VOL 4, NO.2 (2024) 29-78
https://jagst.utm.my

JAGST
Journal of Advanced Geospatial Science & Technology

Harnessing Machine Learning and Ensemble Models for Tourism Potential Zone Prediction for the Assam State of India

Shrinwantu Raha[1]*, Shasanka Kumar Gayen[1], Sayan Deb[2]
[1]Department of Geography, Cooch Behar Panchanan Barma University, Cooch Behar, West Bengal, Pin: 736101
[2]Department of Geography, Bhairab Ganguly College, Belgharia, Kolkata 7,000,56

*Corresponding author: Email: shrinwanturaha1@gmail.com

Abstract – Although several popular tourist destinations exist in Assam, India, its charm remains enigmatic. This research was aimed at predicting the tourism potential zone (TPZ) for the state of Assam using five machine learning models (i.e., Conditional Inference Tree, Bagged CART, Random Forest, Random Forest with Conditional Inference Tree, and Gradient Boosting models) and one ensemble model. A 5-step methodology was implemented to conduct this research. First, a tourism inventory database was prepared using Google Earth Imagery, and a rapid field investigation was performed using the global positioning system and nonparticipant observation technique. A total of 365 tourism points were present in the inventory, 70% (224 points) of which were used for the training set and 30% (124 points) for the validation set. Tourism conditioning factors, such as relief, aspect, viewshed, forest area, wetland, coefficient of variation of rainfall, reserve forest, population density, population growth rate, literacy rate, and road–railway density, were used as independent variables in the modeling process. The TPZ was predicted using the above machine learning models, and finally, a new TPZ ensemble model was proposed by combining all the models. The result showed that all machine learning models performed well in terms of prediction accuracy, and the ensemble model outperformed other models by achieving the highest area under the curve (97.6%), Kappa (0.82), and accuracy (0.93) values. The findings from this research using machine learning and ensemble methods can provide accurate and significant information for decision-makers to develop tourism in the region.

Keywords *– Tourism Potentiality, Analytic Hierarchy Process, ROC-AUC, Conditional Inference Tree, Bagged CART, Random Forest, TPZ Ensemble Model*

Article History: Received 1 May 2024, Accepted 14 August 2024, Published 31 August 2024

How to cite: Raha, S., Gayen, S. K. and Deb. S. (2024). Harnessing Machine Learning and Ensemble Models for Tourism Potential Zone Prediction for the Assam State of India. Journal of Advanced Geospatial Science and Technology. 4(2), 29-78.

1.0 Introduction

In recent years, tourism has been a key driver of economic growth in both underdeveloped and wealthy nations (Manzoor et al., 2019). Tourism helps expand the economy of underdeveloped and developed countries in various ways, such as providing gains in foreign exchange, luring foreign investment, and increasing tax receipts (Zabihi et al., 2020). The economic world and its productivity are facilitated by tourism, which is one of the biggest industries in the world (Puška et al., 2021). Tourism is an important demographic mechanism that solves the unemployment problem in crowded and sparsely populated areas. The United Nations World Trade Organization estimates that only 25 million travellers worldwide were globetrotters in 1950. In a mega decade, this number increased to 1.4 billion people. 2018, the most outstanding tourist growth was registered in the Middle East and Asia Pacific regions (Scarpocchi, 2020).

Assessing tourism potentialities is unquestionably significant and necessary as tourism intrinsically and extrinsically contributes to the socioeconomic development of a place (Kachniewska, 2015). Tourism potentiality is the sum of environmental, ethnographic, cultural, political and social values for establishing tourist activities in a place (Katelieva & Muhar, 2022). The philosophy of tourism potentiality plays a significant role in retaining the personality of a region (Kontogeorgopoulos, 2017). It promotes tourists' conduct, which considerably and favourably aids in preserving the natural and chemical integrity of the ecosystem (Khadka et al., 2021). Tourism potentiality can evaluate a region's capacity for sustainable and inclusive growth (Blapp & Mitas, 2018; Zekan et al., 2022). As a result, it is strongly advised that the touristic potential for the integrated growth of society and culture be evaluated (Banik & Mukhopadhyay, 2020). Although the state of Assam is filled with tourism potentiality, such areas are yet to be explored. Therefore, the distinction and prediction of potential zones (TPZs) are necessary to utilize tourism resources effectively.

In recent decades, advancements in photogrammetry, geographical information systems (GIS), and remote sensing (RS) have led to the creation of numerous new machine learning (ML) models and algorithms by researchers worldwide (Marín-Buzón et al., 2021; Apostolopoulos et al., 2021). Determining the TPZs involves analyzing the physical and socioeconomic parameters to forecast tourism trends in unexplored areas. Globally, various ML models have been employed in research fields such as identifying potential groundwater zones (e.g., Vafadar et al., 2023; Roy et al., 2024), environmental and ecological planning (e.g., Mosebo Fernandes et al., 2020), resource

management (Garg et al., 2022), forestry (Liu et al., 2018; Zhao et al., 2019), urban and regional planning (Chaturvedi & de Vries, 2021; Tekouabou et al., 2022), and natural hazard management (Wang et al., 2021; Linardos et al., 2022). Furthermore, ML models have been applied in diverse sectors of tourism, such as tourism demand forecasting (e.g. Claveria et al., 2016; Bi et al., 2022; Karakitsiou & Mavrommati, 2017; Cankurt & Subasi, 2015; Law et al., 2019), tourist review analysis (Le et al., 2021; Puh & Bagić Babac, 2023), and tourist arrival forecasting analysis (Sun et al., 2019), for which historical time series data are available. But for TPZ identification, the decision-making model, particularly the analytic hierarchy process (AHP), has so far dominated the frame (e.g., Raha et al., 2021, 2022, 2023; Sahani, 2020; Pathmanandakumar et al., 2023). AHP is an objective decision-making process that aids in determining the base information about any project over a specific region (Chowdary et al., 2013).

Generally, expert opinions are utilized to fine-tune the model. However, to accurately model similar geo-environmental factors in the same region, the models rely on long-term data on TPZ and its' causal factors. Regrettably, such data is often unavailable in most cases. Several ML models have played a significant role in this regard. Specifically, advanced ML models could play a substantial role in project viability by analyzing tourist behaviour, accommodation, and destination facilities (Singh et al., 2023; Nath et al., 2020). ML models have been used to examine tourist databases via pattern recognition, data-driven insights, adaptability and flexibility, improved accuracy, scalability, and automation (Danish et al., 2023; Chien et al., 2023). In this research, the following ML models were applied: (i) conditional inference tree (Kuhn & Johnson, 2013), (ii) bagged CART (Hamze-Ziabari & Bakhshpoori, 2018), (iii) random forest model (Gevrey et al., 2003), (iv) random forest with conditional inference tree (Quinlan, 1992), and (v) gradient boosting (Kuhn & Johnson, 2013). Furthermore, an ensemble model was prepared by combining the forecasts of various bootstrapping base models, such as decision trees, support vector machines, bagged CART, and gradient boosting models. This combination eliminated the shortcomings of individual models and provided the benefit of their combined advantages. In addition, for spatial assessments, the ML results were linked to the GIS, which made it possible to represent spatial data in the visual form (maps) to identify areas in which TPZs can be developed. This data integration is expected to facilitate understanding and provide users with added spatial context during decision-making.

Therefore, considering the above perspectives, the research objective is to predict the TPZ using several machine learning and ensemble models and make an inter-model comparison. The remaining manuscript was structured as follows:

- The second section was marked with the introduction of the study area.
- The third section, the materials and method section, was identified.
- The fourth section of the manuscript identifies the 'Result' section of the manuscript.
- The fifth section highlights the 'Discussion' section and
- The last section highlights the conclusion section of the manuscript.

2.0 Materials and Methods

2.1 First Step: Selecting the Study Area

Assam, the largest state in terms of population and the second largest in terms of territory, is bordered by Bhutan and Arunachal Pradesh in the north; Nagaland, Arunachal Pradesh, and Manipur in the east; Bangladesh and Meghalaya in the west; and Tripura, Mizoram, and West Bengal in the south (Assam state portal; assam.gov.in). Initially, the state of Assam had 27 districts; however, in 2016, 5 districts were added. Nonetheless, in this research, only 27 districts were considered for simplification. The fact that Assam is home to three of India's six physiographic divisions, namely, the northern Himalayas (Eastern Hills), northern plains (Brahmaputra plain), and deccan plateau, is an important geographical feature (Karbi Anglong). Assam often experiences a "tropical monsoon rainforest climate," with high humidity and precipitation. Warm summers and mild winters make for a temperate climate that residents can use throughout the year. Spring (March–April) and fall (September–October) are often pleasant, with mild temperatures and rains. According to the Census of India (2011) ("Census Tables | Government of India"), Assam has a total population of 3.12 crores. Assam's population constituted 2.58% of all Indians in 2011. Assam has a population of 31,205,576 people, with 15,939,443 men and 15,266,133 women. The state has a total size of 78,438 km^2. Therefore, Assam has a population density greater than the national average.

Assam is significant for drawing tourists because of its magnificent mountains, biodiversity, plenty of foliage, ethnic diversity (fairs and festivals), and emission zones (Huismann, 2014). Despite the country's popularity for tourists visiting, there are still many opportunities to explore

its scenic beauty, grasslands (Bugyals), caves, bird-watching sites, camping grounds, parks, and wildlife sanctuaries, as well as its skiing areas, river valleys, passes, glaciers, mountain peaks, trekking trails, and river rafting locations (Choden et al., 2018). The Himalayan mountains allow people to escape the pre-monsoon heat because of the good weather and picturesque scenery (Huismann, 2014).

To the best of our knowledge, this research is the first to highlight the tourism potentialities of the entire Assam. Previous research highlighted the tourism potential only in certain pockets of the state. For example, the nature-based tourism potentials of the Tinsukia District of Upper Assam were divulged by Bordoloi and Agarwal (2015). The tourism potential of the Karabi Anglong autonomous council districts were estimated by Ronghang and Sen (2022). However, the previous research activities completely neglected spatial assessments, model building, and their validation. This research emphasized the tourism potentials of the Assam state using machine learning algorithms, making it an appealing reading for academicians and tourism practitioners. Therefore, this research is novel and has implications for assessing tourism potential.

2.2 Second Step: Choosing Tourism Potentiality Causative Factors and Multicollinearity

Tourism potentiality is a multidimensional concept in which several physical and socioecological variables are linked (Raha & Gayen, 2023). The second step in the methodology (Figure 1) involved choosing the causative factors for tourism potentiality and measuring the multicollinearity. Here, 11 criteria were selected based on the recommendation of a five-membered (out of 5 members, three members were male, and the remaining were females) expert panel (Table 1) and in-depth literature reviews (i.e., Raha & Gayen, 2023; Sahani, 2020; Trukhachev, 2015; Schultze et al., 2014; Gourabi & Rad, 2013; Hoang et al., 2018). Three of them are academicians or researchers with at least five years of expertise in travel and tourism. The remaining two are destination managers with ten years of experience in the travel and tourism industry. In a separate consent form, the consulted experts permitted the results to be used purely for academic purposes without revealing their original identity. The following criteria were used in this research: relief (RL), aspect (AS), viewshed (VS), forest area (FA), wetland (WL), coefficient of variation of rainfall (CVR), reserved forest (RF), population density (PD), population growth rate (PGR), literacy rate (LR), and road/railway density (RRD). RL, AS, VS, FA, WL, CVR and RF created an

initial base for tourism activities (Yuxi & Linsberg, 2020). The PD, PGR, LR, and RRD indicated a regional population structure.

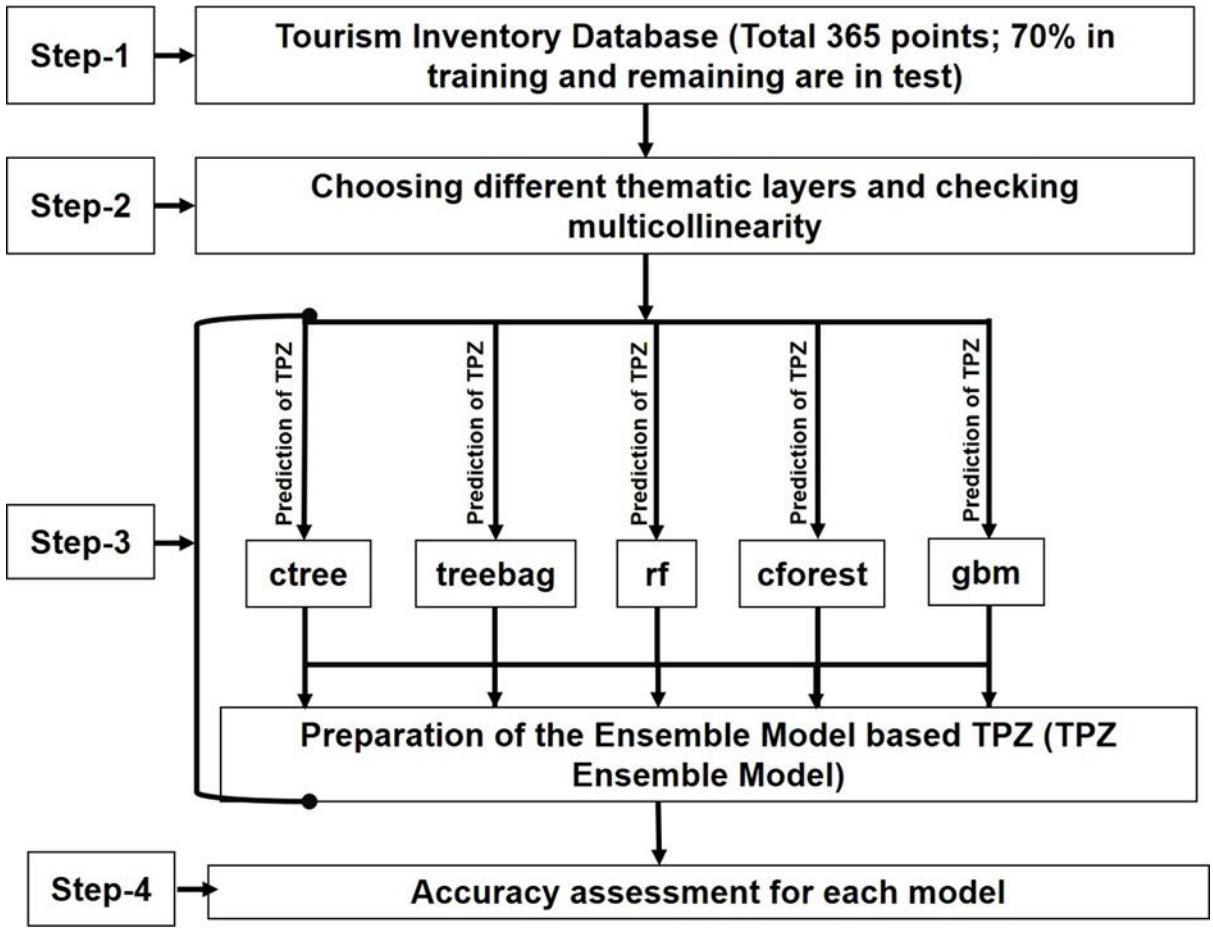

Figure 1. Methodological framework.

