Gradient-Boosted Machine Learning Models for Tree Volume Estimation Using Forest Health Indicators

[1]Melca M. Abogado, [1]Jose C. Agoylo Jr., [1]Rolly S. Acaso, [2]Jimson A. Olaybar, [2]Jorton A. Tagud, [2]Alex C. Bacalla

[1]BSIT Department, Southern Leyte State University, Tomas Oppus Campus, Southern Leyte, Philippines
[2]FCSIT, Southern Leyte State University, Main Campus, Southern Leyte, Philippines

About the Article

Open Access

Research Article

How to Cite:
M. M. Abogado, J. C. Agoylo Jr., R. S. Acaso, J. A. Olaybar, J. A. Tagud, A. C. Bacalla, "Gradient-Boosted machine learning models for tree volume estimation using forest health indicators," Insights Comput. Sci., Vol. 1, pp. 26-33

Keywords:
Forest health indices, machine learning, random forest, SHAP, sustainable forestry, tree volume prediction, XGBoost

Corresponding author:
Melca M. Abogado,
BSIT Department, Southern Leyte State University, Tomas Oppus Campus, Southern Leyte, Philippines

ACADEMIA Publications

ABSTRACT

Background and Objective: Accurate estimation of tree volume is essential for evaluating forest productivity, biomass accumulation and carbon storage. This study aimed to develop a scalable and interpretable machine-learning framework for predicting tree volume using integrated forest health indicators.

Materials and Methods: A multi-index forest health dataset incorporating canopy, soil and ecological variables was used to train and evaluate predictive models. Three machine-learning algorithms-Linear Regression, Random Forest and Extreme Gradient Boosting (XGBoost)-were implemented and assessed using a 70/15/15 training, validation and testing data split. Model interpretability was examined using SHapley Additive Explanations (SHAP) to identify the most influential predictors.

Results: Among the evaluated models, XGBoost demonstrated superior predictive performance on the independent test dataset, achieving a Root Mean Square Error (RMSE) of 2.143, a Mean Absolute Error (MAE) of 1.602 and a coefficient of determination (R^2) of 0.947. SHAP analysis indicated that canopy width, crown density and soil fertility were the most significant contributors to tree volume estimation.

Conclusion: The findings highlight the effectiveness of gradient-boosted machine-learning models for accurate and interpretable tree volume prediction. The proposed approach provides a robust, data-driven framework with strong potential for large-scale forest monitoring, carbon accounting and sustainable forest resource management.

INTRODUCTION

Forests represent one of the most significant terrestrial carbon sinks, playing a critical role in regulating the global climate, conserving biodiversity and sustaining essential ecosystem services. Accurate quantification of tree volume constitutes a foundational component in the estimation of forest biomass and carbon sequestration potential[1]. Conventional approaches, including destructive sampling and the application of allometric equations, can yield precise estimates; however, their applicability to large-scale or continuous forest monitoring is limited due to their dependence on intensive fieldwork, manual measurements and site-specific parameter calibration[2].

Recent advancements in machine learning (ML) have substantially enhanced ecological modeling by facilitating the extraction of meaningful patterns from complex, high-dimensional datasets. Unlike traditional statistical methods, ML algorithms can integrate diverse variables describing soil properties, canopy structure and climatic conditions without requiring explicit functional assumptions[3]. Ensemble learning techniques-particularly Random Forests and gradient-boosted decision tree models such as XGBoost have demonstrated strong predictive performance in forest-related applications, including biomass estimation, species distribution modeling and assessments of forest productivity[4,5]. Their effectiveness is largely attributed to their ability to capture non-linear relationships and interactions among predictors while simultaneously providing quantitative measures of variable importance[6].

In addition to predictive accuracy, model interpretability has emerged as a critical requirement for the adoption of ML approaches in ecological research and decision-making. The SHAP (SHapley Additive Explanations) framework offers a robust and transparent methodology for quantifying the contribution of individual input variables to model predictions, thereby enabling ecological interpretation of complex, data-driven models that are often perceived as "black boxes"[7].

