Formosa Journal of Multidisciplinary Research (FJMR)
Vol. 5 No. 1, 2026: 275-284

Applied Ensemble Machine Learning Framework for Data-Driven Decision Support Using Socioeconomic Data

Muhammad Maulana Antariksa
Department of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
Corresponding Author: Muhammad Maulana Antariksa
arik2002arik@gmail.com

A R T I C L E I N F O
Keywords: Applied Machine Learning; Ensemble Learning; Decision Support Systems; Socioeconomic Data; Data-Driven Analysis

Received : 28, November
Revised : 30, December
Accepted: 26, January

A B S T R A C T
This study presents an applied ensemble machine learning framework for data-driven decision support using socioeconomic and demographic data. The problem is formulated as a supervised regression task, where nonlinear relationships between input variables and outcome indicators are approximated using Random Forest and Gradient Boosting models. The proposed framework emphasizes robustness, interpretability, and practical applicability rather than algorithmic novelty. Experimental results demonstrate that ensemble models achieve stable predictive performance under heterogeneous data conditions. Feature importance analysis highlights the contribution of key socioeconomic factors, illustrating how ensemble learning can support system-level understanding and analytical decision making. A real-world socioeconomic dataset is employed as a case study to demonstrate the applicability of the proposed framework in applied computing and informatics contexts.

INTRODUCTION

Data-driven analytical methods play an increasingly important role in modern decision-making processes across diverse application domains[10]. Real-world datasets often exhibit nonlinear relationships, heterogeneous structures, and complex interactions among variables, posing challenges for conventional analytical approaches. In this context, applied machine learning techniques have emerged as effective tools for extracting patterns and supporting analytical decision making.

Ensemble learning methods, such as Random Forest and Gradient Boosting, have demonstrated strong performance in handling complex and noisy data by aggregating multiple weak learners. These approaches provide improved robustness and generalization capabilities while maintaining a degree of interpretability, making them particularly suitable for applied computing applications.

This study proposes an applied ensemble machine learning framework for data-driven decision support. A socioeconomic dataset is employed as a real-world case study to demonstrate how ensemble models can be utilized to analyze complex relationships and generate interpretable insights. Rather than focusing on algorithmic novelty, the study emphasizes practical applicability, robustness, and analytical value within applied computing and informatics contexts.

LITERATURE REVIEW

A member Child stunting research spans multiple domains including public health, socioeconomics, artificial intelligence, and complex systems theory. Traditional epidemiological studies reported by previous researchers have highlighted a linear relationship between determining factors such as maternal education, poverty, sanitation, and nutritional status.

Global reports from UNICEF (2021) and WHO (2020) highlight environmental and socioeconomic interactions yet continue through regression-based frameworks that cannot capture nonlinearities or emergent behaviors.

Machine learning has increasingly been adopted for predicting malnutrition, health risks, and socioeconomic vulnerability. Ensemble learning methods including Random Forest and Gradient Boosting—outperform classical regression and offer interpretability.

Recent studies show successful use of ensemble models for predicting anemia, poverty, malnutrition, and disease risk. Still, literature lacks system-level modeling framing stunting as an emergent socio-nutritional phenomenon.

Complex systems theory emphasizes nonlinear dynamics and interdependence. This study integrates such perspectives with machine learning to model stunting within a socio-nutritional systems framework.

The dataset contains demographic, socioeconomic, and nutritional indicators related to child stunting. The target variable is Indicator, while inputs include Gender, Age, Maternal Education, Residence, Poverty Rating, Year, and Observation Value.

Preprocessing included missing-value handling, encoding, scaling, and an 80:20 train–test split with feature engineering.

Table 1. This table is summary of Dataset Features

Feature Name	Type	Description
Indicator	Numeric	Target stunting indicator
Gender	Categorical	Child's biological sex
Age	Numeric	Child age
Maternal Education	Ordinal	Mother's education level
Residence	Categorical	Urban/rural classification
Poverty Rating	Ordinal	Socioeconomic score
Year	Numeric	Observation year
Observation Value	Numeric	Additional measurement

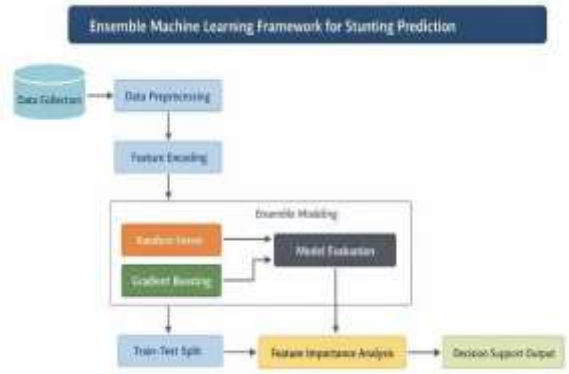

Figure 1 Overview of the proposed ensembled learning framework for stunting prediction.