The sources of data are presented in Table 1. RL, AS, and VS were obtained from the digital elevation model prepared by the Shuttle Radar Topographic Mission (SRTM) (resolution was 1 ARC and pixel depth was 16 bits, published in 2014). The SRTM DEM was downloaded from the NASA Earth Explorer ("EarthExplorer"). Details of the FA were obtained from the India State Forest Report (2019). Data on the WL and RF were acquired from the Assam Project on Forest and Biodiversity Conservation Society (https://apfbcs.nic.in/apfbcs/wetland/annexure2.pdf). The CVR data (2020) was obtained from the Climate Research and Service Unit, Pune (2020). PD, PGR, and LR were sourced from the Statistical Handbook of Assam (2016).

Table 1. Considered thematic layers and their sources of data.

Criteria	Sources of data	Directionality of influence Descriptions
Relief (RL)	SRTM DEM, spatial Resolution 90 m	The moderate to low relief aspect helps to vibrate the tourism potentialities positively (Codrea et al. 2022)
Aspect (AS)	SRTM DEM, spatial Resolution 90 m	
Viewshed (VS)	SRTM DEM, spatial Resolution 90 m	If the viewshed increases, so will tourism potentialities and vice versa (Sahani, 2020)
Forest area (FA)	India State Forest Report 2019	Forest areas, wetlands, rainfall variation, and reserved forest positively enhance tourism potential (Deribew et al., 2022)
Wetland (WL)	Assam Project On Forest and Biodiversity Conservation (Apfbc) Society https://apfbcs.nic.in/apfbcs/wetland/annexure2.pdf	
CV of Rainfall (CVR)	Climate Research and Service Unit, Pune, (2020)	
Reserve Forest (RF)	Assam Project On Forest and Biodiversity Conservation (Apfbc) Society https://apfbcs.nic.in/apfbcs/wetland/annexure2.pdf	
Population Density (PD)	Statistical Hand Book of Assam, 2016	The population density and growth rate indicate a balanced population structure; therefore, it motivates the positive outcomes of tourism potentiality (Fakfare et al., 2020)
Population Growth Rate (PGR)	Statistical Hand Book of Assam, 2016	
Literacy Rate (LR)	Statistical Hand Book of Assam, 2016	The literacy rate and the RRD positively motivate the potential value of tourism
Road & Railway Density (RRD)	Open Street Map	

Multicollinearity is essential before developing a particular model (Memon et al., 2019). A high correlation between two or more dependent variables disturbs the model's predictability. This research used the tolerance and Variance Intrusion Factor (VIF) methods to detect it. The formula was as follows (Eq. 1 and Eq. 2):

$$Tolerence\ of\ i-th\ predictor\ variable\ (T_i) = 1 - R_i^2 \quad (1)$$

$$VIF\ of\ i-th\ predictor\ variable\ (T_i) = \frac{1}{1-R_i^2} \quad (2)$$

where, R_i^2 depicts the coefficient of determination in the regression equation. For independence, the tolerance level should be more than 0.10, and the VIF should be less than 10.0.

3.3 Third Step: Spatial Modelling of TPZ

In the third step (Figure 1), TPZs were predicted using the single classification tree model (ctree model), bagged CART (treebag model), random forest model (rf model), random forest with bagging ensemble algorithms using conditional inference tree (cforest model), and gradient boosting (gbm model) models. Further, combining the ctree, treebag, rf, cforest and gbm models, a new ensemble model was prepared. For this research, we have set the scenario as the 10-fold cross-validation with repeats in train control. Here, the objective with a cross-validated data set was to optimize and determine the size of the tree by tuning the complexity parameters.

3.3.1 Conditional Inference Tree (ctree Model)

The ctree model is a nonparametric class of regression trees embedding tree-structured regression models into a well-defined theory of conditional inference procedures. It applies to all kinds of regression problems, including multiple response scales of covariates (Hothorn et al., 2015; Fu, 2017). The response *Y* is given the status of m covariates employing tree-structured recursive partitioning. The m dimensional covariate vector $X = (X_1, \ldots\ldots\ldots, X_m)$ was taken from the sample space $\aleph = \aleph_{1\ldots\times\ldots\ldots\ldots\ldots} \times \aleph_{m.}$Here, both response and covariates can be measured on arbitrary scales. The conditional function distribution D(Y/X) of the response Y given the covariates X depends on the function *f* of covariates (Eq. 3) (Hothorn et al., 2015).

$$D(Y|X) = D(Y|X_1, \ldots\ldots\ldots\ldots\ldots\ldots X_m) = D(Y|f(X_{1,\ldots\ldots\ldots\ldots}X_m) \quad (3)$$

where we restrict ourselves to portion-based regression relationships, i.e., r disjoint cells, $B_{1,\ldots\ldots\ldots}B_{r,}$partitioning covariate space, i.e., $\aleph = U_{K=1}^{r}B_{K.}$

The regression relationship should be fitted on a learning sample. L_n i.e., learning samples with n independent observations, possibly with the same coordinates X_{ij} missing (Eq. 4).

$$L_n = \{(Y_{i,\ldots\ldots}X_{1i,\ldots\ldots\ldots\ldots,}X_{mi}); i = 1, \ldots\ldots, n\} \quad (4)$$

A genetic algorithm for the recursive binary partitioning for a given learning sample L_n can be formulated using non-negative integer-valued case weights $w = (w_1, \ldots\ldots\ldots\ldots, w_n)$. Each tree node is attached with non-zero or zero weights, as the case may be.

3.3.2 Bagged CART Model (treebag Model)

Bagging, which stands for Bootstrap Aggregating, is a technique used to enhance the stability and accuracy of machine learning algorithms, in particular for decision trees, such as CART (treebag model) (Vrontos et al., 2021). This research uses the following steps: (a) multiple bootstrap samples were created in each iteration. Each sample was a random sample with replacement and had the same size as the original dataset. (b) A decision tree model was trained for hundreds of bootstrap samples for CART. These models differed slightly due to the differences in the training samples. After training all of the decision trees, the predictions for the new data were made by aggregating the predictions of each tree. The classification tasks in this research were done by majority voting – the class out of all that had the most votes. It helps to reduce overfitting because bagging trains many models on slightly different datasets and aggregates all their predictions. So, it lowers the variance of the whole final model (Choi & Hur, 2020; González et al., 2020).

3.3.3 Random Forest Model (rf Model)

One of the most popular supervised learning models is rf, which is used as a classification model in this research. Breiman (2001) first developed the rf model algorithm with the help of decision trees (Zhang et al., 2017; Gayen et al., 2019). Its basis is the ensemble learning method, in which a user can "tie" multiple classifiers to address challenging tasks and enhance the model's performance. One of the assets of Random Forest is essentially a decision tree. The final decision is made by result polling from different trees and selecting the most popular one. Most of the outcomes of each tree generate the end output. The random forest model provides more accuracy than the other model and provides a clear and separate distribution plot of features in each class

(Couronné et al., 2018; Fox et al., 2017). It can handle the missing data effectively; the developed model can be saved with the new data for future use. In this research, the following steps were followed to build a Random Forest model:

- First of all, 'K' features from total m features were randomly selected, where k<m
- Node 'd' was calculated among 'K' features.
- Split the node into several daughter nodes using the best-split method.
- Repeated the previous steps until reaching the 'I' number of nodes.
- Built a forest by repeating all steps for 'n' number of times to create 'n' number of trees.
- The mean squared error of each decision tree with their OOB samples (E_{OOB}) is used to calculate the learning error.

The pros of this approach are – (i) that it can deal with very voluminous data of sufficiently large dimensions and escape the risk of overfitting (Naghibi et al., 2017), (ii) this does not imply additional assumptions about the factors to be manipulated and the result (Youssef et al., 2016), (iii) the analyst needs a working dataset, as there is no transformation and scalability before the procedure (Gayen et al., 2019) and (iv) the model is readily applicable to any regional scale (Pourghasemi & Rahmati, 2019).

3.3.4 Conditional Inference Random Forest (cforest Model)

The cforest (Conditional Inference Random Forest) combines random forest and bagging ensemble algorithm implementation in this research. The main advantage of the Cforest is that it uses conditional inference trees as its base learners (Naghibi et al., 2017). Furthermore, Cforest uses out-of-the-bag (OOB) data; although it means more information and better accuracy, it is slower and can handle less data for the same memory (Thanh et al., 2022). The weighted average of the trees was used in this research to open the final ensemble. The main reason for the enhanced reliability of cforest predictions is that it produces unbiased trees (Strobl et al., 2007; Mogensen et al., 2012). The cforest is always better when the model has computational resources.

3.3.5 Gradient Boosting Model (gbm model)

The gbm offers a powerful technique for tackling regression and classification problems, leveraging ensemble learning by strategically combining multiple weak algorithms into a robust solution (Islam et al., 2024). This method was applied in this research. By employing decision trees as its basic components, this approach incrementally builds a predictive model via an iterative process and its' accuracy is gradually improved as each new element is incorporated into the evolving whole (Zhang et al., 2017; Sachdeva & Kumar, 2021). Basic learners are established in the initial phase, which are shallow decision trees and are a common choice owing to their simplicity, i.e., they serve as a foundation.

Subsequently, via gradient boosting, new weak learners are fitted in further rounds to the residual errors of their predecessors, which reduces the discrepancies between actual and predicted outcomes turn by turn. By minimizing collective error residuals, each additional tree is tuned to those of earlier stages, and the ensemble members are combined in an optimized manner that strengthens overall predictive capabilities with every included model. This model uses gradient descent optimization to minimize the loss function (Ridgeway, 2007). At every iteration, it computes the gradient of the loss function to the predictions of the ensemble, and then it updates the projections in the direction that minimizes the loss (Lu et al., 2020). To fight overfitting, the gbm model also uses shrinkage or learning rate. Instead of adding a smaller number of trees, a contribution of less than one scale of each tree is added to the ensemble. Smaller contributions will make the optimization faster and more conservative, requiring more iterations but protecting the model against overfitting. It can also use some regularization; it can be either tree pruning, which removes some of the splits, producing no positive results, or it can limit the maximum possible depth of the trees.

3.3.6 Ensemble Model

Combining the above models created and applied an ensemble model to predict TPZ. In this research, it was called the *TPZ ensemble model*. An ensemble model is an ML method that combines single models to produce a precise prediction model, which is more robust than any one model individually (Ganaie et al., 2022; Mohammed & Kora, 2023). The rationale for this ensemble model is consistent with the so-called "wisdom of the crowd," meaning that the average belief of numerous models (in this case, ctree, treebag, rf, cforest, and gbm models) is more efficient and

reliable than the view of any given model. This study used stacking or stacked generalization to construct the TPZ ensemble model. Stacking is a popular method in this research to prepare the ensemble model. Initially, base models, such as ctree, treebag, rf, cforest, and gbm, were trained. During the prediction phase, the base models make their predictions on new data, and then the meta-model combines these predictions to deliver the final output (Polikar, 2006; Opitz & Maclin, 1999; Ünlü & Xanthopoulos, 2021). Different base models capture different aspects or patterns within the data.

By using all of them, we can combine these models' strengths and negate their weaknesses. It introduces diversity among the models, which helps it be less vulnerable to overfitting (Gomes et al. 2017). If one model overfits to some patterns within the data, others might capture different patterns or generalize better towards unseen data (Dong et al., 2020; Opitz & Maclin, 1999; Polikar, 2006). Ensemble methods are more robust to noise and outliers (Mienye & Sun, 2022; Rokach, 2010). For example, if a model's prediction becomes unreliable when the input data falls outside the training data, ensembles can enhance this prediction by incorporating reliable predictions from other models based on the same input data (Blockeel, 2011; Rincy & Gupta, 2020; Ganaie et al., 2022). This research computed all models in the RStudio Version 2023.12.1 with Intel(R) Core (TM) i5-9300H CPU @ 2.40GHz 2.40 GHz processor with 8GB RAM and 64-bit operating system.

3.5 Fourth Step: Accuracy Assessments

One of the most crucial aspects of model building is the evaluation of the output models' precision (Das, 2020). The TPZ was validated using the Kappa Coefficient, Accuracy and AUC-ROC curve. The ROC-AUC shows how specificity and sensitivity are traded off. The ROC is a two-dimensional graph in which the x-axis depicts the specificity, and the y-axis depicts the sensitivity. Equations 5, 6 and 7 illustrated the attributes of the x and y axis, where the TN represents true negative, FP represents false positive, TP represents true positive, and FN represents false negative (Eq. 5, Eq. 6 and Eq. 7) (Roodposhti et al., 2017).

$$x = specificity = \left[\frac{TN}{(TN + FP)}\right] \tag{5}$$

$$y = sensitivity = \left[\frac{TP}{(TP + FN)}\right] \tag{6}$$

$$Accuracy = \left[\frac{TP + TN}{(TP + TN + FP + FN)}\right] \tag{7}$$

The Kappa coefficient was defined as (Eq. 8).

$$Kappa = \left[\frac{P_0 - P_{est}}{1 - P_{est}}\right] \tag{8}$$

where, P_0 is defined as the observed agreement and P_{est} is the expected agreement.

The model's performance is quantitatively depicted by the area under the ROC curve (AUC) (Tang et al., 2020). A standard scale of AUC is (i) ≥ 0.9 denotes excellent, (ii) 0.8 to 0.9 denotes accepted, (iii) 0.7 to up to 0.8 denotes excellent or satisfactory, (iv) 0.5 to 0.7 is considerable and (v) less than 0.5 is rejected (Trabelsi et al., 2023; Mitra et al., 2022). It is recommended that the machine learning models should be judged based on the validation or test dataset (Vabalas et al., 2019). The validation sets are commonly used for hyperparameter tuning, where different hyperparameter configurations of the model are tested to find the best-performing one. This ensures the model's performance is optimized for the specific dataset while maintaining its generalization ability.

4.0 Results and Discussion

4.1 Analysis of Multicollinearity

For all criteria, the tolerance level fluctuated from 0.322 (for PD) to 0.977 (for AS), and the VIF varied from 1.217 (for RF) to 2.839 (for FA) (Table 2). The highest VIF assures the lowest tolerance level. Here, all VIF values were <10, and all tolerance level values were <1.0. Therefore, the findings demonstrate the absence of multicollinearity in the 11 investigated preprocessing factors.

Table 2. Tolerance level and Variance Intrusion Factor for different indicators.

