained in heterogeneous predictor variables.

Overall, the comparative analysis demonstrates that SVR is the most suitable algorithm for predicting regional economic growth in North Sumatra. Notably, the socioeconomic dataset achieved superior predictive accuracy compared with both the remote sensing and integrated datasets. Although the highest-performing model produced a relatively low R^2 (0.198), this suggests that regional economic growth is influenced by numerous economic, institutional, and environmental factors beyond the variables considered in this study. The addition of remote sensing variables also did not substantially improve predictive performance, possibly due to the limited sample size (n = 33) and the complexity of regional economic systems. Nonetheless, the integrated dataset remains valuable for contextualizing environmental conditions alongside socioeconomic drivers and supporting more comprehensive regional economic assessments.

Best Hyperparameter Configuration

Hyperparameter optimization was conducted using RandomizedSearchCV to identify the optimal parameter configuration for each machine learning algorithm under the three predictor scenarios. The optimization process evaluated 30 randomly sampled parameter combinations using Leave-One-Out Cross Validation (LOOCV), and the parameter set yielding the lowest cross-validation RMSE was selected as the optimal configuration. The resulting hyperparameters are summarized in Table 4.

Table 3. Best hyperparameter configuration obtained by RandomizedSearchCV

Predictor Scenario	**Model**	**Best Hyperparameter Configuration**
Remote Sensing	SVR	Kernel = Linear, C = 0.01, Gamma = 0.001, Epsilon = 0.10
	Random Forest	n_estimators = 100, max_depth = 2, min_samples_split = 3, min_samples_leaf = 3, max_features = log2
	KNN	n_neighbors = 6, weights = uniform, distance = Manhattan
Socioeconomic	SVR	Kernel = Linear, C = 0.10, Gamma = auto, Epsilon = 0.20
	Random Forest	n_estimators = 100, max_depth = 2, min_samples_split = 10, min_samples_leaf = 5, max_features = None
	KNN	n_neighbors = 5, weights = distance, distance = Euclidean
Integrated	SVR	Kernel = RBF, C = 10, Gamma = 1, Epsilon = 0.10
	Random Forest	n_estimators = 500, max_depth = 3, min_samples_split = 5, min_samples_leaf = 3, max_features = log2
	KNN	n_neighbors = 9, weights = uniform, distance = Manhattan

The optimization results indicate that different predictor scenarios required different hyperparameter configurations to achieve optimal predictive performance. For the remote sensing and socioeconomic datasets, the optimal SVR model employed a linear kernel, suggesting that the relationships between predictors and regional economic growth could be adequately represented in the transformed linear feature space. In contrast, the integrated dataset selected the Radial Basis Function (RBF) kernel, indicating that combining remote sensing and socioeconomic variables introduced more complex nonlinear relationships that were better captured by a nonlinear kernel.

For the Random Forest model, the optimal configurations generally consisted of shallow tree depths and relatively large leaf sizes, indicating that simpler tree structures produced better generalization performance for the limited number of observations. Likewise, the K-Nearest Neighbors model selected different neighborhood sizes and distance metrics across the three predictor scenarios, reflecting differences in the local structure of each dataset. Overall, these results demonstrate that hyperparameter optimization successfully adapted each algorithm to the characteristics of the corresponding predictor scenario, thereby improving the fairness of the comparative evaluation presented in the previous section.

Prediction Performance of the SVR Model Using the Integrated Predictor Set

Although the socioeconomic predictor set achieved the highest predictive accuracy, *of the SVR model using the integrated predictor set* was selected for further interpretation because it incorporates both remote sensing and socioeconomic variables within a single prediction framework. This enables a comprehensive evaluation of prediction performance while simultaneously examining the contribution of all predictor variables. Figure 2 illustrates the relationship between the observed and predicted values of regional economic growth generated by the integrated SVR model.

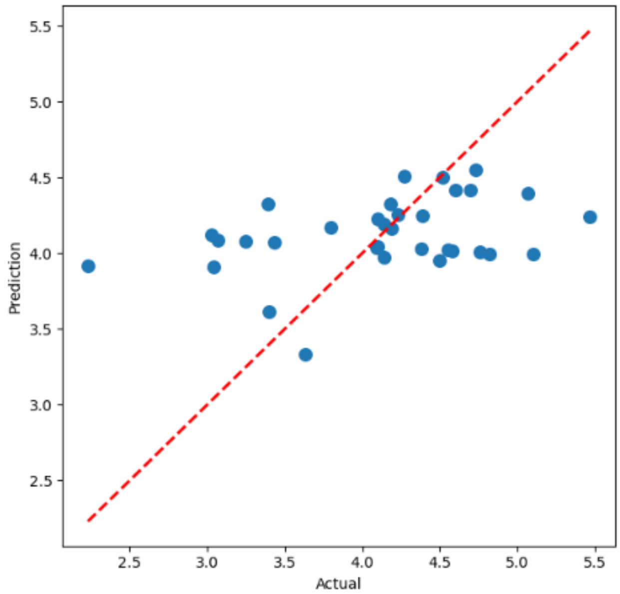

Figure 2. Actual versus predicted regional economic growth using the integrated SVR model.

