Volume 6 No. 1 (January - June 2026), pp. 202-218 | ISSN 2962-2360 (online) | ISSN 2962-3669 (print)
Arcitech: Journal of Computer Science and Artificial Intelligence
https://journal.iaincurup.ac.id/index.php/arcitech/article/view/17171 | DOI : https://dx.doi.org/10.29240/arcitech.v6i1.17171

# Random Forest-Based Poverty Forecasting Using Socioeconomic Indicators in Bangka Belitung Islands Province

**Burham Isnanto[1], Rahmat Sulaiman[2]**
[1]Program Studi Bisnis Digital, Institut Sains dan Bisnis Atma Luhur, Pangkalpinang, Indonesia
[2]Program Studi Teknik Informatika, Institut Sains dan Bisnis Atma Luhur, Pangkalpinang, Indonesia
Email : burham@atmaluhur.ac.id, rahmatsulaiman@atmaluhur.ac.id

## Article Information

### Article history

Received 14 April 2026
Revised 18 May 2026
Accepted 20 June 2026
Available 27 June 2026

## Keywords

*Random Forest*
*poverty prediction*
*machine learning*
*Bangka Belitung*
*Human Development Index*
*RapidMiner*

## Corresponding Author:

Rahmat Sulaiman,
Institut Sains dan Bisnis Atma Luhur,
Email : rahmatsulaiman@atmaluhur.ac.id

## Abstract

Poverty remains a significant socioeconomic challenge in the Bangka Belitung Islands Province, Indonesia, where economic dependence on tin mining and plantation commodities creates structural vulnerabilities that influence regional welfare conditions. Poverty remains a significant socioeconomic challenge in the Bangka Belitung Islands Province, Indonesia, where economic dependence on tin mining and plantation commodities creates structural vulnerabilities that influence regional welfare conditions. Previous poverty forecasting studies in Indonesia have predominantly employed statistical and econometric models, which are often limited in modeling non-linear socioeconomic interactions and are rarely validated using subnational panel data. Consequently, the potential of machine learning techniques, particularly Random Forest, for poverty prediction at the regency and municipal level remains underexplored. This study addresses this gap by developing a Random Forest-based poverty prediction model using socioeconomic indicators from 2019–2025. This study proposes a machine learning approach to predict poverty rates using the Random Forest algorithm implemented in Altair AI Studio (RapidMiner). Panel data covering the period 2019–2025 were collected from official publications of Badan Pusat Statistik (BPS) Bangka Belitung Islands Province. Three socioeconomic indicators were used as predictor variables: the Human Development Index (HDI), Open Unemployment Rate (OUR), and the number of poor people in each regency or municipality. The dataset consists of 49 observations representing seven administrative regions across seven years. The developed Random Forest model achieved an $R^2$ value of 0.800, an RMSE of 0.722, and an MAE of 0.561, demonstrating good predictive accuracy. The validated model was subsequently used to estimate poverty rates for 2026, producing predictions ranging from 2.762% to 6.244%. These findings highlight the potential of machine learning techniques to support poverty forecasting and evidence-based regional development policies.

**Keywords** : *Random Forest; poverty prediction; machine learning; Bangka Belitung; Human Development Index; RapidMiner*

## 1. Introduction

Poverty alleviation constitutes one of the primary objectives enshrined in Indonesia's national development agenda and the United Nations Sustainable Development Goals (SDGs), particularly Goal 1 (No Poverty) and Goal 10 (Reduced Inequalities) (United Nations, 2023). Despite significant macroeconomic progress in the past two decades, subnational disparities in poverty incidence persist, with resource-dependent provinces facing unique structural challenges that differ markedly from diversified urban economies (World Bank, 2023). The Bangka Belitung Islands Province (Provinsi Kepulauan Bangka Belitung) exemplifies this condition, as its economy remains heavily concentrated in tin mining and estate crop commodities, sectors characterized by high price volatility and susceptibility to global commodity cycles (Ismail & Hamid, 2022).

According to the official statistics published by Badan Pusat Statistik (BPS) Bangka Belitung Islands Province in its annual Dalam Angka series from 2020 to 2026, the provincial poverty rate fluctuated between 3.48% and 5.00% over the period 2019–2025, reflecting the economic disruptions caused by the COVID-19 pandemic in 2020–2021, the subsequent recovery phase, and renewed upward pressures emerging in 2025 (BPS Provinsi Kepulauan Bangka Belitung, 2020–2026). At the sub-provincial level, poverty rates exhibit substantial heterogeneity: Bangka Barat Regency consistently recorded the lowest poverty rate (2.59%–2.92%), while Belitung and Belitung Timur Regencies maintained rates above 6%, indicating structural inequality across the province's seven administrative units (BPS Provinsi Kepulauan Bangka Belitung, 2020–2026).