Criteria	Tolerance	VIF
RL	.491	2.036
AS	.977	1.023
CVR	.495	2.022
FA	.352	2.839
VS	.595	1.681
RF	.822	1.217
WL	.756	1.323
PGR	.496	2.017
LR	.458	2.182
PD	.322	3.102
RRD	.814	1.229

4.2 Spatial Interrelationship Between Tourism Location and Causative Factors

The distribution of tourism locations and causative factors was illustrated in this research using the AHP model. The AHP is an unbiased multicriteria decision-making method for selecting the best option from many alternatives (Munier & Hontoria, 2021; Senapati & Das, 2021). The AHP method was proposed by Saaty (1980, 1987), and it attracted the attention of numerous researchers owing to its adaptability and usefulness. Initially, the pairwise comparative matrix, which represents the relative priorities of each criterion and contains an identical number of rows and columns, was prepared. The five-membered expert panel chose the significance of several criteria in this case. The panel included experts and researchers with at least five years of experience in travel and tourism. In a separate consent form, the consulted experts agreed that the scores would be used only for academic purposes without disclosing their identities. The preference of each criterion was estimated using a relative dominance scale of 1–9 (Saaty, 1980). Here, 1, 3, 5, 7, and 9 were marked as equal, moderate, strong, very strong, and extreme or substantial importance, respectively. In contrast, 2, 4, 6, and 8 represented intermediate values. The necessary condition for a considerable AHP matrix is that the consistency ratio should be <0.1.

The RL of Assam varied from 1 meter to 1971 meters (Figure 2a). The study area's Western and Eastern sections had lower elevations (i.e.,1 meter to 300 meters). The southern sections are comparatively high (i.e., 300 meters to 1971 meters). The tourism potentiality decreases with the increasing relief value. As a result, priority rises as the class value of the relief decreases, and vice

versa. The 1 to 300-meter relief class had higher coverage (87.055% area). Moreover, an inverse trend was observed between very high RL and tourist locations. Therefore, tourism potentiality pixels overlapped substantially in the moderate to low RL classes. The AS fluctuated from -1 to 359.458 (Figure 2b).

Therefore, lower AS is gaining increased attention in the case of tourism and its potential. In this study, the AS was divided into four categories, and its importance increased as the class value decreased. The first category (−1 to 89.211) was directed toward flat northern, northeastern, and eastern directions. The Brahmaputra River passes through the middle portion of Assam, and the lowest AS value was marked along the river (1 to 89.211). The second category (89.212–179.424) was marked toward the eastern, southeastern, and southern directions. The third category (179.424–269.636) was directed toward the southern, southwestern, and western directions. The fourth category (269.637– 359.848) was marked toward the western, northwestern, and northern directions. As the distance from the river increased, the AS increased Tourism potentiality pixels overlapped largely in moderate to low AS (i.e., 194 pixels). The FA fluctuated from 4.52% to 86.07% within the study area. Large FAs were marked in the districts of Karbi Anglong, West Karbi Anglong, Dima Hasao, Cachar, Hailakandi, and Karimganj.

Moreover, 30%–50% of FAs were present in the Kakrajhar, Chirang, Golpara, Kamrup, Karimganj, and Tinsukia districts. The remaining districts were marked by comparatively lower FAs (i.e., 4.52%–30%). Picturesque attractiveness is positively accelerated by the FA, attracting many adventure travellers (Karali et al., 2021). The FA was categorized into four groups in this study, and as the category value increased, the weights increased and vice versa (Figure 2c). Moderate to high forest cover attracted more tourism potentiality pixels (i.e., 162 pixels). The number of RFs in Assam fluctuated from 0 to 29 (Figure 2d). A comparatively higher number of RFs (i.e., 16–29) was found in the West Karabi Arlong, Naogaon, Karbi Arlong, Tinsukia, and Kamrup metropolitan districts. The remaining districts were marked by fewer numbers (0–15) of RFs. The number of reserved forests exerts a favourable impact on tourism potential. Here, in this research, the RF was classified into four classes, and the class value increased, weightage increased, and vice versa (Figure 2d). Moderate and high classes of RF incur higher tourism potentiality pixels in the study area (i.e., 234 pixels).

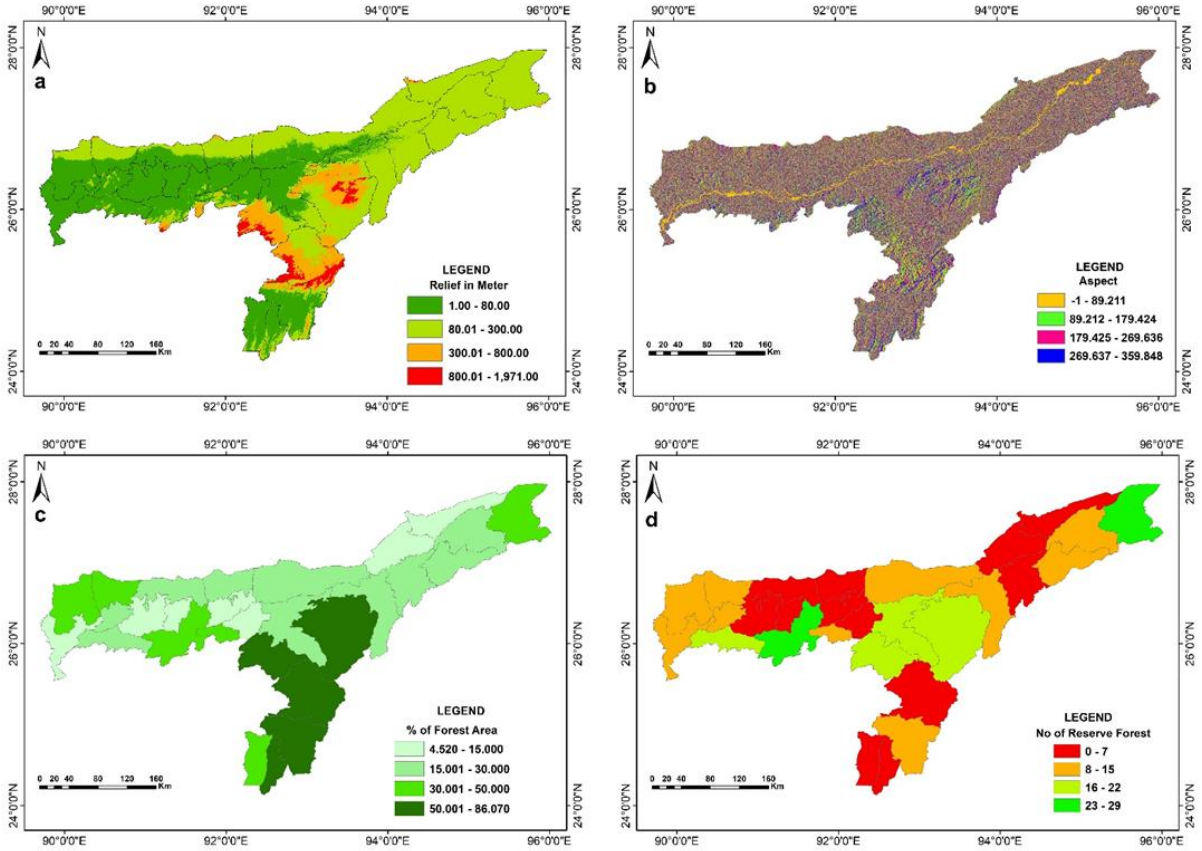

Figure 2. Criteria of the study area: (a) Relief, (b) Aspect, (c) % of forest area, and (d) Number of reserved forests.

The number of WL in the study area varied from 0 to 1790 (Figure 3a). The higher number of WLs (i.e., 251 to 1790 numbers) were identified in the districts of Kakrajhar, Karbi Arlong, Naogaon, West Karbi Arlong, Sontipur, Hailakandi, Karimganj and Tinsukia districts. The remaining portions were noticed to have a lower number of wetlands. The WL favourably vibrates the tourism potential. In this study, the WL was categorized into four classes, and the priority increased with increasing class value. Furthermore, tourism potentiality pixels merged substantially in very high, high, and moderate class values of WLs (i.e., 228 pixels).

The CVR fluctuated from 86.401% to 108.007% (Figure 3b). Most parts of Assam experienced considerable variations in rainfall. Comparatively higher amounts of rainfall variation (92.889%–108.006%) were noted in the districts of Dhubri, Kokrajhar, Golpara, Bongajagaon, Barpeta, Chirang, Baksa, Nalbari, Kamrup, Karabi Arlong, Darang, Kamrup Metropolitan, Morigaon, Sontipur, Lakhimpur, Dhemji, Naogaon, Dima Hasaao, Cachar, Hailakandi, and

Kamrup. Lower amounts of rainfall variation were seen in the remaining districts. The CVR positively enhanced tourism potential (Giorgi and Lionello 2008). Four categories were used to classify the CVR in this instance, and as the value of each category increased, so did the priority and vice versa. In addition, tourism location pixels merged most substantially in very high, high, and moderate class values of CVR (i.e., 228 pixels).

The VS of the study area was marked by identifying 17 major hills (Figure 3c). Here, the VS was categorized into four classes. The Western sections of the study area were more prominent. The VS positively enhances the potential for tourism (Giorgi and Lionello 2008). In this instance, the VS was split into four categories, and priority increased as the class value increased. Tourism potentiality pixels overlapped mainly in very high, high, and moderate classes of VS (i.e., 273 pixels).

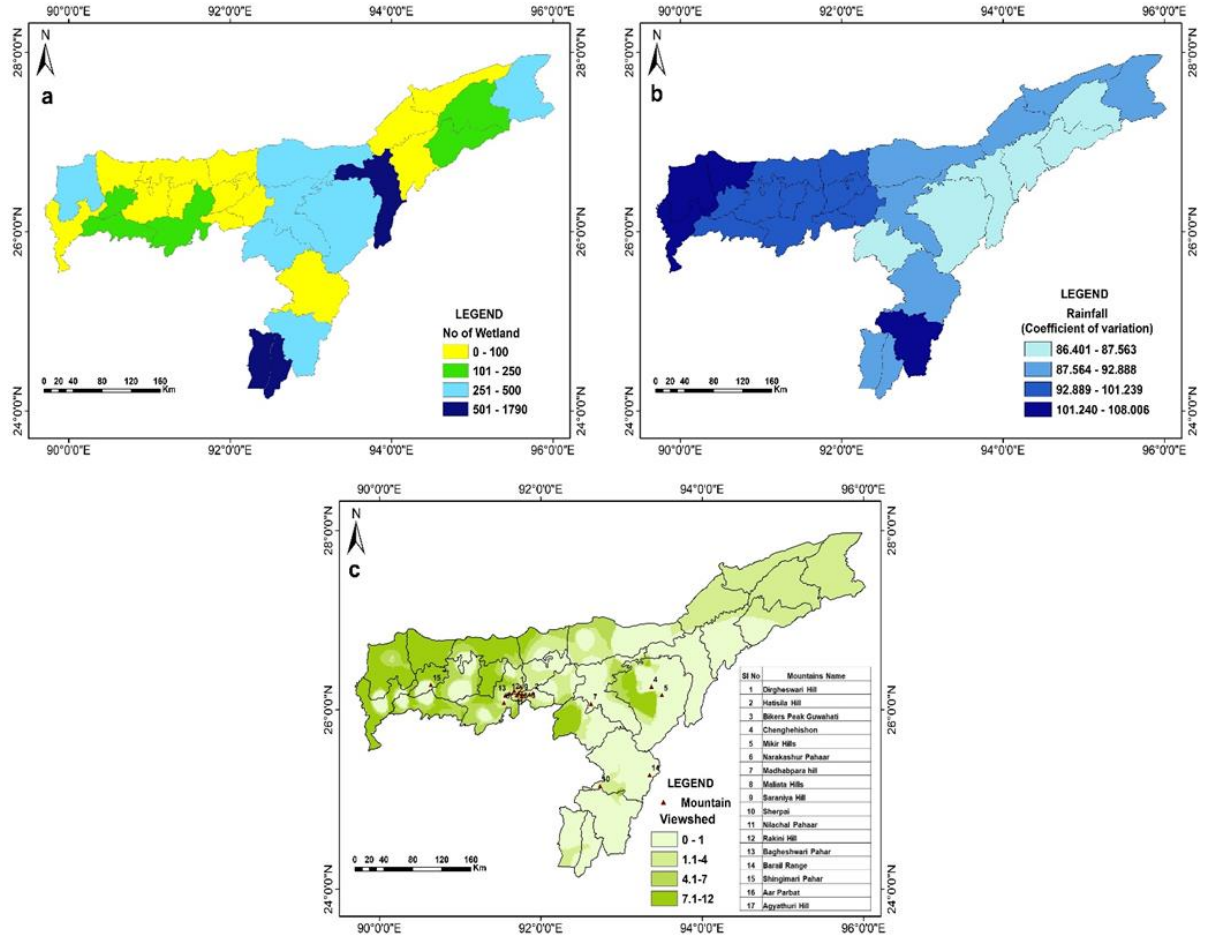

Figure 3. Criteria of the study area: (a) Number of wetlands, (b) Coefficient of variation of rainfall, and (c) Viewshed.

The PGR of the research area varied from 5.210% to 24.440% (Figure 4a). Comparatively higher PGR values (i.e., 20–24%) were identified in the Dhubri, Bongaigaon, Golpara, Barpata, Darrang, Kamrup Metropolitan, Morigaon, Cacchar, Hailakandi, and Karimganj districts. The remaining districts exhibited a growth rate of 5.210%–20%. The PGR, in this instance, was divided into four categories, and the priority increased as the quantity of each class increased and vice versa. High, very high and moderate classes of PGR were noted for a high number of pixel counts of tourism potentiality (i.e., 311 pixels).

The PD of the study area varied from 44 to 1313 individuals per square kilometer. A comparatively higher PD value (i.e., 701–1313 people/sq.km) was observed in the Dhubri, Barpata, Nalbari, Kamrup Metropolitan, and Naogaon districts. The remaining districts had 44–700 people/sq.km. (Figure 4b). High PD negatively affects tourism potentiality. Therefore, the PD, in this instance, was divided into four distinct classes, and its importance increased as the class value decreased and vice versa. Substantial tourism potentiality pixels overlapped in the low and moderate PD classes (i.e., 315 pixels).

The LR varied from 65.37% to 88.71% within the study area of Assam. Comparatively lower LRs were marked in the Dhubri, Chirang, Sontipur, Baksa, Golpara, Jorhat, and Tinsukia districts. On the contrary, higher LRs were identified in the Darrang, Kamrup Metropolitan, Morigaon, Golaghat, Jorhat, Sivsagar, Karbi Anglong, Hailakandi, and Karimganj districts (Figure 4c). As tourism is a tertiary activity, the LR would positively impact tourism activities. Here, the LR was classified into four groups, and as the value of each class increased, so did the priority and vice versa. High, very high, and moderate classes of LR combined achieve a higher number of pixels for tourism potentiality (i.e., 310 pixels).

Road and railway networks connect the tourism destinations quickly and smoothly (Bast et al. 2016). Therefore, the RRD was divided into four categories. As the class value increased, so did the priority and vice versa. Moderate, high, and very high classes of RRD combined overlapped with higher tourism potentiality pixel values (i.e., 249 pixels). The detailed areal coverage of different thematic layers, their classes and weights were marked in Table 3.