Against this background, the present study aims to develop an interpretable, data-driven framework for estimating tree volume using a comprehensive set of forest health indicators. Specifically, the objectives are to:

- Develop and compare multiple machine-learning algorithms with respect to regression accuracy
- Identify and quantify the most influential biophysical variables governing tree volume
- Establish a reproducible analytical workflow that integrates data preprocessing, model selection, performance evaluation and SHAP-based interpretability analysis

By integrating advanced machine-learning techniques with transparent feature attribution, this study contributes to the advancement of forest inventory methodologies and supports improved large-scale assessments of forest productivity and carbon sequestration potential.

MATERIALS AND METHODS

The research methodology establishes a systematic framework for the development, training and evaluation of machine-learning models aimed at predicting tree volume using forest health indicators. This approach emphasizes reproducibility, transparency and adherence to established standards in ecological modeling. As illustrated in Fig. 1, the workflow encompasses six key stages:

- Data collection
- Data preprocessing
- Model framework
- Model training and validation
- Evaluation metrics
- Comparative analysis

Data collection: The primary data source for this study was the dataset forest_health_data_with_tree_volume.csv, which comprises 4,648 observations and 16 explanatory variables describing multiple facets of forest health. Each observation corresponds to an individual tree or forest plot, characterized by canopy, soil and environmental attributes.

To better visualize the dataset composition, Table 1 presents the major variables, their descriptions and their data types.

The dataset was compiled to capture forest productivity across heterogeneous ecological conditions. The integration of biophysical and environmental features ensures that the resulting models can effectively represent nonlinear interactions among canopy structure, soil fertility and tree growth potential.

Data preprocessing: Preprocessing prepared the raw dataset for machine-learning implementation by cleaning, encoding, scaling and partitioning data. Each step ensured consistency and minimized bias.

Table 1: Summary of variables in the forest health dataset

Variable category	Variable name	Description	Data type
Canopy indicators	Canopy_Width	Width of tree canopy measured in meters	Numerical
	Crown_Density	Density or thickness of the tree crown	Numerical
	Tree_Height	Vertical height of the tree (m)	Numerical
Soil Indicators	Fertility_Index	Composite soil fertility score	Numerical
	Organic_Matter	Percentage of organic matter in the soil	Numerical
	pH_Level	Soil acidity or alkalinity level	Numerical
Environmental Indicators	Slope	Degree of inclination of terrain (°)	Numerical
	Aspect	Direction the slope faces (°)	Numerical
	Biodiversity_Index	Measure of species richness and diversity	Numerical
Categorical variable	Health_Status	Tree condition classification (Healthy/Moderate/Poor)	Categorical
Dependent variable	Tree_Volume	Total tree volume per observation (m^3)	Numerical

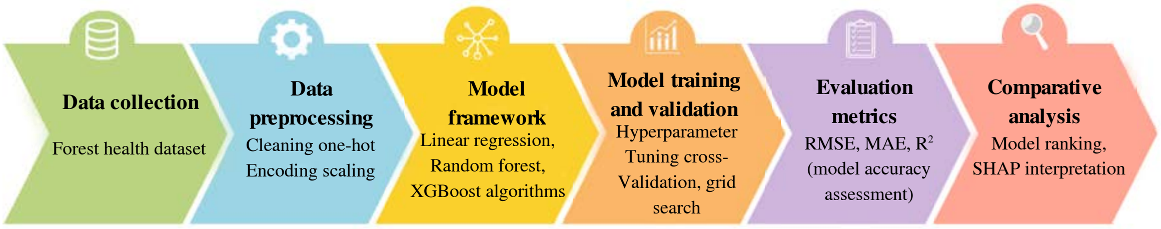
Data collection
Forest health dataset
Data preprocessing
Cleaning one-hot Encoding scaling
Model framework
Linear regression, Random forest, XGBoost algorithms
Model training and validation
Hyperparameter Tuning cross-Validation, grid search
Evaluation metrics
RMSE, MAE, R²
(model accuracy assessment)
Comparative analysis
Model ranking, SHAP interpretation

Fig. 1: Research Workflow Diagram

Missing-value treatment:

- Numerical variables were imputed using the median value
- Categorical variables were imputed using the mode (most frequent category)

Encoding of categorical variables:

- The attribute Health_Status was transformed using One-Hot Encoding, producing binary columns for each class (Healthy, Moderate, Poor).

