Journal of Infrastructure, Policy and Development 2024, 8(7), 4074.
https://doi.org/10.24294/jipd.v8i7.4074

Article

Application of non-parametric learning method in soil suitability assessment in present day economy

Vladislav Kukartsev[1,2,*], Andrei Gantimurov[2], Kirill Kravtsov[1,2], Aleksey Borodulin[2], Yadviga Tynchenko[2,3]

[1] Department of Information Economic Systems, Reshetnev Siberian State University of Science and Technology, 660037 Krasnoyarsk, Russia
[2] Artificial Intelligence Technology Scientific and Education Center, Bauman Moscow State Technical University, 105005 Moscow, Russia
[3] Laboratory of Biofuel Compositions, Siberian Federal University, 660041 Krasnoyarsk, Russia
*** Corresponding author:** Vladislav Kukartsev, vlad_saa_2000@mail.ru

CITATION

Kukartsev V, Gantimurov A, Kravtsov K, et al. (2024). Application of non-parametric learning method in soil suitability assessment in present day economy. Journal of Infrastructure, Policy and Development. 8(7): 4074. https://doi.org/10.24294/jipd.v8i7.4074

ARTICLE INFO

Received: 6 January 2024
Accepted: 3 April 2024
Available online: 1 August 2024

Abstract: This research delves into the urgent requirement for innovative agricultural methodologies amid growing concerns over sustainable development and food security. By employing machine learning strategies, particularly focusing on non-parametric learning algorithms, we explore the assessment of soil suitability for agricultural use under conditions of drought stress. Through the detailed examination of varied datasets, which include parameters like soil toxicity, terrain characteristics, and quality scores, our study offers new insights into the complexities of predicting soil suitability for crops. Our findings underline the effectiveness of various machine learning models, with the decision tree approach standing out for its accuracy, despite the need for comprehensive data gathering. Moreover, the research emphasizes the promise of merging machine learning techniques with conventional practices in soil science, paving the way for novel contributions to agricultural studies and practical implementations.

Keywords: sustainable growth; land cover change; land degradation; land use; soil quality

1. Introduction

The concept of machine learning revolves around the ability of systems to analyze data, discern patterns, and anticipate future events or make decisions based on that information. This learning process encompasses various methodologies, including supervised, unsupervised, and reinforcement learning techniques. Machine learning finds application across diverse domains such as medicine (for disease diagnosis and treatment), finance (in market forecasting and risk management), autonomous navigation (for unmanned vehicles), language processing (enabling automatic translation and text analysis), among others.

As computing power advances and big data becomes increasingly accessible, the capabilities of machine learning expand, reshaping conventional perceptions of computational abilities. This dynamic field remains a focal point of research, offering vast potential for innovation and improvement across multiple sectors.

In the contemporary economy, characterized by a pressing need for sustainable development, the assessment of soil suitability emerges as a pivotal component of strategic planning in agriculture and environmental conservation (Kieliszek, 2017). This task is complicated by numerous factors, including soil attributes, climatic conditions, technological advancements, and consumer preferences. Addressing this multifaceted and dynamic challenge, non-parametric learning methods, such as multi-criteria decision analysis (MCDA), stand out as potent tools for organizing and analyzing data (Bashmur, 2023).

MCDA algorithms facilitate the formalization and classification of diverse agricultural aspects, empowering decision-makers to comprehend the interplay between different criteria and optimize processes (Kieliszek et al., 2017). They enable stakeholders, including farmers and policymakers, to sustainably manage production, ensuring the provision of safe and high-quality food.

Agricultural soils play a pivotal role in this intricate system, serving as a nexus for various pressures, including climate change, population growth, and escalating food demand (Doran, 1994). Science-based and user-friendly tools are imperative for making informed decisions regarding agricultural land use and soil quality management (Schwilch, 2011).

This paper explores the effective utilization of non-parametric learning methods in soil suitability assessment within the contemporary economic landscape. It elucidates how these methodologies contribute to informed decision-making in agriculture, pivotal for advancing sustainable development objectives and safeguarding food security. Furthermore, it outlines how the formulation and testing of hypotheses guide the development of predictive models aimed at addressing the complexities of soil drought suitability assessment. Subsequently, we present hypotheses aimed at determining a more cost-effective and accessible model for predicting soil drought suitability in expert systems.

2. Review of literature

Soil quality is defined as "the ability of the soil to function within the boundaries of an ecosystem, support biological productivity, maintain environmental quality, and contribute to plant, animal, and human health" (Karlen et al., 1997). The physical quality of agricultural soil primarily pertains to its suitability for cultivation, as well as the fluid transmission and storage characteristics of the crop root zone (Topp et al., 1997).

Several conceptual frameworks have been proposed in recent studies for monitoring soil quality (Alawi, 2022). These frameworks select soil characteristics from a minimum data set based on their suitability to assess specific soil functions (Andrews et al., 2004). However, the cost and labor intensiveness of standard procedures for monitoring all soil quality indicators across different areas and land management types remain significant challenges (Cecilion, 2009). Relevant soil physical indicators play a crucial role in determining soil quality status as they reflect the soil's capacity to store and provide water, air, and nutrients necessary for crop growth (Malek et al., 2018).

The combined influence of soil characteristics, such as nutrient content, moisture level, and drought tolerance, determines soil suitability for vegetation growth under drought conditions. New machine learning (ML) methods and algorithms are increasingly applied to automate soil classification processes, aiming to reduce time and cost expenditures (Robertson, 2016). For instance, a recent study from Turkey and Korea identified decision tree (CART) classification as the most successful among various classifiers, significantly reducing the time and cost of soil classification (Aydin et al., 2023).