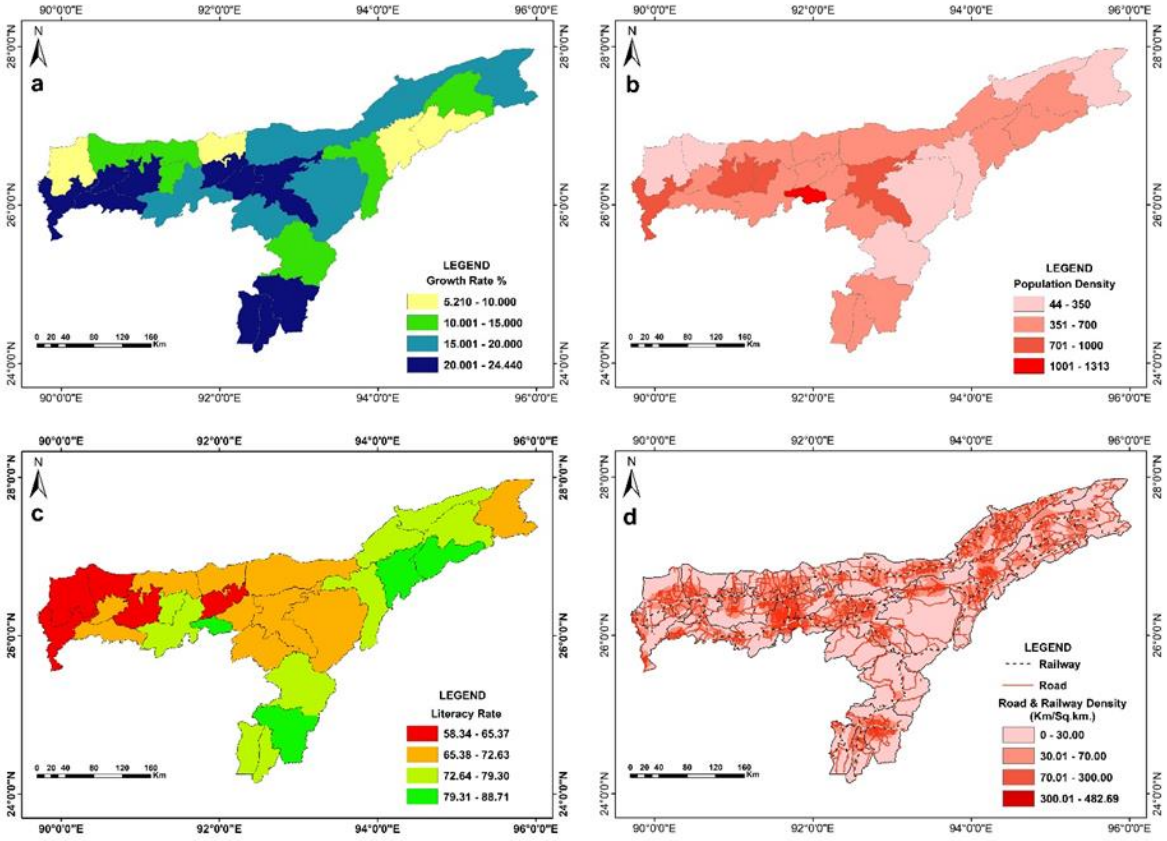

Figure 4. Criteria of the study area are (a) growth rate (%), (b) population density, (c) literacy rate (%), and (d) road and railway density.

Table 3. Relationship between the variables with the help of the AHP method.

Indicators	Class	Category				Priority % (Weightage)	Principal Eigenvalue & Consistency Ratio (CR.)	Number of Pixels in the domain	Number of tourism potentiality pixels in the domain	% Area
RL		**1.00 - 80.00 (Low)**	**80.01 - 300.00 (Moderate)**	**300.01 - 800.00 (High)**	**800.01 - 1971.00 (Very High)**					
	1.00 - 80.00	1	3	4	5	(55.00%) 8	Principal Eigenvalue = 4.057, Consistency Ratio CR = 2.1%	3696933	140	39.234
	80.01 - 300.00	0.33	1	2	2	(21.40) % 6		44485020	178	47.281
	300.01 - 800.00	0.25	0.5	1	2	(14.20%) 4		9533666	37	10.249
	800.01 - 1971.00	0.2	0.5	0.5	1	(9.40%) 2		2508725	10	2.697
AS		**-1 - 89.211 (Low)**	**89.212 - 179.424 (Moderate)**	**179.425 - 269.636 (High)**	**269.637 - 359.848 (Very High)**					
	-1 - 89.211	1	2	5	7	(54.50%) 8	Principal Eigenvalue = 4.072, Consistency Ratio CR = 2.6%	23798802	80	25.583
	89.212 - 179.424	0.5	1	2	3	(24.70%) 6		24023779	114	25.825
	179.425 - 269.636	0.2	0.5	1	3	(14.10%) 5		22381165	95	24.059
	269.637 - 359.848	0.14	0.33	0.33	1	(6.70%) 3		22820598	73	24.532
FA		**50.001 - 86.070 (Very high)**	**30.001- 50.00 (High)**	**15.001 - 30.00 (Moderate)**	**4.520 - 15.000 (Low)**					
	50.001 - 86.070	1	2	5	7	(53.20%) 8		7858	101	25.856
	30.001 - 50.00	0.5	1	3	3	(27.30%) 6		5793	62	19.062

	15.001 - 30.000	0.2	0.33	1	3	(12.80%) 5	Principal Eigenvalue = 4.118, Consistency Ratio CR = 4.3%	12149	134	39.976
	4.520 - 15.000	0.14	0.33	0.33	1	(6.70%) 3		4591	66	15.106
WL		**501 – 1790 (very high)**	**251 – 500 (High)**	**101 – 250 (Moderate)**	**0 – 100 (Low)**					
	501 - 1790	1	3	4	7	(56.30%) 8	Principal Eigenvalue = 4.063, Consistency Ratio CR = 2.3%	2616	30	8.608
	251 - 500	0.33	1	2	3	(22.30%) 6		11800	142	38.827
	101 - 250	0.25	0.5	1	3	(14.80%) 4		4737	56	15.587
	0 - 100	0.14	0.33	0.33	1	(6.70%)3		11238	134	36.978
CVR		**86.401 - 87.563 (Low)**	**87.564 - 92.888 (Moderate)**	**92.889 - 101.239 (High)**	**101.240 - 108.006 (Very high)**					
	86.401 - 87.563	1	2	4	7	(52.50%) 7	Principal Eigenvalue = 4.050, Consistency Ratio CR = 1.8%	4293	87	14.126
	87.564 - 92.888	0.5	1	2	3	(25.40%) 6		6896	124	22.691
	92.889 - 101.239	0.25	0.5	1	3	(15.20%) 5		8816	78	29.009
	101.240 - 108.006	0.14	0.33	0.33	1	(6.90%) 3		10386	73	34.175
VS		**7.1 – 12 (Very High)**	**4.1 – 7 (High)**	**1.1 - 4 (Moderate)**	**0 – 1 (Low)**					
	7.1 - 12	1	3	4	8	(57.20%) 8	Principal Eigenvalue = 4.052,	5051	60	16.669
	4.1 - 7	0.33	1	2	3	(22.00%) 7		2932	37	9.676
	1.1 - 4	0.25	0.5	1	3	(14.50%) 5		8614	176	28.428
	0 - 1	0.12	0.33	0.33	1	(6.30%) 3		13704	92	45.226

							Consistency Ratio CR = 1.9%			
RF		**23-29 (very high)**	**16-22 (High)**	**8-15 (Moderate)**	**0-7 (Low)**					
	23-29	1	3	4	8	(56.40%) 8	Principal Eigenvalue = 4.048, Consistency Ratio CR = 1.7%	2788	31	9.174
	16-22	0.33	1	2	5	(24.20%) 6		2258	26	7.430
	8-15	0.25	0.5	1	3	(13.90%) 4		14813	177	48.741
	0-7	0.12	0.2	0.33	1	(5.40%) 3		10532	128	34.655
PD		**44 – 350 (Low)**	**351 – 700 (Moderate)**	**701 – 1000 (High)**	**1001 – 1313 (Very High)**					
	44 - 350	1	2	4	5	(50.70%) 8	Principal Eigenvalue = 4.021, Consistency Ratio CR = 0.8%	10642	108	35.017
	351 - 700	0.5	1	2	3	(26.40%) 6		15875	207	51.940
	701 - 1000	0.25	0.5	1	2	(14.30%) 4		3651	43	12.013
	1001 - 1313	0.2	0.33	0.5	1	(8.60%) 3		313	04	1.030
PGR		**20.001 - 24.440 (Very high)**	**15.001 - 20.000 (High)**	**10.001 - 15.000 (Moderate)**	**5.210 - 10.000 (Low)**					
	20.001 - 24.440	1	2	4	5	(51.20%) 8	Principal Eigenvalue = 4.047, Consistency Ratio CR = 1.7%	8213	98	27.024
	15.001 - 20.000	0.5	1	2	2	(24.40%) 6		11272	138	37.090
	10.001 - 15.000	0.25	0.5	1	2	(14.60%) 4		6737	75	22.168
	5.210 - 10.000	0.2	0.5	0.5	1	(9.80%) 2		4169	51	13.718

LR		**79.31-88.71 (Very high)**	**72.64-79.30 (High)**	**65.38-72.63 (Moderate)**	**58.34-65.37 (Low)**					
	79.31-88.71	1	2	3	5	(48.80%)8	Principal Eigenvalue = 4.041, Consistency Ratio CR = 1.5%	313	47	1.030
	72.64-79.30	0.5	1	2	2	(25.20%) 6		4704	116	15.478
	65.38-72.63	0.33	0.5	1	2	(16.10%) 4		8539	147	28.097
	58.34-65.37	0.2	0.5	0.5	1	(10.00%) 2		16835	50	55.395
RRD		**300.01 - 482.69 (Very High)**	**70.01 - 300.00 (High)**	**30.01 - 70.00 (Moderate)**	**0 - 30.00 (Low)**					
	300.01 - 482.69	1	2	3	9	(52.70%) 8	Principal Eigenvalue = 4.021, Consistency Ratio CR = 0.8%	156	01	0.513
	70.01 - 300.00	0.5	1	1	3	(21.50%) 6		1337	16	4.399
	30.01 - 70.00	0.33	1	1	3	(19.30%) 5		9701	232	31.921
	0 - 30.00	0.11	0.33	0.33	1	(6.40%) 3		19197	113	63.167

4.3 Analysis of Different Pre-Requisites of Machine Learning Models

Before the initialization of ML models, a 10-fold repeated cross-validation framework was developed. The objective was to optimize the size of the tree. After plotting the 1-P value threshold vs. accuracy (repeated cross-validation), accuracy was maximized into a relatively less complex tree (Figure 5). The final value used was mincriterion 0.01, which was used to tune the model best.

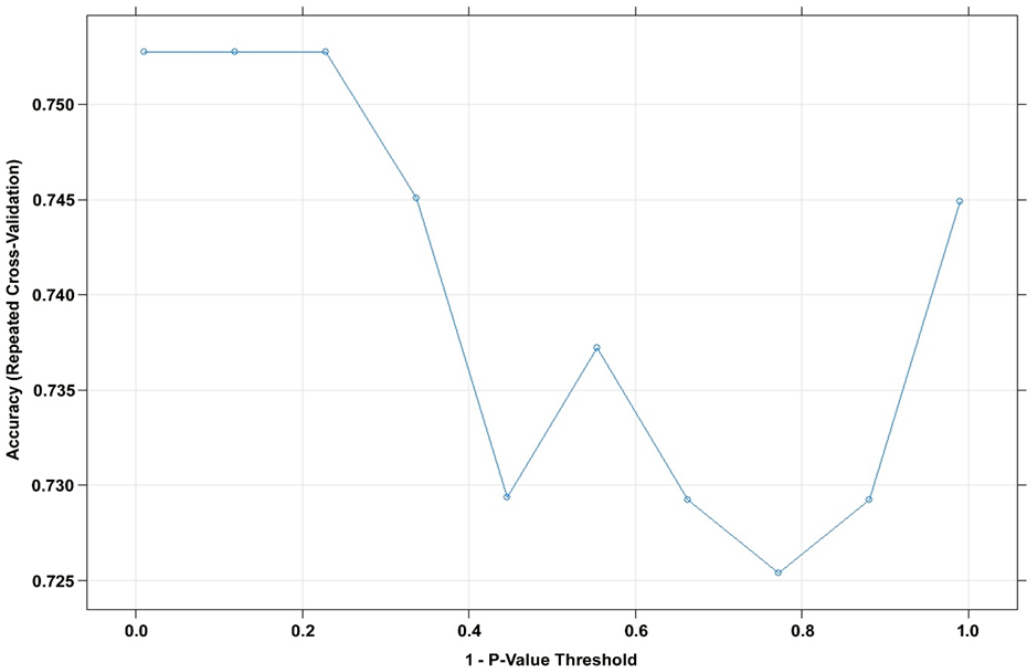

Figure 5. Threshold vs. accuracy plot.

In the final model, 16 terminal nodes were created in the regression trees (Figure 6). The rules for this ctree model are presented in Appendix. 1. Within these terminal nodes; the model predicted criterion 1 (very high to high TPZ) nine times. The final treebag model was created using 25 bootstrap replications. The final accuracy was used for the rf model to select the optimal model using the largest value. The final value for the rf model was mtry = 11 (best tuned). For this final rf model, the number of trees was 500, and the number of variables was split into six parts. The OOB error estimated for the model was 17.25%. For the cforest model, the final accuracy was used to select the optimal model using the largest value mtry = 11 (best tuned) with a total of 500 trees in the final model. The final tuned gradient-boosted model was created using the Bernoulli loss function with 50 iterations. The model had 50 trees, an interaction depth of 2, shrinkage of 0.1, and a minobsinnode of 10.

