Clustering and Artificial Intelligence-based Prediction of Ecologically Sustainable Species Introductions

Shuqiao Liu, Zhao Zhang, Hongyan Zhou, and Xue-Bo Chen

***Abstract*—There is a growing interest in sustainable ecosystem development, which includes methods such as scientific modeling, environmental assessment, and development forecasting and planning. However, due to insufficient survey data in many current development areas, development progress is delayed and stagnant. To address this situation, this paper proposes a SWOT-TOPSIS-K-Means (STK) data analysis and evaluation model to analyze ecological factors, which can realize a comprehensive and complete data analysis with fewer samples. Decision tree (DT), random forest (RF), and multilayer perceptron (MLP) neural network models were constructed from the results of this analysis, and statistical tests such as r-squared, mean absolute error, and cross-validation are used to further confirm the performance efficiency of the computational prediction models to provide real-time prediction research solutions. For this purpose, data from research scholars on species introduction in ecosystem development were selected for testing. The results show that the proposed assessment model and modeling results satisfy all accuracy-related acceptance requirements. Among them, MLP is better than DT and RF. In summary, the STK assessment model and the MLP prediction model can provide a basis for the selection and development of ecological factors.**

***Index Terms*—Data analysis, TOPSIS, Machine learning, K-Means clustering**

I. Introduction

Today, the world faces enormous environmental challenges, such as climate change, biodiversity loss, land degradation, and water scarcity. These challenges have significant implications for human well-being and sustainable development, prompting the search for sustainable ecosystem development solutions [1]. Sustainable eco-environmental development is an approach to ensuring ecological health and sustainability [2]. Among them, data analysis plays an important role in ecological environment research. It can help us understand and reveal the trends, patterns, and laws of changes in the ecological environment. At the same time, data analysis provides a basis for building environmental models and making predictions [3].

By analyzing environmental data, key variables and influences can be identified, and mathematical or statistical models can be built to describe and simulate the dynamic processes of ecosystems. These models can be used to predict future environmental change, assess ecological impacts under different scenarios, and provide predictions for decision-making [4].

SWOT is also often used in conjunction with other data analysis methods [5], Eslamipoor R and Sepehriar A, have used a combination of SWOT matrix and hierarchical analysis method (AHP) [6], for better results in processing the data. At the same time, Bas E proposed a SWOT-Fuzzy TOPSIS combined with AHP synthesis methodology to rank defined SWOT factors and assign strategies by prioritizing them [7]. In terms of the ecological environment. In 2019, Solangi Y A and Tan Q M et al. used an integrated approach of SWOT-AHP and Fuzzy Techniques for Ideal Solution Similarity Ranking Performance(F-TOPSIS) to evaluate energy strategies for sustainable energy planning [8].

In recent years, to further analyze and process the data, Duarte-Duarte J B and Hosseini S M have combined the SWOT matrix with TOPSIS alone, and then used the analyzed and processed data as a basis for the selection of optimization algorithms for processing [9], this approach provides new directions in data processing optimization. This paper is inspired to combine the K-Means clustering algorithm into the SWOT model to further process the data for classification, the advantage of this method is that K-means clustering helps to group and aggregate the data, and through the clustering results, different subsets of data are identified, and then different feature extraction and processing are performed for each subset to better capture the features and patterns of the data, thus improving the performance of the sorting model [10-12]. When the data analysis model is established, to further validate the accuracy of the model and better apply the model to the actual ecological environment analysis and prediction, Big Data Machine Learning (ML) as an important data prediction and analysis technology system is used in practice from time to time [13]. In recent years in the direction of ecological environment prediction research, ML-related researchers integrated the criterion decision-making method TOPSIS with machine learning tools in terms of performance [14], aspects of variable prediction [15], and systematic evaluation

Manuscript received September 8, 2024; revised February 17, 2025. The research work was supported by the Fundamental Research Funds for the Liaoning Universities (LJ212410146025).

Shuqiao Liu is a graduate student at the School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: liushuqiao@qq.com).

Zhao Zhang is an associate professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author, e-mail: zhangzhao333@hotmail.com).

Hongyan Zhou is a doctoral student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: zhou321yan@163.com).

Xue-Bo Chen is a professor of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: xuebochen@126.com)

[16] were obtained. For this purpose, decision tree (DT), random forest (RF), and MLP in ML are used to build the prediction model in this paper. These techniques are well-known and often used in the field of modeling predictions for data models [15, 17, 18].

Considering the incompleteness of databases often encountered in ecological development [19], i.e., it is difficult to predict modeling with a small number of samples. Modeling results can be tested through post-modeling cross-validation (e.g., Cakici N & Fieberg C et al. applied cross-validation to the analysis of data with only 46 sets of samples [20]. The results obtained from the analysis also have higher accuracy. This paper presents a case study of the development of an artificial intelligence-based data analysis model (SWOT-TOPSIS-KMeans; STK) as well as the simulation of a predictive model in the context of a small dataset of sustainable eco-environmental development research and intelligent forecasting. In the authoritative 2020 Global Assessment of Biodiversity and Ecosystem Services (GABES) [21], it is highlighted that ecosystems are still being degraded due to loss of biodiversity and that many of nature's contributions to humanity are being jeopardized [22]. Therefore, in this paper, we contextualize the study under the conditions of species introductions in ecosystem development and select data from a sample of potential receiving sites for the Pacific-mouthed kangaroo rat in the context of species introductions [23] and the data reconstructed from this modeling[24].

Neural networks, linear regression, and cross-validation were used as validated prediction techniques. The proposed data analysis model STK combined with the predictive model was validated by several statistical tests. In addition, the accuracy of the trained Shingo network was tested by simulation with randomly generated datasets. The following are the main objectives of this study.

A. Based on the basic indicators and rational planning of the data, the STK data analysis model was developed to pre-process and rationally categorize the data, and the text was selected from the sample data of species introduced in the thesis.

B. Intelligent forecasting models for DT, RT, and MLP developed and compared through STK analysis results.

C. Parameterization and simulation forecasting studies to analyze the impact on siting options by adjusting the data.

D. Participate in the modeling results of this experiment by reconstructing the data. Provide data support for the results.

II. Research Methodology

The process of developing the STK model and predictive model for data analysis based on artificial intelligence is shown in Fig. 1. The first part is the data analysis model (STK). Firstly, the data after SWOT analysis is classified and organized, then the results are sorted using the TOPSIS model, in which the entropy weighting method is selected, and then the sorted results are randomly classified by K-Means clustering algorithm, and the classification results are set as binary output.