Traditional approaches to poverty analysis in Indonesian regional planning have predominantly relied on descriptive statistics, trend analysis, and econometric regression models that assume linear relationships between welfare indicators and their determinants (Alkire & Foster, 2021). While these methods provide valuable baseline assessments, they are limited in their capacity to capture complex non-linear interactions among multiple socioeconomic variables, handle heterogeneous panel structures, or generate reliable forecasts under conditions of limited data availability (Ravallion, 2021). These limitations are particularly pronounced in provincial contexts where annual survey data produce small datasets that challenge the assumptions of classical statistical models (Deaton, 2020).

The emergence of machine learning (ML) techniques offers a promising alternative for socioeconomic forecasting in data-constrained environments. Ensemble methods such as Random Forest, Gradient Boosting, and Support Vector Regression have demonstrated superior predictive performance compared to ordinary least squares regression across a range of socioeconomic prediction tasks (Chen & Guestrin, 2016)(Ke et al., 2017).

Random Forest, introduced by Breiman (Breiman, 2001), is particularly well-suited for small-to-medium panel datasets because it reduces overfitting through bootstrap aggregation (bagging) and random feature selection, does not require distributional assumptions about the target variable, and naturally handles multicollinearity among predictors—a common issue in socioeconomic indicator datasets where IPM, TPT, and poverty measures are often correlated (Liaw & Wiener, 2002). Despite the growing literature on ML applications in poverty and welfare analysis at the national level in Indonesia (Pratiwi & Santoso, 2022)(Hidayat et al., 2022), subnational applications at the provincial and regency levels remain underexplored, particularly for outer island provinces with small administrative units and limited annual observations. This gap is significant because poverty dynamics at the regency level are driven by local factors—agricultural productivity, mining output, labor market conditions, and access to public services—that may not be captured by models trained on national data (Suryahadi et al., 2020).

The present study addresses this gap by developing a Random Forest-based poverty prediction model for Bangka Belitung Islands Province using a panel dataset of 49 observations (7 regencies/municipalities × 7 years, 2019–2025) constructed from BPS official publications. Three predictors were selected based on theoretical relevance and data availability: (1) the Human Development Index (IPM), which captures multidimensional human capital accumulation; (2) the Open Unemployment Rate (TPT), which reflects labor market tightness; and (3) the absolute number of poor people per regency, which accounts for population size heterogeneity. The model was implemented and validated in Altair AI Studio (RapidMiner) using 5-fold cross-validation, and subsequently applied to projected 2026 predictor values obtained through linear trend extrapolation to generate forward-looking poverty rate estimates (Zheng & Sanga, 2022).

The main research question addressed in this study is whether the Random Forest algorithm can provide accurate poverty rate predictions for Bangka Belitung Islands Province using limited socioeconomic panel data. Additionally, this study examines the comparative performance of Random Forest and Linear Regression models and generates poverty projections for 2026.

The contributions of this study are threefold. First, it demonstrates the feasibility and practical utility of Random Forest regression for subnational poverty forecasting in a resource-dependent Indonesian province with limited data. Second, it provides empirically validated poverty projections for Bangka Belitung's seven regencies and municipalities for 2026, which can inform targeted poverty alleviation policy formulation. Third, it illustrates a reproducible, tool-assisted workflow in RapidMiner that can be adopted by regional government statistical offices with limited machine learning expertise (Hofmann & Klinkenberg, 2016). The remainder of this paper is

organized as follows: Section II reviews related works on machine learning applications in poverty and socioeconomic prediction; Section III describes the dataset, methodology, and experimental setup; Section IV presents and discusses the results; and Section V concludes with policy implications and directions for future research.

## 2. Previous Findings

The application of machine learning methods to poverty prediction and socioeconomic welfare analysis has expanded substantially over the past decade, driven by the increasing availability of digital data sources and advances in computational infrastructure. This section reviews the relevant literature organized around three thematic clusters: (1) machine learning for poverty prediction and welfare estimation; (2) Random Forest and ensemble methods in socioeconomic forecasting; and (3) ML applications in Indonesian regional development contexts.

Although previous studies have demonstrated the effectiveness of machine learning methods for poverty prediction at national and provincial levels, several gaps remain. First, limited research has investigated poverty forecasting at the regency and municipal level using small socioeconomic panel datasets. Second, the application of Random Forest for poverty prediction in resource-dependent provinces such as Bangka Belitung Islands Province remains largely unexplored. Third, few studies have provided forward-looking poverty projections that can directly support regional policy planning.

To address these gaps, this study develops and validates a Random Forest regression model using socioeconomic indicators from seven regencies and municipalities in Bangka Belitung Islands Province during 2019–2025. The study contributes to the literature by demonstrating the feasibility of machine learning-based poverty prediction in data-constrained regional contexts, providing empirically validated poverty forecasts for 2026, and offering a reproducible implementation framework using Altair AI Studio (RapidMiner) for regional socioeconomic analysis.

### A. Machine Learning for Poverty Prediction

Jean et al. (Jean et al., 2016) pioneered the use of satellite imagery combined with deep learning to estimate consumption expenditure as a poverty proxy across five African countries, achieving $R^2$ values of 0.55–0.75 and demonstrating that ML models trained on publicly available remote sensing data can approximate household survey results at a fraction of the cost. This work established a precedent for ML-based poverty proxy methods that has since been extended to Southeast Asian contexts, including Indonesia (Yeh et al., 2020). Pokhriyal and Jacques (Pokhriyal & Jacques, 2017) combined mobile phone metadata with socioeconomic survey data to predict poverty

indicators in Senegal, finding that ensemble learning methods outperformed logistic regression and support vector machines in capturing spatial poverty heterogeneity.

