land

MDPI

Article

Identifying Key Environmental Factors for *Paulownia coreana* Habitats: Implementing National On-Site Survey and Machine Learning Algorithms

Yeeun Shin [1], Suyeon Kim [2], Se-Rin Park [1], Taewoo Yi [3], Chulgoo Kim [3], Sang-Woo Lee [1] and Kyungjin An [1,*]

[1] Department of Forestry and Landscape Architecture, Konkuk University, Seoul 05029, Korea; julie9276@konkuk.ac.kr (Y.S.); serin87@konkuk.ac.kr (S.-R.P.); swl7311@konkuk.ac.kr (S.-W.L.)
[2] Rural Environment & Resource Division, National Institute of Agricultural Sciences, Wanju-gun 55365, Korea; mdln94@korea.kr
[3] National Institute of Ecology, Seocheon-gun 33657, Korea; ytw117@nie.re.kr (T.Y.); ecorest@nie.re.kr (C.K.)
* Correspondence: dorian@konkuk.ac.kr

Abstract: Monitoring and preserving natural habitats has become an essential activity in many countries today. As a native tree species in Korea, *Paulownia coreana* has periodically been surveyed in national ecological surveys and was identified as an important target for conservation as well as habitat monitoring and management. This study explores habitat suitability models (HSMs) for *Paulownia coreana* in conjunction with national ecological survey data and various environmental factors. Together with environmental variables, the national ecological survey data were run through machine learning algorithms such as Artificial Neural Network and Decision Tree & Rules, which were used to identify the impact of individual variables and create HSMs for *Paulownia coreana*, respectively. Unlike other studies, which used remote sensing data to create HSMs, this study employed periodical on-site survey data for enhanced validity. Moreover, localized environmental resources such as topography, soil, and rainfall were taken into account to project habitat suitability. Among the environment variables used, the study identified critical attributes that affect the habitat conditions of *Paulownia coreana*. Therefore, the habitat suitability modelling methods employed in this study could play key roles in planning, monitoring, and managing plants species in regional and national levels. Furthermore, it could shed light on existing challenges and future research needs.

Keywords: typological habitats; habitat monitoring; habitat suitability models; machine learning; Artificial Neural Network (ANN); Decision Tree & Rules

Citation: Shin, Y.; Kim, S.; Park, S.-R.; Yi, T.; Kim, C.; Lee, S.-W.; An, K. Identifying Key Environmental Factors for *Paulownia coreana* Habitats: Implementing National On-Site Survey and Machine Learning Algorithms. *Land* **2022**, *11*, 578. https://doi.org/10.3390/land11040578

Academic Editor: Benjamin Burkhard

Received: 18 February 2022
Accepted: 13 April 2022
Published: 14 April 2022

1. Introduction

Natural habitats are among the most essential ecological bases sustaining the existence and survival of life. The survival of all forms of life is inextricably tied to the status of their habitat. Species distribution modeling is considered an important aspect of various fields, such as biology and ecology; therefore, great attention has been given to species distribution research despite the complexity of the existing environments [1–6]. Such prediction of habitat suitability is necessary for the planning and implementation of forest conservation and management. In fact, habitat suitability models (HSMs) have already drawn great attention for predicting plant environments in various scenarios as a result of climate change. A number of HSMs have recently been created to envisage the environmental changes caused by recent climate issues [7–9]. HSMs can help to comprehensively understand potential habitats. The number and quality of predictive techniques have seen an increase in recent years, with a direct effect on the accuracy of the model.

However, most HSMs are based on remote sensing data, which has led to major validity and credibility issues. Therefore, implementing on-site survey data supplied by

the Korea National Institute of Ecology could play a key role in increasing credibility of habitat suitability modeling. Rather than the discrepancy of HSMs between remote sensing data and actual species distribution, this study focused on identifying the hierarchy of environmental variables that affect HSMs and creating HSMs for individual species to protect and maintain a suitable ecological status.

As in most countries, a number of floral and faunal species have been listed by the Ministry of Environment in Korea as unique and endangered, and with some kind of environmental, biological, and social value [10]. Periodically, various governmental agencies in South Korea carry out local, regional, and national surveys on flora and fauna. In particular, the National Institute of Ecology has spent a large amount of resources to perform extensive species surveying on a national scale to identify and analyze what needs to be conserved in the environment. The survey method is simple for flora. For the past decade, a large number of surveyors worked nationwide on the ground to identify individual plant species [10]. As such, the collected information is solid evidence based on site data, which are often used to create and maintain national ecological grading maps. Unfortunately, it has not yet been used to perform habitat suitability modeling for individual species, nor has it been used for habitat monitoring. *Paulownia coreana* is a protected tree plant species unique to the Korean Peninsula. Therefore, the study focuses on *Paulownia coreana* to identify environmental variables affecting its habitat and create HSMs.

The aim of the presented study is to develop these models for *Paulownia coreana* using National Ecological Survey data. Moreover, the study also aims to assess environmental variables to identify the most affective elements of the species habitat. In order to achieve these aims, the study employs machine learning algorithms of the Artificial Neural Network (ANN) and Decision Tree & Rules. The former is for assessing the hierarchy of environmental variables, while the latter is to develop habitat suitability models of *Paulownia coreana* and other key tree species.

The study is divided into five distinct sections. After the introduction, the literature review illustrates current trends on HSMs with machine learning algorithms. It also gives a clear indication on how HSMs are created with various algorithms along with their respective pros and cons. Moreover, it also expands on the implication of using remote sensing data for habitat suitability modeling. In the methodology section, the study illustrates the environmental variables applied to the National Ecological Survey data and employs two machine learning algorithms to identify environmental variables affecting *Paulownia coreana* habitats and create its HSMs. In the discussion section, the study argues for the impact of environmental variables and accuracy of habitat suitability models. Finally, the study concludes with a short review on the high potential of HSMs created from on-site survey data for monitoring and ecological policy planning.

2. Habitat Suitability Modeling with Machine Learning Algorithms: Literature Reviews

Recently, the growing efficiency of big data solutions and application of machine learning has received a huge amount of attention. With the proliferation of machine learning, many studies have employed machine learning algorithms in the field of ecology [11,12]. Big data from a variety of sources such as weather stations, National Ecological Survey, and public agencies can be used for habitat modeling. By synthesizing assorted big data, complex habitat suitability modeling of various scales becomes possible. However, as discussed later, only some machine learning algorithms are suited for plant species habitat prediction.

Within a broader context, machine learning algorithms have been utilized on ecological data [11]. For instance, they have assisted in the prediction of mass mortality events in the Mediterranean Sea [13]. However, a great number of studies have focused on remote sensing data in the implementation of artificial intelligence [14]. For instance, it is possible to predict plant species habitats through environmental variables derived from remote sensing data and analyze how variables affect habitat prediction.