As shown in Figure 2, most predicted values generally follow the distribution of the observed values, indicating that the model captures the overall pattern of regional economic growth across districts and cities. The points are concentrated around the central region of the plot, although several observations exhibit noticeable deviations, particularly for regions with relatively high or low growth rates. These deviations indicate that prediction errors remain for some observations, reflecting the complexity of regional economic growth and the heterogeneous influence of environmental and socioeconomic factors.

The integrated SVR model provides a reasonable representation of the observed variation in regional economic growth. Although its predictive accuracy is slightly lower than that of the socioeconomic model, it offers a more comprehensive analytical framework by incorporating complementary information from both remote sensing and socioeconomic variables. This integrated representation provides the basis for examining the relative contribution of each predictor, which is discussed in the following section.

Variable Importance

Permutation importance was applied to evaluate the contribution of each predictor variable to the predictive performance of the integrated Support Vector Regression (SVR) model. Variables with higher importance scores have a greater influence on the model predictions. The ranking of predictor importance is presented in Figure 3.

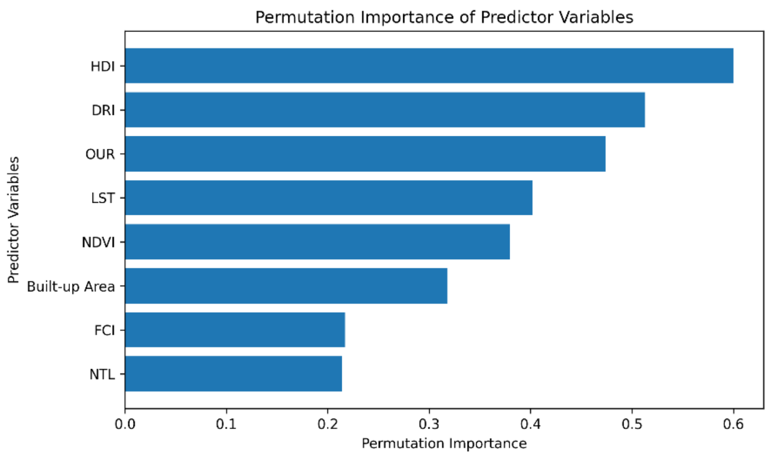

Figure 3. Permutation importance of predictor variables in the integrated SVR model.

Figure 3 shows that the Human Development Index (HDI) was the most influential predictor, with an importance score of 0.600. This was followed by the Disaster Risk Index (DRI) (0.513) and the Open Unemployment Rate (OUR) (0.474). These results indicate that socioeconomic indicators contributed more substantially to the prediction of regional economic growth than the other variables included in the model. Among the remote sensing variables, Land Surface Temperature (LST) exhibited the highest importance score (0.402), followed by NDVI (0.380) and Built-up Area (0.318). Nighttime Light had the lowest importance score (0.214), while the Fiscal Capacity Index (FCI) also showed a relatively limited contribution (0.217). Although these variables ranked below the leading socioeconomic indicators, they provided complementary environmental and spatial information that enhanced the predictive capability of the integrated model.

The dominance of socioeconomic variables is consistent with the comparative model performance presented before, where the socioeconomic dataset achieved the highest prediction accuracy. At the same time, the contribution of several remote sensing variables, particularly LST and NDVI, demonstrates that satellite-derived information can complement conventional socioeconomic indicators by capturing environmental characteristics associated with regional economic conditions.

CONCLUSION

This study compared the predictive performance of remote sensing variables, socioeconomic variables, and their integration for predicting regional economic growth in North Sumatra using four predictive models. The results show that Support Vector Regression (SVR) consistently achieved the best predictive performance across the three predictor scenarios. Although the socioeconomic predictor set produced the highest prediction accuracy, the integrated predictor set provided a more comprehensive representation by combining environmental and socioeconomic information within a single prediction framework.