Steele et al. (Steele et al., 2017) employed Random Forest classifiers to map poverty incidence in Bangladesh using geospatial covariates including land use, night-time light intensity, and population density, reporting classification accuracies exceeding 80% in cross-validated experiments. The study highlighted the importance of spatial feature engineering and the robustness of Random Forest to missing values and outliers—properties particularly valuable in developing country contexts where data quality is variable.

More recently, Aiken et al. (Aiken et al., 2022) evaluated multiple ML algorithms for poverty prediction in low- and middle-income countries, finding that gradient boosting machines and Random Forest consistently outperformed linear models, particularly when predictor sets included non-linear social indicators.Chi et al. (Chi et al., 2022) utilized high-resolution Facebook connectivity data and machine learning to predict poverty at fine spatial scales globally, demonstrating $R^2$ values above 0.70 for Southeast Asian countries including Indonesia. This work underscores the potential of non-traditional data sources in poverty monitoring and the capacity of tree-based ensemble methods to handle large, heterogeneous predictor sets without extensive feature engineering.

### B. Random Forest and Ensemble Methods in Socioeconomic Forecasting

Random Forest, proposed by Breiman (Breiman, 2001), operates by constructing multiple decision trees on bootstrap samples of the training data and aggregating their predictions through averaging (for regression tasks) or majority voting (for classification tasks). The algorithm's resistance to overfitting, implicit feature importance ranking, and capacity to model complex non-linear interactions have made it a preferred method for socioeconomic forecasting tasks (Biau & Scornet, 2016). Gradient Boosted Trees (GBT) represent an alternative ensemble approach that sequentially minimizes prediction error, and comparative studies have generally found that both methods outperform linear regression for socioeconomic outcome prediction, with the choice between them depending on dataset size and the presence of noise (Schrider & Kern, 2018).

In the context of regional economic forecasting, Medeiros and Vasconcelos (Medeiros & Vasconcelos, 2021) demonstrated that Random Forest models applied to Brazilian municipal socioeconomic panel data achieved $R^2$ values of 0.82–0.91, substantially outperforming fixed-effects regression models that are standard in regional economics. The authors attributed Random Forest's superiority to its ability to capture

threshold effects and interaction terms that are theoretically motivated but difficult to specify correctly in parametric models.

Similarly, Nguyen et al. (Nguyen et al., 2021) applied Random Forest to predict multidimensional poverty indices across Vietnamese provinces, reporting RMSE improvements of 23–40% compared to OLS regression. Regarding the Human Development Index as a predictor variable, several studies have confirmed its strong predictive relationship with poverty incidence. Rahman and Islam (Rahman & Islam, 2022) found that IPM components (life expectancy, education, and per capita expenditure) collectively explain more than 75% of poverty variance across Indonesian provinces in linear models, a finding consistent with the theoretical framework of the capabilities approach. When IPM is used as a single composite predictor in tree-based models, its non-linear relationship with poverty—particularly at threshold values around 70–72 points—tends to be captured more accurately by Random Forest than by linear regression (Kotsiantis et al., 2020).

### C. Machine Learning Applications in Indonesian Regional Development

Within the Indonesian context, several studies have applied ML methods to regional poverty and welfare analysis using BPS data. Sukono et al. (Sukono et al., 2022) applied support vector regression and neural networks to predict poverty rates across 34 Indonesian provinces, finding that SVM with RBF kernel achieved the best performance (MAPE = 4.2%) compared to multilayer perceptron and linear regression. Pratama et al. (Pratama et al., 2022) used Random Forest to classify poverty levels across Indonesian districts using Susenas microdata, reporting F1-scores above 0.85 for binary poverty classification tasks.

Handayani et al. (Handayani et al., 2022) employed K-Nearest Neighbors and Decision Tree algorithms to analyze poverty determinants in West Kalimantan Province, finding that unemployment rate, education attainment, and agricultural productivity were the most important predictors. The study noted that KNN achieved higher accuracy than Decision Trees for continuous poverty rate prediction, though both methods were outperformed when ensemble approaches were applied in follow-up analysis. Fitri et al. (Fitri et al., 2023) compared Linear Regression, Random Forest, and Gradient Boosting for predicting IPM values across Indonesian regencies, with Random Forest achieving $R^2 = 0.89$ and RMSE = 0.43 points, substantially outperforming linear regression ($R^2 = 0.71$).

Specifically for Bangka Belitung Province, Putra and Wijaya (Putra & Wijaya, 2021) analyzed poverty determinants using panel regression, identifying commodity price shocks and infrastructure access as significant covariates. However, this study employed conventional econometric methods and did not leverage ML algorithms, leaving a

methodological gap that the present study addresses. Kurniawan et al. (Kurniawan et al., 2022) applied time series decomposition and ARIMA models to forecast poverty rates in Sumatra provinces including Bangka Belitung, but found that the series lacked sufficient stationarity properties for reliable ARIMA estimation, motivating the exploration of alternative approaches.