The second part is to organize the dataset output from the STK model and then preprocess the variables and remove the randomness by selecting the influencing factors. The dataset was then further divided into two halves, one for training (70%) and the other for testing (30%). Next, algorithm development techniques based on artificial intelligence were selected. Finally, the model was validated using various methods to support the reliability and dependability of the model.

III. Combinatorial Model

A. Data Set

Samples of potential acceptance sites for Pacific kangaroo rats about species introductions were obtained from data examined by Rachel Y. Chock's team from actual surveys [23]. Forty-nine categories of indicators from seven addresses were selected for pre-processing in this study and categorized into positive indicators (strengths, opportunities) and negative indicators (threats, weaknesses) through SWOT analysis.

Fig. 2 shows an overview of the numerical analysis of the four categories of indicators (taking the LAX region as an example). To validate the model, two sets of 49-category indicator datasets were also randomly generated for model testing. In this case, the reconstructed dataset in the ablation experiments was derived from the authors' modeled reconstructed data[24].

B. STK Data Analysis Model

This study establishes the STK data analysis model based on the dataset, and the model flow is shown in Fig. 3. Firstly, the SWOT analysis data are sorted by TOPSIS analysis, in which the entropy weight method is used to determine the weights. Then the results obtained from TOPSIS analysis were divided into three categories through the K-Means clustering algorithm, set as binary output, and finally, the input and output data were summarized.

C. TOPSIS Analytical Model

The TOPSIS distance method model of optimal and inferior solutions is a commonly used comprehensive evaluation method, which can make full use of the information of the original data, and its results can accurately reflect the gap between the evaluation solutions [25]. The ranking process of TOPSIS is based on the normalized raw data matrix, using the cosine method to find out the optimal and the worst solutions among the limited methods, and then calculate the distance between each evaluation object and the optimal and the worst solutions respectively [26].

To synthesize the statistics of positive and negative factors, this model is adopted to analyze the indicators by assigning weights.To make the indicators homothetic. The negative indicators (disadvantages, threats) in the evaluation indicators are transformed into positive indicators (advantages, opportunities) by using the difference method (1-X). The transformed data matrix is still noted as X the formula is in equation (1).

Next, the raw data are normalized in equation (2) where n=7.The final data matrix was analyzed to obtain (n=7, p=49).

$$Z_{ij} = -\frac{X_{ij}}{\sqrt{\sum_{k=1}^{n}\left(X_{ij}\right)^2}} \tag{1}$$

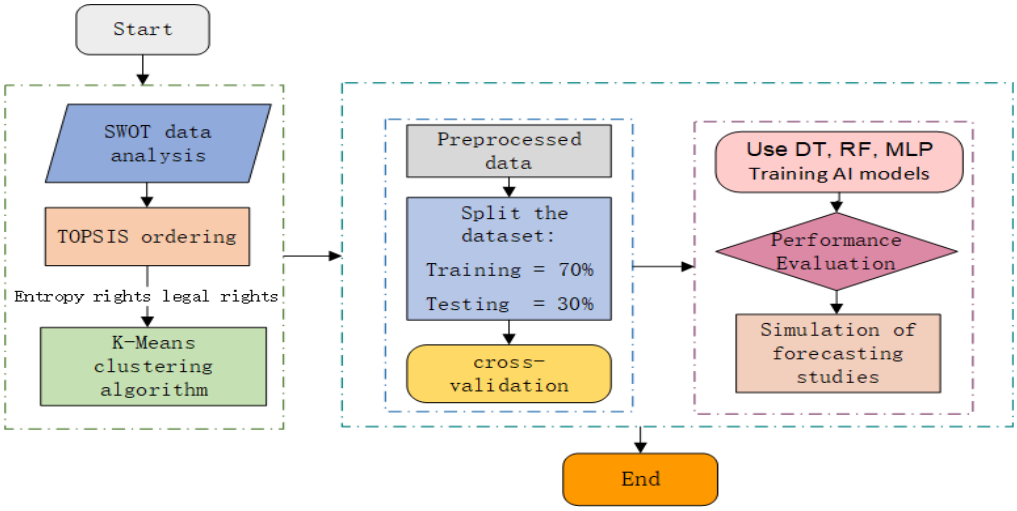


Fig. 1. Flowchart for developing data analysis and predictive models

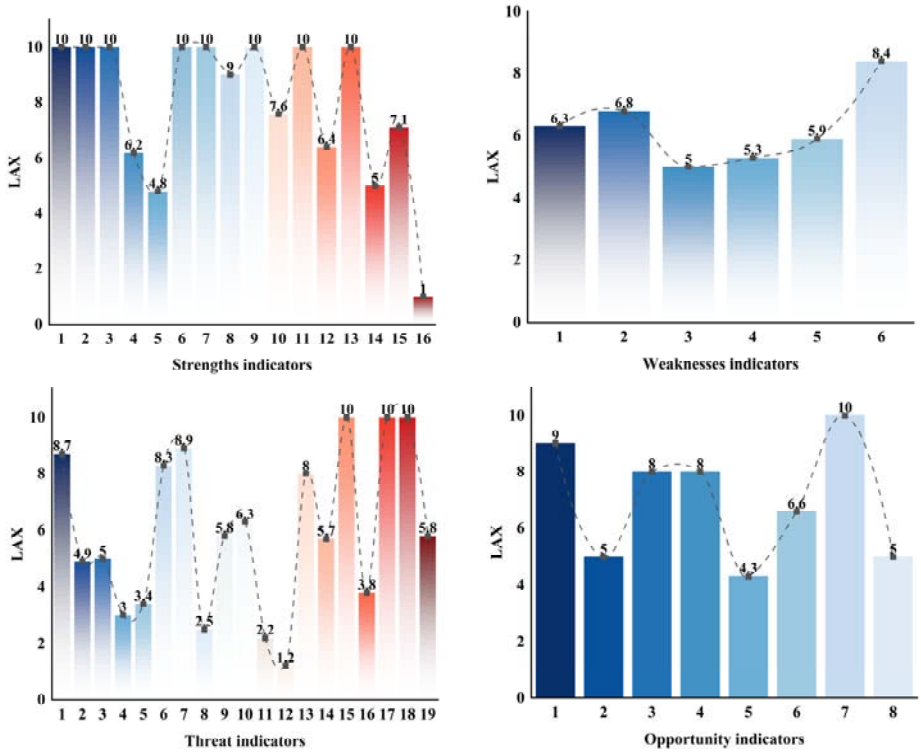


Fig. 2. Indicators for the four categories of SWOT analysis (LAX)

$$Z=\begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1p} \\ Z_{21} & Z_{22} & \cdots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{np} \end{pmatrix} \tag{2}$$