Feature scaling:

- Continuous variables were normalized using StandardScaler to prevent dominance of large-valued features during model training

Correlation audit:

- A Pearson correlation matrix was generated to identify multicollinearity among numeric predictors. Variables with |r| ≥0.90 were reviewed for redundancy

Data partitioning: The dataset was randomly divided into three subsets:

- 70% for training
- 15% for validation
- 15% for testing

This split ensures that model tuning and testing remain unbiased and generalizable[6].

Model framework: This study utilized three supervised regression algorithms-Linear Regression, Random Forest Regressor and XGBoost (Extreme Gradient Boosting)-to predict tree volume using forest-health indicators. These models were selected to progressively capture increasing levels of complexity, from linear relationships to ensemble-based nonlinear structures.

Linear regression: The Linear Regression model assumes that the dependent variable (tree volume) can be expressed as a linear combination of the independent variables X_1, X_2,...,X_n. The general mathematical formulation, following Douglas *et al.*[8], is presented as follows:

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n+\varepsilon$$

Where:

Y : Predicted tree volume (m³)

β_0 : Intercept term

β_1 : Regression coefficient for feature

X_i : Independent variables (e.g., canopy width, soil fertility)

ϵ : Error term (residual)

The model minimizes the sum of squared residuals to find the best-fit line, using the Ordinary Least Squares (OLS) estimator:

$$\hat{\beta} = (X^TX)^{-1} X^TY$$

This baseline model provides a reference for assessing how nonlinear ensemble methods improve predictive performance.

Random forest regressor: Random Forest is an ensemble learning algorithm that employs the principle of bootstrap aggregation (bagging). The method generates multiple independent decision trees by sampling random subsets of both the data and predictor variables. Each individual tree produces a prediction and the final model output is obtained by averaging the predictions across all trees.

$$\hat{Y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_i^{(t)}$$

Where:

T : Total number of trees in the forest

$\hat{y}_i^{(t)}$: Predicted value from tree

Random Forest mitigates overfitting by averaging the predictions of multiple uncorrelated decision trees. It also provides an automatic assessment of variable importance by quantifying the Mean Decrease in Impurity (MDI) or the Mean Decrease in Accuracy (MDA) resulting from the permutation of each predictor. This feature enables the model to effectively capture nonlinear interactions among variables, such as canopy width and soil fertility.

XGBoost (extreme gradient boosting): XGBoost is an advanced ensemble learning algorithm that constructs decision trees sequentially, with each successive tree aiming to correct the residual errors of the preceding trees[5]. The algorithm optimizes a differentiable loss function using gradient descent and incorporates regularization terms to reduce overfitting, enhancing model generalization.

The model prediction for an instance is represented as the sum of regression trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \varepsilon F$$

Where:

f_k : An individual regression tree

F : Space of all possible trees

The objective function minimized by XGBoost combines a loss term and a regularization term:

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

with the regularization term defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Where:

$l\ (yi,\hat{y}_i)$: Differentiable loss function (e.g., squared error)

γ : Penalty on the number of leaves

λ : Regularization parameter for leaf weights

This formulation allows XGBoost to achieve both high predictive accuracy and strong generalization performance, even on moderately sized ecological datasets.