Furthermore, studies discussing the suitability of soil classification and analysis highlight the importance of planting trees in appropriate areas. For example, research on the distribution of natural and anthropogenic forests in the Yanhe River Basin revealed mismatches between forest planting locations and environmental conditions, resulting in low-productivity forests (Shi et al., 2016). Other studies emphasize the potential of machine learning approaches in improving soil property prediction accuracy (Trontelj and Chambers, 2021). Additionally, the estimation of evapotranspiration using deep learning techniques aids in real-time irrigation management, enhancing water resource utilization in agriculture (Mohan and Patil, 2018; Sokolov et al., 2023).

In the realm of big data and ML technologies in agriculture, Hadoop and Apache Spark emerge as prominent tools for data processing and analysis (Cravero et al., 2022; Pandya et al., 2020; Sitokonstantinou et al., 2020). ML models, such as CatBoost, have proven accuracy and effectiveness in various applications, including predicting school performance and disease diagnosis (Bharati, 2022; Chen and Ding, 2023). The CatBoost model, introduced by Yandex engineers in 2017, utilizes gradient-boosted decision trees to handle noisy data and complex relationships (Prokhorenkova et al., 2018).

Moreover, ML-based intelligent diagnosis systems, employing classifiers like random forests and support vector machines, demonstrate high accuracy in diagnosing complex disorders such as polycystic ovary syndrome (Danaei Mehr and Polat, 2022; Tiwari et al., 2022). Ensemble approaches, including the BorutaShap method and random forest models, are effective in identifying significant clinical markers for disease diagnosis (Silva et al., 2022).

Other applications of ML and data analysis techniques include the analysis of road accidents, determination of gas dynamic characteristics in coal mine facilities, and prediction of load measurements in software and hardware systems (Kukartsev et al., 2022; Martyushev et al., 2023; Masich et al., 2022; Shutaleva et al., 2023). Ensemble approaches, particularly those incorporating neural network models, are recognized as powerful tools for solving data analysis problems across various practical applications (Panfilova et al., 2022).

Assessing soil suitability holds paramount importance from an economic perspective due to its direct implications on agricultural productivity, resource allocation, and environmental sustainability. A thorough understanding of soil quality and its impact on crop yield helps optimize resource utilization, minimize input costs, and maximize profits for farmers and stakeholders in the agricultural sector (Andrews et al., 2004; Karlen et al., 1997). Moreover, soil suitability assessment plays a pivotal role in land use planning, ensuring efficient allocation of arable land and mitigating risks associated with soil degradation and erosion, which can have significant economic repercussions (Cecilion, 2009; Malek et al., 2018). By integrating economic considerations into soil suitability assessment, policymakers and agricultural practitioners can make informed decisions regarding land management practices, investment strategies, and policy interventions aimed at promoting sustainable agricultural development and food security (Aydin et al., 2023; Robertson, 2016). Therefore, the literature review in this manuscript seeks to elucidate the economic dimensions of soil suitability assessment and identify gaps in existing research to pave

the way for the development of cost-effective and economically viable predictive models. This integration of economic factors into soil suitability assessment not only facilitates better resource management but also contributes to long-term agricultural sustainability and food security, aligning with broader socioeconomic development objectives (Aydin et al., 2023; Karlen et al., 1997).

In summary, the literature review presented underscores the multifaceted nature of soil suitability assessment, emphasizing its economic significance. Studies such as those by Karlen (1997) and Aydin (2023) demonstrate the potential benefits of constructing predictive models for soil suitability assessment, highlighting the opportunities to optimize resource allocation, minimize input costs, and maximize agricultural productivity. These insights emphasize the urgency of developing advanced predictive models to address contemporary agricultural challenges effectively.

By examining existing research findings and identifying areas requiring further investigation, this review aims to provide insights for the development of robust and economically viable approaches to soil suitability assessment. Such models not only have the potential to enhance agricultural sustainability but also contribute to the broader goals of ensuring food security and promoting socioeconomic development. With these considerations in mind, the subsequent section delves into hypotheses aimed at guiding the formulation and testing of predictive models for soil suitability assessment.

Hypotheses:

1) Factors related to laboratory soil tests (such as root condition, nutrient availability, soil toxicity) may pose the greatest challenge in data collection, yet they could be crucial for accurate forecasting of soil suitability for cultivation.

2) Data concerning terrain features and landscape characteristics (e.g., slope and aspect of the terrain) are publicly available and collected by various organizations for their own purposes, making them more accessible and less resource-intensive for use in soil suitability assessment models.

3) Data on land use assessment in the county (e.g., assessment of land area, crops, and water bodies) may be accessible through models such as PLUS, providing high accuracy and tools for analyzing land use evolution in regions.

These hypotheses are proposed to determine a more cost-effective and accessible model for predicting soil drought suitability in expert systems.

3. Materials and methods

The soil classification dataset used in this study contains records derived from U.S. area-wide drought monitoring, manually created by experts using a wide range of data (Fischer et al., 2008). These data were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program. The U.S. Drought Monitor is produced through a partnership between the National Drought Mitigation Center at the University of Nebraska-Lincoln, the United States Department of Agriculture, and the National Oceanic and Atmospheric Administration. Training and testing of the exploratory analysis model were conducted using deductor studio.

The main objective of the study was to identify factors influencing soil suitability for cultivation under drought conditions. Data object classification, a data mining and value management technique used to group similar data together, was used to conduct the study and identify new factors influencing soil suitability for cultivation under drought conditions.

This study uses the Deductor analytical platform, which is the basis for creating comprehensive application solutions (Kolenchukov et al., 2022). It is also worth noting that the version of the Deductor platform is a training version, which severalfold limits the possibility of data customization and thus leads to a high error rate and incorrect data display.

A decision tree is a tree-like structure similar to a flowchart (Bukhtoyarov, 2022; Vlasov et al., 2022). In this algorithm there is an automatic selection of features to the nodes from the set of features, construction of decision rules in a form understandable to the expert (Priyam et al., 2013).