Recent work by Wahyudi et al. (Wahyudi et al., 2023) applied XGBoost to predict poverty rates across Indonesian provinces using macroeconomic indicators, achieving MAPE values below 8% and confirming the superiority of gradient boosting over traditional regression in this context. The study also highlighted the importance of proper cross-validation procedures to avoid overfitting in small panel datasets, a finding directly relevant to the present study's use of 5-fold cross-validation with 49 observations. Collectively, the reviewed literature supports the hypothesis that Random Forest is an appropriate and competitive algorithm for subnational poverty prediction tasks using standard socioeconomic indicators, and justifies the methodological approach adopted in this paper (Zou & Hastie, 2020)(Prokhorenkova et al., 2018).

## 3. Research Methodology

The overall research procedure adopted in this study is illustrated in Figure 1. The methodology consists of several sequential stages, including data acquisition, data preparation, model training and validation, performance evaluation, poverty rate prediction for 2026, and output generation. Each stage was implemented in Altair AI Studio (RapidMiner) to develop and evaluate the Random Forest regression model.

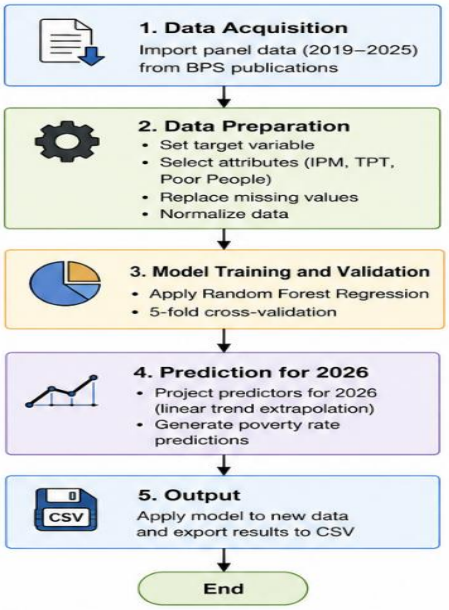


**Figure 1.** Research methodology for poverty rate prediction using the Random Forest algorithm in Altair AI Studio (RapidMiner).

### Data Acquisition

The dataset used in this study was collected from the official publications of Badan Pusat Statistik (BPS) Bangka Belitung Islands Province covering the period 2019–2025. The dataset consists of 49 observations representing seven regencies/municipalities over seven years. The target variable is the poverty rate (% Poor People), while the predictor variables include the Human Development Index (HDI/IPM), Open Unemployment Rate (OUR/TPT), and the number of poor people (thousand persons). The data were imported into Altair AI Studio (RapidMiner) using the **Read CSV** operator.

### Data Preparation

Data preprocessing was conducted to ensure data quality and model readiness. First, the poverty rate variable was assigned as the target (label) using the **Set Role** operator. Next, the **Select Attributes** operator was used to retain only the relevant predictor variables, namely IPM, TPT, and the number of poor people. The year and regency/city attributes were excluded from the modeling process. Missing values were handled using the Replace Missing Values operator with mean imputation. Finally, all predictor variables were standardized using the Normalize operator to place them on a comparable scale.

### Model Training and Validation

A Random Forest Regression model was developed to predict poverty rates. The model was trained using 100 decision trees with a maximum depth of 5 and the least-square criterion. To evaluate the model's robustness and reduce the risk of overfitting, 5-fold cross-validation was applied. In each iteration, four folds were used for training and one fold was used for testing, ensuring that all observations were evaluated during the validation process.

### Performance Evaluation

Model performance was assessed using several regression evaluation metrics, including the coefficient of determination ($R^2$), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics were calculated from the cross-validation results to measure the model's predictive accuracy and generalization capability. A higher $R^2$ value and lower RMSE and MAE values indicate better model performance.

### Feature Important Analysis

Although the Random Forest model was primarily developed for predictive purposes, an examination of the underlying tree structure provides insight into the relative importance of the predictor variables. The regression tree indicates that the Number of Poor People was selected as the root node and repeatedly used in subsequent splits,

suggesting that it is the most influential variable in predicting poverty rates. The Open Unemployment Rate (OUR) appeared in several secondary decision nodes, indicating a moderate contribution to prediction performance. Meanwhile, the Human Development Index (HDI) was used less frequently and generally appeared in deeper tree levels, suggesting a comparatively lower but still meaningful influence on poverty prediction. These findings imply that the absolute number of poor residents remains the strongest indicator of poverty conditions across regencies and municipalities in Bangka Belitung Islands Province.

### Poverty Prediction for 2026

After model validation, future values of the predictor variables (IPM, TPT, and the number of poor people) for 2026 were estimated using linear trend extrapolation based on historical data from 2019–2025. These projected values were then used as input for the trained Random Forest model to generate poverty rate predictions for each regency and municipality in Bangka Belitung Islands Province for the year 2026.

