Contents lists available at *Dergipark*

Journal of Scientific Reports-A

journal homepage: *https://dergipark.org.tr/tr/pub/jsr-a*

E-ISSN: 2687-6167 ***Number 58, September 2024***

RESEARCH ARTICLE

Receive Date: 22.03.2024 *Accepted Date: 06.08.2024*

Forecasting urban forest recreation areas in Turkey using machine learning methods

Mehmet Cüneyt Özbalcı[a,*], Sena Dikici[b], Turgay Tugay Bilgin[c]

[a]*Bursa Technical University, Bursa, 16310, Türkiye, ORCID: 0000-0003-4499-0061*
[b]*Bursa Technical University, Bursa, 16310, Türkiye, ORCID: 0000-0002-1759-6045*
[c]*Bursa Technical University, Bursa, 16310, Türkiye, ORCID: 0000-0002-9245-5728*

Abstract

Recreation is the process of revitalizing and renewing human existence through optional activities, serving as a broad description. It has prominently arisen as a reaction to personal requirements for stress reduction, especially in developed urban areas. Engaging in this recreational activity provides a way to utilize one's spare time, providing refreshment for both the physical and mental aspects, whether done alone or with others, in countryside or city environments. Urban forests are important leisure places within city environments. An expanded presence of urban forest places can greatly enhance the general well-being of society. The estimation of urban forest areas in the future may receive increased attention, leading to measures to extend current areas or prepare for future activities and services. We utilized official statistics from the years 2013 to 2021, sourced from the Republic of Turkey official website. Ministry of Agriculture and Forestry's General Directorate of Forestry. We used statistics that contained information about urban forests, classified as Type D recreational areas, to create a dataset. We performed provincial-level area projections for the year 2021. Using the KNIME platform, we used three different analysis techniques: linear regression analysis, gradient-boosted regression trees and artificial neural networks. It is seen that the results of linear regression and artificial neural networks are close to each other and give good results. The peak performance was attained using artificial neural networks, resulting in an R^2 score of 0.99. This study differs from other similar projects by concentrating on calculating urban forest recreational spaces per province throughout Turkey, using data provided by government agencies. The accomplishments highlight the ability to make reliable predictions about future forest resources by using analogous forecasts in the upcoming years.

Keywords: Recreation areas; urban forests; linear regression analysis; gradient boosted regression trees; artificial neural network.

*Corresponding author.
e-mail: mehmet.ozbalci@btu.edu.tr

1. Introduction

The benefits of nature-based tourism and outdoor recreation are growing. It is commonly known that spending time outside and in natural settings enhances people's health and wellbeing. For reasons like fostering a sense of connection with one's natural and cultural heritage, strengthening social bonds, fostering indigenous identity, and boosting the economy, it is critical to increase public awareness of the need to conserve natural places [1]. As awareness of these benefits has increased over time, natural areas have become a focal point of interest and focus for society. Recreation is the activities that people voluntarily do in their free time in order to reach the fitness of their physical and mental health in order to lead a healthy and productive life. Recreation is considered as an important contribution to physical and mental health. Parks and forest areas are considered to be important recreation areas for societies [2]. Nowadays, it is necessary to increase the organization and services in recreation areas for the entertainment and recreation needs of people living in cities or to intensify the studies depending on the increase in recreation areas. The need for recreation areas is increasing in order to strengthen the human-environment relationship of the society around the city [3]. Because there aren't many places to relax in cities, especially the larger ones, people choose to move to seaside locations. Recreational spaces are essential for satiating emotional and spiritual requirements as well as providing the energy required for daily productivity [4]. Urban forests are valued as significant recreational spaces. Because of this, a rise in recreational places benefits society, but a fall has the opposite effect. The protection of urban recreation areas and our natural assets is of great importance for society. Urban forests and natural resources are vital for our future. Scientific studies should draw attention to this issue and emphasize its importance.

When various regression studies are examined, it is observed that linear regression and artificial neural networks are frequently used and successful results are obtained [5], [6], [7], [8], [9], [10], [11]. It is also thought that the GBRT (Gradient Boosted Regression Trees) algorithm may be another algorithm to be preferred after experimental observations. The study utilized official information obtained from the Ministry of Agriculture and Forestry's General Directorate of Forestry website. The KNIME platform was utilized to predict the area in hectares of Type D (urban forest) regions in Turkey for the upcoming year through the use of linear regression analysis, gradient boosted regression tree, and artificial neural network (ANN) approaches. The evaluation criteria indicated that the approaches successfully forecasted the results. The goal is to enhance the existing literature by predicting recreational areas using linear regression analysis, gradient boosted regression tree, and ANN.

1.1. Literature survey

Data from 399 watersheds in 15 wildfires that burned in Colorado, Idaho, Montana, and New Mexico between 2000 and 2002 were analyzed by Ruppert et al. in their study. The study demonstrated how debris flows in landscapes that have recently seen fires can be predicted using logistic regression. This was achieved by examining more than 35 independent factors related to burn severity, geology, slope, precipitation, and soil properties [5]. Using information from a survey of 2449 persons, a representative sample of the Norwegian population, Bjerke et al. assessed the relationship between interest in 15 outdoor recreational activities and environmental attitudes. It was observed that there was an opposite relationship between age and score. In addition, women were found to score higher than men. In addition, it was determined that there was a complex relationship between predictor variables and recreational interests [6]. Human population density and activity are the primary drivers of land use change. Urban growth modeling has gained a lot of interest since it aids in understanding the processes underlying changes in land use, which in turn aids in the development of pertinent policies. Nong Yu and Du Qingyun used logistic regression to model urban expansion in Jiayu County, Hubei Province, China. The model accuracy is displayed by the relative operating characteristic (ROC), which has a curve of 0.891 and a standard error of 0.001. According to the findings, the model does a good job of simulating urban growth at the county level [7]. Bivariate and regression analysis techniques were used in a study to assess the association between the utilization of recreational places on

