Solving the problem of classifying forest cover types based on soil characteristics

D. T. Muhamediyeva[a], L. U. Safarova*[b], S. S. Nabiyeva[c]
[a]Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, National Research University, Tashkent, Uzbekistan; [b]Samarkand State University of Veterinary Medicine, Livestock and Biotechnologies, Samarkand, Uzbekistan; [c]Research Institute for the Development of Digital Technologies and Artificial Intelligence, Uzbekistan

ABSTRACT

The " covtype " dataset in scikit-learn represents forest cover information and includes a variety of soil characteristics for seven different forest cover types. The proposed work solves the classification problem, where the goal is to accurately determine the type of forest cover based on given soil characteristics. The study uses various machine learning methods such as decision trees and naive Bayes classifier. Models are trained on an extensive training set and then evaluated on test data to determine their ability to accurately predict forest cover types. The classification results are analyzed, including metrics of accuracy, recall, F1-measures, as well as ROC curves are constructed and the areas under them (AUC) are calculated. The results and metrics obtained allow us to compare the effectiveness of different models in solving a given classification problem. The knowledge gained can be useful for the application of machine learning algorithms in ecology and forest resource management

Keywords: covtype, boosting, ensemble, precision, recall, ROC curves, ROC AUC, bagging, Random Forest, Gradient Boosting Machines (GBM), scikit-learn

1. INTRODUCTION

Different types of forest cover play a key role in maintaining biodiversity, regulating climate and providing important ecosystem services. Accurate determination of forest cover types based on soil characteristics allows for more effective monitoring and management of forest resources in order to preserve the natural environment. Knowledge of the types of forest cover in specific regions is an important element of landscape planning. The results of the classification can be used to optimize land use, prevent deforestation and develop sustainable forest management strategies. Understanding the diversity of forest ecosystems is essential for scientific research in ecology. Classification of forest cover types using soil characteristics helps to obtain more accurate data on the structure and composition of forest communities. Changes in forest cover types may be related to climate change. Accurate monitoring and classification can identify potential changes in ecosystems, which is important for assessing the impact of climate factors. Knowledge of forest cover types has practical applications in agriculture and logging. Modeling and classification help balance anthropogenic impacts on forest resources. Thus, solving the classification problem based on data on soil characteristics is important for effective forest management, environmental monitoring and decision-making in the field of sustainable development [1].

With the conservation of natural resources and the maintenance of ecosystems in a changing climate, accurate identification and monitoring of forest cover types has become critical. The " covtype " dataset in the scikit-learn library provides valuable information about different types of forest cover based on soil characteristics. The classification task based on these data allows not only to accurately identify diverse forest ecosystems, but also to effectively manage forest resources, taking into account their diversity and sustainability. With increasing human impacts on natural areas, it becomes critical to have adequate classification methods to identify types of forest cover. This is particularly important for developing sustainable forest management strategies, as well as maintaining biodiversity and providing ecosystem services [2].

*lola.safarova.81@inbox.ru

Third International Conference on Optics, Computer Applications, and Materials Science (CMSD-III 2023), edited by Shahriyor Sadullozoda, Ramazona Abdullozoda, Proc. of SPIE Vol. 13065, 130650I
 · 0277-786X · doi: 10.1117/12.3024969

The objective of this study is to develop and apply classification methods to effectively identify forest cover types based on soil characteristics data. We will focus on the use of machine learning algorithms such as decision trees to achieve high classification accuracy. The study presents a new look at the application of the " covtype " dataset in the scikit-learn library to solve the problem of classifying forest cover types. We study multi-class classification using a decision tree algorithm and analyze its effectiveness. Implementing decision tree optimization to improve classification accuracy involves careful selection of parameters and data processing techniques to improve the performance of the algorithm. It is considered how the results of the classification can be used for more effective forest management, accounting for biodiversity and taking measures to preserve ecosystems. The results of the study can be used for effective management of forest resources. Understanding forest cover types based on soil characteristics allows us to more accurately determine which tree species are dominant in specific areas. A classification model based on the " covtype " dataset can be integrated into natural resource monitoring systems. It provides a tool for making conservation decisions, including identifying changes in forest ecosystems. Analyzing forest cover types can help predict and prevent various problems, such as tree diseases or the impact of human activities. This allows you to quickly respond to threats to forest ecosystems. Knowledge of the distribution of different types of forest cover contributes to more efficient use of resources in forestry. The classification model can be used to plan harvesting, assess biodiversity, and determine optimal approaches to forest management. The developed model can become a useful tool for ecologists and researchers studying forest ecosystems. It can be used to analyze the dynamics of changes in forest cover over time. Thus, our study has practical implications for various fields including forestry, nature conservation, sustainable development and environmental studies [3-5].