The database comprises 29 indicators across 3109 FIPS counties in the USA, encompassing various aspects such as geographical coordinates (latitude and longitude), mean elevation, slope measurements (in different ranges), aspect direction (North, East, South, West), qualitative assessments of water bodies (WAT_LAND), infertile land (NVG_LAND), urbanization (URB_LAND), vegetation presence (GRS_LAND, FOR_LAND), cultivated land (CULTIR_LAND, CULT_LAND), and soil quality parameters (SQ1-SQ7). These parameters encompass factors like soil nutrient availability, retention capacity, root conditions, oxygen availability for roots, salt excess, soil toxicity, and overall suitability for plant growth. Each parameter is assigned a numerical value ranging from 0 to 7, reflecting its respective condition or suitability level.

For instance, the SQ1 parameter represents soil nutrient availability, with values ranging from 0 (no nutrients) to 7 (full nutrient availability). Similarly, SQ2 indicates soil nutrient retention capacity, SQ3 denotes conditions required for rooting and vegetation growth, SQ4 signifies oxygen availability for roots, SQ5 measures salt excess in soil, SQ6 evaluates soil toxicity, and SQ7 assesses the suitability of soil for plant growth, ranging from 0 (not suitable) to 7 (highly suitable).

These indicators collectively provide comprehensive insights into the soil characteristics and environmental conditions of each county, facilitating the assessment of soil suitability for various agricultural purposes.

Statistical data mapping **Figure 1** was generated from the raw data.

Before starting to work with a dataset, processing is necessary to achieve the best sampling conditions. This process is called data normalization. At the data normalization stage, outliers and extreme values were edited. The method of data processing was chosen to fill in the average value. It was also necessary to fill in the missing data, filling occurred for parameters where the pass exceeded 30%. The method was chosen as a fill equal to the median of the values. This method allows you to insert the average of the input value instead of the missing data.

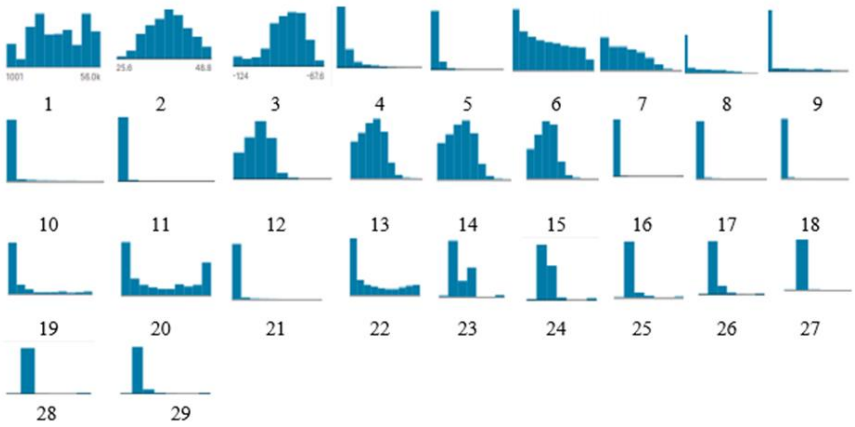

Figure 1. Displaying sample statistics for each parameter in the dataset.

The next step was to configure the decision tree parameters separately for each model. The dataset was divided into 95% of the training set and 5% of the test set.

The data processing setup steps were followed by training and visualization of the results.

It is worth mentioning, before proceeding to the results, what correlation analysis of data is.

Correlation analysis is a statistical method used to study the relationship between two or more variables. It helps to determine if there is a statistically significant relationship between these variables and what the strength of this relationship is.

Correlation analysis uses a correlation coefficient that measures the degree of linear dependence between variables. The correlation coefficient can take values from −1 to 1. A value close to 1 indicates a strong positive correlation, whereas a value close to −1 indicates a strong negative correlation. A value close to zero means there is no correlation.

4. Results

Correlation analysis was used to determine the quality of the data set. At the stage of correlation analysis, the Pearson correlation coefficient was used.

Using the Pearson correlation coefficient, it is possible to determine the strength and direction of the linear relationship between two processes occurring simultaneously and without taking into account the time lag. The value of the correlation parameter indicates the strength of dependence of one factor on another. Such values are categorized into weak (less than 0.29), moderate (0.3–0.49), medium (0.5–0.69) and strong (0.7 or more). From the correlation analysis, it is evident that the grade data for the first and second periods of training have a strong correlation with the output parameter. This analysis allowed us to remove attributes whose significance is less than 0.05, i.e., they have weak dependence.

Correlation analysis of the data was used to identify the extent to which a factor was influenced by the output data (**Table 1**).

Using correlation, factors that were directly dependent on the soil suitability index and factors that were inversely dependent on the soil suitability index were identified. The factors that were directly dependent were: root condition, soil toxicity, salt excess, oxygen availability to roots, ability to retain nutrients in the soil, indicator

of water bodies, and nutrient availability.

Let us start by describing the Decision Tree method. The model is built taking into account the correlation analysis given in **Table 1**.

Table 1. Correlation analysis, the output parameter SQ7 is suitability for growing plants in the soil (workability).