The distance equation (5) and (6) between the evaluated pairs and the optimal and worst values is calculated from the vectors of optimal and worst values given in equation (3) and equation (4), and the distance is compared to determine the ranking of the samples.

$$Z^{+}=\max_{nj}\left(Z_1^{+},Z_2^{+},\cdots,Z_p^{+}\right) \tag{3}$$

$$Z^{-}=\max_{nj}\left(Z_1^{-},Z_2^{-},\cdots,Z_q^{-}\right) \tag{4}$$

$$D_i^{+}=\sqrt{\sum_{j=1}^{m}\left(Z_{ij}-Z_j^{+}\right)^2} \tag{5}$$

$$D_i^{-}=\sqrt{\sum_{j=1}^{m}\left(Z_{ij}-Z_j^{-}\right)^2} \tag{6}$$

The entropy value method assigns weights according to the degree of difference in the sign value of each indicator, to derive the corresponding weight of each indicator equation (7), and the indicator with a large degree of relative change has a larger weight.

$$d_j=1-e_j \tag{7}$$

Subsequently, the composite evaluation value equation (9) for each evaluation sample No. 1 is determined by defining the indicator weights equation (8) corresponding to each sample i.

$$W_j = \frac{d_j}{\sum_{j=1}^{n} d_j} \tag{8}$$

$$F_i = \sum_{j=1}^{n} W_j P_{ij} \tag{9}$$

Continuous iteration of the centroids of each cluster. In this study, the purpose of adding the optimization algorithm after TOPSIS classification is to better capture the features and patterns of the data, thus improving the performance of the ranking model.

Data preprocessing was first performed on the SWOT-TOPSIS collapsed sorted data above. K centers were randomly selected and denoted as $\mu_1^{(0)}, \mu_2^{(0)}, \cdots, \mu_k^{(0)}$

Define the loss function, where for each class of samples, the μ is the center equation (10) between different samples, t=0, 1, 2, ...is the number of iteration steps, and the following process is repeated until J converges. For each of the samples x_i, Assign it to the nearest center equation (11), and finally the center k of each class, and repeat the iteration until the optimal solution equation (12).

$$J(c,\mu) = \min \sum_{i=1}^{M} \left\| x_j - \mu_{c_1} \right\|^2 \tag{10}$$

$$c_i^t < -\arg\min_k \left\| x_i - \mu_k^t \right\|^2 \tag{11}$$

$$\mu_k^{(t+1)} < -\arg\min_\mu \sum_{c_i^t}^{b} \left\| x_i - \mu \right\|^2 \tag{12}$$

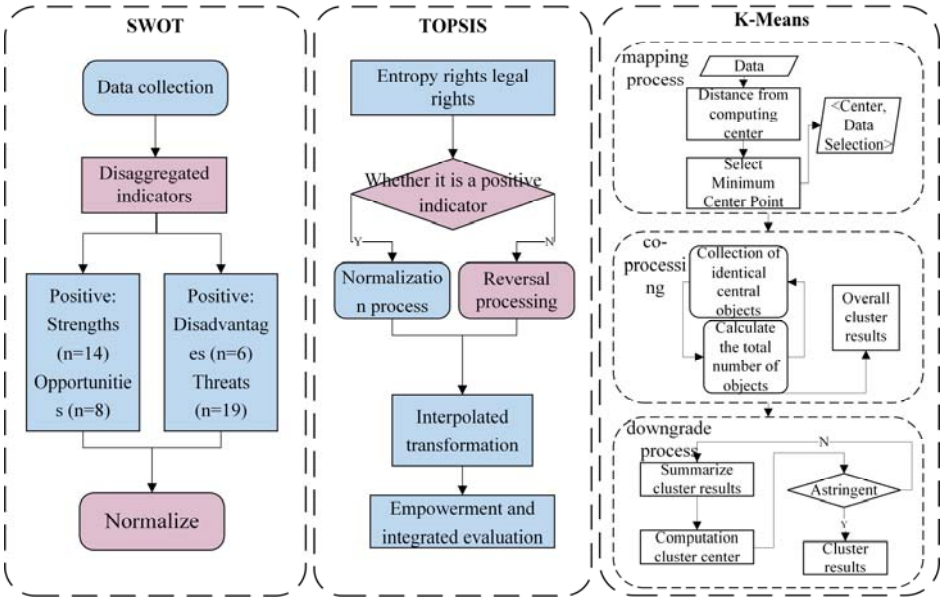


Fig. 3. STK Model Flow Diagram

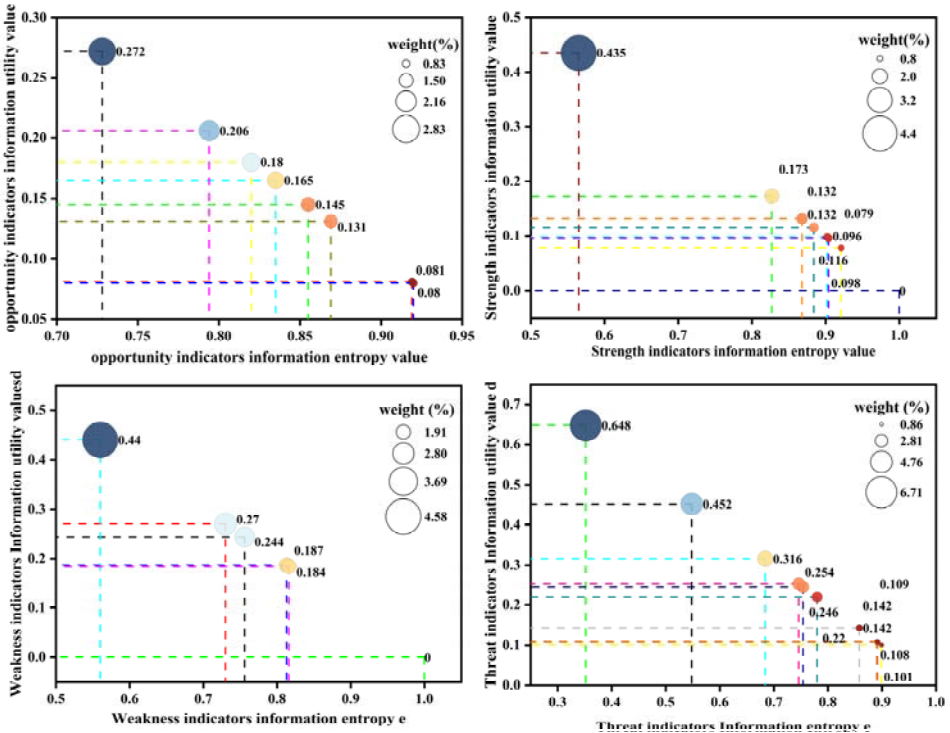


Fig. 4. Comparison of Information Entropy Value and Information Utility Value in Entropy Weight Indicator

IV. Experiment and Analysis

A. STK Model Analysis Results

Fig. 4 shows the relationship between the information entropy value and the information utility value in the entropy weight indicators of TOPSIS weights variation. It can be observed that the two programs have a linear relationship, indicating that even though the sample size of the four types of indicators is different, the weights defined by the entropy weighting method are in line with the linear relationship, which can be used as the next step in the classification and ranking. Then the results of similarity and proximity sorting are shown in Table 1.