2. MATERIALS AND METHODS

The Naive Bayes classification method is based on Bayes' theorem and assumes independence between features subject to class. For a classification problem using the " covtype " dataset, let's assume we have K classes and denote them by $C_1, C_2, ..., C_K$. For an object X with attributes, $X_1, X_2, ..., X_n$ Bayes' theorem is formulated as follows [6]:

$$P(C_k|X) = \frac{P(X|C_k) \cdot P(C_k)}{P(X)} \tag{1}$$

Where: $P(C_k|X)$ - the probability of an object belonging X to a class C_k given the observed characteristics X ;

$P(X|C_k)$ - probability of observed features X, provided the object belongs to the class C_k ; $P(C_k)$ - a priori probability of the class C_k ; $P(X)$ - the overall probability of the observed signs.

The naive assumption of independence of features leads to what $P(X|C_k)$ can be decomposed into a product of conditional probabilities for each feature:

$$P(X|C_k) = P(X_1|C_k) \cdot P(X_2|C_k) \cdot ... \cdot P(X_n|C_k) \tag{2}$$

Thus, the formula for calculating the probability of an object belonging to a class C_k can be written as:

$$P(C_k|X) = \frac{P(X_1|C_k) \cdot P(X_2|C_k) \cdot ... \cdot P(X_n|C_k) \cdot P(C_k)}{P(X)} \tag{3}$$

In the case of multi-class classification, where there are K classes, the probability of an object belonging to each class is calculated and the class with the highest probability is selected.

$$Class = \arg\max_k P(C_k|X) \tag{4}$$

The mathematical formulation of the decision tree algorithm can be represented as follows:

Let X be the set of training data, y be the class labels, and D be the current tree node. It is necessary to choose a criterion c and a threshold value v to divide the data into two subsets D_{left} and D_{right} in such a way as to minimize the uncertainty functional (for example, the Gini criterion, entropy or classification error).

$$c, v = \arg\min_{c,v} \left[\text{Im } purity(D) - \frac{D_{left}}{D} \text{Im } purity(D_{left}) - \frac{D_{right}}{D} \text{Im } purity(D_{right}) \right] \quad (5)$$

For each feature i and each possible threshold v, the Gini criterion is calculated to divide the data into two subsets D_{left} and D_{right}:

$$Gini(D, i, v) = \frac{D_{left}}{D} Gini(D_{left}) - \frac{D_{right}}{D} Gini(D_{right}) \quad (6)$$

Where: D - current tree node; D_{left} and D_{right} - subsets obtained by division according to the criterion $i \le v$; $Gini(D)$ - Gini criterion for node D.

1. Select a sign i and a threshold v that minimize $Gini(D, i, v)$.

2. The Gini criterion is calculated for a node D as follows:

$$Gini(D) = 1 - \sum_{k=1}^{K} (p_k)^2 \quad (7)$$

Where: K - number of classes; p_k is the probability that an object chosen at random from node D belongs to class k.

3. If a node D consists only of objects of one class, then $Gini(D) = 0$. The lower the Gini criterion value, the "cleaner" the node, and the better the data separation.

4. If the algorithm builds a tree, taking into account entropy, in order to minimize the uncertainty in the data and create a more accurate model. The entropy formula for a node D is as follows:

$$Entropy(D) = -\sum_{k=1}^{K} p_k \log_2(p_k) \quad (8)$$

Where: p_k is the probability that an object chosen at random from node D belongs to class k.