campus and the individual variables of university students, such as exercise practices, social physical anxiety, and self-efficacy. The Cronbach's alpha coefficient is the internal consistency coefficient that is determined by calculating the pairwise correlation between test items. The study's evaluation criterion was the Cronbach's alpha coefficient, and a value of 0.91 was attained [8]. The length and timing of a plant's growth stages vary from place to region according on climate, which affects chickpea phenology. It was proposed that a strong prediction might be made jointly about the geographical origin of accession and the local environmental circumstances in the growing region. They put up a novel forecasting model for the blossoming period of wild chickpeas cultivated in Turkey under four distinct environmental circumstances and in 21 distinct seasons. The study employed the nonlinear regression method, and the R^2 value was found to be 0.97 [9]. Another study analyzed many factors to understand the changes in forest acreage in the Mediterranean Region between 2004 and 2019. Data on factors such as precipitation, temperature, land surface temperature, ozone, fire, urban areas and population were obtained. Using this data, the factors that most affect the changes in forest areas were analyzed by multiple linear regression method. The analysis's findings demonstrated that the strongest factor causing loss in forest areas is forest fires [10]. Zelin Liu et al. compared the three machine learning techniques of support vector machines (SVM), ANN and decision trees. Their applications in forest ecology in the last decade were compared. Models of species distribution, carbon cycles, hazard assessment and prediction, and other forest management applications are the items that are compared [11]. In the study by Qingxia Zhao et al. parameters of forests were estimated using various regression methods. Four different machine learning algorithms were used. These are: SVM, ANN, RF, and classification and regression tree (CART) to determine how well they could predict the forest parameters of black locust plantations on the Loess Plateau. The findings demonstrated that, for all forest parameter estimations, RF provided the maximum accuracy, SVM and ANN approaches performed moderately, and the CART method yielded the lowest accuracy. The highest R^2 values was 0.85 and the lowest relative root mean square error (RRMSE) was 0.18 [12]. Three distinct data sets were employed in the investigation of forest degradation detection. BFAST method was used and two machine learning methods, SVM and RF, were used. The findings demonstrate that the SVM approach, when applied to the entire forest dataset, achieves the maximum accuracy. It was found that the latter model was more accurate than the magnitude threshold model [13]. Snezhana G. Gocheva-Ilieva et al. suggest a new method for modeling air pollutant data using RF and AutoRegressive Integrated Moving Average (ARIMA) methods. First, the pollutant's RF model is made and examined in connection to meteorological factors. After that, univariate ARIMA is used to model its residuals and correct it. The method is used for nine years and three months' worth of hourly data for seven air pollutants (CO, NO, NO_2, NOx, O_3, PM_{10} and SO_2) in Dimitrovgrad, Bulgaria. Three time variables and six meteorological variables were included as predictors. High-performing models were produced, with R^2 values between 90% and 98% [14]. Maryam Pourshamsi et al. employed machine learning techniques to improve the estimates of forest canopy height using LiDAR measurements by utilizing baseline polarimetric interferometric SAR (PolInSAR) data. Multiple baseline merging was performed using a support vector. The methodology was created by combining two datasets. The method was analyzed on NASA AfriSAR. The accuracy of the resulting estimated height approximation (R^2 = 0.81, RMSE = 7.1 m) was measured as the previously introduced multiple baseline fusion approach (R^2 = 0.67, RMSE = 9.2 m) [15]. In this study by Xiaobang Liu et al. For the dataset, 30-m Landsat-8 data and merged 1-m GaoFen-2 (GF-2) satellite imagery were used. In the Three North Regions of China, the fractional forest cover was estimated using a variety of machine learning algorithms. Based on the findings of the 10-fold cross-validation, all nonlinear algorithms perform well, with an R^2 value of more than 0.8 and a root-mean-square error of less than 0.05. The RF outperformed the light gradient assisted machine in the boosting ensemble (R^2 = 0.992, RMSE = 0.022) and the bagging ensemble (R^2 = 0.993, RMSE = 0.020). Additionally, the study's findings demonstrate that the RF method is the best choice for estimating fractional forest cover and serve as a guide for the following research. [16]. Manley and Egoh developed a model to predict how non-urban recreation will change due to climate change. The model shows that current patterns will worsen in the future due to climate change, with cottage recreation areas becoming more suitable and unpopular cottage recreation areas becoming less suitable for recreation. At various intervals from 2030 to 2099, recreation areas have been forecasted for some

regions in the state of California [17]. Chen et al. uses logistic regression (LR), decision tree (DT), and random forest (RF) models to analyze landslide susceptibility in forest-covered regions in Lin'an, China. It identifies key predictive factors like forest type (FT), maximum daily rainfall (MDR), distance to road (DTRD), understory vegetation height (UVH), and the normalized differential vegetation index (NDVI). Hickory plantations have the highest susceptibility, and the RF model proves to be the most accurate for predicting high-susceptibility zones. This research helps understand landslide risks, especially when transitioning from natural forests to plantations, and guides disaster mitigation and prevention strategies [18]. Adhikari et al. conducted a comparative study evaluating various modeling methods including Least Squares Regression, Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO), Random Forest, and Generalized Additive Model Selection for predicting attributes of urban forest areas such as volume, basal area, and dominant height. The experimental results showed R^2 values of 88% for volume, 83% for basal area and 83% for dominant height, respectively [19]. Forest land categories typically collide with other land use types during societal development, resulting in environmental repercussions. As a result, study into changes in forest land categories has become a more international priority. As urbanization trends increase, regional land use modeling studies become increasingly significant in identifying possible forest ecological security concerns. This study, which uses land use data from the Ganjiang River basin, looks at the distribution of land use categories and how they've changed from 2000 to 2020. It estimates the land use pattern of the basin in 2040 and looks into the features of forest land types that transfer using the CA-Markov model. The results indicate that the transfer between different subcategories of forest land, which was particularly widespread in the high-altitude mountainous portions of the basin in the south and west, was the most significant trend in the development of land use between 2000 and 2020. With a kappa value of 0.92, the land use pattern prediction model created in this work suggests that the projections are accurate and trustworthy [20].

2. Material and method

2.1. KNIME data analytics tool

KNIME, an open source, modular data mining software, includes many possibilities such as data analysis, statistical evaluation, clustering, classification, and reporting. Operations are carried out using nodes in KNIME, which distinguishes out for its usability and user-oriented interface. The application provides a wide range of data science services and includes statistical analysis modules, learning algorithms, and clustering algorithms. The user can manually handle every stage of the process using the modules, from uploading the raw data to pre-processing, segmenting, and formatting the data to applying the appropriate algorithms. Data from files with the .txt, .csv, .arff and .table extensions can be imported into KNIME. The software offers a wide range of visualization features [21].

2.2. Dataset

The data used in the study were collected from the official website of the General Directorate of Forestry of the Ministry of Agriculture and Forestry of the Republic of Turkey, covering the years 2013 to 2020. The data set contains the total number of hectares of A, B, C, and D category recreational spaces from 2007 to 2021. Types A, B, and C designate recreational spaces, whereas type D designates urban woods. This study examined the transformation of type D data due to the high value attributed to urban forest assets. The dataset contains the number of urban forest and their hectare equivalents. Analysis was conducted on data from 2013 to 2021 to utilize the most recent information. A projection for 2021 was created and accomplishments were assessed based on data up to 2020. The dataset is around 20 kilobytes in size. The file with numerical data for all provinces in our country was converted to xlsx format and prepared for processing. 70% of the data is allocated for training, while the remaining 30% is designated for testing. To ensure reliability in the results, missing recreation data in the provinces were replaced with the mean values of the row data in the dataset. The study could not estimate statistics for Adana,

Aksaray, Kilis, Muş, Yozgat, and Gümüşhane provinces due to unavailability of information [22]. The information of some cities belonging to the dataset is given in the figure below.