Raghukumar and Narayanan have also suggested the comparison of machine learning algorithms for the detection of medicinal plants [12]. Automatic recognition of medicinal

plants has been tested, where their features (shape, texture, and color) were extracted from images of their leaves before being classified using machine learning classification techniques such as K-Nearest Neighbors algorithm (KNN) and Support Vector Machines (SVM). Another study [15] built relationships between field and remote sensing data with Random Forest (RF). The field data included over 15,000 points of data from the assessment, inventory, and monitoring of landscape monitoring framework programs throughout the Western U.S.

This integration between the remote sensing and machine learning process has been utilized often. For example, a study was carried out to predict vegetation classification using environmental data derived from various spatial scales, aerial images, and the classifier RF in Britain [16]. Another research study performed similar modeling and mapping using a bootstrap-aggregating machine learning ensemble with the RF classifier to derive a European forest formation suitability map [17]. The modeling used field data provided by the European Environmental Agency, and forests in 10 categories were classified. The overall accuracy of the model results was 76%, and the influence of environmental factors such as isotherm and precipitations and map applicability were discussed. Zlinszky et al. [18] used airborne laser scanning data for implementing habitat mapping with high resolution.

The marriage between machine learning and remote sensing is not uncommon in the field of ecology. A new methodology was developed [19] using drones (remotely piloted aircraft systems). It was characterized to analyze complex habitat environments and structures with high conservation value. Implementing a machine learning technique with another set of decision rules, the study carried out the discrimination of plant types. Meanwhile, a study that performed object-based image analysis (OBIA) and machine learning algorithm analysis was conducted [20]. The study classified the types of wetland plants in the Ramsar wetland conservation area in China and applied six machine learning algorithms to compare classification accuracy. The classification showed meaningful results, but problems appearing in the pixels and resolution of the image were derived.

There are also a number of ecological models derived from machine learning [1–6]. RF and Rotation Forest were employed for image classification using polarimetric and spatial features [21,22]. In the comparison between those two, the Rotation Forest produced better accuracy while RF calculated faster than Rotation Forest. Another study [23] demonstrated the applicability of machine learning (ANNs, Classification Additionally, Regression Trees, RFs, and SVMs) in habitat quality and its spatial diversity. The study developed the habitat suitability models with the data of *Oryzias latipes*, water depth and flow velocity in agricultural canals. Similar to the paper [23], habitats of a specific species were investigated with machine learning algorithms. For instance, *Pinus sylvestris* in Iberian Peninsula were examined in 2006 [24]. This research integrated several machine learning algorithms (Tree-based Classification, Neural Networks, and RF) within the Geographic Information System and predicted a habitat model. Hematological value references of *Sicalis flaveola* were also tested using machine-learning-based classifiers [25].

In terms of machine learning algorithms, the RF model has been used to classify fine-scale coastal vegetation aerial data [26]. In this study, near infrared imagery and DEM data were used to classify vegetation types, and a total of three scenarios were set to verify the model of the RF algorithm. Together with RF, SVM was implemented [27] to identify invasive plant species such as *Solidago* spp., *Calamagrosties epigejos*, and *Rubus* spp. from aerial images.

Machine learning is also commonly used for classifying plant species. Using SVM, Shobana and Perumal [28] structured and built an astute framework of machine vision that would advance plant development in restricted water conditions. Sukumaran et al. also presented [29] a model for phylogenies with evolution and diversification progression. In another study [30] carried out in 2019, plant diseases were recognized using Convolutional Neural Networks, predicting them by comparing the characteristics and changes of leaves. Meanwhile, a study [31] classified four artificial mangrove species using Decision Trees, SVM, and RF. The study was conducted using two classification systems (pixel-based; object-

based), and both methods resulted in meaningful classification results. Zohmann et al. [32] also successfully modelled habitat suitability applying the object-based classification for alpine rock ptarmigan (Lagoups muta Helvetica).

Meanwhile, this study aimed to predict the habitats of forest species, using the ANN and Decision Tree & Rules algorithms in order to identify environmental attributes affected to their ecological conditions. Within this framework, the study obtained on-site survey data from the National Ecological Survey. It is claimed [33] that remote sensing data have their own drawbacks of spatial, spectral, and radiometric limits on resolutions. However, implementing with field survey information can improve overall credibility and validity of the habitat models by interpolating or extrapolating individual locations. Such importance of survey data on credibility comes from fields. As it implicates on-site survey data for plant species in the duration of the last decade extensively, this study can open a new avenue to create, analyze, and evaluate HSMs along with the accumulation of periodic on-site survey data.

3. Methodology

3.1. Study Area

The study area comprises the South Korean Peninsula, including Jeju Island in the south. Spatial data were obtained from the Ordnance Survey of Korea, which covered the entire nation for a total area of 100,210 square km.

3.2. National Ecology Survey

Since 1986, the Ministry of Environment in Korea has periodically carried out national ecological surveys in the South Korean Peninsula. This survey was commissioned via statutory legislation and has been amended since the commencement. Every five years, flora and fauna are spatially identified by on-site surveyors, while any environmental issues including invasive species are identified and raised.

The study only uses data from the 3rd (2006–2013: eight years) and 4th (2014–2018: five years) National Ecological Survey since only these have been digitized in the database. Since 2006, the survey has utilized 1:25,000 terrain maps which are further divided into 824 cells. Each cell is split into nine grids which are assigned to the surveyors to work on (Figure 1).

3.3. Environmental Variables

A large number of environmental variables can influence the natural habitat. All the available variables that can affect habitat modeling on different scales were collected from various sources including Ministry of Environment and Forestry Commissions in Korea. Then, it was narrowed down to the nine variables most likely to affect plant habitats based on previous research and literature review. These are categorically topographic and climatic environmental variables as shown in Table 1 below.

Furthermore, the main datasets were obtained from the National Ecological Survey, which includes the locations of individual tree species. Because this study is mainly focused on tree species, the prediction included both climatic and topographic variables. Hence, a total of nine environmental variables were considered for modeling (Table 1). The topographic variables were derived from the National Forest Location Soil Maps, which are published by Forestry Commissions at a 1:25,000 scale, and include information such as forest management, maintenance, and evaluation. The annual rainfall attributes were mainly interpolated by means of trend surfaces and National Forest Location Soil Maps. Moreover, annual rainfall volume was extracted from Korea's Met Office, which is divided into 62 locations nationally. Then, the annual rainfall data were merged with tree locations using QGIS program (version 3.14) and adjusted accordingly.