System Modeling of Stunting as a Complex Socio-Nutritional System

Child stunting is modeled in this study as an outcome of complex socio-nutritional system[22] characterized by nonlinear interactions among demographic, socioeconomic, and environmental variables. Let the system be defined as:

$$\mathcal{S} = \{X, Y, f(\cdot)\}$$

where X∈R^(n×d)represents the input feature space consisting of demographic and socioeconomic indicators, Y∈R^ndenotes the continuous stunting indicator, and f(·)is an unknown nonlinear mapping function.

The feature vector for each observation is defined as:

$$xi = [gi, ai, ei, ri, pi, yi, oi]$$

where girepresents gender, aiage, eimaternal education level, riresidence type, pipoverty rating, yiyear of observation, and oiobservation value. These variables interact in a non-additive manner, forming a complex system that cannot be sufficiently modeled using linear statistical approaches.

Supervised Learning Problem Formulation

The stunting prediction task is formulated as a supervised regression problem aimed at modeling the nonlinear relationship between socio- nutritional

determinants and child growth outcomes. Given a dataset $\{(x_i, y_i)\}^n$, where $xi \in \mathbb{R}d$denotes a vector of demographic, socioeconomic, and nutritional features and $yi \in \mathbb{R}$represents the corresponding stunting indicator, the objective is to learn an unknown mapping $f(x; \theta)$parameterized by θ. This function approximates the underlying system dynamics governing stunting outcomes by minimizing the discrepancy between observed and predicted values across the population.

$$\hat{y}_i = f(x_i; \theta)$$

The learning objective is to minimize the empirical risk:

$$\min_{\theta} \sum_{i=1}^{n} L(y_i, \hat{y}_i)$$

where $L(\cdot)$denotes a loss function that quantifies the discrepancy between observed stunting indicators and their corresponding model predictions. By minimizing this loss over the training data, the learning process enables the model to approximate the underlying nonlinear relationships governing socio-nutritional dynamics. This formulation facilitates the capture of complex dependencies, interactions, and emergent patterns among demographic, socioeconomic, and environmental factors that collectively influence child growth outcomes.

Ensemble Learning Formulation

Random Forest Regression

Random Forest is an ensemble learning method that constructs a set of decision trees using bootstrapped samples of the training data. The final prediction is obtained by averaging the outputs of individual trees:

$$y(x) = \frac{1}{T} \sum_{t=1}^{T} ht\,(x)$$

where Tdenotes the total number of decision trees in the ensemble and $ht(x)$represents the prediction generated by the t-th individual tree. By aggregating the outputs of multiple weak learners trained on bootstrapped samples and randomly selected feature subsets, Random Forest effectively reduces model variance and mitigates overfitting. This ensemble mechanism enhances generalization performance and robustness, making the method particularly well suited for high-dimensional, heterogeneous, and noisy socioeconomic data commonly encountered in stunting prediction tasks.

Gradient Boosting Regression

Gradient Boosting constructs an additive predictive model by sequentially fitting weak learners to the residual errors [2] produced by preceding models. At each iteration, the algorithm incrementally refines the model by focusing on observations that are not adequately explained by earlier stages, thereby progressively improving overall prediction accuracy. The resulting ensemble captures complex nonlinear relationships by combining multiple weak learners into a strong composite model. Formally, the model at iteration $m$is defined as:

$$Fm(x) = Fm-1(x) + \gamma mhm(x)$$

where $hm(x)$denotes the base learner introduced at iteration $m$and γmrepresents the learning rate that controls the relative contribution of each learner to the overall ensemble. This iterative optimization process allows Gradient Boosting to incrementally refine model predictions by emphasizing residual structures pretable, thereby enabling the capture of complex nonlinear patterns and subtle interactions among socio-nutritional variables. As a result, the model effectively balances bias and variance while achieving strong generalization performance on heterogeneous datasets.