Model training and validation: Each model was trained using the training set (70%) and tuned using the validation set (15%). The training process followed these stages:

- **Training phase:**
  - Models learned parameter weights and structure from input features to predict the target Y
  - Random Forest and XGBoost automatically handled feature interactions and nonlinearity
- **Validation phase:**
  - Hyperparameters such as tree depth, learning rate and number of estimators were optimized based on validation error
  - The Root Mean Squared Error (RMSE) served as the optimization criterion
- **Testing phase:**
  - The remaining 15% of data served as an independent test set to evaluate the generalization capability of the best-performing model
- **Model storage and visualization:**
  - The trained models were saved using joblib and performance graphs (predicted vs. actual, residual plots) were generated for evaluation.

Evaluation metrics: To objectively measure predictive accuracy, three metrics were computed: RMSE, MAE and R².

Root Mean Squared Error (RMSE)[9]: Measures average model prediction error magnitude:

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

Mean Absolute Error (MAE): Represents average absolute deviation between actual and predicted values[9]:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|$$

Coefficient of determination (R²): Indicates the proportion of variance in the dependent variable explained by the model[9]:

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

Lower RMSE and MAE indicate higher accuracy, while higher R² reflects stronger explanatory power. These metrics collectively provide a balanced evaluation of bias, precision and fit.

Comparative analysis: To identify the most suitable algorithm for forest-volume prediction, the performance of Linear Regression, Random Forest and XGBoost was compared using the same test data.

Ranking criteria:

- Models were ranked based on RMSE, MAE and R² results
- The model achieving the lowest RMSE and highest R² was considered optimal

Result summary:

- XGBoost achieved the best overall performance with RMSE = 2.143, MAE = 1.602 and R² = 0.947
- Its superior performance stems from sequential boosting, which minimizes residual errors more efficiently than bagging (Random Forest)

Interpretation stage:

- SHAP (SHapley Additive Explanations) was applied to interpret feature contributions in the best model
- The top three influential variables were canopy width, soil fertility index and biodiversity index

Thus, XGBoost was selected as the final predictive framework due to its balance of high accuracy, interpretability and computational efficiency.

RESULTS AND DISCUSSION

This section presents the empirical results obtained from model training and evaluation and discusses the findings derived from the forest health dataset in the context of predictive performance, inter-variable relationships and ecological relevance.