Table 1. A fragment of the content of the data in the dataset.

Cities	Number of Areas	Area in Hectares
İstanbul	12	1783,59
Tekirdağ	3	211,90
Edirne	2	81,50
Kırklareli	2	122,87
Balıkesir	1	15,00

2.3. Analysis methods

2.3.1 Regression analysis

The functional depiction of the relationship between the dependent variable and the independent variable or variables is known as regression analysis [23]. In regression analysis, working with as much data as possible increases the likelihood of consistent results. The consistency of the available data within itself is also an important factor for the result to be successful. Regression analysis is used extensively in fields of study that are open to statistical and empirical studies such as finance, marketing and health. In order to obtain the desired results, the data used should not be underfit or overfit. When the learning process fails miserably with the training data, the test data yields inconsistent results, which is known as underfitting. Typically, underfitting is seen when there is an excessive amount of noisy data. When the training set's state is retained by the learning process through memorization, this is known as overfitting, and bad test data findings are the outcome.

2.3.2 Linear regression analysis

Predicting the value of a variable based on the values of other variables is the basis of linear regression analysis. Independent variables are those that are not employed in the prediction of the dependent variable, while the variable or variables to be forecasted form dependent variables [24].

Linear regression analysis consists of two basic concepts. These are simple linear regression and multiple linear regression. In simple regression analysis, it shows the linear relationship between only one explanatory variable and the response variable [25]. In a simple linear regression model, Y is the response variable, Xi is the explanatory variable, β0 and β1 are the unknown parameters of the variable, and e is the error term due to chance. In this case, the simple linear regression model is calculated as given in Eq. 1:

$$Y = \beta_0 + \beta_1 X_{i_1} + e_i \quad i = 1,2,..n \tag{1}$$

Regarding a multivariate linear regression model with n observations and p explanatory variables;

$$Y_i = \beta_0 + \beta_1 X_{i_1} + \beta_2 X_{i_2} + \ldots + \beta_p X_{i_p} + e_i \quad i = 1,2,..n \tag{2}$$

The R^2 value in regression analysis is a metric that indicates the predictive performance of the model. This value is associated with the extent to which the independent variables affect the dependent variable. The values that R^2 can take vary between 0 and 1. The higher this value is, the better the regression model fit [26].

2.3.3 Gradient boosted regression tree (GBRT)

The GBRT technique can be used to solve regression and classification issues. Regression problems show that this approach performs remarkably well. As a result, it has emerged as one of regression analysis's most often used algorithms. GBRT is made up of three main components. These three are the additive model, the weak learner, and the loss function. The goal of the problem solution must be the main emphasis in order to determine the loss function [27]. The GBRT model employs an algorithm that is focused on building individual decision trees. The model is improved through training the decision trees. Training the decision trees results in the generation of an objective function. Here, maximizing the desired features by iterative splitting is the key goal [28]. One popular method of binary classification is to measure node purity during the splitting process. Here, the purity index is referred to as the gini index. Figure 1 illustrates how the proper value will be chosen at each step based on the gini index, which will determine the most appropriate decision.

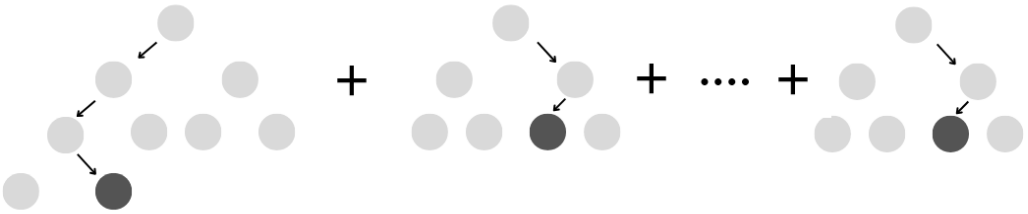

Fig. 1. Structural visualization of the GBRT algorithm with many decision trees [29].

2.3.4 Artificial neural networks (ANN)

The architecture of ANN, essentially imitates that of the human brain. Its relational structure is comparable to human neuronal connections. Neurons are made up of three layers: input, hidden, and output. A collection of interconnected nodes made up of artificial neurons makes up an ANN [30]. Each link has the ability to transmit a signal from one artificial neuron to another, just like synapses do in the brain. After processing the signal, the receiving neuron can communicate with other neurons [31]. Data filtering, interpretation, association, prediction, and classification are all done with ANN [32]. The illustration of the ANN's layers and connections between them is provided below.

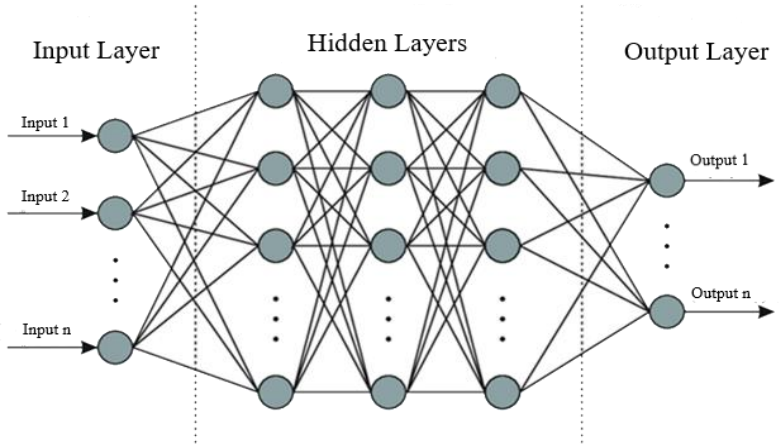

Fig. 2. Layers of ANN and their relational structure [33].

In the initial step of an ANN, a result is obtained by multiplying the input data by the predetermined weights. The impact of inputs on the desired outcomes can be changed in this way. The activation function is the one that takes the weighted sum of the inputs from the preceding layer and outputs a value. Activation functions that are most widely used and favored are sigmoid and ReLU.

2.4 Performance metrics

In this study, R^2, mean absolute error value, square of mean errors, root mean square roots error and mean absolute percentage error were calculated for performance measures [34,35].

2.4.1 R^2 value

The R^2 value is a measure of the success of the regression model fit. The closer this value is to 1, the lower the error rate. The calculation is performed according to Equation 3.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} \tag{3}$$

Here, y_i denotes the actual value of y, $\hat{y}_i$ is denotes the predicted value of y, $\bar{y}_i$ denotes the mean value of y.