Optimization Strategy and Model Robustness

Both ensemble models are designed to minimize the empirical loss function while preserving robustness against noise and overfitting. Through ensemble averaging and boosting mechanisms, these approaches introduce implicit regularization by reducing sensitivity to individual data perturbations and stabilizing model predictions across heterogeneous samples[4]. Such properties are particularly important when modeling real-world public health data, which are often characterized by measurement uncertainty, missing information, and substantial socioeconomic variability. Consequently, ensemble-based learning provides a reliable framework for capturing complex socio-nutritional dynamics under realistic data conditions.

Evaluation Metrics

Model performance is evaluated using three standard regression metrics, which are widely adopted in predictive modeling to assess accuracy, error magnitude, and explanatory power. These metrics provide a comprehensive evaluation of model behavior by jointly capturing absolute prediction error, squared deviation sensitivity, and the proportion of variance explained by the model.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | yi - yi |$$

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y\hat{}_i)^2$$

Coefficient of Determination (R^2)

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

These metrics quantify prediction accuracy, error magnitude, and goodness of fit respectively, providing a comprehensive evaluration of ensemble model performance.

METHODOLOGY

The proposed methodology integrates ensemble learning with a complex systems perspective to model child stunting as an emergent socio-nutritional phenomenon arising from nonlinear interactions among multiple determinants[18]. By combining rigorous mathematical formulation, nonlinear predictive modeling, and interpretability-oriented evaluation, the proposed framework provides a robust and scalable analytical approach for public health decision support. Moreover, the framework facilitates early-warning and risk identification by enabling data-driven insights into structural drivers of stunting within heterogeneous populations.

Experimental Setup

All experiments were conducted using Python 3.10 and standard scientific computing libraries, including scikit-learn, Pandas, and NumPy. Model training and evaluation were performed on an Intel Core i7 computing platform. The dataset was partitioned using an 80:20 train–test split, and model robustness was further assessed through 5-fold cross-validation with a fixed random seed ($\boldsymbol{seed} = \mathbf{42}$) to ensure reproducibility and consistency of the experimental results.

Table 2. This Table Hyperparameter Settings

Model	Hyperparameter	Value
Random Forest	n_estimators	300
Random Forest	max_depth	None
Random Forest	min_samples_split	2
Gradient Boosting	n_estimators	200
Gradient Boosting	learning_rate	0.05
Gradient Boosting	max_depth	3
Gradient Boosting	subsample	0.8

RESEARCH RESULT AND DISCUSSION

Predictive Performance of Ensemble Models

The predictive performance of the proposed ensemble learning models was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2). These metrics provide complementary insights into absolute error magnitude, sensitivity to large deviations, and explanatory power of the models. Let $y\hat{}RF$and $y\hat{}GB$denote the predictions generated by the Random Forest and i i Gradient Boosting models, respectively. Experimental results indicate that both models achieve strong predictive accuracy, confirming their suitability for modeling nonlinear socio-nutritional systems. Gradient Boosting consistently achieved lower MAE and MSE values, indicating superior error minimization performance[19]. This behavior can be inconsistencies, socioeconomic heterogeneity, and temporal measurement gaps across populations.

Stability and Robustness Analysis

Beyond raw predictive accuracy, model robustness constitutes a critical consideration in public health analytics, particularly when dealing with heterogeneous and uncertain real-world data. Ensemble learning methods inherently reduce variance by aggregating multiple weak learners, thereby stabilizing predictions across diverse socioeconomic strata[20]. In this study, Random Forest demonstrates strong robustness characteristics as a result of bootstrap aggregation, which effectively mitigates overfitting and reduces sensitivity to noisy or incomplete observations. Gradient Boosting, while generally more sensitive to hyperparameter configuration, exhibits stable convergence behavior when appropriate learning rates and subsampling strategies are employed. Such robustness is especially important for modeling stunting outcomes, where data uncertainty frequently arises from reporting indicating superior error minimization performance.

Actual Versus Predicted Response Analysis

To further assess model fidelity, the relationship between observed stunting indicators $yi$and their corresponding predicted values $y\hat{}i$was examined using scatter plots. Under ideal predictive conditions, all observations would align along the identity line $y = y\hat{}$, indicating perfect agreement between predicted and actual outcomes. Empirical results reveal that the majority of data points cluster closely around this diagonal reference line, demonstrating strong concordance between model estimates and observed stunting indicators.