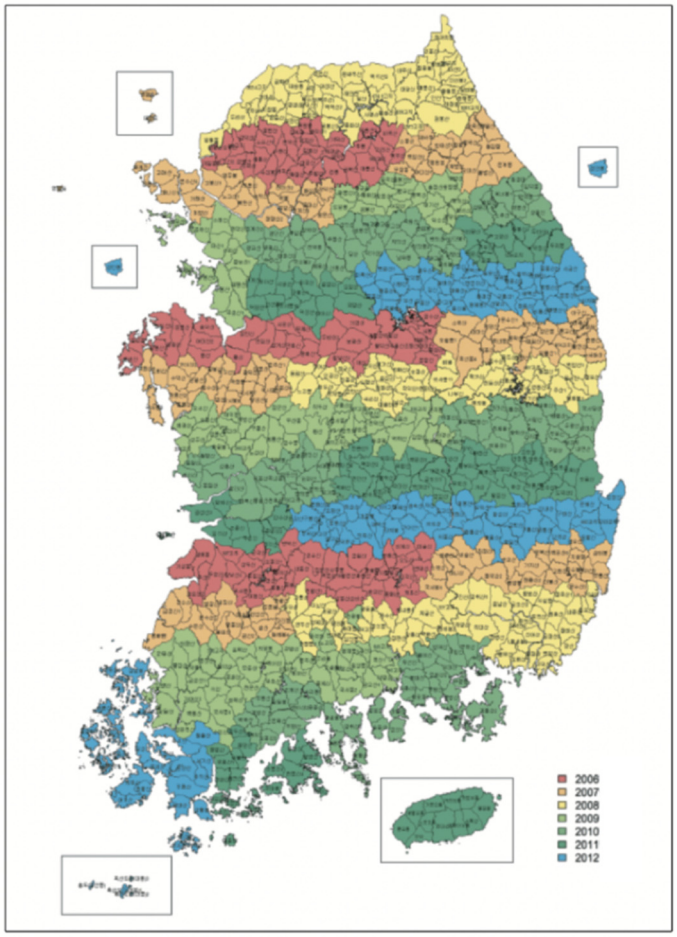

Figure 1. Areas in the 3rd National Ecological Survey. The surveyed elements include landform, vegetation, mammals, birds, amphibians, reptiles, insects, and fish (extracted from National Institute of Ecology's guideline 2019).

Table 1. Environmental variables implemented.

Environmental Variables		Grade Value	Description
Topographic	Soil Drainage Type	1–4	Poor, Normal, Good, Very Good
	Land Slope	001, 015, 020, 025, 030, 999	Below 15 degrees (Mild), 15–20, 20–25 (Steep), 25–30, More than 30 degrees, etc.
	Soil Accumulation	1–3	Residual (static soil), Creep, Colluvial Soil
	Altitude	01–20	Less than 100 m, 100–1900 m, More than 1900 m
	Soil Depth	10–30	Less than 30 cm, 30–60 cm, 61 cm or more
	Erosion	1–3	None, Slight, Heavy
Climatic	Wind Exposure	1–3	Exposed, Normal, Protected
	Annual Rainfall	mm	
	Weathering Effloresces Degree	01–03	Upper, Medium, Lower

3.4. On-Site Surveyed Species Locations

The presence of *Paulownia coreana* was taken from the 3rd and 4th phases of the National Ecological Survey. As the survey records contain ordinance locations of the species, environmental variables were embedded into ordinance survey maps within QGIS. Together with *Paulownia coreana*, other tree species such as *Robinia pseudoacacia*, *Quercus*

variabilis, and *Pinus densiflora* were used for identifying and evaluating the impact of environmental variables on the tree distribution.

3.5. The Modeling Process

A modeling framework was established as shown in Figure 2. Firstly, based on the processed data above, the machine learning algorithm appropriate for the habitat suitability models was selected, and two predictive models were chosen: ANN and Decision Tree & Rules. The software used for analysis was R 1.74 and the necessary packages were downloaded accordingly.

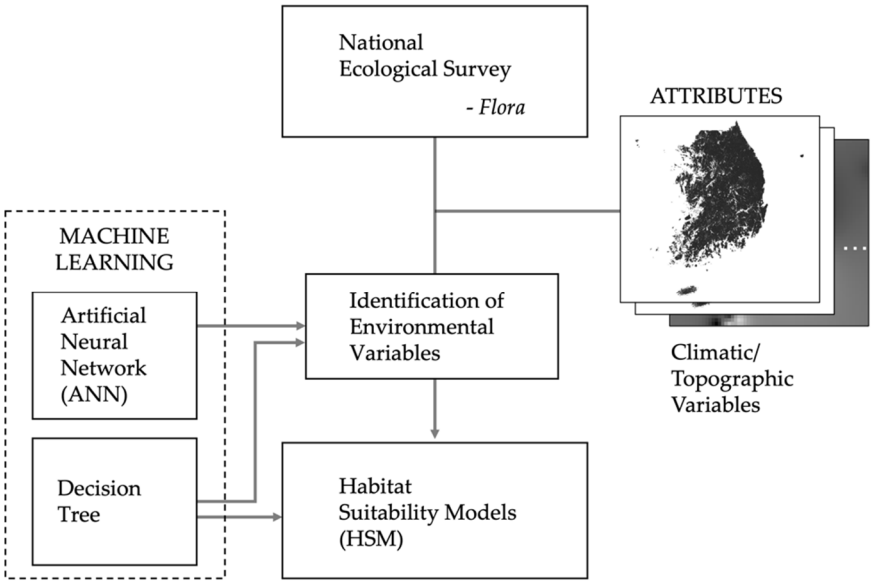

Figure 2. Modeling process.

The ANN models are effective for classification, prediction, and pattern recognition, forming networks according to input information in solving problems and modeling their relationships. Therefore, in this study, the ANN model was applied to derive the relationship between the nine variables affecting the habitat of *Paulownia coreana*. For executing a neural networks predictor, neuralnet package version 1.44.2 is installed in this study.

For the ANN analysis, numerical and factor data need to be comparable; therefore, in order to have them in the same scale, the normalize or standardize function was implemented. This process puts numerical and factor data of individual variables into a 0–1 scale. The function in the R is as below [34]:

```
> normalize <- function(x) {
    return((x − min(x))/(max(x) − min(x)))
}
```

The data were also partitioned into two main groups, training and testing. The prediction models were derived from training data processing; thereafter, the test datasets were evaluated through the predicted model to assess the model performances and their validity. In the case of ANN analysis, the training group contains 75% of the samples, and the testing group has 25% of the samples. However, in order to make valid models, the samples need to be randomly sorted. The training data samples were used to implement the ANN and the testing dataset was used to assess how valid the model is at identifying the effect of environmental variables on each other.

The Decision Tree & Rules models analyze the relationship between input information and results using tree structure. Within this study, a Decision Tree & Rules algo-

rithm was implemented to comprehensively analyze the variables affecting the habitat of *Paulownia coreana* and other tree species. For the Decision Tree & Rules process, the categorical values shown in Table 1 were applied as actual numerical values for more detailed results. Various algorithms have been utilized in performing the Decision Tree & Rules; this study used the C5.0 algorithm version 0.1.4.

As well as the ANN analysis, for the Decision Tree & Rules modeling process, the data were split into two groups. The training sample group was used to formulate Decision Tree models and the testing sample group to assess Decision Tree model performance. For the training group, 80% of all samples was selected, and the remaining 20% was used for the testing group. Additionally, the model boosted 10 trials for the performances. In this study, after performing each of the ANN and Decision Tree & Rules models, the results were compared and analyzed.

4. Results

This study mainly develops two models and scenarios with machine learning algorithms. The two machine learning algorithms, ANN and Decision Tree methods are implemented for plant HSMs. The comparison between two methods and models is shown in Table 2 below.

Table 2. Comparison of the machine learning algorithms.