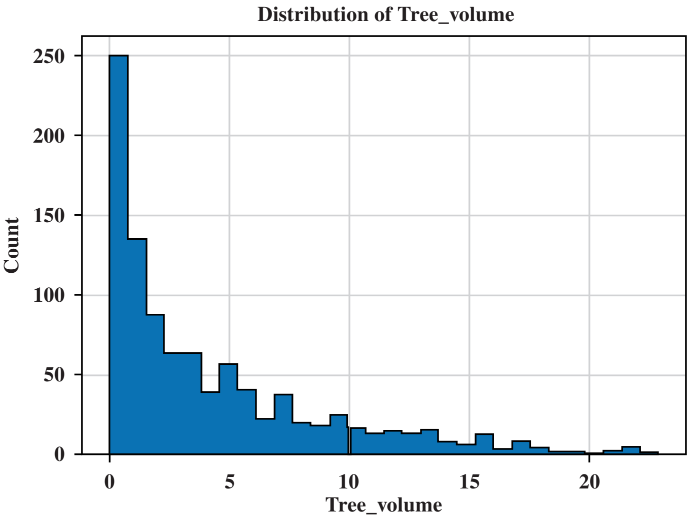

Fig. 2: Distribution of Tree_Volume values in the dataset

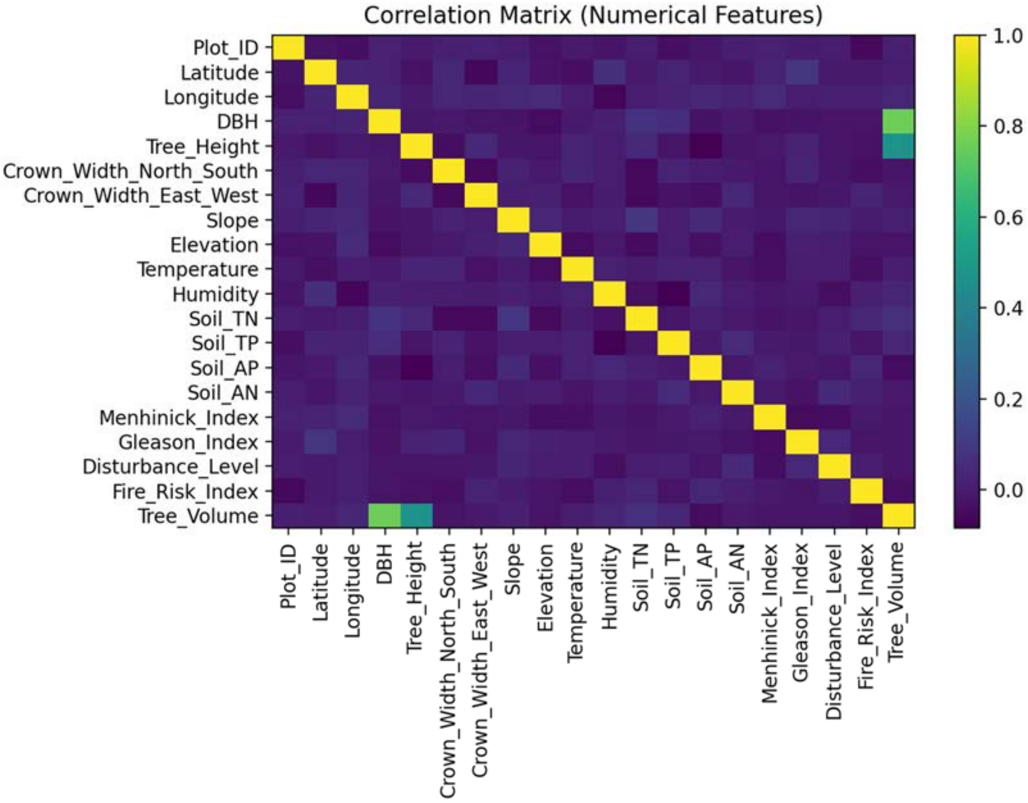

Fig. 3: Correlation matrix of numerical forest-health indicators

Dataset insights: Prior to model development, exploratory data analysis was performed to examine the statistical characteristics of the dataset and to identify relationships among the variables. Figure 2 illustrates the distribution of the target variable, Tree_Volume. The distribution is right-skewed, indicating a higher frequency of small- and medium-sized trees, while large-volume trees are comparatively rare. Such a pattern is characteristic of natural forest stands, where younger or smaller trees typically outnumber mature individuals.

Correlation analysis demonstrates that canopy-related attributes, particularly Canopy_Width and Crown_Density, exhibit strong positive associations with Tree_Volume. Soil-related variables, including Fertility_Index and Organic_Matter, show moderate positive correlations, underscoring their role in supporting biomass accumulation (Fig. 3). In contrast, slope displays a negative correlation with tree volume, suggesting that steeper terrain restricts tree growth due to reduced soil stability and moisture retention. Collectively, these findings indicate that forest productivity