NO.	Attribute name	Correlation, %	Graph as a percentage
1	The state of the roots	0.958	
2	Soil toxicity	0.876	
3	Excess salts	0.863	
4	Oxygen availability for roots	0.763	
5	Ability to retain nutrients in the soil	0.715	
6	Indicator of water bodies	0.672	
7	Nutrient availability	0.571	
8	Total area of cultivated land in the district	−0.191	
9	Evaluation of cultures	−0.178	
10	Slope 2	−0.163	
11	Slope 7	0.114	
12	Slope 8	0.101	
13	Slope 6	0.089	
14	Aspect	−0.086	
15	Slope 4	−0.070	
16	Slope 1	−0.065	
17	Assessment of land cultivation	−0.063	
18	Aspect	−0.060	
19	Aspects	−0.057	
20	Aspects	−0.048	
21	Slope 5	0.039	
22	Assessment of infertility of the district	0.038	
23	Average altitude above sea level	0.035	
24	Width	0.029	
25	Assessment of the presence of forests in the district	−0.019	
26	Assessment of the presence of vegetation in the district	−0.012	
27	Assessment of urbanization in the district	0.011	
28	Longitude	−0.001	

After adjustments, a model was obtained that is several times superior to the models built by the Kohonen Maps and Neural Network methods, as the error of our model was only 1.19%. The following methods were chosen to describe the results: conjugacy table, attribute significance table and decision tree, **Figures 2** and **3**, **Table 2**.

Figure 2. Decisive rules for the distribution of attributes by the decision tree.

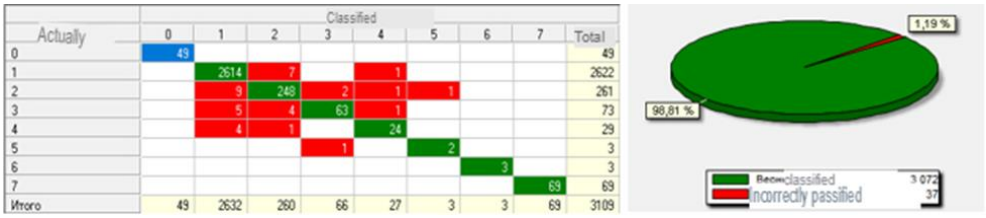

Figure 3. Conjugacy table and description of the error of the constructed model.

Table 2. Significance of attributes, the output parameter SQ7 is suitability for growing plants in the soil (workability).

Attribute	**Significance, %**
Indicator of the state of the roots	54.066
Indicator of the availability of nutrients in the soil	34.554
Slope 1	3.052
Slope 3	2.457
Indicator of oxygen availability for roots in the soil	1.751
Slope 4	1.238
Aspect	1.033
Assessment of land cultivation in the district	0.597
Assessment of the total area of cultivated land in the district	0.531
Slope 6	0.409
Aspect	0.177
Slope 8	0.134
Indicator of excess salts in the soil	0
Indicator of soil toxicity	0
Indicator of the ability to retain nutrients by the soil	0
Aspects	0
Slope 7	0
Assessment of crops in the district	0
Assessment of water bodies in the district	0
Indicator of the state of the roots	54.066
Indicator of the availability of nutrients in the soil	34.554
Slope 1	3.052
Slope 3	2.457
Indicator of oxygen availability for roots in the soil	1.751
Slope 4	1.238
Aspect	1.033

The attribute significance visualizer allows you to determine the significance of the output variables. Based on the table, it could be concluded that the following attributes are significant: root condition indicator, soil nutrient availability indicator and slope ($2\% \leq \text{slope} \leq 5\%$). Further, the obtained models can be applied to the test sets taking into account the possible error percentage.

Since the factors in the constructed model vary in terms of their practical applicability and overall monetary value, it was decided to separate the factors according to their characteristics and practical applicability if such a model were to be applied with other datasets to predict soil suitability for cultivation under drought conditions.

In **Table 2**, there are 19 factors that can be divided into several programmed groups to determine a cheaper and more accessible use of the model in expert systems. The missing factors were excluded in the correlation analysis step. A detailed explanation of each factor is given in the materials and methods section.

Group 1 (Indicator of the state of the roots, Indicator of the availability of nutrients in the soil, Indicator of excess salts in the soil, Indicator of soil toxicity, Indicator of the ability to retain nutrients by the soil), the factors in which refer to laboratory tests, such data are the most difficult to collect, requiring appropriate soil samples in each area.

The next group, number 2 (Slope 1, Slope 3, Slope 4, Aspect W, Slope 6, Aspect E, Slope 8, Slope 7), consists of data on the slopes and terrain aspects of the area where the soil is located, such data being publicly available and collected by various organizations for their own purposes.

The final group 3 (Assessment of land cultivation in the district, Assessment of the total area of cultivated land in the district, Assessment of crops in the district, Assessment of water bodies in the district) consists of land valuation data in the county, such data can also be found, for example, with the PLUS model, which provides support for a highly accurate study of the evolution of land use areas. The PLUS model is a new and improved CA (cellular automata) model based on the FLUS model. It combines a new strategy for analyzing land use expansion and a CA model based on multi-class random seeded areas. You can also find such data with a narrow appeal to an organization in a particular area where such data is collected for crop statistics etc.

In the first group there are such factors as indicator of root condition, indicator of nutrient availability in soil, indicator of oxygen availability for roots in soil, indicator of excess salts in soil, indicator of soil toxicity, indicator about the ability of soil to retain nutrients.

The decision tree was constructed considering the correlation analysis shown in **Table 1**. Its description is presented using visualizers: decision tree, contiguity table and attribute significance table, **Figures 4** and **5**, **Table 3**.

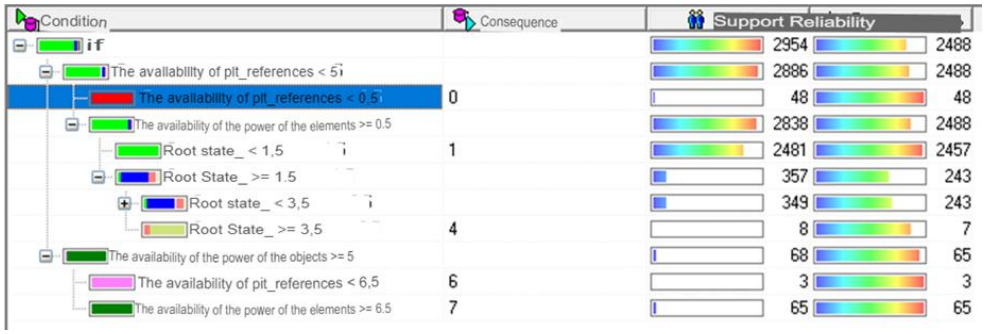

Figure 4. Decisive rules for the distribution of attributes by the decision tree of group No. 1.