2.4.2 Mean squared error (MSE)

MSE is the difference of the mean squares between the actual values and the predicted values. The calculation is performed according to Equation 4.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

2.4.3 Root mean squared error (RMSE)

RMSE is a measure of the average of the square roots of erroneous estimates. The calculation is performed according to Equation 5.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2} \quad (5)$$

2.4.4 Mean absolute error (MAE)

MAE is a quantity that expresses the sum of the absolute error value. The calculation is performed according to Equation 6.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - y| \quad (6)$$

2.4.5 Mean absolute percentage error (MAPE)

MAPE is the error rate of the accuracy of the statistical estimate. It is a measure of the accuracy of the estimate. The calculation is performed according to Equation 7.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (7)$$

2.4.6. Standard deviation (STD. DEV.)

Standard deviation is the square root of variance. The ideal situation is when the standard deviation is the lowest.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(N-1)}} \quad (8)$$

3. Result and discussion

The study employs GBRT, ANN, and linear regression to predict the D-type urban recreation spaces for 2021 based on province-specific data from 2013 and 2020. The selected algorithms were applied to the data using the nodes in the KNIME software. Below are system flow diagrams created using KNIME software. Figures 3, 5, and 7. Hatay province had missing data for 2013 and 2014, while Siirt had missing data for 2013. To address this, missing data was excluded by calculating the average for each technique based on the rows (provinces), ensuring consistent and reliable analysis. The data underwent the decimal scaling normalization technique. Decimal scaling normalization involves dividing the data by a power of 10 to reduce it to a number less than 1. Equation 9 gives the equation for the normalizing process, which involves decimal scaling.

$$A' = \frac{A_i}{10^j} \quad (9)$$

Methods such as Z-score normalization, which require normally distributed data for best performance, are less appropriate given our data's non-normal distribution. Decimal scaling, on the other hand, is more appropriate for datasets with uncertain or non-normal distributions because it does not make any assumptions about the data's distribution. Maintaining the independence of each province's data is also essential because the data from one does not affect the data from another. By normalizing each value according to its own order of magnitude, decimal scaling successfully addresses this issue instead of utilizing group statistical measures like standard deviation and mean, which may cause biases if the data points were interdependent. This approach ensures that the unique characteristics of the forest area data in each province are preserved without the distortion that might occur with other normalization techniques.

The normalized data is represented by A', the value to be normalized is represented by A_i, and the value that reduces A' to 1 is represented by J. The J value for Normalization by Decimal Scaling is determined automatically by KNIME in its current edition. This J value cannot be manually set or determined by researchers. Based on the highest absolute value of every feature in the dataset, the software chooses the best J, making sure that the greatest absolute value is less than 1 after scaling. The efficient standardization of data through this automated approach eliminates the need for user participation in defining scaling parameters, hence streamlining the data preprocessing procedures. This implies, however, that users' control over this step of the KNIME data normalization process is restricted.

The data was divided into training and test sets after normalizing the data. Figures 4, 6, and 7 display comparisons of the training and test data obtained at different rates during the study. The same processes for handling missing data, normalizing data, and dividing data were used to all three approaches. Figure 3 displays the flow diagram of the linear regression approach.

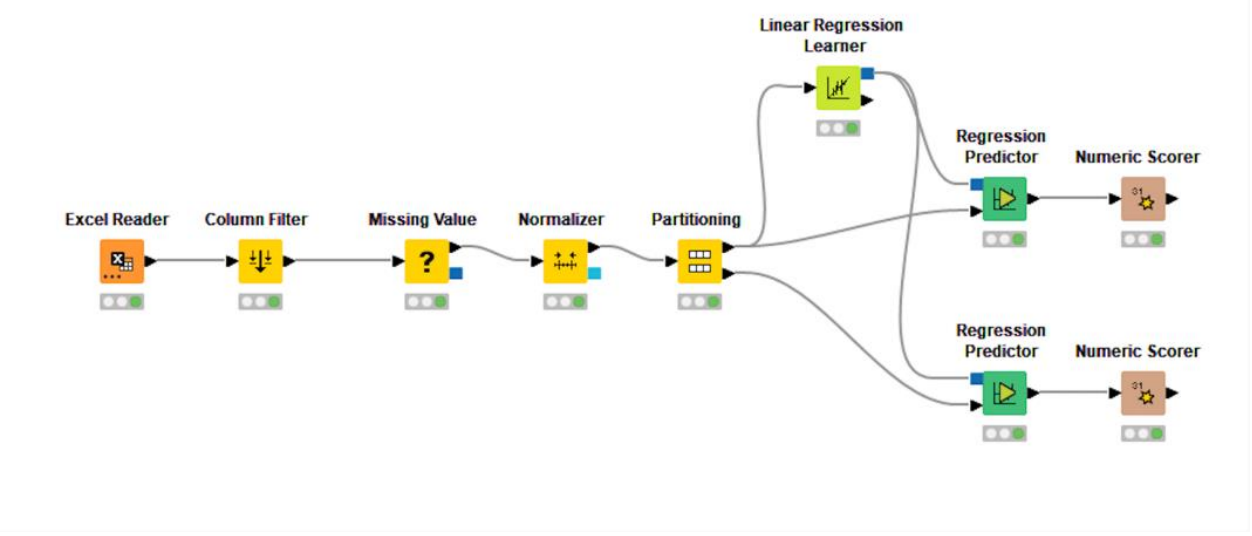

Fig. 3. Flow diagram of linear regression model in KNIME tool.

The flow diagram generated by the Knime tool for the linear regression method is shown in Figure 3. The Knime tool utilizes a node called "Linear Regression Learner" for implementing the linear regression approach. Merely the target column for prediction is chosen in this node. The study focuses on column 2021. The predictor node is named "Regression Predictor". The predictor utilizes data from before 2021 to generate forecast results for 2021. Performance values are determined by the disparities between the expected and actual data.

The study on linear regression involved analyzing training and test data at various proportions. Figure 4 presents a comparison of the data obtained.

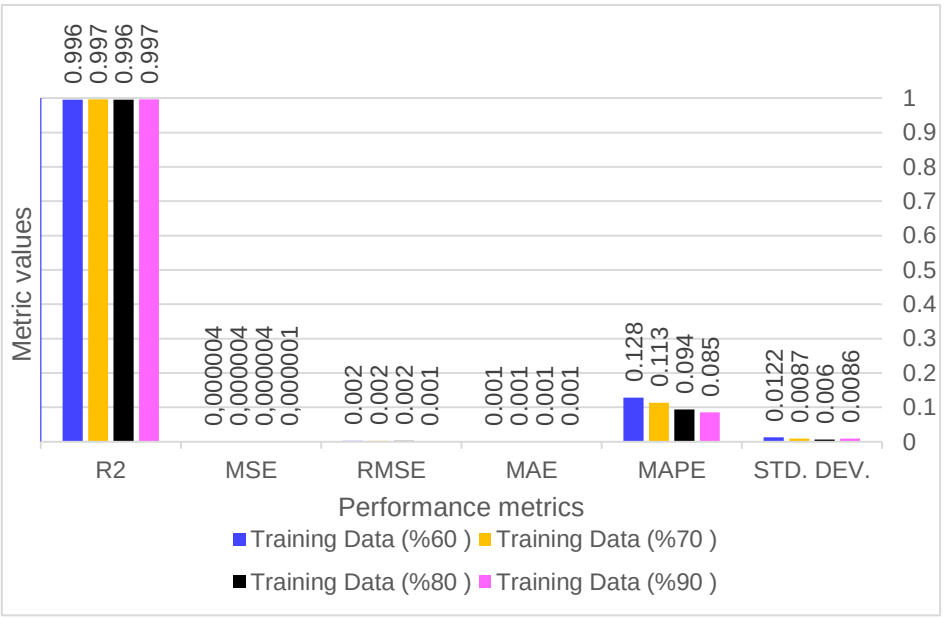

Fig. 4. Results from various training and test data set percentages using the linear regression approach.