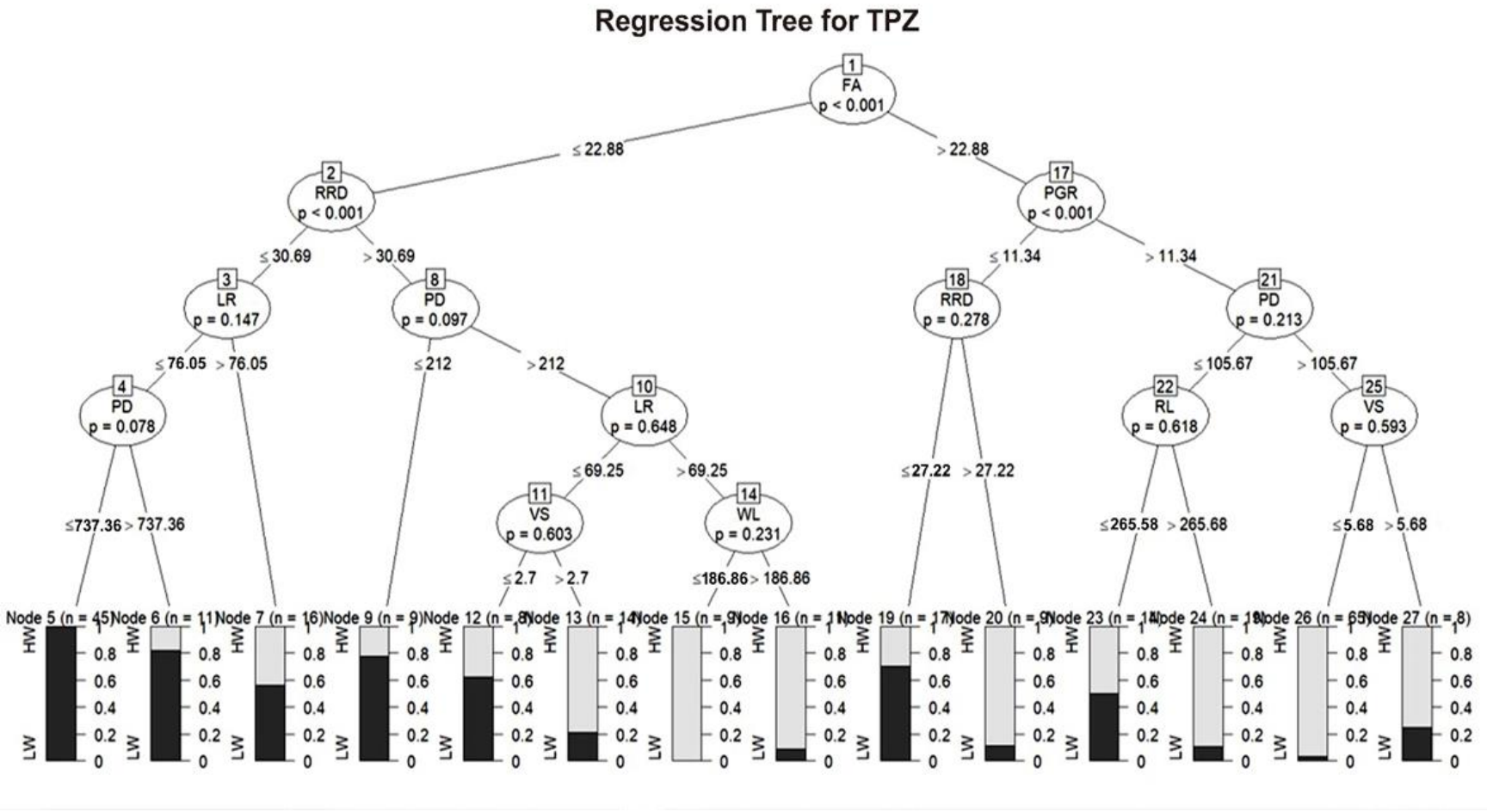

Figure 6. Regression tree for Conditional Inference Tree (ctree) model (HW – high to moderate class; LW – low class).

4.4 Variable Importance Analysis

Tourism Potentiality is a multifaceted concept dependent on several criteria (Raha & Gayen, 2021). Therefore, it is essential to determine the influential factors of tourism potentiality and their contribution. After training each model, the Variable Importance Plot (Figure 7) was assessed. The FA is most important for the ctree model (Figure 7a). The FA was followed by LR, WL, RF, RL, CVR, RRD, AS, VS and PGR (least importance). RRD was marked with the highest importance for the treebag model (Figure 7b), and PD was identified as the least important. LR, FA, RL, AS, CVR, VS, RF, PGR, WL and PD followed RRD. In the rf model (as shown in Figure 7c), the feature RRD was identified as having the most significant importance, with FA, LR, RL, AS, VS, PD, RF, CVR, WL, and PGR following in descending order of importance. For the cforest model (Figure 7d), FA was marked with the highest importance, followed by RRD, LR, PD, RF, RL, PGR, CVR, VS, WL and AS. The FA was also found to be most important for the gbm model (Figure 7e). Here, WL was found to be the least important. RRD and LR were also found with substantial importance for the gbm model. For the TPZ Ensemble Model (Figure 7f), the rf model was the most important, followed by the gbm, treebag, ctree and cforest model.

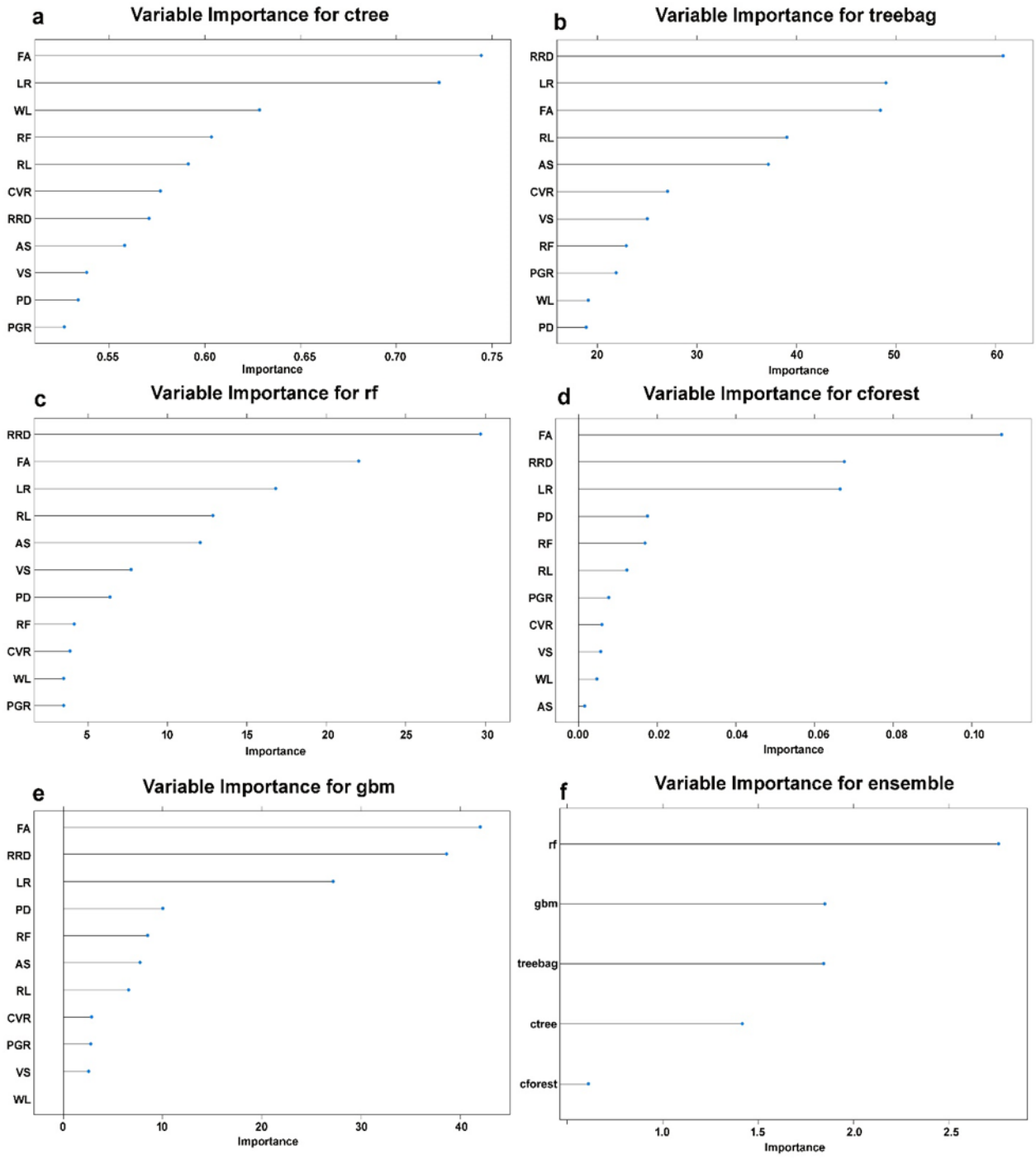

Figure 7. Variable Importance Plot of different models (a) ctree model, (b) treebag model, (c) rf model, (d) cforest model, (e) gbm model, and (f) ensemble model.

4.5 Tourism Potentiality Assessment Models

Tourism potentiality assessment models were applied in this research, which included five ML (i.e., ctree, treebag, rf, cforest, and gbm) and one ensemble model (Figures 8a, 8b, 9a, 9b, 10a, and 10b). All models showed very high, high, and moderate to low tourism potentiality per the natural break strategy (Gayen et al., 2019). Higher values indicate greater tourism potentiality.

Using the ensemble model, 35.42% of the area had high tourism potentiality. Moreover, 64.58% of the area was identified as having moderate to low potentiality. Using the treebag model, 53.61% of the area was found to have high to very high tourism potentiality, whereas 46.39% displayed moderate to low potentiality. Furthermore, using the ctree model, 50.62% of the area was restricted as having high tourism potentiality and the remaining 49.38% was marked as having moderate to low potentiality. Using the cforest model, almost 40.52% and 59.48% were identified to have high to very high and moderate to low tourism potentiality, respectively. On the contrary, using the rf model, 28.24% of the area exhibited high tourism potentiality, and the remaining 71.76% had moderate to low potentiality. Finally, using the gbm model, 36.5% of the area was restricted as having high tourism potentiality and 63.5% as displaying moderate to low potentiality (Table 4).

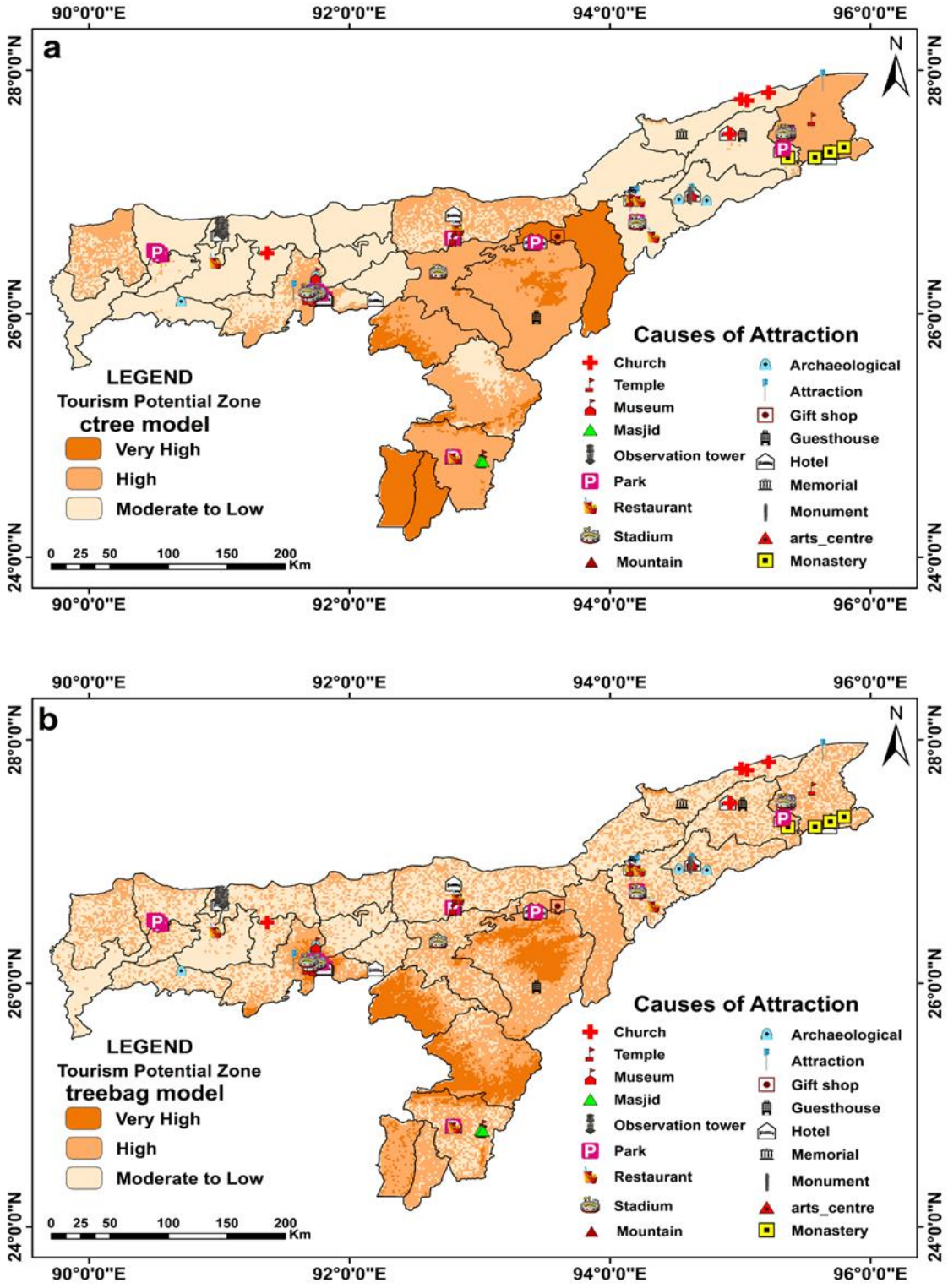

Figure 8. Tourism Potential Zone (TPZ) is predicted by (a) Conditional Inference Tree Model (ctree model) and (b) Bagged CART model (treebag model).

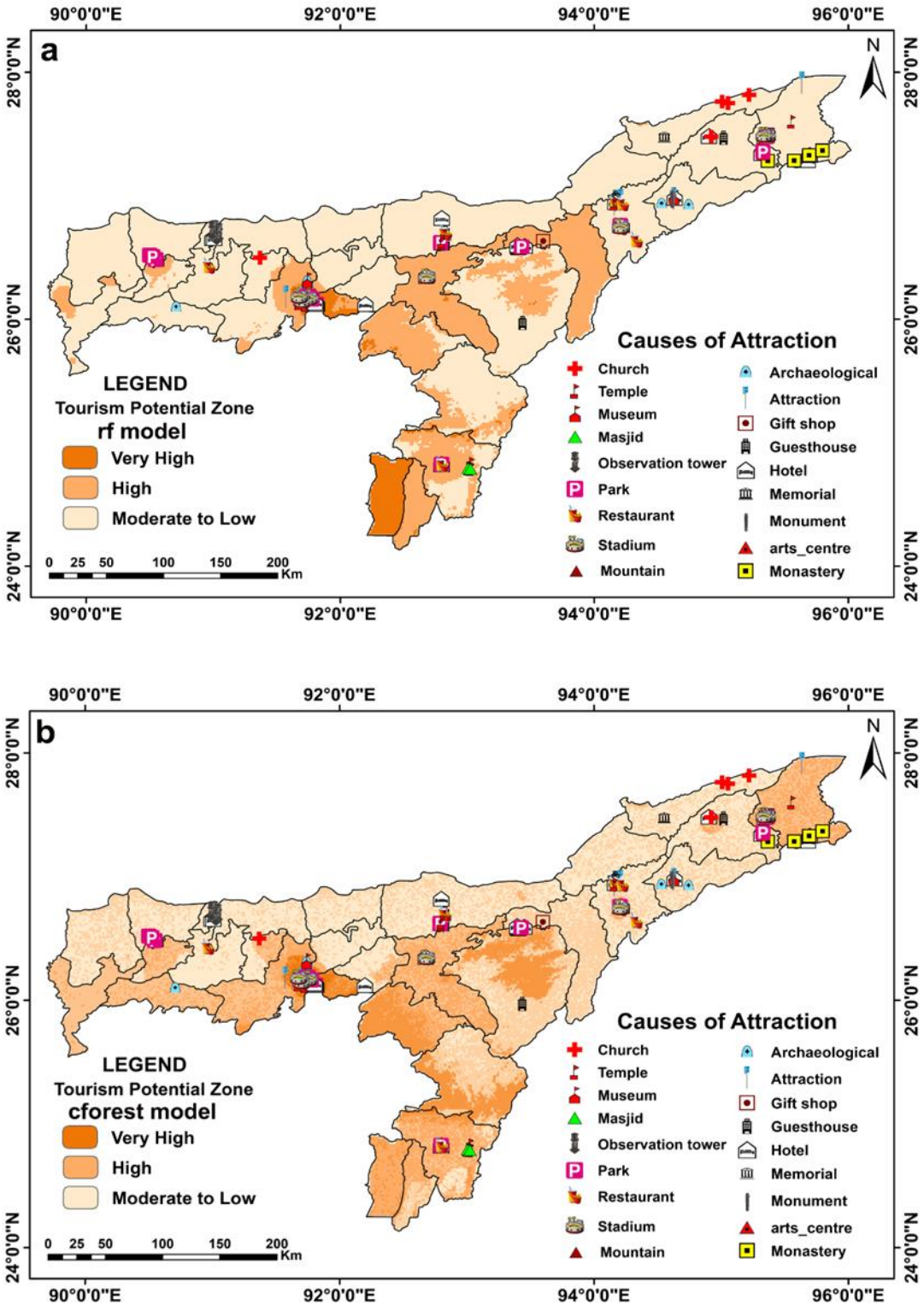

Figure 9. Tourism Potential Zone (TPZ) is predicted by (a) rf model and (b) cforest model.