Machine Learning Types	Algorithms	
Artificial Neural Network (ANN)	neuralnet	Identification of environmental variables affecting *Paulownia coreana* habitats
Decision Tree	C5.0	Habitat suitability models of *Paulownia coreana*, *Quercus variabilis*, *Pinus densiflora*, *Robinia pseudoacacia*

For the ANN analysis, 301 locations of *Paulownia coreana* were used from the National Ecological Survey. Therefore, the 301 rows of a data frame were created together with environmental variables indicated above, such as annual rainfall. Among the environmental variables implemented, Altitude and Slope are critical attributes affecting the habitat of *Paulownia coreana*; therefore, these are set up as output nodes. As illustrated in Figure 3 below, Altitude is used as an output node and is run through the multi-layer feedforward perception with a single node first.

Species_model <- neuralnet(Altitude ~ Drainage + AnnualRainfall + Slope + SoilAccumulation + ErosionLevel + WindExposure + WeatheringLevel + SoilDepth, data = species_train)

This network has one input for the nine variables. With a hidden node, an output node predicts altitude-focused habitats. It also shows individual weights for each connection. The numeric constants show biased value as in a linear equation. The negative signs within the network indicate inverse proportional relationships rather than literally negative effects. The number of training steps and an error measure are given at the bottom of the figure. A lower number indicates better projecting performance.

The multi-layer forward network with a single node to Altitude (Figure 3) indicates that Slope is the strongest element affecting *Paulownia coreana's* habitats with Altitude (7.19). This is followed by other environmental variables such as Soil Depth (3.26), Drainage (2.21), Soil Accumulation (−2.12), Wind Exposure (1.84), Erosion Level (−0.91), Annual Rainfall (0.41), and Weathering Level (0.33) in descending order of effect.

The network topology diagram provides the ANN's black box feature but does not indicate much about the model's future performance; however, the single-layered perception model (Figure 3) is performing well showing only a hidden node, the study increased the number of hidden nodes to five in order to improve the model's performance. As a result, the study illustrates much improved models in Figure 4 as below.

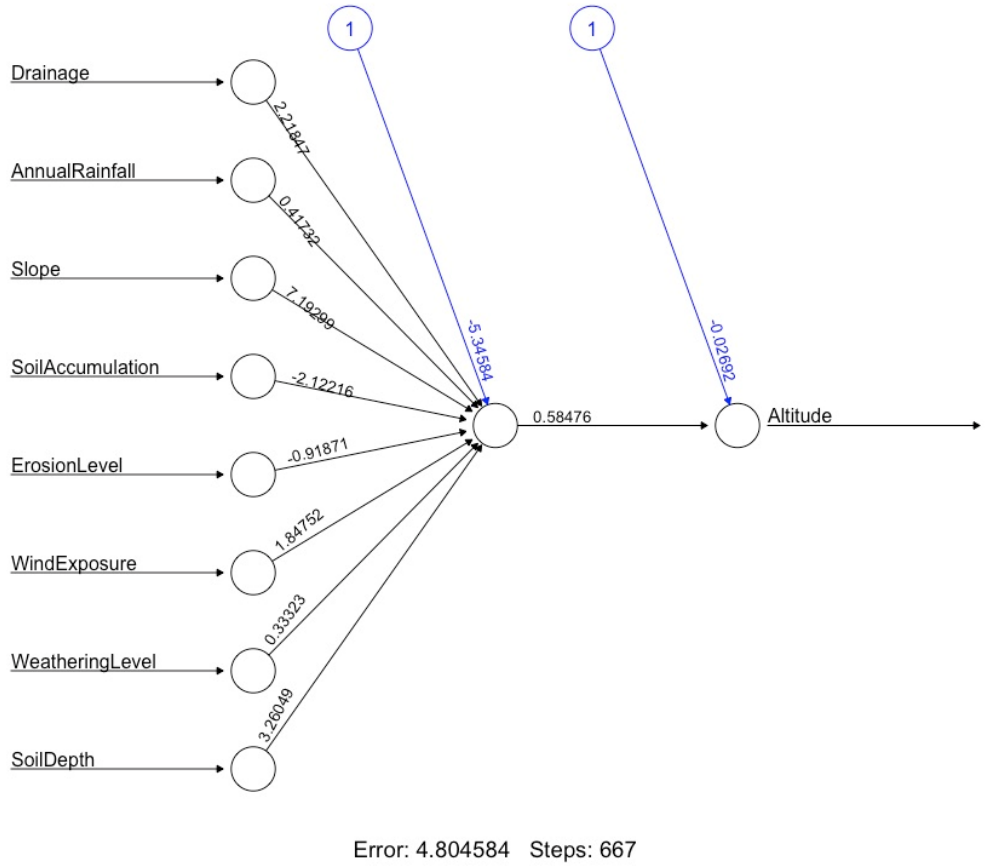

Figure 3. Multi–layer forward network with a single node to Altitude.

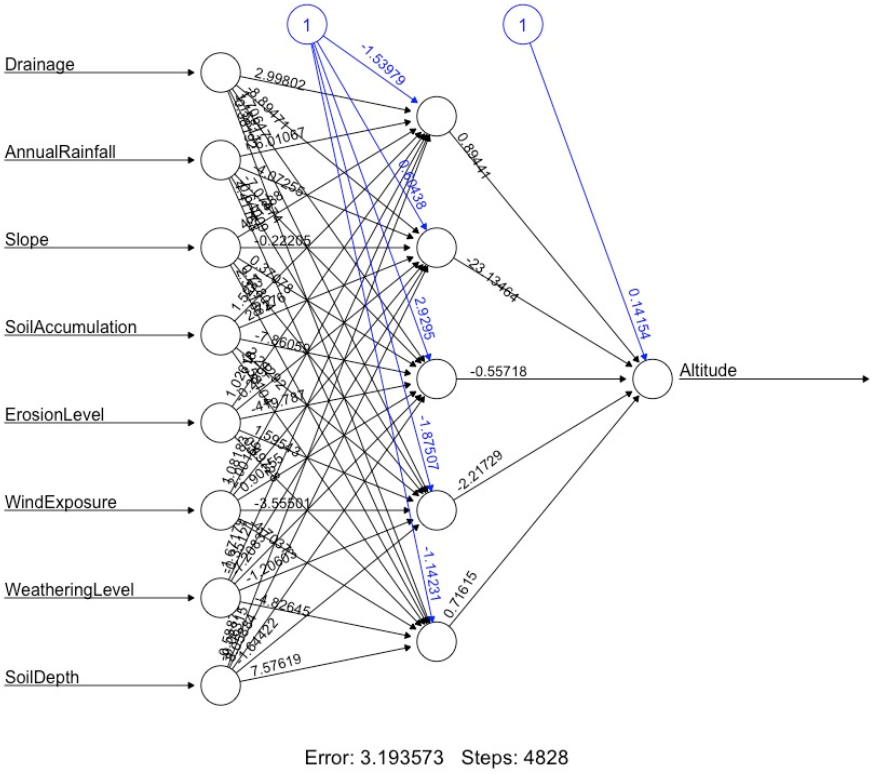

Figure 4. Multi–layer forward network with five hidden nodes to Altitude.