Experimentation utilizing various percentages of training and test data showed that the best results for the linear regression technique were obtained when 90% of the data was utilized for training. The results display almost perfect explanatory power across all training set percentages, with R^2 values close to 1, indicating that the model excellently captures the variance of the dependent variable regardless of the data set size.

Looking at the mean squared error (MSE) and the root mean squared error (RMSE), the values are exceptionally low across all training data scenarios, ranging from 0.0002 to 0.0001 for MSE and consistently at 0.001 for RMSE. These metrics signify that the model's predictions are very close to the actual values, highlighting a high degree of accuracy and consistency. MAE values are between 0.0007 and 0.0008 and MAPE values are between 0.085 and 0.128. The standard deviation value is between 0.006 and 0.008. The low levels of these values indicate the success of the model.

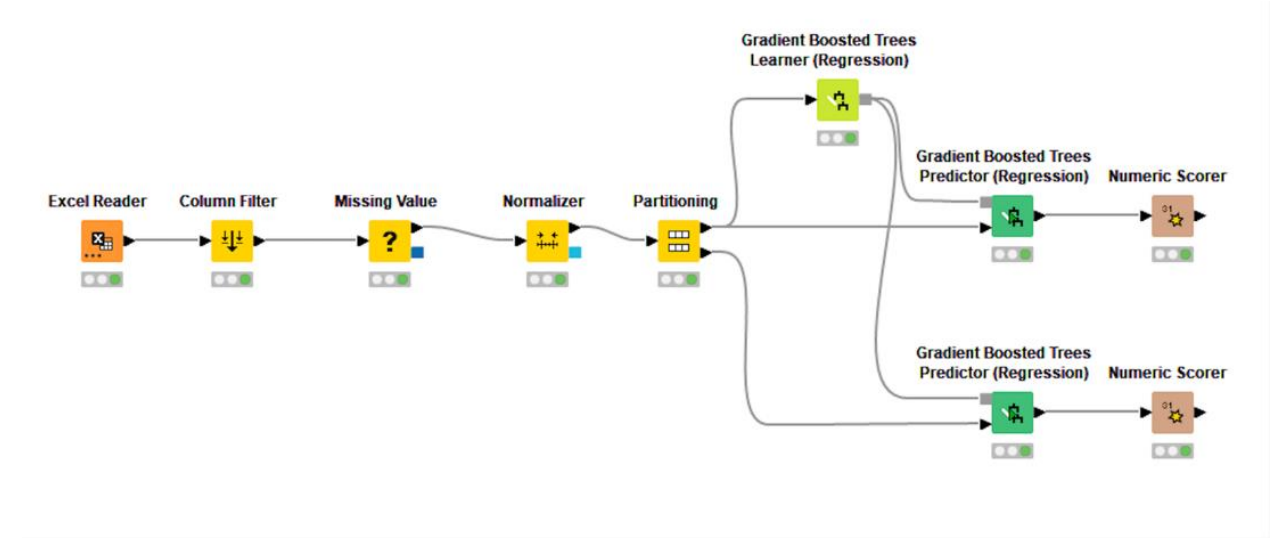

Fig. 5. Flow diagram of GBRT model in Knime program.

Figure 5 illustrates the flow diagram of the gradient boosted regression tree technique generated with the Knime tool. To enable the gradient boosted regression Tree to understand the dataset, the "Gradient Boosted Tree Learner" node is necessary. This node has a maximum tree depth set at four levels. The learning rate is established at 0.05 and the number of models is determined at 100. The target column is designated as the latter part of 2021. The "Gradient Boosted Trees Predictor" node was utilized by the predictor. The R^2, MSE, RMSE, MAE, and MAPE values were calculated by comparing the predicted values from the predictor node with the actual 2021 data. The values were acquired using the "Numeric Scorer" node. Figure 6 displays the outcomes of the gradient boosted regression tree approach using various percentages of training and test data. While conducting the experiments, tree depth was determined as 4. Learning rate was taken as 0.05. Alpha value was 0.95.

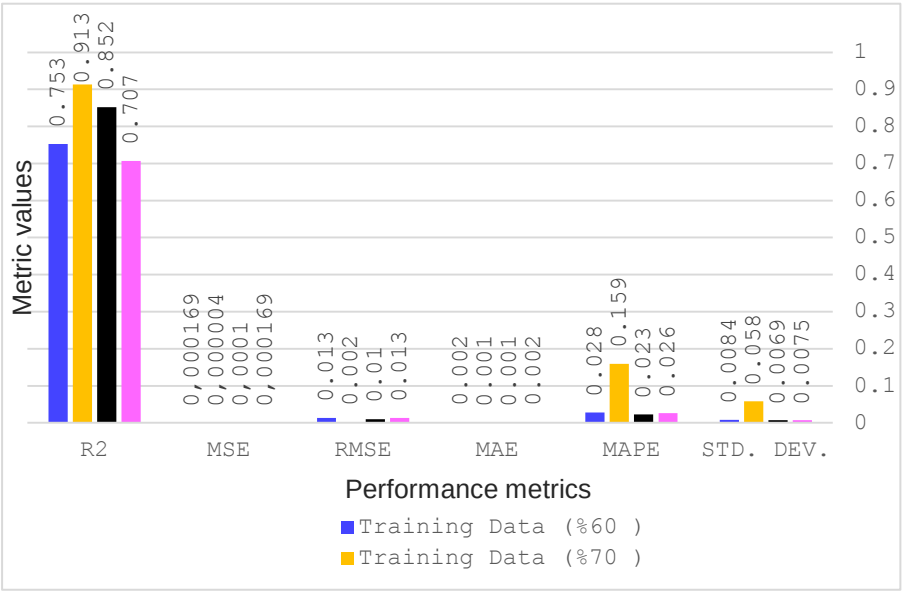

Fig. 6. Results from various training and test data set percentages using the gradient boosted regression tree approach.