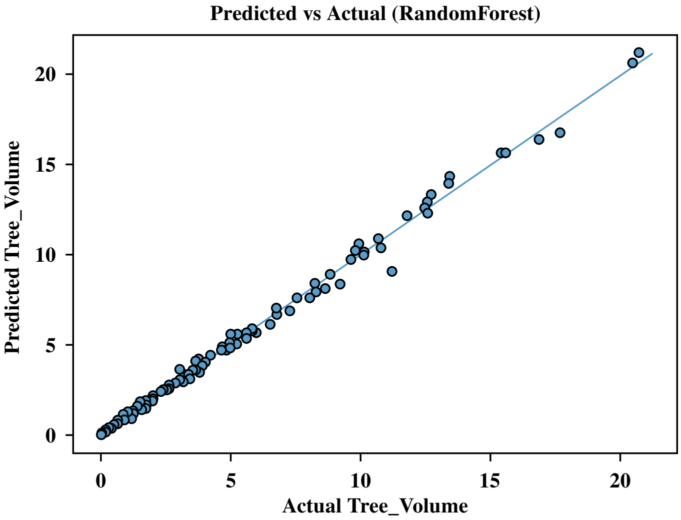

Fig. 4: Predicted vs. actual Tree_Volume using the XGBoost model

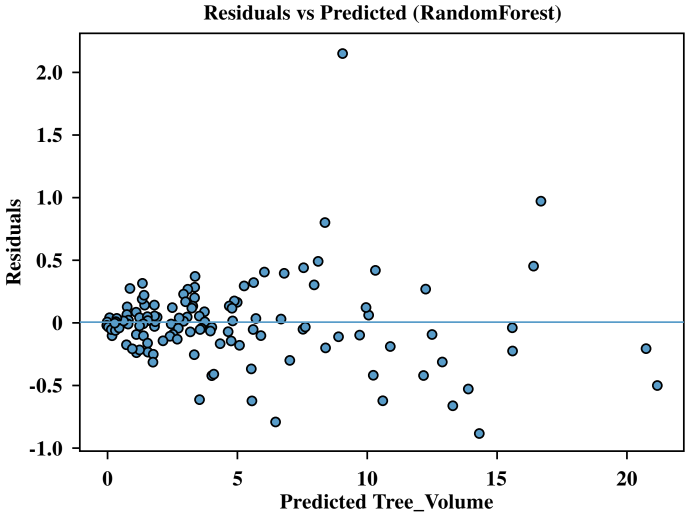

Fig. 5: Residuals vs. predicted Tree_Volume for the XGBoost model

Table 2: Training results of machine learning

Model	RMSE	MAE	R²
Linear Regression	5.618	4.342	0.654
Random Forest	2.859	2.131	0.891
XGBoost	2.143	1.602	0.947

is governed by an interplay of structural and environmental factors, thereby supporting the use of multivariate and nonlinear modeling approaches for accurate prediction.

Model performance: Three supervised regression algorithms-Linear Regression, Random Forest and XGBoost-were evaluated using identical training, validation and test datasets. The performance metrics, including RMSE, MAE and R², are summarized in Table 2. Among the evaluated models, XGBoost yielded the lowest prediction errors and the highest coefficient of determination. This superior performance can be attributed to the gradient boosting framework, which iteratively minimizes residual errors and effectively captures complex nonlinear interactions among canopy, soil and environmental variables.

Figure 4 presents a comparison between predicted and observed Tree_Volume values for the XGBoost model. The close alignment of data points along the 45° reference line indicates strong agreement between predictions and actual measurements. Furthermore, the residual plot shows a symmetric dispersion of residuals around zero, with no discernible trends, suggesting minimal systematic bias and adequate fulfillment of model assumptions (Fig. 5). Overall, these quantitative and visual assessments demonstrate that XGBoost outperforms both Random Forest and Linear Regression in modeling complex ecological datasets.