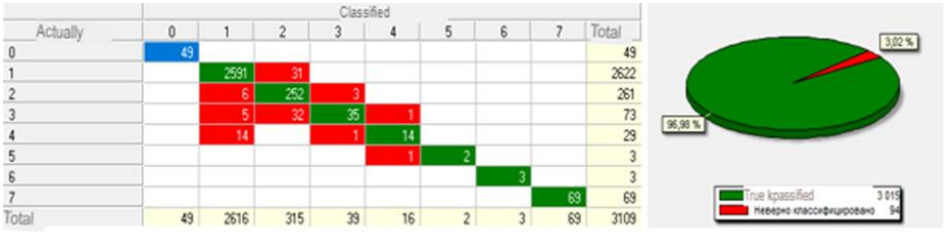

Figure 5. Conjugacy table and description of the error of the constructed model for group No. 1.

Table 3. Significance of attributes for group No. 1.

Attribute	**Significance, %**
Indicator of the state of the roots	59.1
Indicator of the availability of nutrients in the soil	37.9
Indicator of oxygen availability for roots in the soil	2.93
Indicator of soil toxicity	0.065
Indicator of excess salts in the soil	0
Indicator of the ability to retain nutrients by the soil	0

In the first group, there are factors such as root condition indicator, soil nutrient availability indicator, soil oxygen availability indicator for roots in soil, soil salt excess indicator, soil toxicity indicator, soil nutrient retention capacity indicator.

Based on the obtained data, the error of the constructed model was 3.02%, and the most significant factor in such a model is the root condition indicator.

In the second group there are such factors as slope 1, slope 3, slope 4, aspect, slope 6, slope 8, slope 7, aspects.

The decision tree is constructed considering the correlation analysis shown in **Table 1**. Its description is presented using visualizers: decision tree, conjugacy table and attribute significance table, **Figures 6** and **7**, **Table 4**.

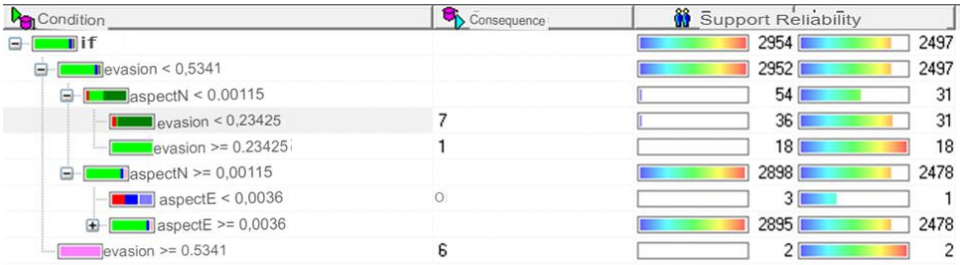

Figure 6. Decisive rules for the distribution of attributes by the decision tree of group No. 2.

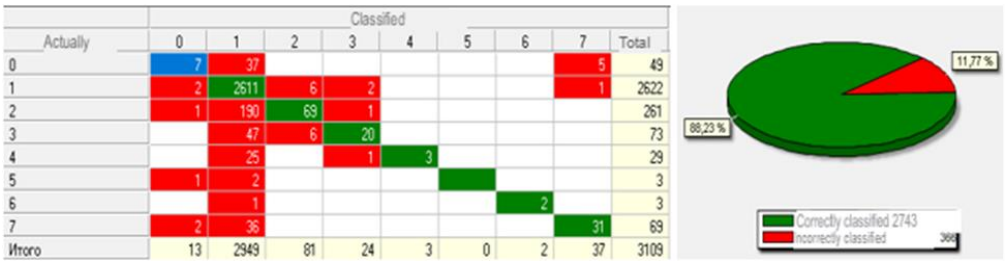

Figure 7. Conjugacy table and description of the error of the constructed model for group No. 2.

Table 4. Significance of attributes for group No. 2.

Attribute	**Significance, %**
Aspect	24.015
Slope 7	16.765
Slope 6	15.082
Aspects	14.263
Slope 1	10.329
Slope 8	9.369
Slope 3	7.335
Slope 4	2.843

Based on the obtained data, the error of the constructed model was 11.77% and the most significant factor in such model is aspects.

The third group includes such factors as assessment of cultivated land in the district, assessment of total cultivated land in the district, assessment of crops in the district, assessment of water bodies in the district.

The decision tree is constructed considering the correlation analysis given in **Table 1**. Its description is presented using visualizers: decision tree, contiguity table and attribute significance table, **Figures 8** and **9**, **Table 5**.

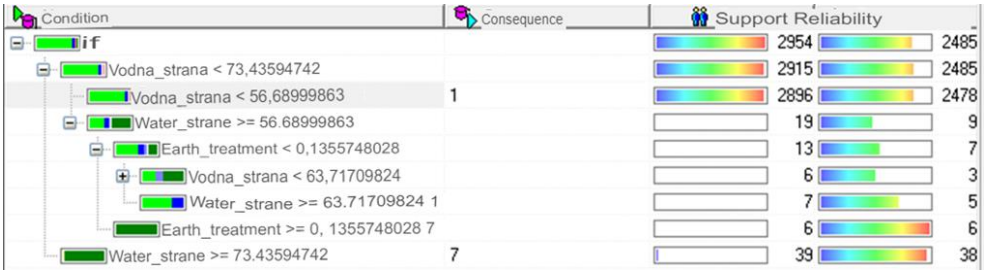

Figure 8. Decisive rules for the distribution of attributes by the decision tree of group No. 3.