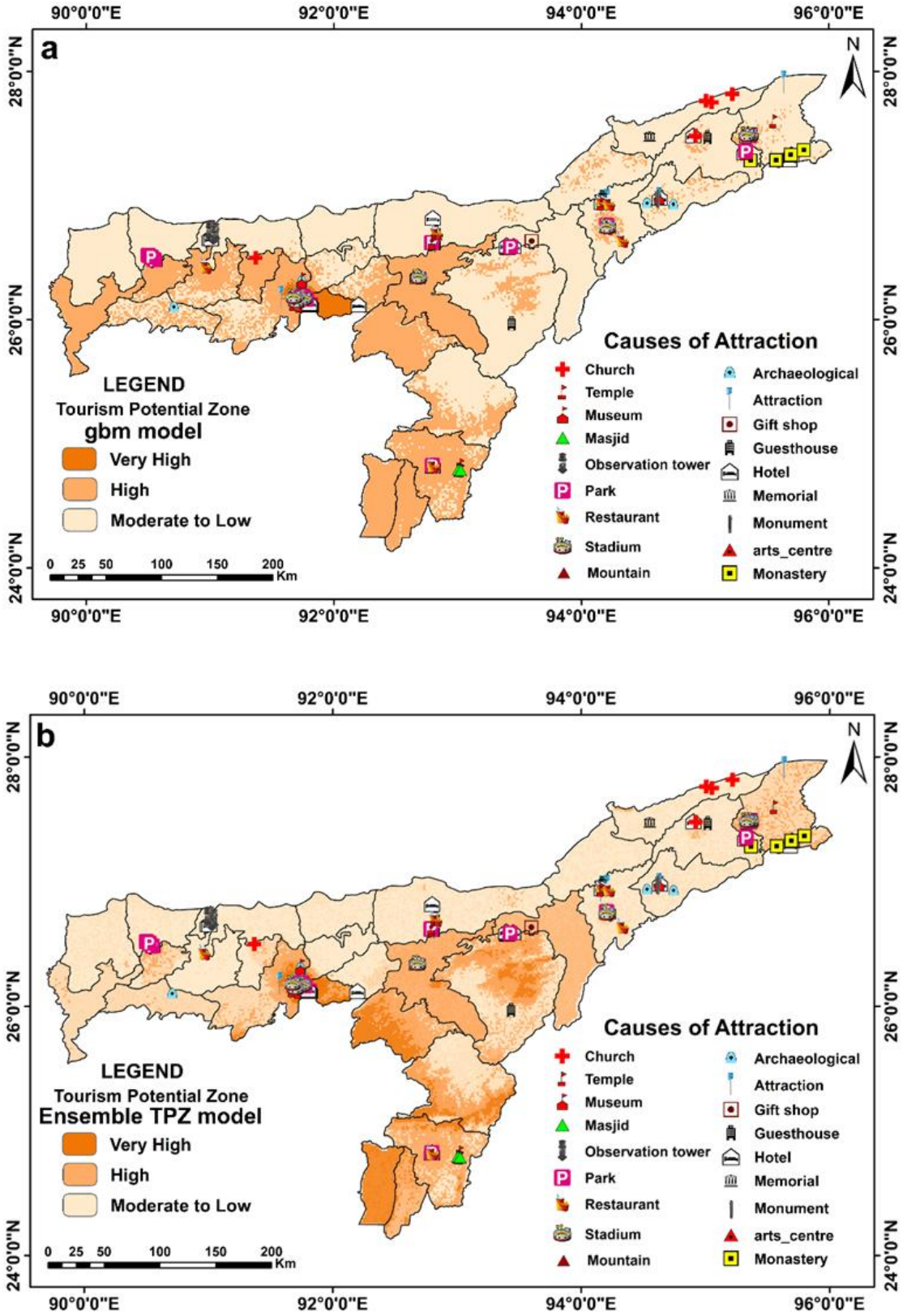

Figure 10. Tourism Potential Zone (TPZ) is predicted by (a) gbm model and (b) Ensemble model.

Table 4. Pixel count with percentage area by different models.

	Pixel Count				% Area			
Models	**Moderate to Low**	**High**	**Very High**	**Total**	**Moderate to Low**	**High**	**Very High**	**Total**
Ensemble	19339	8353	2253	29945	64.58	27.89	7.53	100
Bagged CART	13891	13152	2902	29945	46.39	43.92	9.69	100
Conditional Inference Random Forest	17811	11608	526	29945	59.48	38.76	1.76	100
Conditional Inference Tree	14787	11716	3442	29945	49.38	39.13	11.49	100
Random Forest	21489	7350	1106	29945	71.76	24.55	3.69	100
Gradient Boosting	19015	10488	442	29945	63.5	35.02	1.48	100

4.5.1 Very High (VH) and High (H) TPZ

The detailed characteristics of this zone were briefly discussed as follows:

- VH and H tourism potentialities dominated Assam's middle, southeastern, and southern sections. Over 50% of the area of Golpara, Hailakandi, Jorhat, Kamrup, Kamrup Metropolitan, Karbi Anglong, Karimganj, Lakhimpur, Naogaon, Sivasagar, Tinsukia and West Karbi Anglong districts were marked with Very high potentials for tourism (Figures 8a, 8b, 9a, 9b, 10a, 10b).
- These sections were marked with the moderate to high relief structure (80.01 meters to 1971 meters) (Figure 2a), which creates a broader viewshed of the location and attracts tourists conveniently and easily (Huff & Tingley, 2015).
- Higher number of reserved forests (8 to 29 numbers, Figure 2d) and larger forest area (30.001% to 86.07% area, Figure 2c) in these sections created an amazing ambience for the tourism activities.
- The region is dominated by the wetlands (i.e., 251 to 1790 Figure 3a) and rainfall variation (i.e., 92.889% to 108.006%, Figure 3b). A wetland is an area where the ground is continuously or periodically flooded by water, whether salty, pure, or a combination of both. The seasonal variation of rainfall accelerates a particular region's vegetation pattern and forest cover. The seasonal variation creates the rhythmic diversity of forest cover (Ghazoul, 2016).
- This section has a higher road and rail density (greater than 70, Figure 4d), accelerating the region's connectivity. Tourists can reach their destination more relatedly and easily through accessible roads and rails (Holloway & Humphreys, 2022).
- The population density (351 to 1000 persons/ sq.km., Figure 4b), population growth rate (10% to 24%, Figure 7a) and literacy rate (72.64% to 88.71%, Figure 7c) are higher in these sections of the study area. The literacy rate creates awareness about a balanced population structure, creating a favourable environment for tourism activities (Getz & Page, 2019).

Irrespective of the above issues, these portions are affected by comparatively low pollution levels, as more reserved forests and forest areas dominate them (Singh et al., 2020). As a result, these portions are marked with a higher potential for tourism.

4.5.2 Moderate (M) to Low (L) TPZ

The detailed characteristics of this zone were briefly discussed as follows:

- The upper Assam and the upper northeastern portions were marked with low tourism potential. Over 50% of the areas of Baksa, Barpeta, Bongaigaon, Chirang, Darrang, Dhemaji, Dhubri, Dibrugarh, Golaghat, Kokrajhar, Morigaon, Nalbari, Sontipur, and Udayguri were identified under the moderate to low tourism potentials (Figures 8a, 8b, 9a, 9b, 10a, 10b).
- Comparatively flat terrain (1 to 300 meters, Figure 2a, 2b, 2c) was marked in these sections, which does not create any picturesque beauty and does create a low viewshed of a particular region.
- Near about 5% to 30% forest area (Figure 2d) and a comparatively low number of reserved forests (0 to 15 numbers). These factors lower the ambience, vibrations, and motivations for tourist activities.
- These sections are scarce wetlands (i.e., 0 to 250 numbers in Figure 3a), and these sections are marked with a very low amount of rainfall variation (i.e., 86.401% to 92.888%, Figure 3b).
- Moderate to high population density (i.e., 351 to 1000 persons/square km., Figure 4b) and growth rate (i.e., 10.001 % to 20.000%) are also marked in these portions of the study area (Figure 4a). Literacy rates (i.e., 58.34% to 72.63%, Figure 4c) are also comparatively poor.
- These portions have low to moderate RRD values (0 to 70) (Figure 4d).

Apart from the above-specified issues, these portions are affected by comparatively high pollution levels, as they are crowded with different types of industries. These sections are over-congested. Therefore, these sections do not attract tourists and have comparatively low potential.

4.6 Validation

The Ensemble model appeared with the highest Kappa (0.81), accuracy (0.93 value) and AUC-ROC (96.9% area) values. Based on AUC-ROC measurement, the Ensemble model was followed by the cforest model (89.7% AUC), rf model (79.9% AUC), gbm model (78.8% AUC values), ctree model (74.9% AUC values) and treebag model (74.01% AUC-ROC). According to the quality

criteria of AUC, the performance of the ensemble model appeared to be ‘excellent’. The performance of the cforest model was ‘accepted’. The remaining models’ performances (i.e., ctree, treebag, gbm, and rf models) are ‘good’ or ‘satisfactory’. Based on the accuracy and Kappa measurement, the Ensemble model outperformed other models. The Ensemble model was followed by the rf, gbm, cforest, ctree, and treebag models. For both cases, ctree and treebag are the worst performers. On the contrary, rf, cforest and gbm models performed well. ROC-AUC and accuracy plots were outlined in Figure 11, Figure 12, and Table 5.

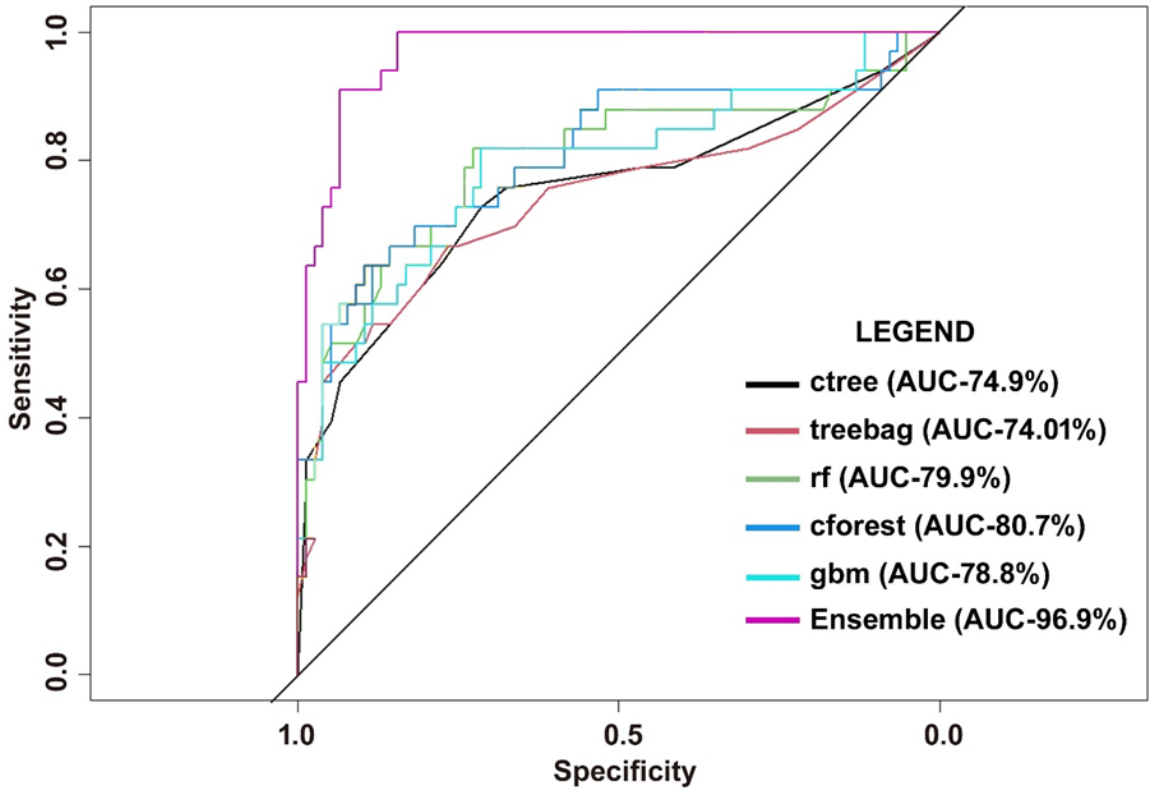

Figure 11. AUC_ROC by different machine learning models (detailed results are presented in Table 5).