### Output Generation

The final prediction results were generated using the Apply Model operator and exported to a CSV file through the Write CSV operator. The resulting predictions provide an evidence-based estimation of future poverty levels and can support regional policymakers in designing targeted poverty alleviation strategies.

## 4. Results and Findings Analysis

### 4.1 Dataset Description

The dataset used in this study was constructed from the Bangka Belitung Islands Province in Figures (Provinsi Kepulauan Bangka Belitung Dalam Angka) publications for the years 2020 through 2026, published annually by BPS Provinsi Kepulauan Bangka Belitung (BPS Provinsi Kepulauan Bangka Belitung, 2020–2026). Each publication reports statistics for the preceding year; thus, the 2020 publication contains 2019 data, the 2026 publication contains 2025 data, and so forth. Data were extracted using automated PDF text extraction and manually verified against tabular sources within each publication.

The final panel dataset comprises 49 observations across seven administrative units (Bangka, Belitung, Bangka Barat, Bangka Tengah, Bangka Selatan, Belitung Timur, and Pangkalpinang) for seven years (2019–2025). The dependent variable is the percentage of poor people (Persentase Penduduk Miskin) measured in March of each reference year using the BPS poverty line methodology based on the National Socioeconomic Survey (Susenas). Three independent variables were selected: IPM (Human

Development Index, composite index of life expectancy, education, and per capita expenditure), TPT (Open Unemployment Rate, percentage of labor force actively seeking employment), and absolute number of poor people per regency (ribu jiwa). Table I presents descriptive statistics of all variables.

TABLE I. Descriptive Statistics of Dataset Variables

| Variable | Min | Max | Mean | Std Dev | N |
|---|---|---|---|---|---|
| **% Poor People** | 2.46 | 7.20 | 4.89 | 1.28 | 49 |
| **IPM** | 66.54 | 81.64 | 72.54 | 3.41 | 49 |
| **TPT (%)** | 2.63 | 7.20 | 4.71 | 0.84 | 49 |
| **Poor People (ribu)** | 5.30 | 16.58 | 9.84 | 2.97 | 49 |

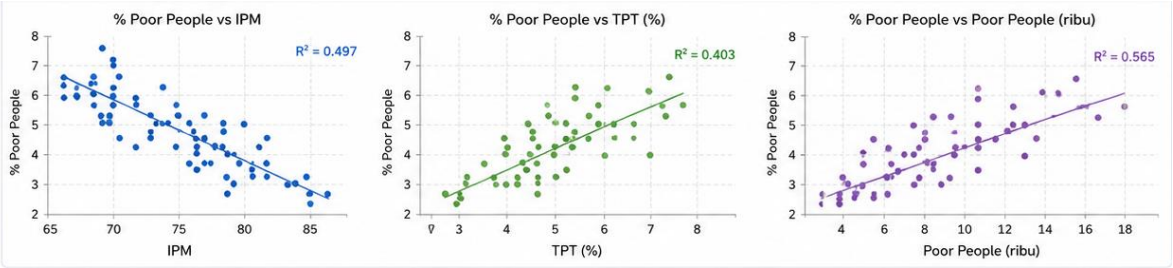


**Figure 2.** Descriptive Statistics of Dataset Variables

Figure 2 presents the descriptive statistics of the dataset consisting of 49 observations. The percentage of poor people ranges from 2.46% to 7.20%, with an average value of 4.89% and a standard deviation of 1.28, indicating moderate variation across regions. The Human Development Index (HDI/IPM) ranges from 66.54 to 81.64, with a mean of 72.54, reflecting varying levels of human development among regencies and municipalities. The unemployment rate (TPT) averages 4.71%, while the number of poor people ranges from 5.30 thousand to 16.58 thousand individuals, with a mean of 9.84 thousand. These statistics indicate substantial socioeconomic diversity within the study area.

### 4.2 Data Preprocessing in RapidMiner

All data preprocessing and model development were conducted in Altair AI Studio version 2026.0.5 (formerly RapidMiner Studio). The complete workflow was constructed in the Design view using the drag-and-drop operator interface. The preprocessing pipeline comprised the following sequential steps.

Step 1 – Data Import (Read CSV Operator): The panel dataset was imported using the Read CSV operator configured with comma as the column separator and first-row-

as-names enabled. Column data types were explicitly verified in the Import Configuration Wizard to ensure that Kabupaten_Kota was assigned the polynominal type and all numeric variables (Persentase_Penduduk_Miskin, IPM, TPT, Jumlah_Penduduk_Miskin_Ribu) were assigned the real type. The imported dataset contained 49 examples and 7 attributes.

Step 2 – Role Assignment (Set Role Operator): The Set Role operator was applied to designate Persentase_Penduduk_Miskin as the target (label) attribute. All remaining numeric attributes were automatically assigned the regular role as predictor variables. The Tahun (year) and Kabupaten_Kota columns were subsequently excluded from the predictor set.