In this result, while the number of training steps increased from 667 to 4828, the error decreased much from 4.80 to 3.19. This made the model significantly more complex; generally, more complex models take more weights into account. The Slope variable is also analyzed in the same process as Altitude. In Figure 5, as well as in the case of Altitude, Slope and the rest of the nine variables were used, with a hidden node and an output calculating the slope-focused habitats.

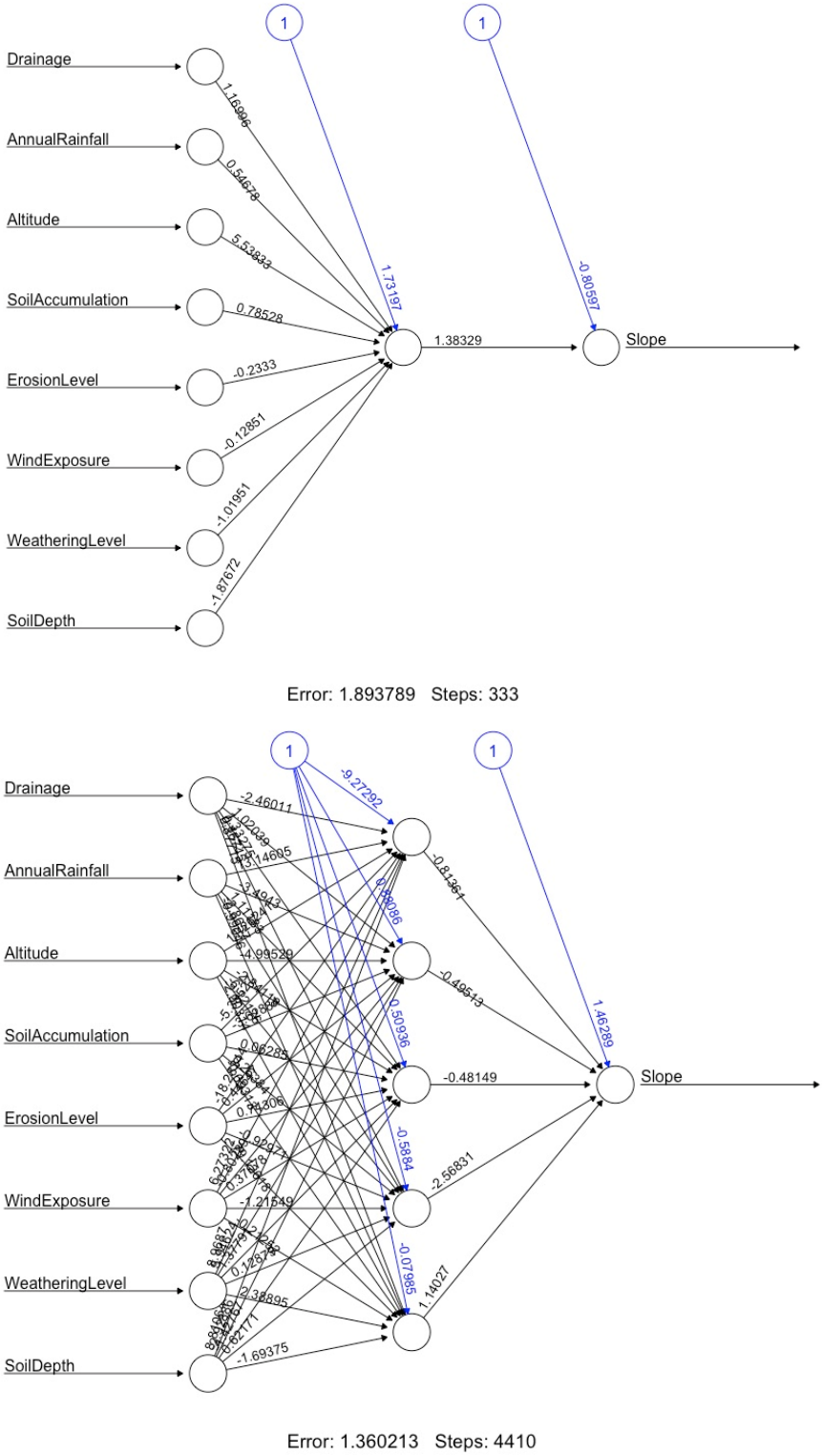

Figure 5. Multi–layer forward network with a single and five hidden nodes to Slope.

The multi-layer forward network with a single node to Slope (Upper, Figure 5) indicates that Altitude is the strongest element affecting *Paulownia coreana* habitats with Slope (5.53). Other environmental variables such as Soil Depth (−1.87), Drainage (1.16), Weathering Level (−1.01), Soil Accumulation (0.78), Annual Rainfall (0.54), Erosion Level (−0.23), and Wind Exposure (−0.12) follow in descending order of their effects (Table 1).

In our comparison of a single node and five hidden nodes, while the number of training steps increased from 333 to 4410, the reported error decreased from 1.89 to 1.36. Similar to the altitude analysis, this result explains that more complicated models provide better performance outcomes owing to the optional weights added to complicated models.

Based on the cross-analysis of environmental variables between Altitude and Slope, we thus conclude that Altitude and Slope are the main attributes affecting the habitat suitability modeling of *Paulownia coreana*.

While the environmental variables were assessed in conjunction with ANN algorithms, Decision Tree algorithms were implemented to create HSMs. For the Decision Tree modeling, additional tree species were introduced in order to carry out comparisons of habitat suitability. Moreover, as well as the ANN process, the Decision Tree models with the additional species enables us to identify the environmental variables which significantly affect the habitat suitability. Therefore, additional tree species of *Robinia pseudoacacia*, *Quercus variabilis*, and *Pinus densiflora* were selected. Decision Tree algorithms are commonly implemented in the ecology field since they produce accurate and statistical outcomes. The HSMs for individual tree species will be created based on environmental variables; when all available information surveyed is included, more accurate HSMs for particular species habitats can be created. This study developed a simple HSM especially implementing C5.0 DT packages.

The dataset used includes 2497 examples on nine topographic and climatic environmental variables (Table 1) previously identified from four species (*Paulownia coreana*, *Robinia pseudoacacia*, *Quercus variabilis*, *Pinus densiflora*). A class variable indicates which species are which out of the four species. The ANN process R 1.74 was employed for the Decision Tree procedure.

In addition to the ANN process conducted previously, the study randomized the data sample order to increase the validity of Decision Tree algorithms. Within the four different plant species in the datasets, the species needed to be shuffled well for better performance outcomes.