Deviations observed at extreme values are primarily attributable to inherent data noise, measurement uncertainty, and unobserved contextual factors, highlighting the intrinsic complexity of modeling real-world socio-nutritional systems.

Feature Importance and System Interpretability

One of the primary advantages of ensemble tree-based models lies in their inherent interpretability through feature importance analysis [30]. Let Ijdenote the relative importance score associated with the j-th feature, which reflects its contribution to the predictive performance of the model. Empirical results indicate that Poverty Rating and Maternal Education consistently exhibit the highest importance scores across both ensemble models [10], suggesting that economic capacity and maternal knowledge function as dominant structural drivers within the socio-nutritional system. Additional variables, including Age and Residence, also demonstrate meaningful contributions, reflecting demographic and environmental influences on child growth patterns. These findings align with established epidemiological evidence while simultaneously providing a data-driven quantification of variable influence within a complex systems modeling framework.

System-Level Interpretation and Nonlinear Interactions

From a complex systems perspective, stunting emerges not as the result of isolated determinants, but as a system-level outcome driven by nonlinear and

interdependent interactions among socioeconomic variables. Ensemble learning models implicitly capture such interactions by recursively partitioning the feature space and aggregating decision boundaries across multiple learners, thereby approximating the underlying structure of the socio-nutritional system[21]. In this context, the superior performance of Gradient Boosting can be attributed to its ability to model higher-order interactions through sequential residual learning, allowing it to capture subtle dependencies among variables, particularly within populations characterized by overlapping socioeconomic vulnerabilities. Random Forest, by contrast, provides a more conservative yet robust approximation of system behavior, emphasizing stability and variance reduction through bootstrap aggregation.

Implications for Public Health Decision Support

The results demonstrate that ensemble machine learning models can serve as effective analytical components of early-warning systems in public health attributed to its sequential residual optimization mechanism, which incrementally corrects prediction errors and enhances model sensitivity to complex interaction patterns contexts. By identifying dominant predictors and quantitatively capturing nonlinear interactions among socio- nutritional variables, the proposed framework supports evidence-based intervention planning, risk stratification, and strategic resource prioritization. From an engineering perspective, the framework provides a scalable and interpretable modeling pipeline capable of integrating heterogeneous socio-nutritional data across populations and temporal settings. This positions ensemble learning as a viable methodological bridge between data science, applied mathematics, and epidemiological system analysis, enabling robust analytical support for complex public health decision-making processes.

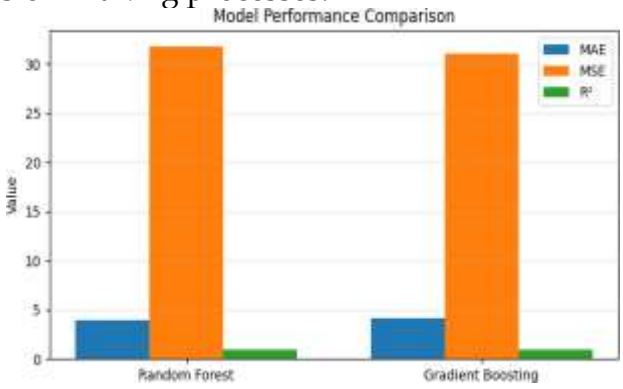

Figure 2 Placeholder: Model Performance comparison

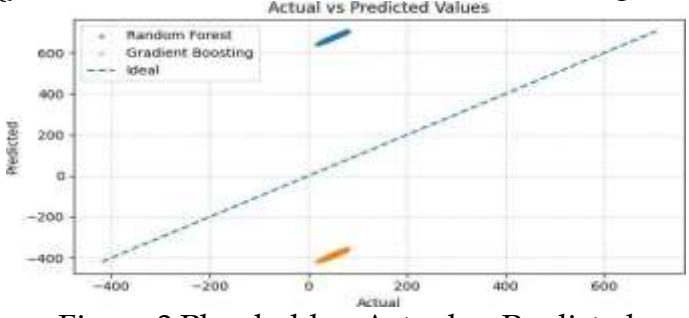

Figure 3 Placeholder: Actual vs Predicted

CONCLUSIONS AND RECOMMENDATIONS

This This study presented an applied ensemble machine learning framework for data-driven decision support using socioeconomic data as a real-world case study. By integrating Random Forest and Gradient Boosting models, the proposed approach demonstrates robust predictive performance and interpretability under heterogeneous data conditions.