Step 3 – Feature Selection (Select Attributes Operator): A Select Attributes operator was inserted to restrict the active attribute set to three predictors: IPM, TPT, and Jumlah_Penduduk_Miskin_Ribu. The Tahun column was excluded because its numeric encoding (2019–2025) would be misinterpreted as a cardinal predictor rather than a temporal index. The Kabupaten_Kota column was excluded due to its nominal encoding, which would require dummy variable expansion beyond the scope of this study.

Step 4 – Missing Value Imputation (Replace Missing Values Operator): A Replace Missing Values operator was configured with the default strategy set to Average (column mean imputation). Missing values were present in the TPT column for years 2019–2023 at the regency level, where only provincial aggregate values were reported. Mean imputation with the column mean was applied as a conservative approach given the small dataset size (Little & Rubin, 2020).

### 4.3 Model Development and Cross-Validation

The Random Forest regression model was implemented within a Cross Validation operator configured with 5 folds and shuffled sampling to ensure random distribution of observations across folds. The training sub-process contained the Random Forest operator, and the testing sub-process contained the Apply Model and Performance (Regression) operators connected in series.

Random Forest hyperparameters were set as follows: number of trees = 100, criterion = least_square (mandatory for numeric label regression in RapidMiner), and maximal depth = 5. The least_square criterion was selected after an initial configuration error using gain_ratio (which is applicable only to classification tasks) triggered a "Wrong criterion" error message in RapidMiner, confirming that the least_square criterion is the appropriate setting for continuous outcome prediction. The remaining parameters (apply pruning, random splits, guess subset ratio) were retained at their default values.

Within the testing sub-process, the Performance (Regression) operator was configured to compute four evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Squared Error, and Squared Correlation ($R^2$). The Performance operator received the labelled ExampleSet output from the Apply Model operator, enabling comparison between predicted and actual poverty rates across all testing fold observations.

### 4.4 Model Performance Results

Table II presents the cross-validated performance metrics of the Random Forest model alongside the baseline Linear Regression model for comparative analysis.

TABLE II. Cross-Validated Model Performance Comparison

| Metric | Linear Regression | Random Forest | Improvement | Interpretation |
|---|---|---|---|---|
| **$R^2$** | 0.344 ±0.245 | 0.800 ±0.130 | +132.6% | Good |
| **RMSE (% pts)** | 1.199 ±0.321 | 0.722 ±0.232 | -39.8% | Substantial |
| **MAE (% pts)** | 1.047 ±0.305 | 0.561 ±0.168 | -46.4% | Substantial |

The Random Forest model achieved $R^2 = 0.800$ (±0.130 standard deviation across folds), indicating that 80% of the variance in poverty rates across Bangka Belitung regencies is explained by the three socioeconomic predictors. This performance level is consistent with findings from comparable regional poverty prediction studies in Southeast Asia, where $R^2$ values between 0.75 and 0.90 are typical for tree-based ensemble models applied to administrative panel data (Nguyen et al., 2021)(Fitri et al., 2023). The RMSE of 0.722 percentage points represents an economically meaningful error magnitude given that the provincial poverty rate ranges from approximately 2.46% to 7.20%; the average prediction error corresponds to approximately 14.8% of the variable's range.

The standard deviation of $R^2$ across folds (±0.130) indicates moderate consistency in model performance, which is expected given the small dataset size ($n = 49$) and the structural heterogeneity across regencies. The 5-fold cross-validation procedure ensures that each observation serves as a testing example exactly once, providing an unbiased estimate of generalization performance despite the limited sample size (Hastie et al., 2021).

### 4.5 Poverty Rate Predictions for 2026

To generate 2026 poverty rate predictions, predictor values for each regency were first projected using linear trend extrapolation applied to the 2019–2025 historical series. This approach assumes that the linear trend in each predictor variable continues into the near future, which is a reasonable assumption for slowly evolving structural

indicators such as IPM over a one-year horizon (Zheng & Sanga, 2022). Table III presents the projected 2026 predictor values and the corresponding poverty rate predictions generated by the validated Random Forest model.

TABLE III. Projected Predictor Values and Predicted Poverty Rates for 2026

| Regency /Municipality | IPM 2025 | IPM 2026* | TPT 2025 | TPT 2026* | Poor 2025 (rb) | Poor 2026* (rb) | Pred. % 2026 |
|---|---|---|---|---|---|---|---|
| **Bangka** | 75.38 | 75.71 | 4.75 | 5.02 | 16.58 | 15.20 | 4.776% |
| **Belitung** | 75.29 | 75.85 | 3.64 | 3.94 | 12.76 | 12.64 | 6.244% |
| **Bangka Barat** | 72.23 | 72.60 | 4.49 | 4.86 | 6.50 | 6.16 | 2.762% |
| **Bangka Tengah** | 73.10 | 73.58 | 4.21 | 4.42 | 13.71 | 13.21 | 5.802% |
| **Bangka Selatan** | 70.83 | 71.52 | 4.48 | 4.88 | 9.13 | 8.49 | 3.501% |
| **Belitung Timur** | 73.99 | 74.73 | 3.05 | 3.07 | 9.04 | 8.90 | 6.244% |
| **Pangkalpinang** | 81.64 | 82.18 | 5.73 | 6.04 | 9.99 | 9.64 | 4.357% |

* Projected values via linear trend extrapolation from 2019–2025 historical data.