We implemented the C5.0 DT algorithm and the basic model, species_model was created. The species_model reveals 267 (tree size), which this means that the tree has 267 decisions. This resulted in the following (Algorithm 1):

Algorithm 1 Result of C5.0 DT algorithm

```
C5.0 [Release 2.07 GPL Edition]    Tue Nov 16 17:20:35 2021
Class specified by attribute 'outcome'
Read 2000 cases (10 attributes) from undefined.data
Decision Tree:
Slope <= 21:
:...ErosionLevel > 1:
:   :...SoilAccumulation > 1: Pinus densiflora (3/1)
:   :   SoilAccumulation <= 1:
:   :   :...AnnualRainfall > 1506:
:   :       :...SoilDepth <= 10: Quercus variabilis (2/1)
:   :       :   SoilDepth > 10: Pinus densiflora (16/6)
:   :       AnnualRainfall <= 1506:
:   :       :...WeatheringLevel <= 2: Pinus densiflora (3/1)
:   :           WeatheringLevel > 2:
:   :           :...SoilDepth <= 20: Robinia pseudoacacia (9/4)
:   :               SoilDepth > 20: Pinus densiflora (2/1)
```

In the results above, the first three lines can be interpreted as follows: when the Slope is no more than 21 deg, Erosion Level is none, and Soil Accumulation is more than 1 (static), the species is likely *Pinus densiflora* (3/1). However, when Slope is no more than 21 deg, Erosion Level is none, and Soil Accumulation is no more than 1 (creep and colluvial), species are likely dependent on Annual Rainfall volume. This continues as this Decision Tree model is 267 decisions deep. After the tree, the species_model output indicates a confusion matrix with cross tabulation of the model's performance data as shown below (Table 3).

Table 3. Confusion matrix of the species model.

Decision Tree				
Size	*Errors*			
267	*744 (37.2%)*	*<<*		
(a)	*(b)*	*(c)*	*(d)*	
100	*87*	*28*	*21*	*(a): class Paulownia coreana*
8	*609*	*75*	*57*	*(b): class Pinus densiflora*
11	*186*	*308*	*38*	*(c): class Quercus variabilis*
22	*151*	*60*	*239*	*(d): class Robinia pseudoacacia*

Within the model results, 744 sample instances out of 2000 were classified rightly, which indicates a 37.2% error rate. Then, the results also illustrate the weight of environmental variables taken into account as per Table 4 below.

Table 4. Attribute usage of Decision Tree algorithm.

	Percentage	**Environmental Variables**
1	100.00	Slope
2	96.05	Altitude
3	82.50	Weathering Level
4	73.75	Annual Rainfall
5	72.30	Drainage
6	62.70	Soil Accumulation
7	62.35	Soil Depth
8	57.55	Erosion Level
9	51.75	Wind Exposure

In order to implement these Decision Tree results to the test data, the study used the *predict()* script in R and the results are shown in Table 5 below.

Within the 497 testing species samples, the output model only rightly calculated *Paulownia coreana* (10), *Pinus densiflora* (88), *Quercus variabilis* (35), and *Robinia pseudoacacia* (25), indicating 49.7% accuracy and 50.3% error rate. The performance is not good enough for training outcomes; however, this result is not surprising under the circumstances wherein environmental variables often overlap without clear distinctive features. Hence, the study found that the error rate of environmental models is often too high for evaluating the suitability of habitats. Through the C5.0 DT algorithm, the study boosted the performance by combining further trial parameters, with the 10 trials added here.

> species_boost10 <- C5.0(species_train[−10], species_train$Species, trials = 10)

The prediction performance was not improved even after the boost, as shown by the 49.7% accuracy. Only attribute usage revealed any changes as shown in Table 6 below, which indicates overall improvement in the usage of environmental variables.

Table 5. Cross tables of predicted and actual species.

	Predicted Species				
Actual Species	*Paulownia coreana*	*Pinus densiflora*	*Quercus variabilis*	*Robinia pseudoacacia*	Row Total
Paulownia coreana	10 0.020	27 0.054	15 0.030	13 0.026	65
Pinus densiflora	15 0.030	88 0.177	51 0.103	26 0.052	180
Quercus variabilis	16 0.032	56 0.113	35 0.070	28 0.056	135
Robinia pseudoacacia	5 0.010	61 0.123	26 0.052	25 0.050	117
Column Total	46	232	127	92	497

Table 6. Attribute usage after the boost.

	Percentage	**Environmental Variables**
1	100.00	Slope
2	100.00	Altitude
3	93.45	Weathering Level
4	92.70	Drainage
5	88.45	Annual Rainfall
6	79.80	Wind Exposure
7	79.25	Soil Depth
8	74.75	Soil Accumulation
9	63.45	Erosion Level

5. Discussion

This study introduced a framework for handling environmental variables and creating HSMs with machine learning algorithms. The procedures are based on on-site surveys and this approach has opened substantial possibilities for future dealings with the National Ecological Survey, creating HSMs and enabling future species predictions, which has real applicability in protecting and managing forests. This approach can be particularly implemented to model plants proliferated owing to recent climate issues. Moreover, it forms a new methodology relating to the various scenarios represented in the literature review, particularly for the incorporation of the ANN and Decision Tree algorithm in the assessment of environmental variables. The results obtained from the Decision Tree method used to predict habitat suitability were not very promising, showing the lowest accuracy among the various machine learning algorithms used in this study. However, the authors had already predicted that environmental variables such as Annual Rainfall, Slope, and Altitude would not play distinctive roles in creating HSMs. As tree species, they would show minor differences in the conditions of habitats. We, however, successfully identified the individual environmental variables affecting the habitats. We implemented nine variables of Slope, Altitude, Weathering Level, Drainage, Annual Rainfall, Wind Exposure, Soil Depth, Soil Accumulation, and Erosion Level as initial attributes. We managed to successfully identify the Slope and Altitude as the most influential environmental variables for HSMs.

As claimed previously, the study employed two types of machine learning algorithms to identify environmental variables and build HSMs. The first of these are the ANNs, which have been gradually implemented for plant distribution modeling. However, the main weakness of ANN is that it is a black box approach whose process and results are difficult to explain. Furthermore, a lot of variables must be considered, such as the number

of hidden layers and neurons, weight decay, learning parameters, and primary connections between the node weights. Hence, high predictive accuracy can only be achieved through the avoidance of overfitting effects in the process.

The study's ANN demonstrated robust relationships among pre-selected environmental variables affecting plant habitats. The nine environmental variables of topographic and climatic attributes (Table 1) were carefully selected based on general plant research and literature reviews. The ANN results based on the National Ecological Survey suggests that Slope and Altitude are the most critical variables for the habitats of *Paulownia coreana*. The remaining six environmental variables related to the habitats, in order of decreasing influence, are: Annual Rainfall, Drainage, Soil Accumulation, Soil Depth, Erosion Level, and Wind Exposure. Furthermore, the Decision Tree algorithm also showed similar results in identifying critical environmental attributes affecting tree habitats (*Paulownia coreana*, *Pinus densiflora*, *Quercus variabilis*, and *Robinia pseudoacacia*). The HSMs in this study using Decision Tree algorithms indicated that the most critical variables within the tree habitats were Slope and Altitude, followed by Weathering Level, Drainage, Annual Rainfall, Wind Exposure, Soil Depth, Soil Accumulation, and Erosion Levels, in order of influence. Therefore, this study was able to successfully identify the most critical environmental predictors to form habitat suitability ecologically.