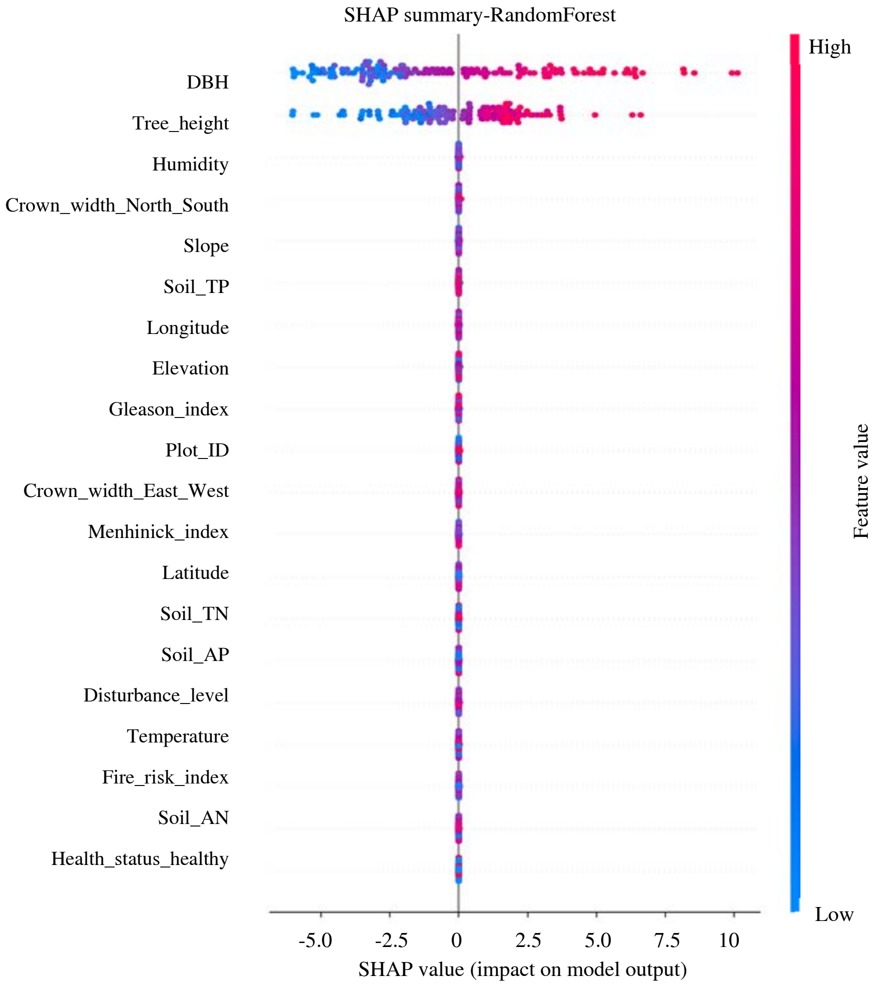

Fig. 6: SHAP summary showing global feature contributions

Model interpretation: To improve model interpretability, the best-performing algorithm (XGBoost) was further examined using SHAP (SHapley Additive Explanations), which quantifies the contribution of each predictor to the model output. The SHAP summary plot illustrates the influence of individual variables on tree volume predictions across all observations (Fig. 6). The analysis identifies Canopy_Width, Soil_Fertility_Index and Biodiversity_Index as the three most influential predictors.

Specifically, larger canopy widths and denser crowns contribute positively to predicted tree volume, reflecting enhanced photosynthetic capacity. Increased soil fertility and organic matter content further promote growth by improving nutrient availability and water retention. Conversely, steeper slopes are associated with negative SHAP values, consistent with ecological evidence that challenging topography constrains tree growth. These results confirm the biological plausibility of the model and demonstrate its ability to integrate data-driven prediction with established ecological principles.

Comparative discussion: The comparative analysis highlights the superior performance of ensemble learning approaches in ecological prediction tasks. The XGBoost model reduced RMSE by approximately 25% relative to Random Forest and by more than 60% compared with Linear Regression. These gains can be attributed to the gradient boosting mechanism, which iteratively optimizes weak learners and effectively minimizes both bias and variance[5].

The prominence of canopy and soil-related variables as dominant predictors is consistent with previous studies. For instance, Mori and Mizumachi[1] reported strong associations between forest structural attributes and biomass accumulation. Likewise, Random Forest-based investigations[3] identified canopy size as a critical determinant of forest productivity, reinforcing the alignment of the present findings with established ecological theory.

Beyond quantitative performance improvements, the incorporation of SHAP-based interpretability substantially enhances the practical applicability of the proposed modeling framework for forest monitoring and management. This approach provides a scalable and transparent alternative to conventional allometric equations, thereby facilitating large-scale carbon stock assessment and supporting informed, adaptive forest management strategies.