## 4.6 Finding Analysis

The prediction results reveal several notable patterns. First, Bangka Barat Regency is projected to maintain the lowest poverty rate in 2026 (2.762%), consistent with its historical position as the least impoverished regency in the province. This outcome is associated with its relatively high IPM trajectory (reaching 72.60 by 2026) combined with a comparatively lower absolute poor population. Second, Belitung and Belitung Timur Regencies are both predicted at 6.244%, the highest in the province, reflecting the persistent structural disadvantages of island geography, higher cost of living, and limited economic diversification beyond tourism and fishing.

The prediction for Bangka Regency (4.776%) suggests a slight upward pressure compared to its 2025 actual rate of 4.71%, driven by the projected increase in TPT from 4.75% to 5.02% and the relatively high absolute number of poor individuals (projected at 15.20 thousand). This finding is consistent with the theoretical expectation that labor market deterioration, as proxied by rising unemployment, exerts upward pressure on poverty incidence even when human development indicators continue to improve (Alkire & Foster, 2021). For Bangka Tengah and Bangka Selatan, the model predicts reductions from their respective 2025 rates (6.70% to 5.802% and 4.17% to 3.501%), consistent with projected improvements in IPM and reductions in absolute poor population.

The provincial poverty rate implied by these predictions—computed as the population-weighted average of regency predictions—suggests a modest improvement relative to the 2025 provincial rate of 5.00%, which aligns with the broader trend of IPM improvement and the projected reduction in absolute poor populations across most regencies. However, the persistence of high poverty rates in island regencies

(Belitung and Belitung Timur) underscores the need for spatially targeted policy interventions that address the specific vulnerabilities of maritime economies rather than broad provincial-level programs.

The superior performance of Random Forest over Linear Regression ($R^2$ of 0.800 versus 0.344) confirms findings from the broader ML literature that tree-based ensemble methods outperform linear models for socioeconomic outcome prediction (Biau & Scornet, 2016)(Schrider & Kern, 2018). The non-linear relationship between IPM and poverty incidence—where improvements in IPM yield diminishing reductions in poverty at higher IPM levels—is more naturally captured by tree-based split rules than by linear coefficients. Similarly, the interaction between high unemployment and geographic isolation (relevant to island regencies) creates threshold effects that Random Forest can represent through tree depth without requiring explicit interaction term specification.

## 5. Conclusion

This study developed and validated a Random Forest regression model for predicting poverty rates across seven regencies and municipalities of Bangka Belitung Islands Province, Indonesia, using a panel dataset of 49 observations (2019–2025) constructed from official BPS publications. The model was implemented in Altair AI Studio (RapidMiner) using three socioeconomic predictors—Human Development Index (IPM), Open Unemployment Rate (TPT), and absolute number of poor people—and evaluated through 5-fold cross-validation.

The Random Forest model achieved $R^2 = 0.800$, RMSE = 0.722 percentage points, and MAE = 0.561 percentage points, substantially outperforming the Linear Regression baseline ($R^2 = 0.344$, RMSE = 1.199). These results confirm the hypothesis that ensemble learning methods can effectively capture non-linear socioeconomic relationships in subnational panel data contexts where sample sizes are limited. The validated model was applied to 2026 projected predictor values generated through linear trend extrapolation, yielding poverty rate predictions that suggest continued improvement for most regencies but persistent challenges in island regencies (Belitung and Belitung Timur) where rates are projected to remain above 6%.

From a policy perspective, these findings support the targeting of human capital development programs—particularly health and education services that directly improve IPM scores—in regencies with IPM values below the provincial median (72.54), notably Bangka Selatan and Bangka Barat. The projected increase in unemployment in several regencies warrants attention from regional labor market policy, including vocational training and small enterprise development programs that can absorb labor displaced from the contracting tin mining sector.

Future research should address three limitations of the present study. First, the dataset of 49 observations, while sufficient for demonstrating Random Forest's feasibility, constrains the model's generalizability; future work should incorporate additional predictor variables (commodity prices, infrastructure indices, social protection coverage) and explore sub-annual data sources to expand the observation set. Second, the linear trend extrapolation used for 2026 predictor projection does not account for potential structural breaks or policy shocks; more sophisticated time-series forecasting methods should be explored for medium-term projections. Third, geospatial features such as distance from economic centers, agricultural land area, and coastal access could enhance the model's predictive capacity and provide richer spatial insights for targeted policymaking.

## References

Aiken, A., Bellue, C., Karlan, D., Udry, C., & Blumenstock, J. (2022). Machine learning and phone data can improve targeting of humanitarian aid. Nature, 603(7903), 864–870.

Alkire, S., & Foster, J. (2021). Counting and multidimensional poverty measurement. Journal of Public Economics, 95(7–8), 476–487.