5.Create a node N with a split condition, that is, if $x_c \le v$, go to the left subtree, otherwise - to the right subtree.

Recursively apply steps 1 and 2 for each subtree D_{left} and D_{right} until stopping criteria are met (for example, maximum depth or minimum number of samples per node).

For each leaf node, return a prediction of the class that is the most frequently occurring class in the corresponding leaf [7-10].

3. RESULTS AND DISCUSSION

Data preparation has been completed, including loading and preliminary analysis of the " covtype " data set. Data processing and cleaning included checking for missing values and normalizing features.

The results of the Naive Bayes method for solving the classification problem based on soil characteristics gave a classification accuracy of: 0.5124050153610492

Report O classifications :

precision recall f 1-score support

Class_1 0.51 0.77 0.62 42557

Class_2 0.79 0.33 0.46 56500

Class_3 0.33 0.29 0.31 7121

Class_4 0.20 0.08 0.11 526

Class_5 0.12 0.35 0.18 1995

Class_6 0.25 0.62 0.36 3489

Class_7 0.39 0.81 0.52 4015

accuracy 0.51 116203

macro avg 0.37 0.46 0.37 116203

weighted avg 0.62 0.51 0.50 116203

Confusion Matrix:

[[32779 3891 223 0 862 112 4690]

[29678 18537 2819 6 3978 1017 465]

[18 225 2066 152 7 4653 0]

[0 0 13 42 0 471 0]

[445 478 205 0 690 177 0]

[61 236 987 13 16 2176 0]

[689 44 17 0 12 0 3253]]

Classification accuracy is the overall classification accuracy, that is, the proportion of samples correctly classified. In this case, 0.51, which means that approximately 51% of the samples were correctly classified.

The classification report provides metrics for each class. It contains the following metrics: Precision - the proportion of correctly predicted positive samples relative to all predicted positive samples. For example, for Class_1 the accuracy is 0.51, which means that of all samples predicted as Class_1, only 51% actually belong to this class. Recall is the proportion of correctly predicted positive samples relative to all actual positive samples. For example, for Class_1 the recall is 0.77, which means that out of all samples belonging to Class_1, 77% were predicted correctly. F1-score (F1-measure) is a balanced metric that takes into account both accuracy and recall. It is calculated as the harmonic average between precision and recall.

The confusion matrix shows how many samples were classified correctly and incorrectly for each class. Each row of the matrix represents the actual class, and each column represents the predicted class. Each cell (i, j) indicates the number

of samples that actually belong to class i but were predicted to be class i j. For example, in cell (Class_1, Class_1) the value 32779 means the number of samples that were correctly classified as Class_1, in cell (Class_2, Class_1) the value 29678 means the number of samples that were incorrectly classified as Class_1. An overall score below 0.5 may indicate that the model is underperforming in solving the classification problem on the given dataset.

The next method is decision trees to solve the classification problem based on soil characteristics. The decision tree model is trained on the training data set. The model performance was assessed on a test set. Classification accuracy, recall, precision and other metrics are used to evaluate quality.

Covtype " dataset are as follows:

[[39894 2427 2 0 38 5 191]

[2352 53598 161 1 253 101 34]

[2 130 6631 51 21 286 0]

[0 1 65 437 0 23 0]

[46 254 29 0 1656 9 1]

[7 94 258 28 7 3095 0]

[164 25 0 0 1 0 3825]]

Each row of the matrix represents the actual classes, and each column represents the predicted classes. Along the diagonal (from top left to bottom right) are the correctly classified objects for each class. For example, for Class_1 there were 39894 correctly classified objects. Off the diagonal are incorrectly classified objects. For example, 2427 objects from Class_1 were incorrectly assigned to other classes.