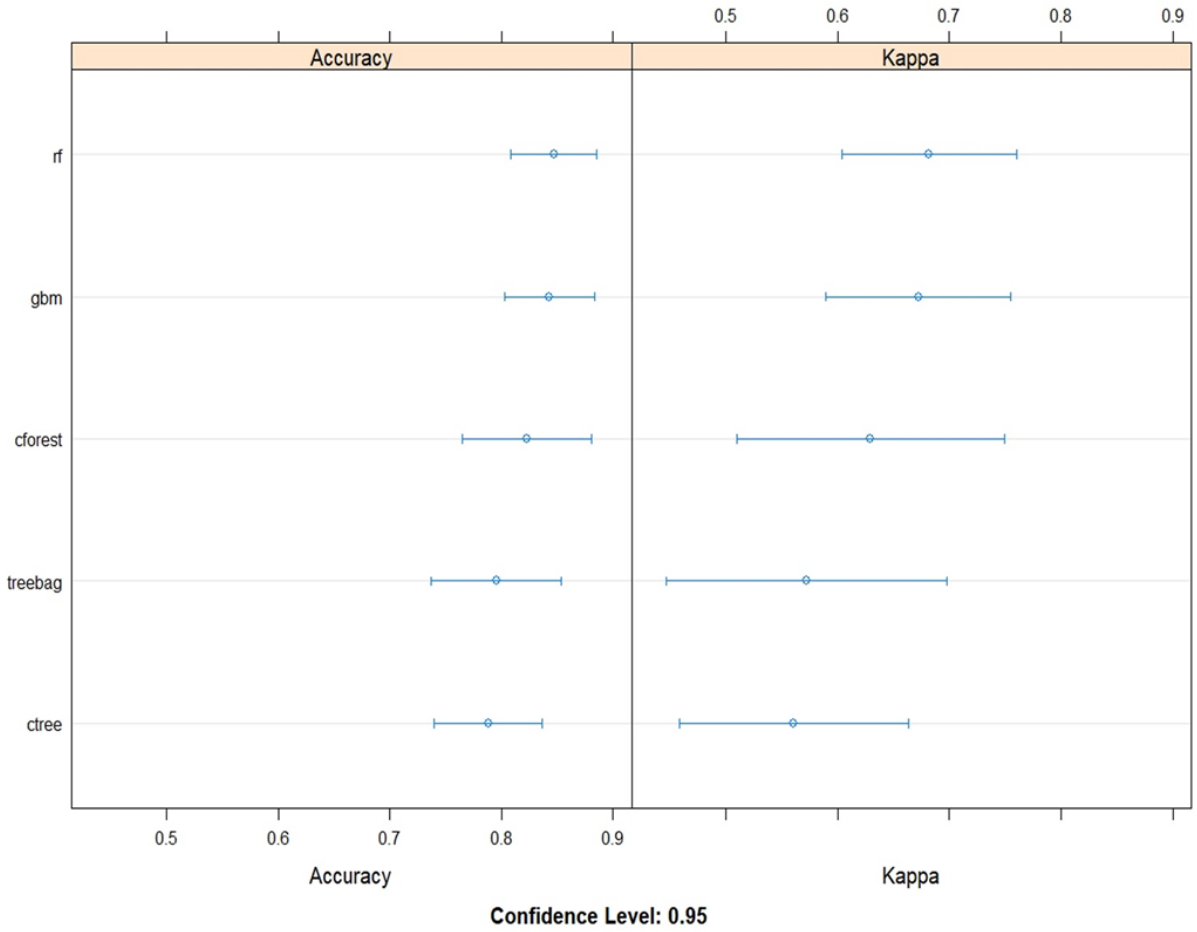

Figure 12. Accuracy and Kappa plots for different models.

s

Table 5. Accuracy of different models (predictive accuracy).

Model names	Accuracy	Kappa	Rank (Based on Accuracy and Kappa)	Model Names	AUC_ROC Values	95% Confidence Interval (For AUC)		Rank (Based on AUC)
						Upper Bound	Lower Bound	
Ensemble	0.92	0.81	1	Ensemble	96.9%	0.850	0.962	1
Conditional Inference Tree (cforest)	0.81	0.62	4	Conditional Inference Tree (cforest)	89.7%	0.723	0.878	2
Random Forest (rf)	0.85	0.68	2	Random Forest (rf)	79.9%	0.713	0.870	3
Stochastic gradient Boosting (gbm)	0.84	0.67	3	Stochastic gradient Boosting (gbm)	78.8%	0.693	0.854	4
Conditional Inference Tree (ctree)	0.80	0.60	5	Conditional Inference Tree (ctree)	74.9%	0.806	0.896	5
Bagged CART (treebag)	0.79	0.57	6	Bagged CART (treebag)	74.01%	0.693	0.895	6

5.0 Discussion

According to the accuracy assessments, the ensemble method is the most accurate approach in this research. By combining the forecasts of various bootstrapping base models, such as decision trees, random forests, gradient boosting, and ensemble methods, the shortcomings of individual models can be eliminated, and the advantages of these methods can be exploited. The ensemble method draws on the combined wisdom of multiple models to improve predictive performance. Therefore, the 'TPZ ensemble model' formulation represents a new direction in tourism research. The TPZ ensemble model outperforms individual models and demonstrates good predictive ability across various accuracy metrics. Rigorous scrutiny revealed that the TPZ ensemble model significantly outperformed its constituent parts, achieving an exceptional area under the curve (AUC) value of 96.9%. Additionally, it exhibited strong performance with a Kappa coefficient of 0.81, indicating substantial agreement beyond chance. The most remarkable finding is that while maintaining its generalization capability, the ensemble model achieved an accuracy of 92%. Furthermore, ensemble methods aid in the decision-making process by helping to select the best model from various machine learning models.

Furthermore, the rf and cforest models handled large datasets without variable deletion, which may have contributed to their strong performance in this study. According to Catani et al. (2013), the RF model can handle nonlinearities between dominant factors, providing good performance in the current context. With respectable performances, this model has also been shown to be helpful in other research domains, including mapping groundwater potential, predicting wildfires, modeling sediment yield (Masselink et al., 2017), and mapping landslide susceptibility (LSM) (Taalab et al., 2018; Zhao et al., 2017). The gradient boosting models also performed well in prediction accuracy, possibly because they combine the predictions of several base estimators, typically decision trees.

According to the present study's results, the ensemble model's accuracy was much greater than any individual model's. The high performance of these methods is scaled further by their ability to support large datasets; therefore, they can be used in applications with large datasets. However, decision trees (such as the conditional inference tree and the Bagged CART models) can isolate outliers in different leaves, which prevents them from substantially impacting the performance of the model as a whole. Additionally, ctree and the Bagged CART model cannot efficiently handle large amounts of data. These methods are also sensitive to outliers. Although the

performance of these models was quite satisfactory in this study, the prediction accuracy, Kappa coefficient and AUC-ROC were relatively low compared to those of other models.

6.0 Conclusion

This research used an ensemble model and various ML algorithms (i.e., ctree, treebag, rf, cforest, and gbm) to predict the TPZ for the state of Assam. Initially, a comprehensive tourism inventory database was created from field research and Google Earth imagery, which produced 365 tourism points. This study produced encouraging results by utilizing a wide range of tourism conditioning factors as independent variables, such as RL, AS, VS, FA, and WL, as well as socioeconomic criteria, such as PD, LR, and RRD. The results clearly showed the good quality of the maps generated, with the best agreement between the ML models and tourism inventory data points. All spatially assessed TPZ maps demonstrated high tourism potentialities, which were dominant in Assam's southeastern and southern sections. High tourism potentialities were associated with moderate to low RL, higher VS, FA, WL, rainfall variation, higher LR, and better communication networks (here, RRD).

Over 50% of the areas of Golpara, Hailakandi, Jorhat, Kamrup, Kamrup Metropolitan, Karbi Anglong, Karimganj, Lakhimpur, Naogaon, Sivasagar, Tinsukia, and West Karbi Anglong districts were marked by very high to high tourism potentiality, and over 50% of the areas of Baksa, Barpeta, Bongaigaon, Chirang, Darrang, Dhemaji, Dhubri, Dibrugarh, Golaghat, Kokrajhar, Morigaon, Nalbari, Sontipur, and Udayguri were marked by low tourism potentiality. All models performed admirably in prediction accuracy after a thorough analysis using Kappa, accuracy, and AUC-ROC methods. The TPZ ensemble model was proposed by combining other base models, and interestingly, this model emerged as the frontrunner, showcasing superior metrics, including the highest AUC (97.6%), Kappa (0.82), and accuracy (0.93) values.

The findings from this research emphasize the potential of ML and ensemble methods in predicting TPZs, furnishing valuable insights for decision-makers tasked with spearheading tourism development in Assam. By offering robust and nuanced predictions, these findings are expected to contribute to informed decision-making processes aimed at harnessing the rich tourism potentiality of the region, thereby fostering sustainable growth and prosperity.

Acknowledgement

The authors thank all Department of Geography and Bhairab Ganguly College faculty members who supported this research wholeheartedly.

Reference

Apostolopoulos, D., & Nikolakopoulos, K. (2021). A review and meta-analysis of remote sensing data, GIS methods, materials and indices used for monitoring the coastline evolution over the last twenty years. *European Journal of Remote Sensing*, *54*(1), 240-265.

Banik, S., & Mukhopadhyay, M. (2020). Model-based strategic planning for the development of community-based tourism: a case study of Ayodhya Hills in West Bengal, India. *GeoJournal*, 1-17.

Bast, H, Delling, D, Goldberg, A, Müller-Hannemann, M, Pajor, T, Sanders, P, … Werneck R F (2016). *Route Planning in Transportation Networks. In L. Kliemann & P. Sanders (Eds.), Algorithm Engineering: Selected Results and Surveys* (pp. 19–80). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-49487-6_2

Bi, J. W., Han, T. Y., & Li, H. (2022). International tourism demand forecasting with machine learning models: The power of the number of lagged inputs. *Tourism Economics*, *28*(3), 621-645.

Blapp, M., & Mitas, O. (2020). Creative tourism in Balinese rural communities. In *Current Issues in Asian Tourism* (pp. 219-246). Routledge.

Blockeel, H. (2011). Hypothesis space. *Encyclopedia of Machine Learning*, *1*, 511-513.

Bordoloi, A. K., & Agarwal, B. K. (2015). Tourism Potentiality in Tinsukia district of Upper Assam: An Analysis. *International Journal of Management and Development Studies*, *4*(4), 375-383.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Cankurt, S., & Subasi, A. (2015). Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components. *Balkan Journal of Electrical and Computer Engineering*, *3*(1), 42-49.

Catani, F., Lagomarsino, D., Segoni, S., & Tofani, V. (2013). Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Natural Hazards and Earth System Sciences*, *13*(11), 2815-2831.

Chaturvedi, V., & de Vries, W. T. (2021). Machine learning algorithms for urban land use planning: A review. *Urban Science*, *5*(3), 68.

Chien, C. F., Ku, C. C., & Lu, Y. Y. (2023). Ensemble learning for demand forecast of After-Market spare parts to empower a data-driven value chain and an empirical study. *Computers & Industrial Engineering*, *185*, 109670.

Choden, K., & Wangchuk, D. (2018). *Bhutan-Culture Smart!: The Essential Guide to Customs & Culture*. Kuperard.

Choi, S., & Hur, J. (2020). An ensemble learner-based bagging model using past output data for photovoltaic forecasting. *Energies*, *13*(6), 1438.

Chowdary, V. M., Chakraborthy, D., Jeyaram, A., Murthy, Y. K., Sharma, J. R., & Dadhwal, V. K. (2013). Multicriteria decision-making approach for watershed prioritization using analytic hierarchy process technique and GIS. *Water resources management*, *27*, 3555-3571.

Claveria, O., Monte, E., & Torra, S. (2016). Combination forecasts of tourism demand with machine learning models. *Applied Economics Letters*, *23*(6), 428-431. https://doi.org/10.1080/13504851.2015.1078441

Codrea, P. M., Bilașco, Ș., Roșca, S., Irimuș, I. A., Iuliu, V., Rusu, R., ... & Sestras, P. (2022). The integrated assessment of degraded tourist Geomorphosites to develop sustainable tourism: A case study of Grădina Zmeilor Geomorphosite, North-West Region, Romania. *Applied Sciences*, *12*(19), 9816.

Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, *19*, 1-14. https://doi.org/10.1186/s12859-018-2264-5

Danish, M. S. S. (2023). AI and expert insights for sustainable energy future. *Energies*, *16*(8), 3309. https://doi.org/10.3390/en16083309

Das, S. (2020). Flood susceptibility mapping of the Western Ghat coastal belt using multi-source geospatial data and analytical hierarchy process (AHP). *Remote Sensing Applications: Society and Environment*, *20*, 100379.

Deribew, K. T., Mihretu, Y., Abreha, G., & Gemeda, D. O. (2022). Spatial analysis of potential ecological sites in the northeastern parts of Ethiopia using multicriteria decision-making models. *Asia-Pacific Journal of Regional Science*, *6*(3), 961-991.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241-258.

Fakfare, P., Talawanich, S., & Wattanacharoensil, W. (2020). A scale development and validation on domestic tourists' motivation: the case of second-tier tourism destinations. *Asia Pacific Journal of Tourism Research*, *25*(5), 489-504. https://doi.org/10.1080/10941665.2020.1745855

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment*, *189*, 1-20.

Fu, W. (2017). *Nonparametric Methods in Statistical Learning: Unbiasedness in Regression Trees, Survival Trees for Nonstandard Data and Estimating the Number of Clusters* (Doctoral dissertation, New York University, Graduate School of Business Administration).

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, 105151.https://doi.org/10.1016/j.engappai.2022.105151

Gayen, A., Pourghasemi, H. R., Saha, S., Keesstra, S., & Bai, S. (2019). Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. *Science of the Total Environment*, *668*, 124-138.

Getz, D., & Page, S. J. (2019). *Event studies: Theory, research and policy for planned events*. Routledge.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, *160*(3), 249-264.

Ghazoul, J. (2016). *Dipterocarp biology, ecology, and conservation*. Oxford University Press.

Garg, S., Sinha, S., Kar, A. K., & Mani, M. (2022). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management*, *71*(5), 1590-1610.

Giorgi, F., & Lionello, P. (2008). Climate change projections for the Mediterranean region. *Global and Planetary Change*, *63*(2-3), 90-104. https://doi.org/10.1016/j.gloplacha.2007.09.005

Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. (2017). A survey on ensemble learning for data stream classification. *ACM Computing Surveys (CSUR)*, *50*(2), 1-36.

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, *64*, 205-237.

Gourabi, B. R., & Rad, T. G. (2013). The analysis of ecotourism potential in Boujagh wetland with AHP method. *Life Science Journal*, *10*(2s), 251-258.

Hamze-Ziabari, S. M., & Bakhshpoori, T. (2018). Improving the prediction of ground motion parameters based on an efficient bagging ensemble model of M5′ and CART algorithms. *Applied Soft Computing*, *68*, 147-161. https://doi.org/10.1016/j.asoc.2018.03.052

Hoang, H. T., Truong, Q. H., Nguyen, A. T., & Hens, L. (2018). Multicriteria evaluation of tourism potential in the central highlands of Vietnam: Combining geographic information system (GIS), analytic hierarchy process (AHP) and principal component analysis (PCA). *Sustainability*, *10*(9), 3097. https://doi.org/10.3390/su10093097

Holloway, J. C., & Humphreys, C. (2022). *The business of tourism*. Sage.

Hothorn, T., Hornik, K., & Zeileis, A. (2015). ctree: Conditional inference trees. The comprehensive R archive network. *8*. 1-33.

https://apfbcs.nic.in/apfbcs/wetland/annexure2.pdf

https://tourism.assam.gov.in/portlets

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3), 2053168015604648. https://doi.org/10.1177/2053168015604648.

Huismann, W. (2014). *Pandaleaks: The Dark Side of the WWF*. Nordbook.