The comparison of the three predictor scenarios indicates that socioeconomic variables play a dominant role in predicting regional economic growth, achieving higher accuracy than both the remote sensing and integrated models. While integrating remote sensing data does not automatically improve predictive accuracy, likely due to sample size constraints, it offers valuable complementary spatial and environmental context (such as LST and NDVI) that cannot be captured by socioeconomic indicators alone. Therefore, data integration provides a more holistic analytical perspective, even though socioeconomic indicators remain the stronger primary predictors in small-sample subnational contexts.

The proposed framework provides a practical approach for comparing different predictor scenarios in regional economic prediction using machine learning. It also provides empirical evidence that remote sensing data can complement socioeconomic indicators, supporting the development of more comprehensive data-driven models for regional economic assessment and development planning.

Several limitations should be acknowledged. The analysis was based on a cross-sectional dataset covering only 33 districts and municipalities in North Sumatra and included observations from a single year. In addition, only a limited number of remote sensing and socioeconomic variables were considered. Future research may incorporate multi-year observations, additional predictor variables, and broader geographic coverage to further evaluate the potential of integrating remote sensing and socioeconomic information for regional economic prediction

ACKNOWLEDGMENT

The authors acknowledge the financial support provided by Universitas Negeri Medan through the PNBP Research Grant. The authors also thank all individuals who contributed to this research.

REFERENCES

[1] BPS-Statistics Sumatera Utara Province, *GDRP OF REGENCY/MUNICIPALITY IN SUMATERA UTARA BY EXPENDITURE, 2021–2025*, vol. 10. BPS-Statistics Sumatera Utara Province, 2026.

[2] D. Donaldson and A. Storeygard, “The View from Above: Applications of Satellite Data in Economics,” *Journal of Economic Perspectives*, vol. 30, no. 4, pp. 171–198, Nov. 2016, doi: 10.1257/jep.30.4.171.

[3] M. Burke, A. Driscoll, D. B. Lobell, and S. Ermon, “Using satellite imagery to understand and promote sustainable development,” *Science (1979).*, vol. 371, no. 6535, Mar. 2021, doi: 10.1126/science.abe8628.

[4] X. Li, Y. Zhou, M. Zhao, and X. Zhao, “A harmonized global nighttime light dataset 1992–2018,” *Sci. Data*, vol. 7, no. 1, p. 168, 2020, doi: 10.1038/s41597-020-0510-y.

[5] M. Zhao *et al.*, "Applications of Satellite Remote Sensing of Nighttime Light Observations: Advances, Challenges, and Perspectives," *Remote Sens. (Basel).*, vol. 11, no. 17, p. 1971, Aug. 2019, doi: 10.3390/rs11171971.

[6] J. Proville, D. Zavala-Araiza, and G. Wagner, "Night-time lights: A global, long term look at links to socio-economic trends," *PLoS One*, vol. 12, no. 3, p. e0174610, Mar. 2017, doi: 10.1371/journal.pone.0174610.

[7] J. Gibson, O. Alimi, and G. Boe-Gibson, "Lost in Translation? A Critical Review of Economics Research Using Nighttime Lights Data," *Remote Sens. (Basel).*, vol. 17, no. 7, p. 1130, Mar. 2025, doi: 10.3390/rs17071130.

[8] J. A. Pagaduan, "Do higher-quality nighttime lights and net primary productivity predict subnational GDP in developing countries? Evidence from the Philippines," *Asian Economic Journal*, vol. 36, no. 3, pp. 288–317, Sep. 2022, doi: 10.1111/asej.12278.

[9] P. Kurnia, P. Pangi, K. D. Astuti, S. W. Mispaki, and Y. K. Nugraha, "The Prediction Modelling Analysis of Regional Economic Activity Based on Nighttime Light in Central Java Province," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1551, no. 1, p. 012028, Nov. 2025, doi: 10.1088/1755-1315/1551/1/012028.

[10] F. Y. Kamal, M. I. Sari, M. F. G. U. Utami, and F. Kartiasih, "PENGGUNAAN REMOTE SENSING DAN GOOGLE TRENDS UNTUK ESTIMASI PRODUK DOMESTIK BRUTO INDONESIA," *Equilibrium: Jurnal Penelitian Pendidikan dan Ekonomi*, vol. 21, no. 02, pp. 37–59, Jul. 2024, doi: 10.25134/equi.v21i02.9455.

[11] Z. Wang *et al.*, "A comprehensive assessment approach for multiscale regional economic development: Fusion modeling of nighttime lights and OpenStreetMap data," *Geography and Sustainability*, vol. 6, no. 2, p. 100230, 2025, doi: https://doi.org/10.1016/j.geosus.2024.08.009.