The results of the final K-Means classification iterations are shown in Table 2. Where a total of six iterations were performed and the data was finally categorized into three classes. Those ranked 1 and 2 are in one category, belonging to cluster 1 and set to 000; those ranked 3 and 4 are in one category, belonging to cluster 2 and set to 010; and those ranked 5, 6, and 7 are in one category, belonging to cluster 3 and set to 100. the classification results are shown in Table 3.

B. Data Pre-processing

The SKT output sorted classification results were used as labels (Table 3-Output), and the four categories of metrics Strengths (n=16), Weaknesses (n=6), Threats (n=19), and Opportunities (n=8) were used as inputs into each of the three types of AI-based predictive model modeling. An indicator to assess the performance of the regression model was used to evaluate the interdependence between the variables considered for modeling purposes [27, 28].

R values ranging from -1 to +1 were used to assess the strength of the correlation. Positive and negative indications show increasing and decreasing trends, respectively. Zero in the Pearson correlation matrix indicates that there is no association between the two variables. Similarly, a value close to 1 indicates a high degree of correlation [29].

TABLE I
Calculation results of the Topsis Evaluation Method

Norm	Positive ideal solution distance (D+)）	Negative ideal solution distance (D-)	Composite score index	Arrange in order
LAX	0.5758	0.7385	0.5619	1
Tijuana Estuary	0.6083	0.7002	0.5351	2
Alta Vicente	0.6005	0.6893	0.5344	3
Torrey Pines	0.6653	0.6392	0.4900	4
Dilley	0.7283	0.5314	0.4219	5
Laguna Coast	0.7794	0.5346	0.4069	6
Turtle Ridge	0.7749	0.4889	0.3868	7

C. Developmental of Predictive Models

In this study, DT, RF, and MLP were applied to the study of the predictive ability of potential receiving sites of Pacific pocket gophers after STK modeling treatment, respectively. The models developed were validated and verified through parametric studies, statistical indicators, and simulated predictions.

TABLE II
K-means Classification Process

Number of iterations	fasciculus (botany)	type 2 center (botany)	type 3 center (botany)
1	0.6084; 0.7002; 0.5351.	0.6653; 0.6392; 0.4900.	0.7749; 0.4889; 0.3868.
2	0.5949; 0.7096; 0.5432.	0.6653; 0.6392; 0.4899.	0.7615; 0.5513; 0.4048.
3	0.5949; 0.7096; 0.5432.	0.6653; 0.6392; 0.4900.	0.7615; 0.5513; 0.4048.
4	0.5949; 0.7096; 0.5431.	0.6653; 0.6392; 0.4900.	0.7615; 0.5513; 0.4048.
5	0.5949; 0.7096; 0.5431.	0.6653; 0.6392; 0.4900.	0.7615; 0.5513; 0.4049.
6	0.5949; 0.7096; 0.5431.	0.6653; 0.6392; 0.4900.	0.7615; 0.5513; 0.4048.

TABLE III
Post-Classification Results

Development Addresses	Rankings	Clusters	Experts
LAX	1	1	000
Tijuana Estuary	2	1	000
Alta Vicente	3	2	010
Torrey Pines	4	2	010
Dilley	5	3	100
Laguna Coast	6	3	100
Turtle Ridge	7	3	100

Decision Tree (DT)

A regression decision tree is a commonly used machine learning algorithm for solving regression problems. It is based on a variant of the decision tree algorithm, which constructs a tree structure by partitioning the input features to make predictions about continuous-type target variables.

The decision tree construction process begins at the root node, where an optimal feature and corresponding threshold are selected for partitioning, dividing the dataset into two subsets. The process is then repeated recursively on each subset until a predefined stopping condition is met, such as reaching a predefined tree depth or insufficient number of samples in the node. At the leaf node, the predicted value is estimated by counting the mean or other statistics of the target variable in that node [30].

Algorithmically, Regression DT mainly refers to the CART algorithm, the internal node features take the values of "yes" and "no", and are a binary tree structure. Let X and Y be the input and output variables, respectively, given a training set of D, where x_j is the input instance (feature vector), where n=49 for this study is the number of features, and N is the number of samples [31]. A heuristic method is used for the division of feature n. Each division examines all the values of all the features in the current set one by one and selects the optimal one of them as the cut-off point according to the squared error minimization criterion [32]. For $x^{(j)}$ in the feature variable and its value s in the training set, as the

cut-off variable and cut-off point, and define two regions equation (13) and (14), find the optimal j and s and then solve equation (15). Where c_1 , c_2 is the output value fixed in the two regions after division. Finally, the corresponding output value equation (16) is determined by the selected pair of (j, s) divided regions, and the DT is generated by loop iteration.

$$R_1(j,s)=\left\{x \mid x^{(j)} \le s\right\} \tag{13}$$

$$R_2(j,s)=\left\{x \mid x^{(j)} > s\right\} \tag{14}$$

$$\min_{j,s}\left[\min_{c_i}\sum R_1\left(x_i-e_i\right)^2+\min_{c_2}\sum R_2\left(y_i-c_2\right)^2\right] \tag{15}$$

$$c_m^{\wedge}=\frac{1}{N_m}\sum xj \in R_{m(j,s)} y_i, x \in R_m, m=(1,2) \tag{16}$$

Due to its applicability to small datasets and tuning parameter enrichment adaptability, this study adopts the regression decision tree technique, which is a machine learning technique that combines the ideas of decision tree and regression analysis, and the regression decision tree framework diagram is shown in Fig. 5. Therefore, the experimental results compared can be extended to other small datasets or datasets with missing parameters, the following experimental environments are Jupyter lab.

Random Forest (RF)

It belongs to the integrated algorithm, which synthesizes multiple decision trees, and the process is based on Fig. 4 to continuously iterate new trees until the optimal solution. Iterating new trees until the optimal solution. The decision tree construction process starts at the root node and divides the dataset into two or more subsets by selecting an optimal feature and a corresponding threshold for division. The process is then repeated recursively on each subset until predefined stopping conditions are met, such as reaching a predefined tree depth, insufficient number of samples in a node, or the impurity of a node is below a certain threshold. Among them, on top of the DT-based learner (evaluator), random feature selection is further introduced in the training process of the DT, which can well avoid the problem of overfitting or underfitting a single DT [33]. This is the reason for modeling RF in this study, to further validate the accuracy of machine learning's modeling accuracy in the case of fewer samples. For random sampling, this study uses the self-sampling method in the bagging integration algorithm. For each node of the base DT, a subset containing k features is randomly selected from the feature set (n=49) of that node.