The gradient boosted regression tree approach yielded optimal results by allocating 70% of the dataset for training and 30% for testing. Gradient enhanced regression tree method shows a significant disparity in results between training and test data compared to linear regression. In addition, GBRT has been observed that the performance of the model decreases as the training set size increases. Low values obtained from the metrics indicate that the generalizability of the model decreases. When the training set size increases significantly, the model may overfit the training data. This leads to low performance on unseen test sets. Figure 7 displays the flow diagram for the ANN technique using the Knime tool. Cross-validation was used to calculate the results. Cross-validation is performed with the X-Partitioner node.

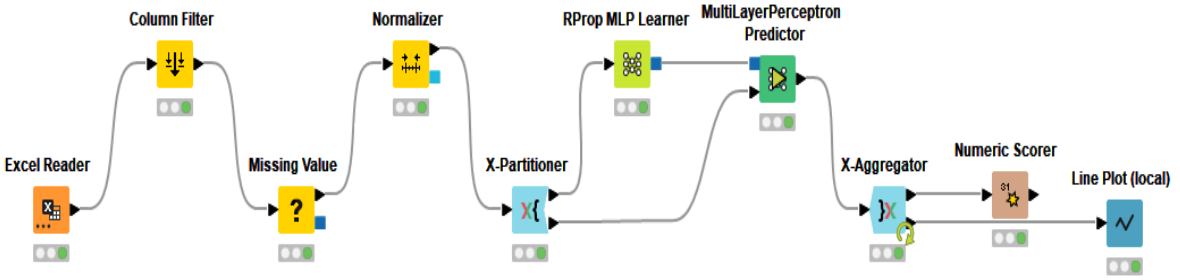

Fig. 7. Flow diagram of MLP model in Knime tool.

The neural network results were confirmed using cross-validation. The cross-validation method made use of the "X-Partitioner" node included in the Knime tool. The data was divided into ten divisions and random sampling was conducted. The "RProp MLP Learner" node was utilized in the ANN method to analyze the dataset in the learning phase. The parameters were as follows: 100 iterations were performed using a neural network with three hidden layers, each containing 100 neurons, and the goal class column was set to 2021. Figure 8 shows the results of the cross-validation process.

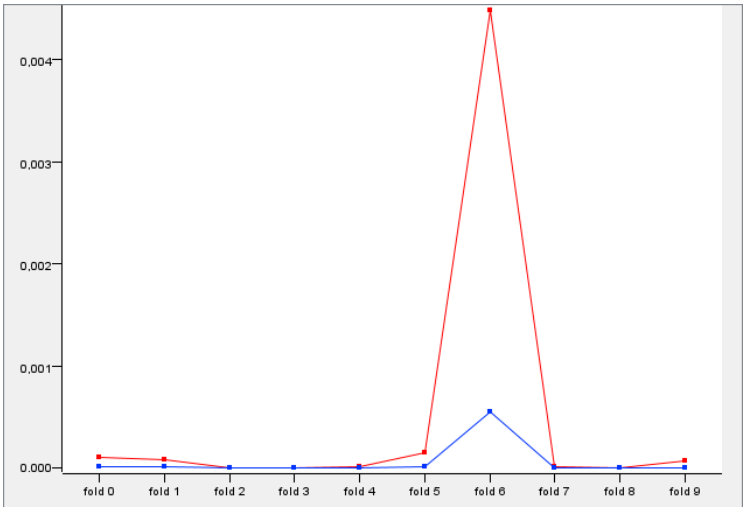

Fig.8. MSE values obtained as a result of cross-validation on multilayer perceptron (red is total MSE, blue is average MSE).

The variance in MSE values among folds is deemed acceptable given the scarcity of provincial data in Turkey and the limited historical data available on the official website. Cluster 6 had the lowest MSE value among the sub-datasets, whereas cluster 8 had the highest MSE value. MSE value suggests a favorable outcome based on the average value.

Figure 9 displays a comparison of the three approaches in the study based on R^2, MSE, RMSE, MAE, MAPE and STD. DEV. values using a bar chart. The Knime nodes used for training are set to repeat a maximum of 50 times with random seed. The node named Numeric scorer gives the best results and their standard deviations when displaying the results.

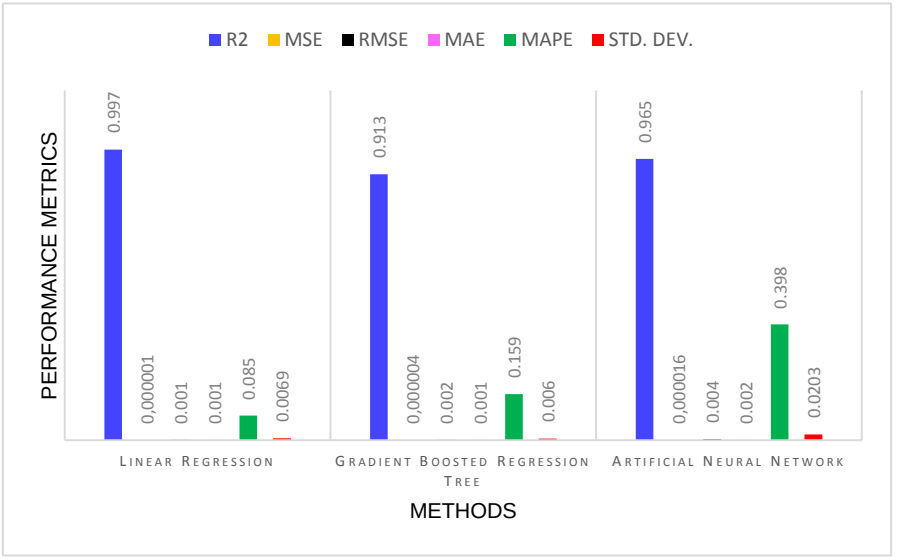

Fig. 9. Comparison of three methods based on R^2, MSE, RMSE, MAE, MAPE and STD. DEV. values.

The analysis findings from the training and test data sets indicate that the predictions made by each approach are within an acceptable range. The linear regression technique yielded the highest R^2 score in the experiments. The ANN approach had the highest rate for estimating the variance of the dependent variable from the independent factors. Performance evaluation of the model is clearer when the R^2 analysis is combined with additional techniques, especially MSE. As the data points approach the regression line, the MSE decreases and eventually approaches zero. Models that predict with less error are more accurate. In both the training and test datasets of this study, the MSE value for each of the three models is approximately 0. The RMSE value represents the square root of the MSE value. It simplifies the comprehension of exceedingly large numbers. Optimal RMSE values are those that approach zero. RMSE values between 0.001 and 0.004 are considered ideal for our study. An ideal range of 0.001-0.002 was identified for the MAE values, which indicate the average absolute difference between the observed data values. MAPE was 0.398 for the ANN, 0.159 for the gradient boosted regression tree, and 0.085 for linear regression. The linear regression approach yielded the best result when all assessment factors were taken into account, similar to the MAPE value.