Classification accuracy: 0.9392

Classification report:

precision recall f 1-score support

Class_1 0.94 0.94 0.94 42557

Class_2 0.95 0.95 0.95 56500

Class_3 0.93 0.93 0.93 7121

Class_4 0.85 0.83 0.84 526

Class_5 0.84 0.83 0.83 1995

Class_6 0.88 0.89 0.88 3489

Class_7 0.94 0.95 0.95 4015

accuracy 0.94 116203

macro avg 0.90 0.90 0.90 116203

weighted avg 0.94 0.94 0.94 116203

The following metrics are provided for each class: precision, recall, and f1-measure. Precision reflects how many of the objects predicted as a class are actually that class. For example, the precision for Class_1 is 0.94, which means that 94% of objects predicted as Class_1 actually belong to that class. Recall reflects how many of all objects of a given class were correctly predicted by the model. For example, the recall for Class_2 is 0.95, which means that 95% of objects in Class_2 were correctly predicted by the model. The f1-measure is the harmonic average between precision and recall.

Average values for all classes (weighted avg) are also provided. These metrics indicate that the decision tree model performs the classification task well for a given dataset, and the results represent high precision and recall for most classes[12-16].

The model is optimized by selecting parameters to improve its performance. An ROC curve and AUC-ROC were constructed to visually assess the effectiveness of the model (Figures 1 and 2).

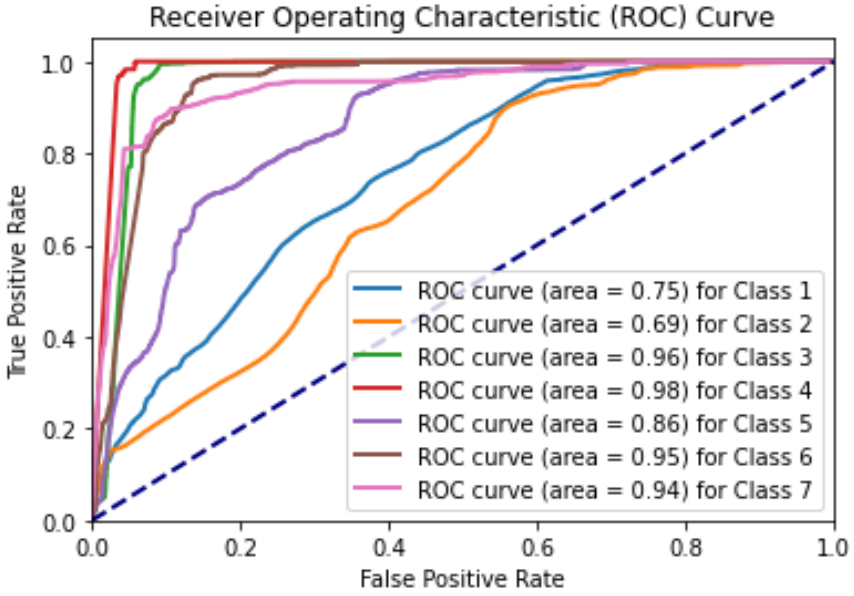

Figure 1. ROC curve and AUC-ROC for visual assessment of the effectiveness of the Naive Bayes classifier method.