The second algorithm type used in the study to create habitat suitability models were Decision Tree algorithms. Here, the model was created with a tree size (267), i.e., containing 267 decisions. The number in parentheses shows the number of samples satisfying and dissatisfying the classification. After the tree structure, the model generated a confusion matrix with cross tabulation (Table 3). This shows incorrect classification in the training datasets. The output error annotates the correct classification; however, it showed a 37.2% error rate, with 744 out of 2000 instances wrongly classified. Out of the 497 test species records, the model only correctly predicted *Paulownia coreana* (10), *Pinus densiflora* (88), *Quercus variabilis* (35), and *Robinia pseudoacacia* (25), indicating about 49.7% accuracy, i.e., 50.3% error rate. Compared to the training dataset, this is not a very good performance; however, it is not unexpected. Since environmental variables cannot be definite, such attributes for plants do not have distinctive value. In particular, the climatic and topographic attributes in the study are generally suggestive of overall ranges rather than dichotomic values. Hence, the model's performance could be quite low in evaluating the suitability of habitats. The study further included 10 boosted trials but the prediction performance was still not improved, showing only 49.7% accuracy. Therefore, in future research, diversification of plant species and environmental variables is required to increase accuracy.

Habitat suitability modeling is affected by a number of environmental variables such as plants' colonization and fragmentation. Therefore, there have been several attempts to model plant habitats. These predictions have gained attention because climate change could affect species distributions. Recently, there have been a great number of challenges to create models for habitat suitability; however, no model has been able to integrate the environmental factors that influence plant habitats so far. Most attempts are only based on remote sensing information. Therefore, the results of this study should be read in light of these unavoidable constraints. Accommodating for these weaknesses, the modeling methods established in our paper could open new possibilities for modeling with on-site survey data rather than remote sensing with greater validity, as well as the analysis of potential plant species at the various scales within various climatic and topographic areas. The modeling methods are essential for planners to make decisions, manage resources, and conserve forests. It is fundamental to study shift patterns of plants habitats within the era of climate crisis.

Moreover, the HSMs for plants would aid in removing any uncertainties regarding certain species managements. Future models will need to consider additional information such as ecological and physical data. Furthermore, different tendencies can be revealed depending on the scales; therefore, it is critical to combine research at various scales and

attributes, and in diverse areas. Even though the entire Korean Peninsula was covered in this study, the habitats assessed here may illustrate highly specific tendencies owing to their individual geographic characters.

6. Conclusions

The study aimed to explore habitat suitability models and environmental variables related to *Paulownia coreana* in conjunction with the National Ecological Survey. Together with carefully selected environmental attributes, the National Ecological Survey information was fed into machine learning algorithms such as the ANN and Decision Tree & Rules. While the ANN algorithm was applied to identify the impact of individual variables, the Decision Tree algorithm was used to create habitat suitability models for *Paulownia coreana* and other relevant tree species. The study utilized periodic on-site survey information which enhanced the credibility of the habitat suitability models overall. Moreover, localized environmental resources such as topographic and climatic attributes were taken into account to predict habitat suitability.

One limitation of the habitat suitability modeling is the fact that environmental variables for plants are not distinctive. The climatic and topographic attributes mentioned in the paper suggest overall ranges rather than dichotomic values.

Despite the fact that the habitat suitability modeling framework presented here provided sub-optimal results, the novelty of this work is that machine learning algorithms (particularly ANN and Decision Tree) were implemented for the identification of environmental variables and habitat suitability modeling using on-site survey information. Moreover, it would be an effective means for monitoring, planning, and managing not only individual species but entire forests at the regional and national levels. Furthermore, it can also shed light on existing challenges and future research needs.

Author Contributions: Conceptualization, Y.S., S.K., S.-R.P. and K.A.; resources, T.Y. and C.K.; writing—original draft preparation, Y.S. and K.A.; writing—review and editing, Y.S., S.K. and K.A.; supervision, K.A. and S.-W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This research was supported by National Institute of Ecology (NIE-A-2021-01) and Korea Forest Service (Korea Forestry Promotion Institute) (FTIS 2021331A00-2223-AA01). This research was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1059444).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Miller-Coleman, R.L.; Dodsworth, J.A.; Ross, C.A.; Shock, E.L.; Williams, A.J.; Hartnett, H.; McDonald, A.I.; Havig, J.R.; Hedlund, B.P. Korarchaeota Diversity, Biogeography, and Abundance in Yellowstone and Great Basin Hot Springs and Ecological Niche Modeling Based on Machine Learning. *PLoS ONE* **2012**, *7*, e35964. [CrossRef] [PubMed]
2. Mouton, A.; Alcaraz-Hernández, J.D.; De Baets, B.; Goethals, P.; Martínez-Capel, F. Data-driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. *Environ. Model. Softw.* **2011**, *26*, 615–622. [CrossRef]
3. Patel, B.; Sharaff, A. Feature Fusion based Growth Analysis of Chhattisgarh Rice Plants using Machine Learning Technique. In Proceedings of the 7th International Conference on Signal Processing and Integrated Networks, Delhi, India, 27–28 February 2020.
4. Paz-Kagan, T.; Silver, M.; Panov, N.; Karnieli, A. Multispectral Approach for Identifying Invasive Plant Species Based on Flowering Phenology Characteristics. *Remote Sens.* **2019**, *11*, 953. [CrossRef]
5. Pitkänen, T.P.; Skånes, H.; Käyhkö, N. Detecting subpixel deciduous components to complement traditional land cover classifications in Southwest Finland. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *42*, 97–105. [CrossRef]
6. Pouteau, R.; Meyer, J.-Y.; Taputuarai, R.; Stoll, B. Support vector machines to map rare and endangered native plants in Pacific islands forests. *Ecol. Inform.* **2012**, *9*, 37–46. [CrossRef]