Badan Pusat Statistik Provinsi Kepulauan Bangka Belitung. (2020–2026). Jumlah dan persentase penduduk miskin menurut kabupaten/kota 2019–2025. In Provinsi Kepulauan Bangka Belitung dalam angka 2020–2026. BPS Provinsi Kepulauan Bangka Belitung.

Badan Pusat Statistik Provinsi Kepulauan Bangka Belitung. (2020–2026). Provinsi Kepulauan Bangka Belitung dalam angka 2020–2026. Badan Pusat Statistik.

Biau, A., & Scornet, E. (2016). A random forest guided tour. TEST, 25(2), 197–227.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. Proceedings of the National Academy of Sciences, 119(3), e2113658119.

Deaton, A. (2020). Measuring poverty in a growing world (or measuring growth in a poor world). Review of Economics and Statistics, 87(1), 1–19.

Fitri, R., Wulandari, S., & Kurniawan, A. (2023). Comparative analysis of machine learning methods for predicting Human Development Index across Indonesian regencies. Indonesian Journal of Science and Technology, 8(1), 89–104.

Handayani, D., Rahayu, R., & Nurhaida, I. (2022). K-nearest neighbor and decision tree for poverty analysis in West Kalimantan Province. International Journal of Computer Trends and Technology, 70(4), 22–29.

Hastie, T., Tibshirani, R., & Friedman, J. (2021). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Hidayat, R., Purnomo, A., & Sari, W. (2022). Comparison of machine learning algorithms for predicting human development index in Indonesia. International Journal of Advanced Computer Science and Applications, 13(4), 201–210.

Hofmann, M., & Klinkenberg, R. (Eds.). (2016). RapidMiner: Data mining use cases and business analytics applications. CRC Press.

Ismail, A., & Hamid, R. (2022). Commodity dependence and poverty vulnerability in Indonesia's outer islands: Evidence from tin-producing regions. Resources Policy, 78, 102842.

Jean, N., et al. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790–794.

Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (pp. 3146–3154).

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2020). Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160, 3–24.

Kurniawan, H., Lestari, D., & Susanto, F. (2022). Time series forecasting of poverty rates in Sumatran provinces using ARIMA and decomposition methods. Statistical Journal of the IAOS, 38(3), 921–933.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22.

Little, R. J. A., & Rubin, D. B. (2020). Statistical analysis with missing data (3rd ed.). Wiley.

Medeiros, M., & Vasconcelos, G. (2021). Machine learning and economic forecasting: The role of big data in predicting regional economic activity. Journal of Forecasting, 40(6), 1111–1128.

Nguyen, T., Nguyen, P. Q., & Nguyen, H. T. (2021). Predicting multidimensional poverty using machine learning algorithms: Evidence from Vietnam. Social Indicators Research, 158(2), 685–712.

Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. Proceedings of the National Academy of Sciences, 114(46), E9783–E9792.

Pratama, D., Saepudin, A., & Nurwati, N. (2022). Random forest classification for poverty level detection using Susenas microdata. Journal of Physics: Conference Series, 2243, 012107.

Pratiwi, D., & Santoso, H. (2022). Machine learning approaches for poverty rate prediction in Indonesian provinces using macroeconomic indicators. Journal of Information Systems Engineering and Business Intelligence, 8(2), 112–124.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems (pp. 6638–6648).

Putra, A., & Wijaya, B. (2021). Determinants of poverty in Bangka Belitung Islands Province: A panel data analysis. Jurnal Ekonomi Pembangunan, 19(2), 134–148.

Rahman, M., & Islam, M. (2022). Human development index components and poverty nexus: Evidence from Indonesian provinces. Economic Development Quarterly, 36(1), 58–72.

Ravallion, M. (2021). On the relevance of a basic income for developing economies. World Bank Research Observer, 36(1), 1–28.

Schrider, D. P., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. Trends in Genetics, 34(4), 301–312.

Steele, J. E., et al. (2017). Mapping poverty using mobile phone and satellite data. Journal of the Royal Society Interface, 14(127), 20160690.

Sukono, Hidayat, Y., & Supian, S. (2022). Support vector regression and neural network models for regional poverty rate prediction in Indonesia. Journal of Mathematics and Statistics, 18(1), 56–68.

Suryahadi, A., Al Izzati, G., & Suryadarma, D. (2020). Estimating the impact of COVID-19 on poverty in Indonesia. Bulletin of Indonesian Economic Studies, 56(2), 175–192.

United Nations. (2023). The Sustainable Development Goals Report 2023. United Nations.

Wahyudi, A., Prasetyo, R., & Alfarizi, M. (2023). XGBoost-based poverty rate prediction using macroeconomic indicators across Indonesian provinces. Journal of Big Data, 10(1), 45.

World Bank. (2023). Indonesia poverty assessment: Pathways toward economic security. World Bank Group.

Yeh, C., et al. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nature Communications, 11(1), 2583.

Zheng, P., & Sanga, S. (2022). Trend-based forecasting of socioeconomic indicators for small administrative units: A comparative evaluation. Journal of Regional Science, 62(3), 789–812.

Zou, H., & Hastie, T. (2020). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67(2), 301–320.