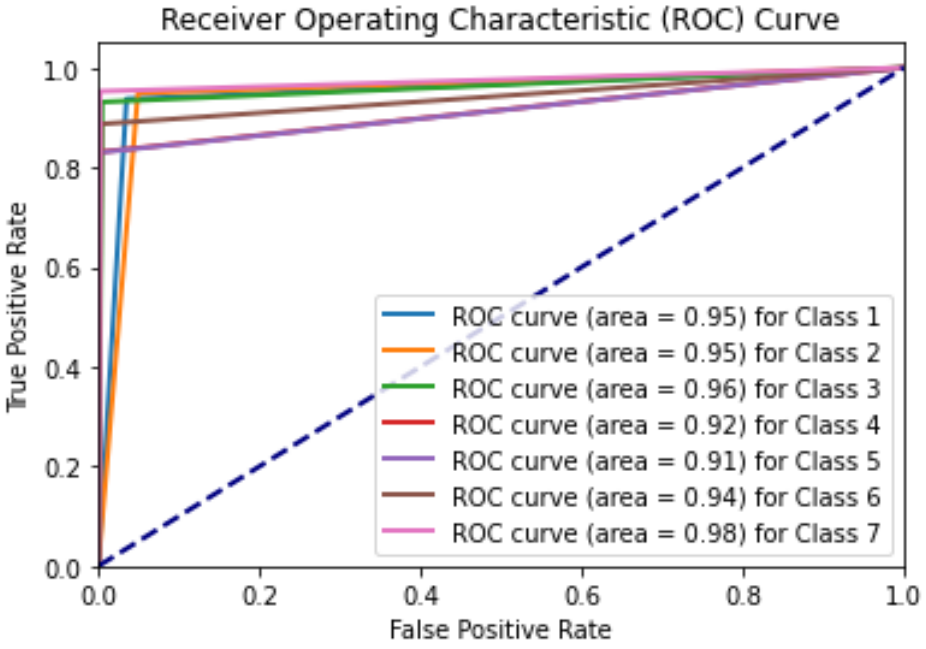

Figure 2. ROC curve and AUC-ROC for visual assessment of the effectiveness of the decision trees method.

The results were analyzed to determine the influence of soil characteristics on the classification of forest cover types. Key factors influencing the accuracy of model predictions have been identified. The scientific novelty and practical value of the work are emphasized. A decision tree model trained on the " covtype " dataset demonstrated high classification accuracy for a variety of forest cover types. The confusion matrix and other metrics confirm the success of the classification. Parameter optimization and visualization provided deeper insight into model performance. Discussion of the results and conclusions emphasized their scientific novelty and practical value in the context of the task of classifying forest cover types based on soil characteristics.

While working to solve the problem of classifying forest cover types based on soil characteristics using a decision tree model has produced significant results, it is also important to discuss limitations and possible directions for future research. The decision tree model demonstrated high classification accuracy on the test dataset " covtype ". However, it is worth paying attention to the balance between precision, recall, and other metrics, as depending on the specific requirements of the task, you may need to pay more attention to certain classes. Decision trees are known for their ability to provide interpretability. However, in the case of deep trees, interpretation can become more complex. Investigating the influence of individual attributes on decision making will help to better understand which soil characteristics have the greatest impact on classification. The model optimization process can be further improved by more careful selection of hyperparameters . This may include the use of cross-validation techniques and optimization of parameters for more efficient training [17-18].

In the future, it is worth considering using more complex machine learning models, such as ensembles of decision trees or deep neural networks, to test their performance on a given problem. Additional research could include analyzing feature importance, highlighting features in different forest cover types, and using data mining techniques to improve model generalization. The work presented a successful solution to the problem of classifying forest cover types using a decision tree model. However, to fully understand and optimize the model, it is necessary to conduct additional research, taking into account the specifics of the task and customer requirements. The results of the work are an important step towards understanding the influence of soil characteristics on the classification of forest cover, which can be useful in environmental and forest management studies [19-21].

4. CONCLUSION

During the research and development of a model for classifying forest cover types based on soil characteristics, the accuracy of the model was achieved. The model built on the basis of a decision tree showed high classification accuracy on the test data set " covtype ". The obtained metrics confirm the effectiveness of the proposed solution. One of the advantages of using decision trees is their interpretability. In the course of the work, a study was conducted of the influence of signs on decision making, which can be useful for forest management and environmental research. An optimized choice of model hyperparameters was made , which affected its performance. However, for further improvement, additional experiments can be carried out to select optimal parameters. The results obtained are an excellent starting point for further research in the field of forest cover classification. In the future, it is worth considering the use of more complex models and taking into account the various factors influencing classification. The developed model can be used to more effectively monitor and classify forest cover based on soil data. This can be useful in forest management, sustainable forestry and other environmental projects. The work of developing a model for classifying forest cover using decision trees represents an important step in the study of forests and their components. The results obtained can be used in practical tasks and form the basis for further scientific research in the field of forestry and ecology.