7. Guerra-Coss, F.A.; Badano, E.I.; Cedillo-Rodríguez, I.E.; Ramírez-Albores, J.E.; Flores, J.; Barragán-Torres, F.; Flores-Cano, J.A. Modelling and validation of the spatial distribution of suitable habitats for the recruitment of invasive plants on climate change scenarios: An approach from the regeneration niche. *Sci. Total Environ.* **2021**, *777*, 146007. [CrossRef]
8. Ksiksi, T.S.; Remya, K.; Mousa, M.T.; Al-Badi, S.K.; Al Kaabi, S.K.; Alameemi, S.M.; Fereaa, S.M.; Hassan, F.E. Climate change-induced species distribution modeling in hyper-arid ecosystems. *F1000Research* **2019**, *8*, 978. [CrossRef]
9. Mohammady, M.; Pourghasemi, H.R.; Yousefi, S.; Dastres, E.; Edalat, M.; Pouyan, S.; Eskandari, S. Modeling and Prediction of Habitat Suitability for Ferula gummosa Medicinal Plant in a Mountainous Area. *Nonrenew. Resour. Res.* **2021**, *30*, 4861–4884. [CrossRef]
10. Kim, C.-H.; Kang, J.-H.; Kim, M. Status and Development of National Ecosystem Survey in Korea. *J. Environ. Impact Assess.* **2013**, *22*, 725–738. [CrossRef]
11. Flach, P.A. On the state of the art in machine learning: A personal review. *Artif. Intell.* **2001**, *131*, 199–222. [CrossRef]
12. Raghukumar, A.M.; Narayanan, G. Comparison of Machine Learning Algorithms for Detection of Medicinal Plants. In Proceedings of the 4th International Conference on Computing Methodologies and Communication, Erode, India, 11–13 March 2020.
13. Crisci, C.; Ghattas, B.; Perera, G. A review of supervised machine learning algorithms and their applications to ecological data. *Ecol. Model.* **2012**, *240*, 113–122. [CrossRef]
14. Bradley, B.A.; Olsson, A.D.; Wang, O.; Dickson, B.G.; Pelech, L.; Sesnie, S.E.; Zachmann, L.J. Species detection vs. habitat suitability: Are we biasing habitat suitability models with remotely sensed data? *Ecol. Model.* **2012**, *244*, 57–64. [CrossRef]
15. Zhang, J.; Okin, G.S.; Zhou, B. Assimilating optical satellite remote sensing images and field data to predict surface indicators in the Western U.S.: Assessing error in satellite predictions based on large geographical datasets with the use of machine learning. *Remote Sens. Environ.* **2019**, *233*, 111382. [CrossRef]
16. Bradter, U.; Thom, T.J.; Altringham, J.D.; Kunin, W.E.; Benton, T.G. Prediction of National Vegetation Classification communities in the British uplands using environmental data at multiple spatial scales, aerial images and the classifier random forest. *J. Appl. Ecol.* **2011**, *48*, 1057–1065. [CrossRef]
17. Casalegno, S.; Amatulli, G.; Bastrup-Birk, A.; Durrant, T.H.; Pekkarinen, A. Modelling and mapping the suitability of European forest formations at 1-km resolution. *Eur. J. For. Res.* **2011**, *130*, 971–981. [CrossRef]
18. Zlinszky, A.; Schroiff, A.; Kania, A.; Deák, B.; Mücke, W.; Vari, A.; Székely, B.; Pfeifer, N. Categorizing Grassland Vegetation with Full-Waveform Airborne Laser Scanning: A Feasibility Study for Detecting Natura 2000 Habitat Types. *Remote Sens.* **2014**, *6*, 8056–8087. [CrossRef]
19. Díaz-Varela, R.A.; Iglesias, S.C.; Castro, C.C.; Varela, E.D. Sub-metric analisis of vegetation structure in bog-heathland mosaics using very high resolution rpas imagery. *Ecol. Indic.* **2018**, *89*, 861–873. [CrossRef]
20. Dronova, I.; Gong, P.; Clinton, N.E.; Wang, L.; Fu, W.; Qi, S.; Liu, Y. Landscape analysis of wetland plant functional types: The effects of image segmentation scale, vegetation classes and classification methods. *Remote Sens. Environ.* **2012**, *127*, 357–369. [CrossRef]
21. Du, P.; Samat, A.; Waske, B.; Liu, S.; Li, Z. Random Forest and Rotation Forest for fully polarized SAR image classification using polarimetric and spatial features. *ISPRS J. Photogramm. Remote Sens.* **2015**, *105*, 38–53. [CrossRef]
22. Peters, J.; De Baets, B.; Verhoest, N.E.C.; Samson, R.; Degroeve, S.; De Becker, P.; Huybrechts, W. Random forests as a tool for ecohydrological distribution modelling. *Ecol. Model.* **2007**, *207*, 304–318. [CrossRef]
23. Fukuda, S.; Tanakura, T.; Hiramatsu, K.; Harada, M. Assessment of spatial habitat heterogeneity by coupling data-driven habitat suitability models with a 2D hydrodynamic model in small-scale streams. *Ecol. Inform.* **2015**, *29*, 147–155. [CrossRef]
24. Garzón, M.B.; Blazek, R.; Neteler, M.; de Dios, R.S.; Ollero, H.S.; Furlanello, C. Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula. *Ecol. Model.* **2006**, *197*, 383–393. [CrossRef]
25. Jacob, M.L.P.; Nunes, C.S.M.; Borba, P.C.D.O.; Ribeiro, G.; De Andrade, T.U.; Endringer, D.C.; Lenz, D. Hematological value references for free-living saffron finch (*Sicalis flaveola*) using a machine-learning-based classifier. *Comp. Clin. Pathol.* **2018**, *28*, 937–941. [CrossRef]
26. Juel, A.; Groom, G.B.; Svenning, J.-C.; Ejrnæs, R. Spatial application of random forest models for fine-scale coastal vegetation classification using object based analysis of aerial orthophoto and DEM data. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *42*, 106–114. [CrossRef]
27. Sabat-Tomala, A.; Raczko, E.; Zagajewski, B. Comparison of Support Vector Machine and Random Forest Algorithms for Invasive and Expansive Species Classification Using Airborne Hyperspectral Data. *Remote Sens.* **2020**, *12*, 516. [CrossRef]
28. Shobana, K.B.; Perumal, P. Plants Classification Using Machine Learning Algorithm. In Proceedings of the 6th International Conference on Advanced Computing and Communication Systems, Coimbatore, India, 6–7 March 2020.
29. Sukumaran, J.; Economo, E.P.; Lacey Knowles, L. Machine Learning Biogeographic Processes from Biotic Patterns: A New Trait-Dependent Dispersal and Diversification Model with Model Choice by Simulation-Trained Discriminant Analysis. *Syst. Biol.* **2016**, *65*, 525–545. [CrossRef]
30. Ullah, M.R.; Dola, N.A.; Sattar, A.; Hasnat, A. Plant Diseases Recognition Using Machine Learning. In Proceedings of the 8th International Conference on System Modelling & Advancement in Research Trends, Moradabad, India, 22–23 November 2019.
31. Wang, D.; Wan, B.; Qiu, P.; Su, Y.; Guo, Q.; Wu, X. Artificial Mangrove Species Mapping Using Pléiades-1: An Evaluation of Pixel-Based and Object-Based Classifications with Selected Machine Learning Algorithms. *Remote Sens.* **2018**, *10*, 294. [CrossRef]

32. Zohmann, M.; Pennerstorfer, J.; Nopp-Mayr, U. Modelling habitat suitability for alpine rock ptarmigan (*Lagopus muta helvetica*) combining object-based classification of IKONOS imagery and Habitat Suitability Index modelling. *Ecol. Model.* **2013**, *254*, 22–32. [CrossRef]
33. Veettil, B.K.; Ward, R.D.; Lima, M.D.A.C.; Stankovic, M.; Hoai, P.N.; Quang, N.X. Opportunities for seagrass research derived from remote sensing: A review of current methods. *Ecol. Indic.* **2020**, *117*, 106560. [CrossRef]
34. Lantz, B. *Machine Learning with R: Expert Techniques for Predictive Modeling*; Packt Publishing Ltd.: Birmingham, UK, 2019.