Then an optimal feature is selected from this subset for segmentation. Weight average (WA) is used to determine the weights equation (17), where is the weight of the individual learner.

$$H(x)=\frac{1}{T}\sum_{i=1}^{T} w_i h_i(x) \tag{17}$$

Multilayer perception (MLP)

Multilayer Perceptron is a basic artificial neural network model with a multilayer structure consisting of multiple neurons [34]. It consists of multiple neuron layers, each of which is fully connected to the neurons of the previous layer. The MLP usually consists of an input layer, a hidden layer, and an output layer, where there can be more than one hidden layer. Each neuron of MLP has an activation function for introducing nonlinear properties. Using forward propagation, each neuron weights and sums the input signals and generates the output signal after the activation function. Then, based on the error between the output of the network and the actual target, a backpropagation algorithm is used to update the network parameters to minimize the error [35].

It is a feed-forward neural network that is commonly used to solve classification and regression modeling problems, and the modeling process is shown in Fig. 6. The basic structure of an MLP consists of an input layer, an output layer, and at least one or more hidden layers. In this study, an MLP neural network model containing four fully connected layers and trained using the mean square error as a loss function is used for the regression task.

First, the input layer is determined by m samples n features and the output labels are three categories 000, 001, 100. where the input and output of the hidden layer are equations (18) and (19).

$$H = XW_h + b_h \tag{18}$$

$$O = XW_h + b_h \tag{19}$$

To improve the model performance as well as the nonlinear fitting ability of the neural network, the ReLu activation function, i.e., as a nonlinear transformation, is used in modeling and the outputs obtained in modeling are in equation (20). The Adam optimizer is used in compiling the model, which gives a definite range of learning rates at each iteration after bias correction, making the parameters smoother.

$$output = \max\left(W^T X + B\right) \tag{20}$$

V. Evaluation Results

A. Experimental Indicators

Modeling DT, RF, and constructing a neural network MLP with the sklearn database in Python (2022), respectively, the model outputs are summarized in Fig. 7, and the following experimental environments are all Jupyter lab.

The selection of a predictive model depends on a variety of factors. Error plots and statistical measures are often used for model selection. In this work, the model selection process considered statistical tests such as MAE and R^2 provided by equations (21) and (22). Predictive models with R^2 values close to 1 and low RMSE and MAE values are considered the most accurate [29].

The results in Fig. 7 show that both DT and MLP models achieve good prediction results, and RF is not so well fitted. However, from the above, it can be seen that RF is a classifier based on the extension of DT by combining multiple DTs into a whole through specific combinations [36]. Since Ho first introduced the concept of Randomized Decision Forests [37], and later [33] provided a comparison of RF and DT on classification and regression tasks, the results showed that a large number of trees are required to obtain stable estimates of variable importance and proximity if prediction is required for more than 2-classification problems.

For the triple classification problem in this study, the RF model results are supposed to be better than the DT, but precisely because of the dataset explored in this study in the case of fewer samples or the case of overfitting and so on as mentioned above, to verify whether the model results are not inaccurate due to the small number of samples, we cross-validated the DT and the RF models separately, with the cross-validation formula equation (23), and the results are shown in Fig. 8.

The results show that after cross-validation, the actual curves of the RF model overlap with the predicted curves to some extent, but the predictions of the DT model are opposite. This result proves that our hypothesis is correct, and the DT model accuracy is likely to be misjudged with a small number of samples.

Next, the respective model accuracies were examined by MAE and the test and training set results are shown in Fig. 9. The MLP has the highest model accuracy, with the sample test set distribution and its lognormal curve on the left side of the image, and the sample training set and its lognormal curve on the top side, with the center of the fit marked by the red line. 0.9989, MAE/10=0.2), followed by RF (training set: R^2=0.984, MAE=0.016; test set: R^2=0.906, MAE/10=0.212), and lastly, DT which was verified to have an inaccurate model accuracy (training set: R^2=0.979, MAE=0.2; test set: R^2=1.0, MAE/10 = 0.0).MLP achieved very good results on both R^2 and MAE results, close to an accuracy of 1.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{21}$$

$$R^2 = 1 - \frac{\sum i\left(\hat{y}_i - y_i\right)^2}{\sum i\left(\bar{y}_i - y_i\right)^2} \tag{22}$$

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \tag{23}$$

B. MLP Model

The MLP model chosen for this study contains three hidden layers and one output layer, with the ReLU activation function chosen for each layer, and the last layer is a single node fully connected layer for the regression task. After training the model, two new sets of data samples data 1 and 2 (1×49) are then generated for model prediction and the randomly generated data are categorized according to the proportion of the model training at the beginning: advantage (A1-A16); disadvantage (B1-B6); threat (C1-C19); and opportunity (D1-D8). Now to verify the accuracy of the model prediction, then change the individual data in these two sets of data samples (adjust data one and two), and analyze whether the model can correctly predict, the prediction results are shown in Table 4.

As can be seen in Table 4, when the values of the two advantages of data (1) A1 and A2 are reduced, the predicted output is increased by 0.3096. When the values of the two disadvantages of data (2) B4, and B5 are reduced, the predicted output is decreased by 1.003. The larger the data, the lower the overall model score is, so the change in the output of the pre-regulation results can prove the accuracy of the predictions of the MLP model of the present study.