[12] N. Jean, M. Burke, M. Xie, W. M. Alampay Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science (1979).*, vol. 353, no. 6301, pp. 790–794, Aug. 2016, doi: 10.1126/science.aaf7894.

[13] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016, doi: https://doi.org/10.1016/j.isprsjprs.2016.01.011.

[14] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: an applied review," *Int. J. Remote Sens.*, vol. 39, no. 9, pp. 2784–2817, May 2018, doi: 10.1080/01431161.2018.1433343.

[15] R. I. Tange, M. A. Rasmussen, E. Taira, and R. Bro, "Benchmarking support vector regression against partial least squares regression and artificial neural network: Effect of sample size on model performance," *J. Near Infrared Spectrosc.*, vol. 25, no. 6, pp. 381–390, 2017, [Online]. Available: https://opg.optica.org/jnirs/abstract.cfm?URI=jnirs-25-6-381

[16] N. D. Muchisha, N. Tamara, A. Andriansyah, and A. M. Soleh, "Nowcasting Indonesia's GDP Growth Using Machine Learning Algorithms," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 2, pp. 355–368, Jun. 2021, doi: 10.29244/ijsa.v5i2p355-368.

[17] P. McSharry and J. Mawejje, "Estimating urban GDP growth using nighttime lights and machine learning techniques in data poor environments: The case of South Sudan," *Technol. Forecast. Soc. Change*, vol. 203, p. 123399, 2024, doi: https://doi.org/10.1016/j.techfore.2024.123399.

[18] H. Zhang *et al.*, "Developing an annual global Sub-National scale economic data from 1992 to 2021 using nighttime lights and deep learning," *International Journal of Applied Earth Observation and Geoinformation*, vol. 133, p. 104086, 2024, doi: https://doi.org/10.1016/j.jag.2024.104086.

[19] M. Amani *et al.*, "Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5326–5350, 2020, doi: 10.1109/JSTARS.2020.3021052.

[20] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning*. Cham: Springer International Publishing, 2023. doi: 10.1007/978-3-031-38747-0.

[21] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY: Springer US, 2021. doi: 10.1007/978-1-0716-1418-1.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

[23] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA: O'Reilly Media, 2022. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/

[24] C. Molnar, *Interpretable Machine Learning : A Guide for Making Black Box Models Explainable*. Lean Publishing, 2019.

NOMENCLATURE

y	Observed regional economic growth
y	Predicted regional economic growth
$\bar{y}$	Mean of the observed regional economic growth
x_i	Predictor variable
$\phi(x)$	Nonlinear feature mapping function
w	Weight vector in Support Vector Regression
b	Bias (intercept) term
β_0	Regression intercept
β_i	Regression coefficient
ε	Random error term
γ	Kernel coefficient in the RBF kernel
$T_b(x)$	Prediction from the (b)-th decision tree
B	Total number of decision trees
k	Number of nearest neighbors
n	Number of observations
$L(\cdot)$	Loss function
CV_{LOO}	Leave-One-Out Cross Validation error
E	Original prediction error
E_j^{perm}	Prediction error after permuting predictor (j)

AUTHOR(S) BIOGRAPHY

Dian Septiana
Dian Septiana is a Lecturer in the Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, Indonesia. She received her M.Sc. in Mathematics from Universitas Gadjah Mada, Indonesia. Her research interests include computational statistics, data science, machine learning, econometrics, and spatial data analysis. Her research focuses on developing statistical and machine learning methods for regional socioeconomic analysis and decision support.

Sisti Nadia Amalia
Sisti Nadia Amalia is a lecturer in the Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, Indonesia. She earned her Master of Statistics (M.Stat.) degree from Universitas Padjadjaran, Indonesia. Her academic and research interests include applied statistics, multivariate analysis, time series analysis, statistical modeling, machine learning, and data analysis. She has published and collaborated on research involving statistical methods, spatial analysis, machine learning, environmental and socioeconomic data, and statistical applications in education. Her recent research focuses on the application of multivariate statistical methods, machine learning, and spatial approaches to address regional and socioeconomic issues.

Fahmi Ashari S Sihaloho

Fahmi Ashari Sihaloho is a Lecturer in the Economics Education Study Program, Faculty of Economics, Universitas Negeri Medan, Indonesia. He earned his Doctoral degree in Economics Education from Universitas Negeri Malang, Indonesia. His research interests include economics education, behavioral economics, financial literacy, educational management, and economic policy. His research focuses on applying quantitative methods and systematic literature reviews to educational, economic, and public policy issues, particularly in the areas of regional development and sustainable economic growth

.