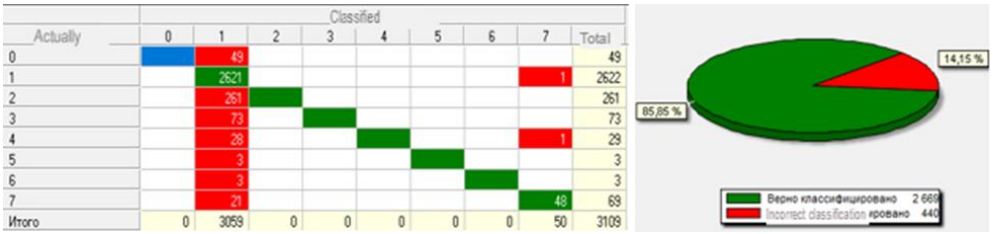

Figure 9. Conjugacy table and description of the error of the constructed model for group No. 3.

Table 5. Significance of attributes for group No. 3.

Attribute	**Significance, %**
Assessment of water bodies in the district	95.817
Assessment of land cultivation in the district	2.994
Assessment of crops in the district	1.190
Assessment of the total area of cultivated land in the district	0

Based on the obtained data, the error of the constructed model was 14.15%, and the most significant factor in such model is the assessment of water bodies in the district.

5. Discussion

In general, our model turned out to have an error of only 1.19%, but it is a very expensive model and requires huge resource costs for its operation, so during the experiment, 3 more models were proposed, consisting not of their 29 factors, but much less. The models were built based on the input parameters combined according to different practical characteristics. The details are summarized in **Table 6**.

Thus, model No. 1 has the lowest error of 3.02%. The most significant parameter is the root condition parameter, but it is the most expensive model as mentioned in the experimental part.

Model No. 2 has an error of 11.77%, the most significant parameter is aspect C. This model is much cheaper than the first model because data collection for this model is much cheaper and less labour intensive.

Model No. 3 has the largest error of all three models, namely 14.15%, the most significant parameter is the assessment of water bodies in the district, it is worth noting that this parameter can only be used in predicting land suitability under drought conditions, data collection using modern technology is quite easy.

Table 6. Correlation analysis.

Attribute	Model No. 1	Model No. 2	Model No. 3	Correlation analysis
	Significance, %	Significance, %	Significance, %	Correlation value
Indicator of the state of the roots	59.100	-	-	0.958
Indicator of soil toxicity	0.065	-	-	0.876
Indicator of excess salts in the soil	0	-	-	0.863
Indicator of oxygen availability for roots in the soil	2.943	-	-	0.763
Indicator of the ability to retain nutrients by the soil	0	-	-	0.715
Assessment of water bodies in the district	-	-	95,817	0.672
Indicator of the availability of nutrients in the soil	37.901	-	-	0.571
Total area of cultivated land in the district	-	-	0	−0.191
Assessment of crops in the district	-	-	1.190	−0.178
Slope 7	-	16.765	-	0.114
Slope 8	-	9.369	-	0.101
Slope 6	-	15.082	-	0.089
Aspect	-	-	-	-0.086
Slope 3	-	7.335	-	
Slope 4	-	2.843	-	−0.070
Slope 1	-	10.329	-	−0.065
Assessment of land cultivation in the district	-	-	2.994	−0.063
Aspect	-	24.015	-	−0.060
Aspects	-	14.263	-	−0.057
Aspects	-	-	-	−0.048
Slope 5	-	-	-	0.039
Assessment of infertility of the district	-	-	-	0.038
Average altitude above sea level	-	-	-	0.035
Width	-	-	-	0.029
Assessment of the presence of forests in the district	-	-	-	−0.019
Assessment of the presence of vegetation in the district	-	-	-	−0.012
Assessment of urbanization in the district	-	-	-	0.011
Longitude	-	-	-	−0.001

However, despite some success of the developed models, it is important to recognize certain limitations. First, the high resource costs associated with the model create practical difficulties for widespread implementation. In addition, the complexity of collecting data on certain features, such as laboratory tests of soil properties, increases operating costs and feasibility problems.

In light of these limitations, future research endeavors should focus on optimizing model performance while minimizing resource requirements. Strategies such as feature selection and data augmentation could help streamline the model development process and enhance cost-effectiveness. Additionally, exploring alternative modeling approaches and integrating advanced technologies could offer new avenues for improving soil suitability assessment under drought conditions.

6. Conclusion

The In this study, we aimed to identify the factors influencing soil suitability for cultivation under drought conditions and propose more cost-effective and accessible models for predicting soil suitability in expert systems. Our findings shed light on the significance of different variables in predicting soil suitability and offer insights into the practical application of these models in agricultural decision-making.

Our analysis revealed several key findings:

1) Model selection: Our results indicate that decision tree models outperform other methods in forecasting land suitability under drought conditions. Furthermore, we observed that subsequent models built upon the initial decision tree model exhibit varying levels of accuracy and resource requirements. Model No. 1, albeit the most accurate, demands substantial resources for data collection, while Model No. 3 presents a more practical approach with slightly higher error rates but minimal resource expenditure.

2) Variable significance: Laboratory soil tests, including root condition, nutrient availability, and soil toxicity, emerged as critical factors influencing soil suitability. Despite the challenges associated with data collection for these variables, their inclusion is vital for accurate forecasting.

3) Accessibility of terrain data: Terrain features and landscape characteristics, such as slope and aspect, proved to be readily available from existing datasets maintained by various organizations. Leveraging these publicly accessible data sets can significantly reduce the resource intensity of soil suitability assessment models.

4) Utility of land use assessment data: Data on land use assessment, including land area, crops, and water bodies, offer valuable insights into soil suitability. Models like PLUS provide a robust framework for analyzing land use evolution and can enhance the accuracy of soil suitability predictions.