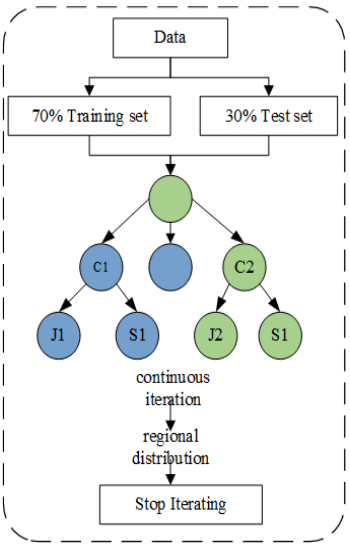


Fig. 5 Regression decision tree framework diagram

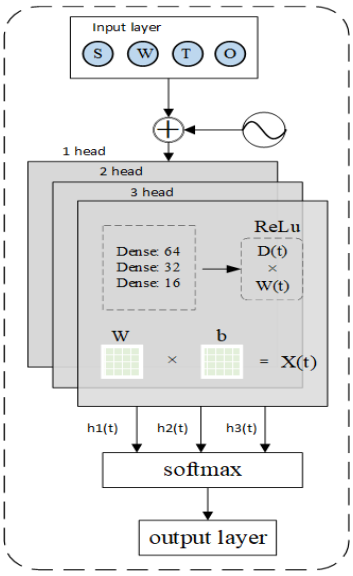


Fig. 6 Structure of MLP prediction

C. Simulation Forecast

To further verify the superiority of this study's model, the models and algorithms in it are combined separately to construct DT, RF, and MLP neural network models in turn, and the results are shown in Table 5. By comparing the values of and MAE, it can be seen that when the SWOT model acts alone as well as when this model is not involved, neither can be predicted by modeling for simulation and when two of these models are combined alone, the model accuracy is inferior to that of this study's model, where SWOT is combined with TOPSIS alone. The dataset without the K-Means clustering algorithm to classify the samples showed a negative performance in the mean squared error of the three types of models; while the SWOT data without the TOPSIS normalized sorting could not be classified correctly, and the modeling accuracy was not sufficient and accurate.

D. Supplementary Experiment

To further verify the accuracy of the experimental prediction results, we did data reconstruction and published it in the journal shown in III-A. Combining the experimentally reconstructed data with the original data, the experimental results are shown in Table 6.

With the increased amount of data, the results are still predicted with some accuracy by the STK model as well as the MLP modeling.

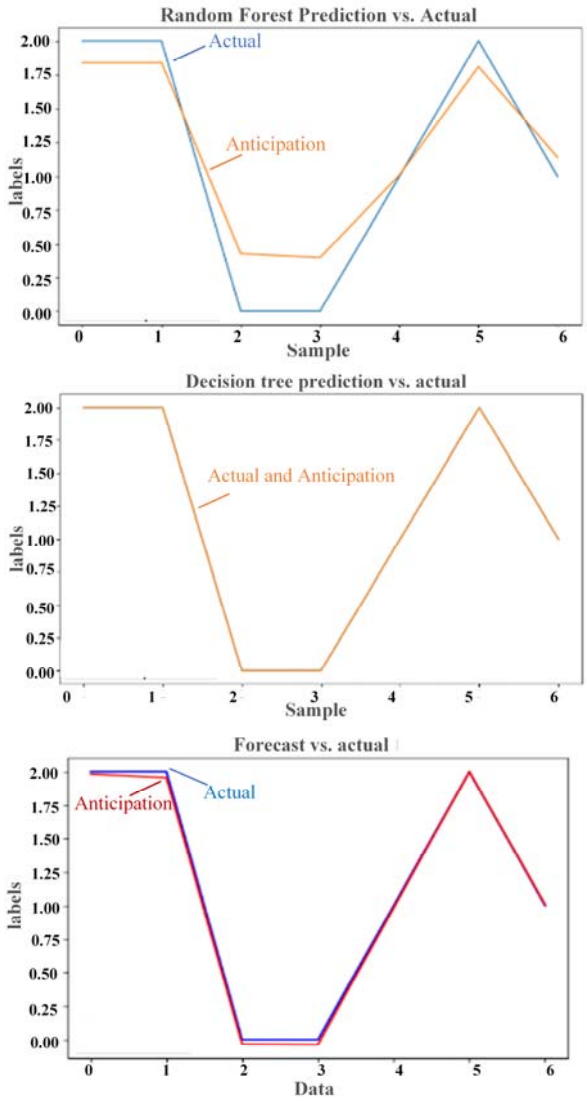
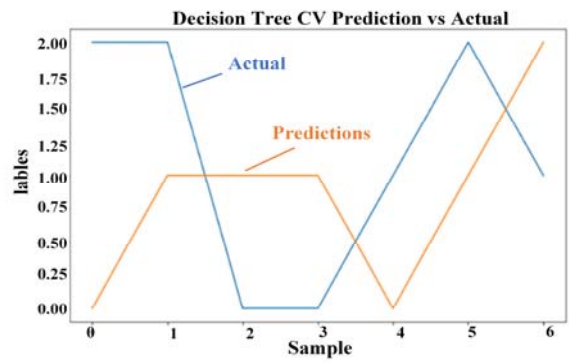