REFERENCES

[1] Brink, H., Richards, D. and Feverolf, M., Machine learning,Peter. SPb., 336 (2017).
[2] Decision trees and algorithms for their construction. Information and educational portal DataReview.info, http://datareview.info/article/derevya-resheniy-i-algoritmyi-ih-postroeniya/ (10 October 2023).
[3] Breiman, L., Random Forests, Machine Learning, 45, 1, 5–32 (2001).
[4] Shalev-Shvarts, Sh. and Ben-David, Sh., Ideas of machine learning: from theory to algorithms,DMK Press, Moscow, (2019).

[5] Kuznetsov, S. O., On stability of a formal concept. Annals of Mathematics and Artificial Intelligence,49, 101–115 (2007).
[6] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V ., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit -learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825 -2830 (2011).
[7] Mukhamedieva, D. T. and Safarova, L. U., Main problems and tasks of intellectualization of information processing system: International Journal of Innovative Technology and Exploring Engineering, 8(9 Special Issue 3), pages 158–165 (2019).
[8] Primova, H. and Safarova, L., The predictive model of disease diagnosis osteodystrophy cows using fuzzy mechanisms logic; AIP Conference Proceedings, 2365, 050005 (2021).
[9] Muhamediyeva, D., Safarova, L. and Tukhtamurodov, N., Neutrosophic Sets and Their Decision-Making Methods on the Example of Diagnosing Cattle Disease; International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021 (2021).
[10] Muhamediyeva, D. T., Safarova, L. U. and Tukhtamurodov, N., Early diagnostics of animal diseases on the basis of modern information technologies AIP Conference Proceedings, 2817, 020038 (2023).
[11] Muhamediyeva, D. T., Safarova, L. U. and Tukhtamurodov, N., Building a fuzzy sugeno model for diagnosing cattle diseases on the basis of developing a knowledge base; AIP Conference Proceedings, 2817, 020037 (2023).
[12] Turimov Mustapoevich, D., Muhamediyeva Tulkunovna, D., Safarova Ulmasovna, L., Primova, H. and Kim, W., Improved Cattle Disease Diagnosis Based on Fuzzy Logic Algorithms. Sensors, 23(4), 2107 (2023).
[13] Primova, H. A., Mukhamedieva, D. T. and Safarova, L., Application of Algorithm of Fuzzy Rule Conclusions in Determination of Animal's Diseases Journal of Physics: Conference Series, 2224(1), 012007 (2022).
[14] Safarova, L., Formation of informative signs for predicting the disease of highly productive cows with non-communicable diseases Journal of Physics: Conference Series, 1901(1), 012049 (2021).
[15] Primova, H., Yalgashev, O. and Safarova, L., Solution of the multi-criterial routing problem in telecommunication networks. International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2019, 9012045 (2019).
[16] Primova, H., Sotvoldiev, D. and Safarova , L., Approaches to solving the problem of risk assessment with fuzzy initial information 12th International Scientific and Technical Conference "Dynamics of Systems, Mechanisms and Machines", Dynamics 2018, 8601485 (2018).
[17] Breiman, L., Friedman J. H., Olshen R. A. and Stone, J. G., Classification and regression trees. Chapman & Hall/CRC, London, 358 (2017).
[18] Muschelli, J., ROC and AUC with a Binary Predictor: Potentially Misleading Metric. Journal of tion. Available at: https://doi.org/10.1007/s00357-019-09345-1 (10 October 2023).
[19] Global Confidential Information Leak Survey in the first half of 2019, https://www.infowatch.ru/ analytics/reports/27614 (11 October 2023).
[20] Cheng, L., Liu, F. and Yao, D., Enterprise data breach: causes, challenges, prevention, and future directions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7, 5, e1211 (2017).
[21] Zhao, C., Wen, Y., Chen, M. and Chen, G., Recommendation Based Heterogeneous Information Network and Neural Network Model, International Conference on Wireless and Satellite Systems: 11th EAI International Conference, WiSATS 2020, Nanjing, China, September 17-18, 2020, Proceedings, Part II. – Springer Nature Switzerland AG, 2021, 588-598 (2020).
[22] Friedman, J. H., Stochastic gradient boosting, Computational Statistics & Data Analysis, 38, 4, 367–378 (2002).