In conclusion, our study underscores the importance of considering both the accuracy and resource requirements of soil suitability assessment models. By strategically selecting variables and leveraging publicly available data sets, researchers and policymakers can develop more cost-effective and accessible models for predicting soil suitability, thereby facilitating sustainable agricultural practices and enhancing food security.

Furthermore, our findings contribute to the ongoing discourse on climate-resilient agriculture and underscore the need for comprehensive development programs to support the adoption of environmentally sustainable farming practices. Future research endeavors should focus on refining existing models, incorporating additional variables, and evaluating their performance across diverse geographic regions to ensure the scalability and applicability of soil suitability assessment frameworks.

In article (Rahman et al., 2021) underscores the critical importance of implementing climate-optimized soil treatments in agricultural practices to mitigate the adverse effects of climate change and promote sustainable agriculture. Similarly, our study emphasizes the need for cost-effective and accessible models for predicting soil suitability under drought conditions, aligning with the broader goal of enhancing agricultural sustainability.

In light of the identified research gaps and the evolving landscape of agricultural sustainability, future studies should explore innovative methodologies, leverage emerging technologies, and foster interdisciplinary collaborations to advance our understanding of soil suitability dynamics and inform evidence-based policymaking in agriculture. By integrating insights from diverse fields such as climatology, agronomy, and data science, researchers can develop holistic approaches to address the complex challenges facing modern agriculture and promote resilient and sustainable food systems.

Author contributions: Conceptualization, VK and AB; methodology, YT; software, KK; validation, KK, VK and YT; formal analysis, AG; investigation, KK; resources, AG; data curation, YT and AG; writing—original draft preparation, KK; writing—review and editing, KK; visualization, VK; supervision, VK; project administration, VK; funding acquisition, AB. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

Alaoui, A., Hallama, M., Bär, R., et al. (2022). A New Framework to Assess Sustainability of Soil Improving Cropping Systems in Europe. Land, 11(5), 729. https://doi.org/10.3390/land11050729

Albahar, M. (2023). A Survey on Deep Learning and Its Impact on Agriculture: Challenges and Opportunities. Agriculture, 13(3), 540. https://doi.org/10.3390/agriculture13030540

Andrews, S. S., Karlen, D. L., & Cambardella, C. A. (2004). The Soil Management Assessment Framework. Soil Science Society of America Journal, 68(6), 1945–1962. https://doi.org/10.2136/sssaj2004.1945

Aydın, Y., Işıkdağ, Ü., Bekdaş, G., et al. (2023). Use of Machine Learning Techniques in Soil Classification. Sustainability, 15(3), 2374. https://doi.org/10.3390/su15032374

Bashmur, K. A., Kolenchukov, O. A., Bukhtoyarov, V. V., et al. (2022). Biofuel technologies and petroleum industry: Synergy of sustainable development for the Eastern Siberian Arctic. Sustainability, 14(20), 13083.

Bondarenko, V. L., Ilyinskaya, D. N., Kazakova, A. A., et al. (2022). Digitalization of Determining the Basic Properties of Hydrogen. Chemical and Petroleum Engineering, 58(1–2), 47–51. https://doi.org/10.1007/s10556-022-01053-9

Bukhtoyarov, V. V., Nekrasov, I. S., Tynchenko, V. S., et al. (2022). Application of machine learning algorithms for refining processes in the framework of intelligent automation. SOCAR Proceedings, SI1. https://doi.org/10.5510/ogp2022si100665

Bystrzanowska, M., & Tobiszewski, M. (2018). How can analysts use multicriteria decision analysis? TrAC Trends in Analytical Chemistry, 105, 98–105. https://doi.org/10.1016/j.trac.2018.05.003

Cabała, P. (2010). Using the Analytic Hierarchy Process in Evaluating Decision Alternatives. Operations Research and Decisions, 1: 1–23.

Cécillon, L., Barthès, B. G., Gomez, C., et al. (2009). Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). European Journal of Soil Science, 60(5), 770–784. https://doi.org/10.1111/j.1365-2389.2009.01178.x

Chen, S., & Ding, Y. (2023). A Machine Learning Approach to Predicting Academic Performance in Pennsylvania's Schools. Social Sciences, 12(3), 118. https://doi.org/10.3390/socsci12030118

Cravero, A., Pardo, S., Sepúlveda, S., et al. (2022). Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. Agronomy, 12(3), 748. https://doi.org/10.3390/agronomy12030748

Danaei Mehr, H., & Polat, H. (2021). Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. Health and Technology, 12(1), 137–150. https://doi.org/10.1007/s12553-021-00613-y

Doran, J. W., Coleman, D. C., Bezdicek, D. F., & Stewart, B. A. (1994). Defining Soil Quality for a Sustainable Environment. In: Soil Science Society of America and American Society of Agronomy. SSSA Special Publications. https://doi.org/10.2136/sssaspecpub35

Fischer, G., Nachtergaele, F., Prieler, S., et al. (2008). Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008).

IIASA, Laxenburg, Austria and FAO, Rome, Italy.