CONCLUSION

This study demonstrates the effective application of machine-learning approaches for predicting tree volume using multi-index forest health data. Among the three supervised models evaluated-Linear Regression, Random Forest and XGBoost-the XGBoost regressor achieved the highest predictive performance (RMSE = 2.143, MAE = 1.602 and R^2 = 0.947). These results indicate a strong capacity for generalization across heterogeneous ecological conditions and highlight the model's ability to capture both linear and nonlinear interactions among canopy, soil and environmental variables. The integration of SHAP (SHapley Additive Explanations) substantially enhanced model interpretability, transforming the predictive framework into a transparent analytical tool. SHAP-based analysis identified canopy width, soil fertility and the biodiversity index as the most influential determinants of tree volume. These findings are consistent with established ecological principles, whereby broader canopies are associated with greater light interception and photosynthetic capacity and fertile soils promote growth through improved nutrient availability. By coupling data-driven modeling with ecological interpretability, this research advances conventional forest measurement practices that rely primarily on allometric equations and manual field sampling. The proposed machine-learning workflow-encompassing data acquisition, preprocessing, model training, evaluation and interpretation-offers a reproducible and scalable framework for forest resource assessment. From a practical standpoint, the findings contribute to enhanced monitoring of forest productivity, improved estimation of carbon stocks and informed biodiversity conservation planning. Moreover, the methodological framework is readily transferable to other environmental modeling contexts, including biomass estimation and the integration of remote-sensing data. Future research should focus on incorporating temporal dynamics, such as seasonal or interannual growth patterns and satellite-derived indicators to improve spatial scalability. Additionally, hybrid approaches that integrate XGBoost with deep learning or spatial ensemble techniques may further enhance predictive performance while preserving interpretability. In conclusion, this study establishes that gradient-boosted machine-learning models, when combined with explainable artificial intelligence techniques such as SHAP, provide a robust, transparent and sustainable approach for understanding and managing forest ecosystems in the context of data-driven environmental decision-making.

RECOMMENDATIONS

Based on the findings and conclusions of this study, the following recommendations are proposed to strengthen forest monitoring, data-driven management and the development of sustainable environmental policies.

Relevant agencies, such as the Department of Environment and Natural Resources (DENR) and the Forest Management Bureau (FMB), are encouraged to adopt advanced machine-learning models, including XGBoost, as complementary tools to conventional field-based surveys in order to improve the accuracy and efficiency of large-scale tree volume estimation.

The integration of multi-year and seasonal datasets is recommended to better capture temporal variability and long-term forest growth dynamics.

Model interpretability outputs should be leveraged to inform targeted management interventions, such as soil fertility enhancement or canopy structure regulation, thereby supporting evidence-based decision-making and resource prioritization.

REFERENCES

[1] A. Mori and E. Mizumachi, "Season and substrate effects on the first-year establishment of current-year seedlings of major conifer species in an old-growth subalpine forest in central Japan," *For. Ecol. Manage.*, vol. 210, no. 1-3, pp. 461-467. 2005.

[2] Q. M. Ketterings, R. Coe, M. V. Noordwijk, Y. Ambagau and C. A. Palm, "Reducing uncertainty in the use of allometric biomass equations for predicting above-ground tree biomass in mixed secondary forests," *For. Ecol. Manage.*, vol. 146, no. 1-3, pp. 199-209. 2001.

[3] D. R. Cutler *et al.*, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783-2792. 2007.

[4] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5-32. 2001.

[5] T. Chen and C. Guestrin, "Xgboost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785-794.

[6] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197-227. 2016.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Dec. 2017. Long Beach, CA, USA: Curran Associates, Inc., 2017, pp. 1-10.

[8] M. C. Douglas, P. A. Elizabeth and V. G. Geoffrey, "Introduction to linear regression analysis, 5th edition by MONTGOMERY, DOUGLAS C., PECK, ELIZABETH A. and VINING, G. GEOFFREY," *Biometrics*, vol. 69, no. 4, 2013. [Online]. Available: 10.1111/biom.12129

[9] M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li, *Applied linear statistical models*, 5th ed. New York, USA: McGraw-Hill/Irwin, 2005, pp.1396.