Fig. 7 DT, RF, and MLP model prediction results



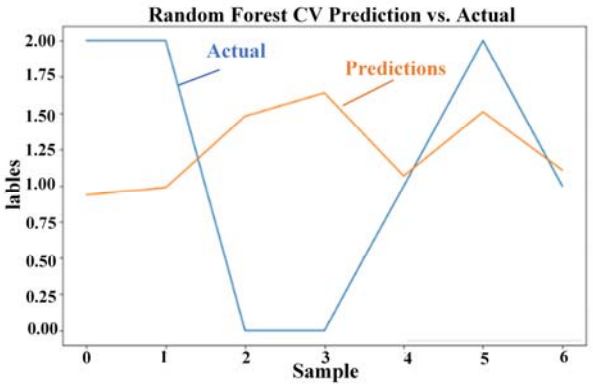


Fig. 8 RF and DT cross-validation

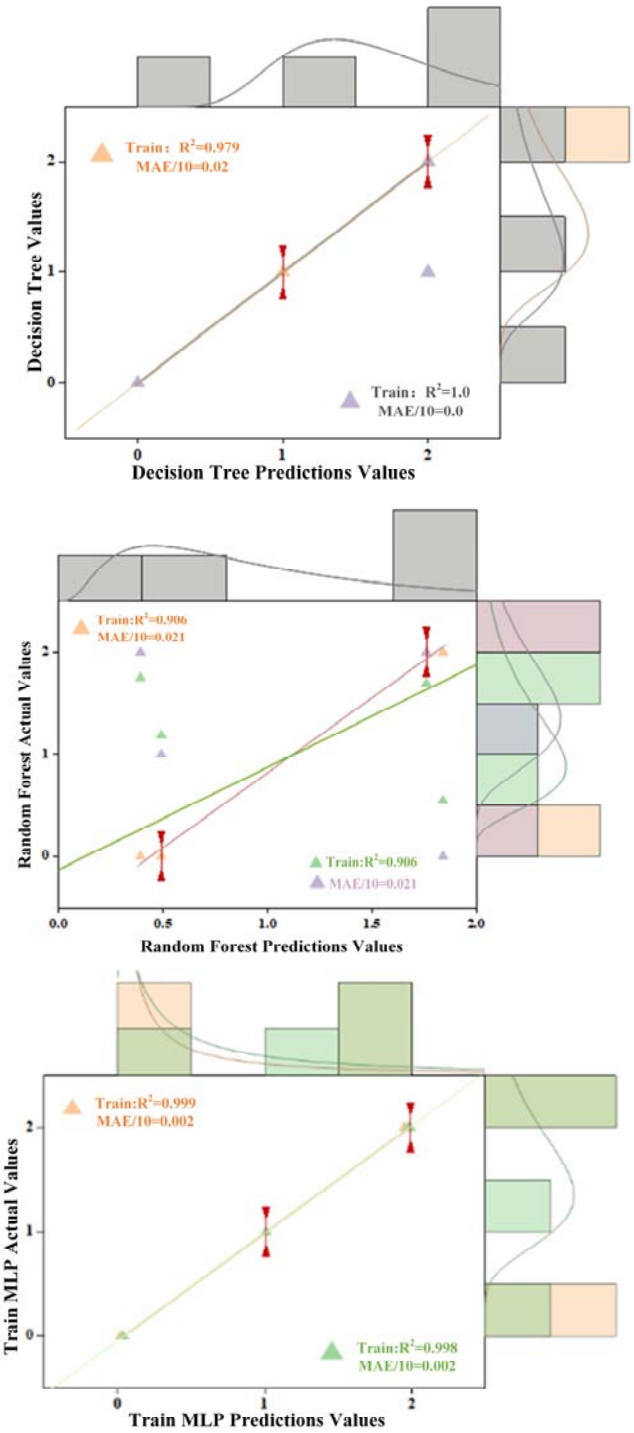


Fig. 9 Plot of R^2 vs. MAE (DT, RF, MLP)

TABLE IV
SAMPLE PROJECTIONS FOR NEW DATA

Random generation of new data	A1-A16; B1-B6; C1-C19; D1-D8	Projected results
Data 1	5, 4, 10, 1, 3, 1, 1, 8, 9, 4, 2, 5, 8, 2, 6, 8, 10, 6, 1, 5, 3, 3, 5, 9, 1, 5, 5, 1, 9, 9, 3, 9, 1, 9, 9, 4, 9, 4, 3, 2, 3, 5, 10, 4, 4, 6, 6, 6, 10	2.2805
Adjusted data 1	**1, 1,** 10, 1, 3, 1, 1, 8, 9, 4, 2, 5, 8, 2, 6, 8, 10, 6, 1, 5, 3, 3, 5, 9, 1, 5, 5, 1, 9, 9, 3, 9, 1, 9, 9, 4, 9, 4, 3, 2, 3, 5, 10, 4, 4, 6, 6, 6, 10	2.5901
Data 2	2, 2, 5, 1, 8, 10, 5, 3, 10, 4, 7, 9, 2, 10, 1, 6,2, 5, 9, 8, 8, 9, 9, 10, 1, 9, 8, 3, 6, 5, 7, 2, 10, 1, 7, 9, 4, 4, 6, 6, 5, 2, 6, 5, 2, 3, 6, 1, 3	1.6867
Adjusted data 2	2, 2, 5, 1, 8, 10, 5, 3, 10, 4, 7, 9, 2, 10, 1, 6, 2, 5, 9, **1, 1,** 9, 9, 10, 1, 9, 8, 3, 6, 5, 7, 2, 10, 1, 7, 9, 4, 4, 6, 6, 5, 2, 6, 5, 2, 3, 6, 1, 3	0.6837

Where the first three are R^2 and the last three are MAE. A comprehensive comparison shows that the model in this study can integrate and categorize the dataset completely and comprehensively. The prediction accuracy is also better than separate modeling. Meanwhile, further prediction and analysis were performed by MLP neural network modeling.

TABLE V
ABLATION DATA

Mold	DT(R^2)	RF(R^2)	MLP(R^2)	DT	RF	MLP
SWOT	0	0	0	0	0	0
TOPSIS+K-Means	0	0	0	0	0	0
SWOT+TOPSIS	-1.5384	-0.6153	0.1874	0.2666	0.22	0.1127
SWOT+K-Means	-0.5	-2	0.4791	0.3333	0.6666	0.3365
SWOT+TOPSIS+K-Means	0.979	0.906	0.9998	0.2	0.212	0.2

TABLE VI
SUPPLEMENTARY EXPERIMENT

Characteristic data	+22	+23	+38	+78
R^2	0.9580	0.8389	0.9576	0.9308
MAE	0.0882	0.1814	0.0858	0.1002

VI. CONCLUSION

In this study, we have developed an artificial intelligence-based data analysis model STK and prediction model using a decision tree, random forest, and MLP for prediction. This approach allows for a full categorical ranking in case of insufficient data and can provide predictive research solutions. The following conclusions can be drawn from this research work.

A. The ablation experiments show that the STK model proposed in this study has better performance and demonstrates better model accuracy in model prediction compared to the model without added optimized clustering.

B. The proposed prediction model can be useful in the case of insufficient ecological data, while we add reconstructed data to further supplement the experiment. The results show that MLP can still achieve a certain level of accuracy in STK model prediction when the amount of data increases.

C. The performance of the proposed predictive models was evaluated using statistical tests (e.g., R2, MAE, and cross-validation). Among them, the accuracy of the MLP model is higher than the other two models.

D. The validation results of simulated predictions on random datasets show that SKL combined with the MLP prediction model has a higher accuracy rate.

E. With this approach, we can make better use of sample data to provide referable predictions and decision support for ecologically sustainable environments and other domains with insufficient data. In the future, more advanced and complex frameworks can be combined to act on complex environmental data.

REFERENCES

[1] C. Folke, S. Polasky, etc, "Our future in the Anthropocene biosphere," *Ambio,* vol. 50, no. 4, pp. 834-869, Apr 2021.

[2] S. Díaz, J. Settele, E. S. Brondízio, et al., "Pervasive human-driven decline of life on Earth points to the need for transformative change," Science, vol. 366, no. 6471, pp. eaax3100, 2019.

[3] W. Ning, Y. Hu, S. Feng, et al., "Ecological risk assessment and transmission of soil heavy metals in pastoral areas of the Tibetan plateau based on network environment analysis," Sci. Total Environ., vol. 905, pp. 167197, 2023/12/20/, 2023.

[4] M. Francisco, "Artificial intelligence for environmental security: national, international, human and ecological perspectives," Current Opinion in Environmental Sustainability, vol. 61, Apr, 2023.

[5] T. Kaya, and C. Kahraman, "Multicriteria renewable energy planning using an integrated fuzzy VIKOR & AHP methodology: The case of Istanbul," *Energy,* vol. 35, no. 6, pp. 2517-2527, 2010/06/01/, 2010.