The findings highlight the practical applicability of ensemble learning techniques for applied computing and informatics applications, particularly in scenarios requiring robust analysis and system-level understanding rather than algorithmic complexity. Future work may extend the framework to incorporate temporal dynamics and additional data modalities to further enhance analytical capabilities. From an applied computing perspective, the results demonstrate that ensemble learning models can function as reliable analytical tools for decision support in complex data environments. The stability observed across different evaluation metrics highlights the robustness of the proposed framework when applied to heterogeneous socioeconomic datasets.

ADVANCED RESEARCH

Based on the findings of this study, future (advanced) research may focus on developing a more adaptive ensemble machine learning framework by incorporating temporal and spatial dimensions to better capture the dynamic nature of socioeconomic change. Further studies could also explore the integration of multimodal data sources, such as geospatial data, policy text, or real-time data streams, to enhance analytical depth and decision-making accuracy. In addition, strengthening the application of explainable AI (XAI) within ensemble models is essential to ensure that predictive outcomes are not only robust but also interpretable, thereby supporting evidence-based decision support systems in complex and heterogeneous data environments.

REFERENCES

A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly, 2019.

A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," Frontiers in Neurorobotics, vol. 7, 2013.

A. Rajkomar et al., "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.

Advances in Neural Information Processing Systems (NeurIPS), pp. 4765– 4774, 2017.

C. G. Victora et al., "Maternal and child undernutrition: consequences for adult health," The Lancet, vol. 371, no. 9608, pp. 340–357, 2008.

C. Molnar, Interpretable Machine Learning. 2nd ed., 2022.

C. Sammut and G. I. Webb, Eds. Springer, 2011, pp. 312–320.

E. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.

G. Brown, "Ensemble learning," in Encyclopedia of Machine Learning,

H. Kantz and T. Schreiber, Nonlinear Time Series Analysis. Cambridge University Press, 2004.
J. Beam and I. Kohane, "Big data and machine learning in health care," JAMA, vol. 319, no. 13, pp. 1317–1318, 2018.
J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
J. H. Holland, Complexity: A Very Short Introduction. Oxford University Press, 2014.
L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
L. Rokach, "Ensemble-based classifiers," Artificial Intelligence Review, vol. 33, no. 1–2, pp. 1–39, 2010.
M. de Onis and F. Branca, "Childhood stunting: A global perspective," Maternal & Child Nutrition, vol. 12, pp. 12–26, 2016.
M. Kuhn and K. Johnson, Applied Predictive Modeling. Springer, 2013.
M. M. Islam et al., "Machine learning approaches for predicting child malnutrition," BMC Medical Informatics and Decision Making, vol. 20, no. 1, 2020.
M. Newman, "Complex systems: A survey," American Journal of Physics, vol. 79, no. 8, pp. 800–810, 2011.
M. Ruel and H. Alderman, "Nutrition-sensitive interventions and programmes," The Lancet, vol. 382, no. 9891, pp. 536–551, 2013.
N. Rathi et al., "Predicting malnutrition using ensemble learning," Journal of Infection and Public Health, vol. 14, no. 8, pp. 1046–1053, 2021.
P. Vega et al., "Prediction of childhood stunting using machine learning models," Scientific Reports, vol. 11, no. 1, 2021.
R. E. Black et al., "Maternal and child undernutrition and overweight," The Lancet, vol. 382, no. 9890, pp. 427–451, 2013.
R. M. May, "Simple mathematical models with very complicated dynamics," Nature, vol. 261, pp. 459–467, 1976.
S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions,"
S. Prendergast and J. Humphrey, "The stunting syndrome in developing countries," Paediatrics and International Child Health, vol. 34, no. 4, pp. 250–265, 2014.
S. Strogatz, Nonlinear Dynamics and Chaos. Boulder, CO, USA: Westview Press, 2015.
T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD, pp. 785–794, 2016.
T. G. Dietterich, "Ensemble methods in machine learning," in Proc. Int. Workshop on Multiple Classifier Systems, pp. 1–15, 2000.
UNICEF, WHO, and World Bank, "Levels and trends in child malnutrition," UNICEF, New York, NY, USA, 2021.
Y. Bar-Yam, Dynamics of Complex Systems. Westview Press, 2003.
Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, FL, USA: CRC Press, 2012.