The high R^2 score obtained with the linear regression technique shows that this method is quite effective for this dataset. However, the artificial neural network (ANN) approach, despite achieving a slightly lower R^2 score, showed a strong ability to capture the variability in the data and proved to be a reliable alternative especially for more complex data patterns. The results obtained are also consistent with previous studies where artificial neural networks were preferred due to their flexibility and accuracy.

Unlike other studies, this research emphasizes how machine learning can be used in environmental management by using these methods to predict urban forest recreation areas. The results obtained demonstrate the effectiveness of these methods and provide a basis for investigating ways to combine them with other advanced techniques to provide greater accuracy.

The impact of the findings is not limited to statistical estimates alone, but also provides valuable information for urban planners and policy makers. It can contribute to making more informed decisions on urban development and environmental protection by accurately predicting recreation area needs. It also highlights the importance of

integrating machine learning tools into traditional urban planning approaches to deal with the contemporary problems of our rapidly urbanizing world.

4. Conclusion

This study aimed to forecast the provincial recreational areas inside urban forests in Turkey and analyze future forest resources using official statistics from 2013 to 2020. Three methods, namely linear regression, gradient boosted regression tree, and ANN, were applied to the data using the KNIME platform. The metrics used to evaluate the performance of the methods are R^2, MSE, RMSE, MAE, MAPE, respectively. The most successful method was linear regression. Inferences from this study can be used in the planning and conservation of forest assets. It is believed that this study makes an important contribution to the literature in terms of protecting existing forest assets and preserving the ecosystem in its natural state. In this study, which emphasizes the protection of forest assets, it has been shown that computer science can be used functionally in this regard. If larger and more detailed data sets can be obtained, much more detailed analyzes will be possible. The limitations of this study are that the study was carried out within the scope of specified areas and within certain parameters.

It has been seen that this study is important in terms of making predictions about future forest assets. Such studies are important in order to take measures to protect our natural assets, which are extremely important for us. There are similar studies in the literature. However, this study has a special importance as it is a country-based study and draws attention to the issue of urban recreation areas, which is one of the main problems of the country.

Accessing larger data sets of the study will enable more in-depth analyses. Efficient actions can be taken to implement the required strategies for forest resources. Accessing greater data sets will enable more intricate analyses. This will enable efficient strategies to be implemented for managing forest resources.

The findings from the study are instructive in the area of the use of machine learning techniques in the prediction of urban forest recreation areas in Turkey. The efficacy of linear regression, gradient-assisted regression tree and artificial neural network methods in this field has been demonstrated.

Using three different methods and various performance metrics, this study shows the effectiveness of artificial neural networks with the highest performance. In addition to the satisfactory results obtained, it is predicted that the methods used in the study will give successful results in more complex and detailed data sets. This study, which can be a guide for future studies, stands out especially as it is a study on the characteristic structures of the cities of countries.

Another critical importance of the study is that it provides guidance for conservation policies related to forest areas. Based on the results of the study, it is possible to take some measures to protect and increase our forest areas in the future. It is also considered to be of particular importance in terms of encouraging the protection and development of our natural resources. The main limitations of this study are that there are some missing data in the dataset and it does not contain too much detail. This situation can be improved and pave the way for more comprehensive studies. For future studies, existing methods can be improved or different machine learning algorithms can be used.

While our study provides a solid basis for estimating urban forest recreation areas, it is important to recognize several limitations. The biggest limitation is that the dataset only covers the years 2013-2020. Not including data from older years or additional variables such as socio-economic factors may limit the generalizability of the findings. Moreover, focusing only on provincial level data may miss differences at the local level (e.g. district or neighborhood level).

For future research, expanding the dataset to include a wider range of variables and more recent data could significantly improve the predictive performance of the model. Furthermore, applying different machine learning algorithms such as Random Forests, Support Vector Machines, or hybrid models combining multiple approaches can provide a more comprehensive understanding of the factors influencing urban forest recreation areas.

Another area to explore in the future is the integration of spatial analysis techniques such as Geographic Information Systems (GIS) with machine learning models. This can enable more precise spatial predictions and a better understanding of the geographical distribution of urban forest resources. Furthermore, long-term studies that track changes over time can provide important insights into the dynamic nature of urban recreation areas and the long-term impact of urbanization on these areas. By addressing these limitations and exploring these issues, future studies can build on existing findings to provide more detailed and actionable insights for urban forest management and policy development.

Author contribution

M.C.Ö., S.D. and T.T.B. actively participated in conducting the experimental studies and writing the manuscript.

Acknowledge

The authors declare that they have no conflict of interest.