[6] R. Eslamipoor, and A. Sepehriar, "Firm relocation as a potential solution for environment improvement using a SWOT-AHP hybrid method," *Process Safety and Environmental Protection,* vol. 92, no. 3, pp. 269-276, May 2014.

[7] E. Bas, "The integrated framework for analysis of electricity supply chain using an integrated SWOT-fuzzy TOPSIS methodology combined with AHP: The case of Turkey," *International Journal of Electrical Power & Energy Systems,* vol. 44, no. 1, pp. 897-907, Jan 2013.

[8] Y. A. Solangi, Q. Tan, N. H. Mirjat, *et al.*, "Evaluating the strategies for sustainable energy planning in Pakistan: An integrated SWOT-AHP and Fuzzy-TOPSIS approach," *Journal of Cleaner Production,* vol. 236, pp. 117655, 2019/11/01/, 2019.

[9] S. M. Hosseini, M. M. Paydar, and C. Triki, "Implementing sustainable ecotourism in Lafour region, Iran: Applying a clustering method based on SWOT analysis," *J. Clean Prod.,* vol. 329, 2021.

[10] A. Jain, and R. Dubes, *Algorithms for Clustering Data*, 1988.

[11] Y. Li, and H. Wu, "A Clustering Method Based on K-Means Algorithm," *Physics Procedia,* vol. 25, pp. 1104-1109, 2012.

[12] R. Nishant, M. Kennedy, and J. Corbett, "Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda," *International Journal of Information Management,* vol. 53, pp. 102104, 2020/08/01/, 2020.

[13] X. Li, Z. Huang, and W. Ning, "Intelligent manufacturing quality prediction model and evaluation system based on big data machine learning," *Computers and Electrical Engineering,* vol. 111, pp. 108904, 2023/10/01/, 2023.

[14] E. Rafiei-Sardooi, A. Azareh, B. Choubin, *et al.*, "Evaluating urban flood risk using a hybrid method of TOPSIS and machine learning," *International Journal of Disaster Risk Reduction,* vol. 66, pp. 102614, Dec, 2021.

[15] S. Kim, Y. Hong, J.-T. Lim, *et al.*, "Improved prediction of shale gas productivity in the Marcellus shale using geostatistically generated well-log data and ensemble machine learning," *Computers & Geosciences,* vol. 181, pp. 105452, 2023/12/01/, 2023.

[16] Y.-D. Xue, W. Zhang, etc, "Serviceability evaluation of highway tunnels based on data mining and machine learning: A case study of continental United States," *Tunnelling and Underground Space Technology,* vol. 142, pp. 105418, 2023/12/01/, 2023.

[17] D. M. Camacho, K. M. Collins, R. K. Powers, *et al.*, "Next-Generation Machine Learning for Biological Networks," *Cell,* vol. 173, no. 7, pp. 1581-1592, Jun 14, 2018.

[18] J. B. Duarte-Duarte, L. H. Talero-Sarmiento, and D. C. Rodriguez-Padilla, "Methodological proposal for the identification of tourist routes in a particular region through clustering techniques," *Heliyon,* vol. 7, no. 4, pp. e06655, Apr 2021.

[19] X. Li, F. Li, X. Min, *et al.*, "Embracing eDNA and machine learning for taxonomy-free microorganisms biomonitoring to assess the river ecological status," *Ecological Indicators,* vol. 155, pp. 110948, 2023/11/01/, 2023.

[20] N. Cakici, C. Fieberg, D. Metko, *et al.*, "Machine learning goes global: Cross-sectional return predictability in international stock markets," *Journal of Economic Dynamics and Control,* vol. 155, pp. 104725, 2023/10/01/, 2023.

[21] E. Brondízio, J. Settele, S. Diaz, *et al.*, *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*, 2021.
[22] Y. Himeur, B. Rimal, A. Tiwary, *et al.*, "Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives," *Information Fusion,* vol. 86-87, pp. 44-75, 2022/10/01/, 2022.
[23] R. Y. Chock, W. B. Miller, etc, "Quantitative SWOT analysis: A structured and collaborative approach to reintroduction site selection for the endangered Pacific pocket mouse," *J. Nat. Conserv.,* vol. 70, pp. 6-15, Dec 2022.
[24] S. Liu, Z. Zhang, H. Zhou, *et al.*, "A Generative Model-Based Network Framework for Ecological Data Reconstruction," *Computers, Materials & Continua,* vol. 0, no. 0, pp. 1-10, 2024.
[25] H.-S. Shih, H.-J. Shyur, and E. S. Lee, "An extension of TOPSIS for group decision making," *Math. Comput. Model.,* vol. 45, no. 7-8, pp. 801-813, Mar, 2007.
[26] S. Opricovic, and G.-H. Tzeng, "Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS," *Eur. J. Oper. Res.,* vol. 156, no. 2, pp. 445-455, Jul, 2004.
[27] K. R. Dayal, S. Durrieu, K. Lahssini *et al.*, "Enhancing Forest Attribute Prediction by Considering Terrain and Scan Angles From Lidar Point Clouds: A Neural Network Approach," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 16, pp. 3531-3544, 2023.
[28] H. Xu, "Prediction on Bundesliga Games Based on Decision Tree Algorithm," in 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2021, pp. 234-238.
[29] P. Rebuschat, and J. Williams, "Introduction: Statistical learning and language acquisition," *Statistical Learning and Language Acquisition*, P. Rebuschat, and J. N. Williams, eds., pp. 1-12: De Gruyter Mouton, 2011.
[30] G. Wei, J. Zhao, Y. Feng, *et al.*, "A novel hybrid feature selection method based on dynamic feature importance," *Applied Soft Computing,* vol. 93, pp. 106337, 2020/08/01/, 2020.
[31] J. Savoy, "Machine Learning Models," *Machine Learning Methods for Stylometry*, J. Savoy, ed., pp. 109-151, Cham: Springer International Publishing, 2020.
[32] Y. Y. Song, and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch Psychiatry,* vol. 27, no. 2, pp. 130-5, Apr 25, 2015.
[33] A. Liaw., and M. Wiener, "Classification and Regression by randomForest."
[34] A. Sharifi, A. Sharafian, and Q. Ai, "Adaptive MLP neural network controller for consensus tracking of Multi-Agent systems with application to synchronous generators," *Expert Systems with Applications,* vol. 184, 2021.
[35] G. E. Hinton, and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science,* vol. 313, no. 5786, pp. 504-7, Jul 28, 2006.
[36] Z. Sun, G. Wang, P. Li *et al.*, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Systems with Applications,* vol. 237, pp. 121549, 2024/03/01/, 2024.
[37] H. Tin Kam, "Random decision forests." pp. 278-282 vol.1.