Abraham, A., Gandhi, N., Hanne, T., et al. (2022). Intelligent Systems Design and Applications. In: Lecture Notes in Networks and Systems. Springer International Publishing. https://doi.org/10.1007/978-3-030-96308-8

Karlen, D. L., Mausbach, M. J., Doran, J. W., et al. (1997). Soil Quality: A Concept, Definition, and Framework for Evaluation (A Guest Editorial). Soil Science Society of America Journal, 61(1), 4–10. https://doi.org/10.2136/sssaj1997.03615995006100010001x

Kieliszek, M., Kot, A. M., Bzducha-Wróbel, A., et al. (2017). Biotechnological use of Candida yeasts in the food industry: A review. Fungal Biology Reviews, 31(4), 185–198. https://doi.org/10.1016/j.fbr.2017.06.001

Kolenchukov, O. A., Bashmur, K. A., Bukhtoyarov, V. V., et al. (2022). The experimental research of n-butane pyrolysis using an agitator. SOCAR Proceedings, SI1. https://doi.org/10.5510/ogp2022si100685

Kukartsev, V. V., Zamolotsky, S. A., & Khramkov, V. V. (2023). Identification of factors influencing heart failure mortality using machine learning methods. News of the Tula State University. Sciences of Earth, 3(1), 101–111. https://doi.org/10.46689/2218-5194-2023-3-1-101-111

Malek, Ž., Verburg, P. H., R Geijzendorffer, I., et al. (2018). Global change effects on land management in the Mediterranean region. Global Environmental Change, 50, 238–254. https://doi.org/10.1016/j.gloenvcha.2018.04.007

Martyushev, N. V., Bublik, D. A., Kukartsev, V. V., et al. (2023). Provision of Rational Parameters for the Turning Mode of Small-Sized Parts Made of the 29 NK Alloy and Beryllium Bronze for Subsequent Thermal Pulse Deburring. Materials, 16(9), 3490. https://doi.org/10.3390/ma16093490

Masich, I. S., Tynchenko, V. S., Nelyub, V. A., et al. (2022). Prediction of Critical Filling of a Storage Area Network by Machine Learning Methods. Electronics, 11(24), 4150. https://doi.org/10.3390/electronics11244150

Meier, R. K. (2018). Polycystic Ovary Syndrome. Nursing Clinics of North America, 53(3), 407–420. https://doi.org/10.1016/j.cnur.2018.04.008

Mohan, P., & Patil, K. (2018). Deep Learning Based Weighted SOM to Forecast Weather and Crop Prediction for Agriculture Application. International Journal of Intelligent Engineering and Systems, 11(4), 167–176. https://doi.org/10.22266/ijies2018.0831.17

Pandya, A., Odunsi, O., Liu, C., et al. (2020). Adaptive and Efficient Streaming Time Series Forecasting with Lambda Architecture and Spark. In: Proceedings of the 2020 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/bigdata50022.2020.9377947

Panfilova, E. V., Ibragimov, A. R., & Shramko, D. Y. (2022). The practice of using artificial intelligence algorithms to adjust the parameters of nanostructures study by the tapping mode of atomic force microscopy. Modeling in Engineering, 2020. https://doi.org/10.1063/5.0075106

Priyam, A., Abhijeeta, G. R., Rathee, A., et al. (2013). Comparative analysis of decision tree classification algorithms. International Journal of current engineering and technology, 3(2), 334–337.

Prokhorenkova, L., Gusev, G., Vorobev, A, et al. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems. NeurIPS Proceedings.

Rahman, M. M., Aravindakshan, S., Hoque, M. A., et al. (2021). Conservation tillage (CT) for climate-smart sustainable intensification: Assessing the impact of CT on soil organic carbon accumulation, greenhouse gas emission and water footprint of wheat cultivation in Bangladesh. Environmental and Sustainability Indicators, 10, 100106. https://doi.org/10.1016/j.indic.2021.100106

Robertson, P. K. (2016). Cone penetration test (CPT)-based soil behaviour type (SBT) classification system—an update. Canadian Geotechnical Journal, 53(12), 1910–1927. https://doi.org/10.1139/cgj-2016-0044

Schwilch, G., Bestelmeyer, B., Bunning, S., et al. (2010). Experiences in monitoring and assessment of sustainable land management. Land Degradation & Development, 22(2), 214–225. https://doi.org/10.1002/ldr.1040

Shi, H., Wen, Z., Paull, D., et al. (2016). Distribution of Natural and Planted Forests in the Yanhe River Catchment: Have We Planted Trees on the Right Sites? Forests, 7(12), 258. https://doi.org/10.3390/f7110258

Shutaleva, A., Martyushev, N., Nikonova, Z., et al. (2023). Sustainability of Inclusive Education in Schools and Higher Education: Teachers and Students with Special Educational Needs. Sustainability, 15(4), 3011. https://doi.org/10.3390/su15043011

Silva, I. S., Ferreira, C. N., Costa, L. B. X., et al. (2022). Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models. Journal of Endocrinological Investigation, 45(3), 497–505.

Sitokonstantinou, V., Drivas, T, Koukos, A., et al. (2020). Scalable distributed random forest classification for paddy rice mapping. In: Proceedings of the 40th Asian Conference on Remote Sensing (ACRS 2019); 14–18 October 2019; Daejeon, Korea.

Sokolov, A. A., Orlova, L G., Bashmur, K. A., et al. (2023). Ensuring uninterrupted power supply to mining enterprises by developing virtual models of different operation modes of transformer substations. MIAB. Mining Inf. Anal. Bull., 2023, (11-1), 278–291.

Tiwari, S., Kane, L., Koundal, D., et al. (2022). SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning. Expert Systems with Applications, 203, 117592. https://doi.org/10.1016/j.eswa.2022.117592

Topp G. C., Reynolds W. D., Cook F. J., et al. (1997). Physical characteristics of soil quality: Achievements in the field of soil science. Elsevier, 25, 21–58. https://doi.org/10.1016/S0166-2481(97)80029-3

Trontelj, M. L. J., & Chambers, O. (2021). Machine Learning Strategy for Soil Nutrients Prediction Using Spectroscopic Method. Sensors, 21(12), 4208. https://doi.org/10.3390/s21124208

Vlasov, A. I., Artemiev, B. V., Selivanov, K. V., et al. (2022). Predictive Control Algorithm for A Variable Load Hybrid Power System on the Basis of Power Output Forecast. International Journal of Energy Economics and Policy, 12(3), 1–7. https://doi.org/10.32479/ijeep.12912