References

[1] P. L. Winter, S. Selin, L. Cerveny and K. Bricker, "Outdoor recreation, nature-based tourism, and sustainability," *Sustainability*, vol. 12, no. 1, pp. 81, 2019, doi: 10.3390/su12010081.
[2] Ç. Kılıçaslan, "Ortaca kenti rekreasyon alanlarının mevcut durumu ve Muğla Üniversitesi Ortaca Meslek Yüksekokulu öğrencilerinin rekreasyon alanlarına yönelik beklentileri," *Düzce Üniversitesi Ormancılık Dergisi*, vol 4, no. 1-2, pp. 3-16, 2008.
[3] S. Uzun and H. Müderrisoğlu, "Kirsal ve kentsel alanlardaki parklarda kullanici memnuniyeti; Gölcük orman içi dinlenme alani ve İnönü Parkı örneği," *Düzce Üniversitesi Orman Fakültesi Ormancılık Dergisi*, vol. 3, no. 2, pp. 84-101, 2007.
[4] H. Akyüz, M. Kul and F. Yaşartürk, "Rekreasyon açısından ormanlar ve çevre," *International Journal of Sport Culture and Science*, vol 2, no. (Special Issue 1), pp. 881-890, 2016.
[5] M. G. Rupert, S. H. Cannon and J. E. Gartner, "Using logistic regression to predict the probability of debris flows occurring in areas recently burned by wild land fires," *US Geological Survey Open-File Report*, vol. 500, no. 1, 2003.
[6] T. Bjerke, C. T. and J. Kleiven, "Outdoor recreation interests and environmental attitudes in Norway," *Managing leisure*, vol. 11, no. 2, pp. 116-128, 2006, doi: 10.1080/13606710500520197.
[7] Y. Nong and Q. Du, "Urban growth pattern modeling using logistic regression," *Geo-spatial Information Science*, vol. 14, no. 1, pp. 62-67, 2011, doi: 10.1007/s11806-011-0427-x.
[8] H. M. Shaikh, M. S. Patterson, B. Lanning, M. R. Umstattd Meyer and C. A. Patterson, "Assessing college students 'use of campus recreation facilities through individual and environmental factors," *Recreational Sports Journal*, vol. 42, no. 2, pp. 145-159, 2018, doi: 10.1123/rsj.2017-0.
[9] K. Kozlov, A. Singh, J. Berger *et al.* "Non-linear regression models for time to flowering in wild chickpea combine genetic and climatic factors," *BMC Plant Biol*, vol. 19, no. 94, pp. 1-14, 2019.
[10] N. Başaran, D. K. Matcı and U. Avdan, "Using multiple linear regression to analyze changes in forest area: the case study of Akdeniz Region," *International Journal of Engineering and Geosciences*, vol. 7, no. 3, pp. 247-263, 2022, doi: 10.26833/ijeg.976418.
[11] Z. Liu, C. Peng, T. Work, J. N. Candau, A. DesRochers and D. Kneeshaw, "Application of machine-learning methods in forest ecology: recent progress and future challenges," *Environmental Reviews*, vol. 26, no 4, pp. 339-350, 2018, doi: 10.1139/er-2018-0034.
[12] Q. Zhao, S. Yu, F. Zhao, L. Tian and Z. Zhao, "Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments," *Forest Ecology and Management*, vol. 434, pp. 224-234, 2019, doi: 10.1016/j.foreco.2018.12.019.
[13] J.V. Solórzano and Y. Gao, "Forest disturbance detection with seasonal and trend model components and machine learning algorithms," *Remote Sensing*, vol. 14, no. 3, pp. 803, 2022, doi: 10.3390/rs14030803.
[14] S. G. Gocheva-Ilieva, A. V. Ivanov and I. E. Livieris, "High performance machine learning models of large scale air pollution data in urban area," *Cybernetics and Information Technologies*, vol. 20, no. 6, pp. 49-60, 2020, doi: 10.2478/cait-2020-0060.
[15] M. Pourshamsi, M. Garcia, M. Lavalle and H. Balzter, "A machine-learning approach to PolInSAR and LiDAR data fusion for improved tropical forest canopy height estimation using NASA AfriSAR Campaign data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 10, pp. 3453-3463, 2018, doi: 10.1109/JSTARS.2018.2868119.
[16] X. Liu, S. Liang, B. Li, H. Ma and T. He, "Mapping 30 m fractional forest cover over China's Three-North Region from Landsat-8 data using ensemble machine learning methods," *Remote Sensing*, vol. 13, no. 13, pp. 2592, 2021, doi: 10.3390/rs13132592.
[17] K. Manley, and B. N. Egoh," Mapping and modeling the impact of climate change on recreational ecosystem services using machine learning and big data", *Environmental Research Letters*, vol. 17, no. 5, pp 054025, 2022.
[18] C. Chen, Z. Shen, Y. Weng, S. You, J. Lin, S. Li, and K. Wang, "Modeling Landslide Susceptibility in Forest-Covered Areas in Lin'an, China, Using Logistical Regression, a Decision Tree, and Random Forests", *Remote Sensing*, vol. 15 no. 18, pp 4378, 2023.

[19] A. Adhikari, C. R. Montes, and A. Peduzzi, "A comparison of modeling methods for predicting forest attributes using LiDAR metrics", *Remote Sensing*, vol. 15, no. 5, pp 1284, 2023.
[20] Y. Zhou, J. Hu, M. Liu, and G. Xie, "Predicting Sub-Forest Type Transition Characteristics Using Canopy Density: An Analysis of the Ganjiang River Basin Case Study", *Forests*, vol. 15, no. 2, pp 274, 2024.
[21] M. Kaya and S. A. Özel, "Açık kaynak kodlu veri madenciliği yazılımlarının karşılaştırılması," *Akademik Bilişim*, pp. 1-8, 2014.
[22] Republic of Turkey Ministry of Agriculture and Forestry General Directorate of Forestry, "Official statistics." ogm.com, https://www.ogm.gov.tr/tr/e-kutuphane/resmi-istatistikler (accesed Feb. 1, 2023).
[23] A.O. Sykes, "An introduction to regression analysis," *Coase-Sandow Working Paper Series in Law and Economics*, 1993.
[24] S. Kılıç, "Doğrusal regresyon analizi," *Journal of Mood Disorders*, vol. 3, no. 2, pp. 90-92, 2013, doi: 10.5455/jmood.20130624120840.
[25] S. Dörterler, "Developing a prediction model with the Battle Royale Optimization Algorithm," in *International Research in Engineering Sciences III*, vol. 1, M. Kamanlı, Eds. Konya, Turkey: Egitim Publishing, 2022, pp. 5-19.
[26] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140-147, 2020, doi: 10.38094/jastt1457.
[27] Ö. G. Uzut and S. Buyrukoğlu, "Veri madenciliği algoritmalari ile gayrimenkul fiyatlarinin tahmini," *Euroasia Journal of Mathematics, Engineering, Natural & Medical Sciences*, vol. 7, no. 9, pp. 77-84, doi: 10.38065/euroasiaorg.81.
[28] H. Alshari, A. Saleh and A. Odabas, "CPU performansı için gradyan artırımlı karar ağacı algoritmalarının karşılaştırılması," *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*, vol. 37, no. 1, pp. 157-168, 2021.
[29] Y. Shin, "Application of boosting regression trees to preliminary cost estimation in building construction projects," *Computational intelligence and neuroscience*, vol. 2015, pp. 1-1, 2015, doi: 10.1155/2015/149702.
[30] İ. Pençe, A. Kalkan and M. Ş. Çeşmeli, "Turkey sanayi elektrik enerjisi tüketiminin 2017-2023 dönemi için yapay sinir ağları ile tahmini," *Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi*, vol. 3, no. 2, pp. 206-228, 2019, doi: 10.31200/makuubd.538878.
[31] R. M. Sadek, S. A. Mohammed, A. R. K. Abunbehan, A. K. H. A. Ghattas, M. R. Badawi, M. N. Mortaja and S. S. Abu-Naser, "Parkinson's disease prediction using artificial neural network," *International Journal of Academic Health and Medical Research*, vol. 3, no. 1, pp. 1-8, 2019.
[32] K. Öztürk and M. E. Şahin, "Yapay sinir ağları ve yapay zekâ'ya genel bir bakış," *Takvim-i Vekayi*, vol. 6, no. 2, pp. 25-36, 2018.
[33] K. Y. Lee, K. H. Kim, J. J. Kang, S. J. Choi, Y. S. Im, Y. D. Lee, and Y. S. Lim, "Comparison and analysis of linear regression & artificial neural network*"*, *International Journal of Applied Engineering Research*, vol. 12, no. 20, pp. 9820-9825, 2017.
[34] D. Özdemir, S. Dörterler and D. Aydın, "A new modified artificial bee colony algorithm for energy demand forecasting problem," *Neural Computing and Applications*, vol. 34, no. 20, pp.17455-17471, 2022.
[35] D. Özdemir and S. Dörterler, "An adaptive search equation-based artificial bee colony algorithm for transportation energy demand forecasting," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 30, no. 4, pp. 1251-1268, 2022.