Islam, M. T., Ayon, E. H., Ghosh, B. P., MD, S. C., Shahid, R., Rahman, S., ... & Nguyen, T. N. (2024). Revolutionizing Retail: A Hybrid Machine Learning Approach for Precision Demand Forecasting and Strategic Decision-Making in Global Commerce. *Journal of Computer Science and Technology Studies*, *6*(1), 33-39.https://doi.org/10.32996/jcsts.2024.6.1.4

Kachniewska, M. A. (2015). Tourism development as a determinant of quality of life in rural areas. *Worldwide Hospitality and Tourism Themes*, *7*(5), 500-515. https://doi.org/10.1108/WHATT-06-2015-0028

Karakitsiou, A., & Mavrommati, A. (2017). Machine learning methods in tourism demand forecasting: Some evidence from Greece. *MIBES transactions*, *11*(1), 92-105. https://www.researchgate.net/publication/323427035_Machine_learning_methods_in_tourism_demand_forecasting_some_evidence_from_Greece

Karali, A., Das, S., & Roy, H. (2024). Forty years of the rural tourism research: Reviewing the trend, pattern and future agenda. *Tourism Recreation Research*, *49*(1), 173-200. https://doi.org/10.1080/02508281.2021.1961065.

Katelieva, M., & Muhar, A. (2022). Heritage tourism products based on traditional nature-related knowledge: assessment of cultural, social, and environmental factors in cases from rural Austria. *Journal of Heritage Tourism*, *17*(6), 631-647. https://doi.org/10.1080/1743873X.2022.2098040

Khadka, D., Chaudhary, A., Karki, R., Sharma, B., & Bhatta, S. (2021). Ecotourism in Ghoda Ghodi Wetland Sukhad, Kailali, Nepal. *Journal of Tourism and Hospitality Education*, *11*, 22-42. https://doi.org/10.3126/jthe.v11i0.38237

Kontogeorgopoulos, N. (2017). Finding oneself while discovering others: An existential perspective on volunteer tourism in Thailand. *Annals of Tourism Research*, *65*, 1-12. https://doi.org/10.1016/j.annals.2017.04.006

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.

Law, R., Li, G., Fong, D. K. C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research*, *75*, 410-423. https://doi.org/10.1016/j.annals.2019.01.014

Le, T. H., Arcodia, C., Novais, M. A., & Kralj, A. (2021). Proposing a systematic approach for integrating traditional research methods into machine learning in text analytics in tourism and hospitality. *Current Issues in Tourism*, *24*(12), 1640-1655. https://doi.org/10.1080/13683500.2020.1829568

Linardos, V., Drakaki, M., Tzionas, P., & Karnavas, Y. L. (2022). Machine learning in disaster management: recent developments in methods and applications. *Machine Learning and Knowledge Extraction*, *4*(2). 446-473. https://doi.org/10.3390/make4020020

Liu, Z., Peng, C., Work, T., Candau, J. N., DesRochers, A., & Kneeshaw, D. (2018). Application of machine-learning methods in forest ecology: recent progress and future challenges. *Environmental Reviews*, *26*(4), 339-350. https://doi.org/10.1139/er-2018-0034

Lu, H., Karimireddy, S. P., Ponomareva, N., & Mirrokni, V. (2020, June). Accelerating gradient boosting machines. In *International conference on artificial intelligence and statistics* (pp. 516-526). PMLR.

Manzoor, F., Wei, L., Asif, M., Haq, M. Z. U., & Rehman, H. U. (2019). The contribution of sustainable tourism to economic growth and employment in Pakistan. *International journal of environmental research and public health*, *16*(19), 3785. https://doi.org/10.3390/ijerph16193785.

Marín-Buzón, C., Pérez-Romero, A., López-Castro, J. L., Ben Jerbania, I., & Manzano-Agugliaro, F. (2021). Photogrammetry as a new scientific tool in archaeology: Worldwide research trends. *Sustainability*, *13*(9), 5319. https://doi.org/10.3390/su13095319

Masselink, R. H., Temme, A. J. A. M., Giménez Díaz, R., Casalí Sarasíbar, J., & Keesstra, S. D. (2017). Assessing hillslope-channel connectivity in an agricultural catchment using rare-earth oxide tracers and random forests models. *Cuadernos de Investigación Geográfica 2017, nº 43 (1), pp. 19-39*. https://doi.org/10.18172/cig.3169

Memon, M. A., Cheah, J. H., Ramayah, T., Ting, H., Chuah, F., & Cham, T. H. (2019). Moderation analysis: issues and guidelines. *Journal of Applied Structural Equation Modeling*, *3*(1), 1-11.

Mogensen, U. B., Ishwaran, H., & Gerds, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, *50*(11), 1. https://doi.org/10.18637%2Fjss.v050.i11

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*, *35*(2), 757-774.

Mosebo Fernandes, A. C., Quintero Gonzalez, R., Lenihan-Clarke, M. A., Leslie Trotter, E. F., & Jokar Arsanjani, J. (2020). Machine learning for conservation planning in a changing climate. *Sustainability*, *12*(18), 7657. https://doi.org/10.3390/su12187657

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, *10*, 99129-99149.

Mitra, R., Saha, P., & Das, J. (2022). Assessment of the performance of GIS-based analytical hierarchical process (AHP) approach for flood modelling in Uttar Dinajpur district of West Bengal, India. *Geomatics, Natural Hazards and Risk*, *13*(1), 2183-2226. https://doi.org/10.1080/19475705.2022.2112094

Munier, N., & Hontoria, E. (2021). *Uses and Limitations of the AHP Method*. Cham: Springer International Publishing.

Naghibi, S. A., Ahmadi, K., & Daneshi, A. (2017). Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resources Management*, *31*, 2761-2775. https://doi.org/10.1007/s11269-017-1660-3

Nath, B., Wang, Z., Ge, Y., Islam, K., P. Singh, R., & Niu, Z. (2020). Land use and land cover change modeling and future potential landscape risk assessment using Markov-CA model and analytical hierarchy process. *ISPRS International Journal of Geo-Information*, *9*(2), 134. https://doi.org/10.3390/ijgi9020134

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, *11*, 169-198.

Pathmanandakumar, V., Goh, H. C., & Chenoli, S. N. (2023). Identifying potential zones for ecotourism development in Batticaloa district of Sri Lanka using the GIS-based Ahp spatial analysis. *GeoJournal of Tourism and Geosites*, *46*(1), 252-261. https://doi.org/10.30892/gtg.46128-1022

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21-45.

Pourghasemi, H. R., & Rahmati, O. (2018). Prediction of the landslide susceptibility: Which algorithm, which precision? *Catena*, *162*, 177-192.

Puh, K., & Bagić Babac, M. (2023). Predicting sentiment and rating of tourist reviews using machine learning. *Journal of hospitality and tourism insights*, *6*(3), 1188-1204. https://doi.org/10.1108/JHTI-02-2022-0078

Puška, A., Pamucar, D., Stojanović, I., Cavallaro, F., Kaklauskas, A., & Mardani, A. (2021). Examination of the sustainable rural tourism potential of the brčko District of Bosnia and Herzegovina using a fuzzy approach based on group decision making. *Sustainability*, *13*(2), 583. https://doi.org/10.3390/su13020583

Quinlan, J. R. (1992, November). Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence* (Vol. 92, pp. 343-348).

Raha, S., & Gayen, S. K. (2023). Application of Analytic Hierarchy Process and weighted sum techniques for green tourism potential mapping in the Gangetic West Bengal, India. *GeoJournal*, *88*(Suppl 1), 197-240. https://doi.org/10.1007/s10708-022-10619-2

Raha, S., & Gayen, S. K. (2022). Tourism Potentiality Zone Mapping by Using the AHP Technique: A Study on Bankura District, West Bengal, India. *Journal of Geographical Studies*. 6. 58-85. https://doi.org/10.21523/gcj5.22060201

Raha, S., Mondal, M., & Gayen, S. K. (2021). Ecotourism potential zone mapping by using analytic hierarchy process (AHP) and weighted linear algorithm: A study on West Bengal, India. *Journal of Geographical Studies*, *5*(2), 44-64. https://doi.org/10.21523/gcj5.21050201

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. *Update*, *1*(1), 2007.

Rincy, T. N., & Gupta, R. (2020). Ensemble learning techniques and its efficiency in machine learning: A survey. In *2nd International Conference on Data, Engineering And Applications (IDEA)* (pp. 1-6). IEEE.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*, 1-39.

Ronghang, S., & Sen, S. (2022). Tourism potentials in the Karbi Anglong autonomous council districts (KAAC) of Assam. *International Journal of Geography, Geology and Environment. 4*.112–117. https://doi.org/10.22271/27067483.2022.v4.i2b.123.

Roodposhti, M. S., Safarrad, T., & Shahabi, H. (2017). Drought sensitivity mapping using two one-class support vector machine algorithms. *Atmospheric research*, *193*, 73-82. https://doi.org/10.1016/j.atmosres.2017.04.017

Roy, S. K., Hasan, M. M., Mondal, I., Akhter, J., Roy, S. K., Talukder, S., ... & Karuppannan, S. (2024). Empowered machine learning algorithm to identify sustainable groundwater potential zone map in Jashore District, Bangladesh. *Groundwater for Sustainable Development*, 101168. https://doi.org/10.1016/j.gsd.2024.101168

Sachdeva, S., & Kumar, B. (2021). Comparison of gradient boosted decision trees and random forest for groundwater potential mapping in Dholpur (Rajasthan), India. *Stochastic*

Environmental Research and Risk Assessment, *35*(2), 287-306. https://doi.org/10.1007/s00477-020-01891-0

Sahani, N. (2020). Application of analytical hierarchy process and GIS for ecotourism potentiality mapping in Kullu District, Himachal Pradesh, India. *Environment, Development and Sustainability*, *22*(7), 6187-6211.https://doi.org/10.1007/s10668-019-00470-w

Saaty, T. L. (1980) The analytic hierarchy process: Planning, priority setting, resource allocation. McGraw Hill.

Saaty, R. W. (1987) The analytic hierarchy process—What it is and how it is used. Math Modelling 9:161–176. https://doi.org/10.1016/0270-0255(87)90473-8.

Scarpocchi, C. (2020). Where are people going? In *Place Branding*. Routledge.

Schultze, J., Gärtner, S., Bauhus, J., Meyer, P., & Reif, A. (2014). Criteria to evaluate the conservation value of strictly protected forest reserves in Central Europe. *Biodiversity and Conservation*, *23*(14), 3519-3542.

Senapati, U., & Das, T. K. (2021). Assessment of basin-scale groundwater potentiality mapping in drought-prone upper Dwarakeshwar River basin, West Bengal, India, using GIS-based AHP techniques. *Arabian Journal of Geosciences*, *14*(11), 960. https://doi.org/10.1007/s12517-021-07316-8

Singh, K. R., Goswami, A. P., Kalamdhad, A. S., & Kumar, B. (2020). Assessment of surface water quality of Pagladia, Beki and Kolong rivers (Assam, India) using multivariate statistical techniques. *International Journal of River Basin Management*, *18*(4), 511-520. https://doi.org/10.1080/15715124.2019.1566236

Singh, S., Rao, M. J., Baranval, N. K., Kumar, K. V., & Kumar, Y. V. (2023). Geoenvironment factors guided coastal urban growth prospect (UGP) delineation using heuristic and machine learning models. *Ocean & Coastal Management*, *236*, 106496. https://doi.org/10.1016/j.ocecoaman.2023.106496

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 1-21.

Sun, S., Wei, Y., Tsui, K. L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, *70*, 1-10. https://doi.org/10.1016/j.tourman.2018.07.010

Taalab, K., Cheng, T., & Zhang, Y. (2018). Mapping landslide susceptibility and types using Random Forest. *Big Earth Data*, *2*(2), 159-178.

Tang, Z., Zhao, W., Xie, X., Zhong, Z., Shi, F., Liu, J., & Shen, D. (2020). Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. *arXiv preprint arXiv:2003.11988*.

Tekouabou, S. C. K., Diop, E. B., Azmi, R., Jaligot, R., & Chenal, J. (2022). Reviewing the application of machine learning methods to model urban form indicators in planning decision support systems: Potential, issues and challenges. *Journal of King Saud University-Computer and Information Sciences*, *34*(8), 5943-5967. https://doi.org/10.1016/j.jksuci.2021.08.007

Thanh, N. N., Chotpantarat, S., Trung, N. H., & Ngu, N. H. (2022). Mapping groundwater potential zones in Kanchanaburi Province, Thailand by integrating of analytic hierarchy process, frequency ratio, and random forest. *Ecological Indicators*, *145*, 109591. https://doi.org/10.1016/j.ecolind.2022.109591

Trabelsi, F., Bel Hadj Ali, S., & Lee, S. (2022). Comparison of Novel Hybrid and Benchmark Machine Learning Algorithms to Predict Groundwater Potentiality: Case of a Drought-Prone Region of Medjerda Basin, Northern Tunisia. *Remote Sensing*, *15*(1), 152. https://doi.org/10.3390/rs15010152

Trukhachev, A. (2015). Methodology for evaluating the rural tourism potentials: A tool to ensure sustainable development of rural settlements. *Sustainability*, *7*(3), 3052-3070.

Ünlü, R., & Xanthopoulos, P. (2021). A reduced variance unsupervised ensemble learning algorithm based on modern portfolio theory. *Expert Systems with Applications*, *180*, 115085.

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., & Al-Katheeri, M. M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, *13*, 839-856.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Vafadar, S., Rahimzadegan, M., & Asadi, R. (2023). Evaluating the performance of machine learning methods and Geographic Information System (GIS) in identifying groundwater potential zones in Tehran-Karaj plain, Iran. *Journal of Hydrology*, *624*, 129952. https://doi.org/10.1016/j.jhydrol.2023.129952

Vrontos, S. D., Galakis, J., & Vrontos, I. D. (2021). Modeling and predicting US recessions using machine learning techniques. *International Journal of Forecasting*, *37*(2), 647-671. https://doi.org/10.1016/j.ijforecast.2020.08.005

Wang, C., Yu, Q., Law, K. H., McKenna, F., Stella, X. Y., Taciroglu, E., ... & Cetiner, B. (2021). Machine learning-based regional scale intelligent modeling of building information for natural hazard risk management. *Automation in Construction*, *122*, 103474. https://doi.org/10.1016/j.autcon.2020.103474

Yuxi, Z., & Linsheng, Z. (2020). Identifying conflicts tendency between nature-based tourism development and ecological protection in China. *Ecological Indicators*, *109*, 105791.

Zabihi, H., Alizadeh, M., Wolf, I. D., Karami, M., Ahmad, A., & Salamian, H. (2020). A GIS-based fuzzy-analytic hierarchy process (F-AHP) for ecotourism suitability decision making: A case study of Babol in Iran. *Tourism Management Perspectives*, *36*, 100726. https://doi.org/10.1016/j.tmp.2020.100726

Zekan, B., Weismayer, C., Gunter, U., Schuh, B., & Sedlacek, S. (2022). Regional sustainability and tourism carrying capacities. *Journal of Cleaner Production*, *339*, 130624. https://doi.org/10.1016/j.jclepro.2022.130624

Zhang, K., Wu, X., Niu, R., Yang, K., & Zhao, L. (2017). The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China. *Environmental Earth Sciences*, *76*, 1-20.

Zhao, Q., Yu, S., Zhao, F., Tian, L., & Zhao, Z. (2019). Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments. *Forest Ecology and Management*, *434*, 224-234.https://doi.org/10.1016/j.foreco.2018.12.019.