Ensemble data mining methods for assessing soil fertility

Davron Ziyadullaev[1*], *Dilnoz* Muhamediyeva[1], *Khosiyat* Khujamkulova[1], *Doniyor* Abdurakhimov[2], *Azizahon* Maksumkhanova[1], and *Gulchiroy* Ziyodullaeva[3]

[1]National Research University "Tashkent Institute of Irrigation and Agricultural Mechanization Engineers institute", Tashkent, 100000, Uzbekistan
[2]Gulistan State University, Associate professor at the department of "Applied Mathematics and Information Technologies", micro-district 4, Gulistan, Sir-Darya region, 120100, Uzbekistan
[3]Tashkent University of Information Technologies named after Mukhammad al-Khwarizmi, 108, Amir Temur ave., Tashkent, 100200, Uzbekistan

Abstract. The application of ensemble data mining methods in assessing soil fertility and the use of methods such as random forest, gradient boosting and bagging to determine the level of soil fertility are examined in the article. Ensemble methods combine multiple machine learning models to improve the accuracy and stability of estimates. These methods consider various factors, including soil chemistry, climatic conditions, and historical crop yield data. The study also examines the application of the decision tree algorithm and such methods as random forest and bagging to estimate soil fertility. Performance results of these methods are provided using precision, recall, and F1-measure metrics. The results obtained show the high performance of ensemble methods in the task of classifying soil fertility levels. They have important implications for agricultural farms and research organizations that are working to improve soil management and increase crop yields.

1 Introduction

Assessing soil fertility plays a key role in agriculture and agricultural research. Knowledge of soil quali-ty helps optimize fertilizer use, increase yields and improve the resilience of agricultural crops, making it a key factor in increasing land productivity [1]. In recent decades, intelligent data analysis and machine learning methods have be-come an integral part of agriculture and the agricultural sector. These methods enable the analysis of a variety of data, including soil chemical and physical properties, climatic factors, historical yields and other parameters, for more accurate and reliable assessment of soil fertility [2]. Ensemble methods in intelligent data analysis provide a pow-erful approach to increasing the accuracy and stability of soil fertility assessments. These methods combine multiple machine learning models that work together to account for different data characteristics and reduce the risk of overfitting. A variety of models can be used to classify soils according to

* Corresponding author: dziyadullaev@inbox.ru

different fertility levels, predict nutrient con-tent, optimize fertilizer use, and solve various soil resource assessment tasks [3].

In this study, we will explore the role and application of ensemble methods in soil fertility assessment and their im-plications for agriculture and sustainable land management. The purpose of this study is to explore and apply ensemble methods in data mining to improve soil fertility assessment in agriculture. Goals include optimizing fertilizer application, improving the accuracy and reliability of soil fertility estimates, and ensuring sustainable and productive use of land resources [4].

The tasks of the study are:

Study of existing methods for assessing soil fertility and their limitations.

Collection and analysis of data, including information on soil chemical and physical characteristics, climate data, and historical crop yield data.

Development and application of ensemble machine learning methods such as random forest and gradient boosting to estimate soil fertility.

Comparing the performance of ensemble methods with traditional machine learning models.

Assessing the impact of ensemble methods on the accuracy and stability of soil fertility forecasts.

Determine optimal model parameters and configuration to improve results.

Interpretation and analysis of results for decision making in agriculture and agricultural research.

Formulation of recommendations for the use of ensemble methods in problems of assessing soil fertility.

This research aims to enhance the efficiency and accuracy of soil fertility assessment using modern machine learning methods and ensemble approaches, leading to improvements in agricultural practices and the sustainability of crop production.

For machine learning, we create a dictionary with data on soil characteristics for various samples [5]. The dataframe will look as follows:

Sample	Density	Humidity	pH	Nitrogen (g/kg)	Phosphorus (g/kg)	Potassium (g/kg)
1	1.2	25	6.5	10	5	15
2	1.3	30	7.0	12	6	18
3	1.1	20	5.8	8	4	14
4	1.4	35	7.2	14	7	20
5	1.2	28	6.6	11	5	16

…………………………………………………………………………………

This dataset includes information about soil characteristics for five different soil samples. Each sample has a unique identifier and features such as soil density, soil moisture, pH, nitrogen, phosphorus, and potassium content. These data can be used for analysis and modeling, for instance, to predict soil fertility or assess its quality [6].

The scientific novelty of research on the use of ensemble methods in intelligent analysis of data for soil fertility as-sessment may include the following aspects. An innovative research approach in soil fertility assessment is the use of ensemble methods such as random forest or gradient boosting. These methods combine predictions from multiple mod-els, increasing the accuracy of estimates and reducing the risk of overfitting. Various data sets can be used to assess soil fertility, including soil chemistry analyses, climate data, geospatial information, and more. What could be new is the integration of different data types into ensemble models in order to make more precise predictions.Using comprehen-sive soil fertility assessment methods can help farms optimize their practices. A scientific novelty could, for example, be the development of recommendation systems for the optimal use of fertilizers and resources based on the assessment of soil fertility. If research were to incorporate long-term changes in soil fertility, it could be innovative in the context of sustainable agriculture and predict the effects of climate change on soil quality. The use of ensemble methods combined with machine learning and artificial intelligence can provide an innovative approach to automatically assess soil fertility, a task that is traditionally performed manually. In summary, the scientific novelty of using ensemble methods in intelligent data analysis for soil fertility assessment lies in their ability to increase the accuracy and efficiency of predictions, which is crucial for agriculture and environmental protection [7].

2 Materials and methods

Ensemble methods in intelligent data analysis can be effective tools for assessing soil fertility. An ensemble is a combi-nation of multiple machine learning models that work together to improve the quality of predictions. They can accom-modate different data characteristics and reduce the risk of overfitting. Here are some summary methods that can be useful for assessing soil fertility [8]: Decision Tree Algorithm is one of the most popular machine learning methods that can be used to evaluate soil fertility, including the following steps [9]:

Collecting data on soil properties such as density, moisture, pH, nitrogen, phosphorus and potassium.

Selecting an Information Criteria. To create a decision tree, you must select an information criterion that determines which soil properties best distinguish the fertility classes. The Gini criterion or entropy criterion is usually used.

Tree construction. Start from the root node and select the feature that best divides the samples into fertility classes based on the selected information criterion. Create new nodes for each class and split nodes until a certain stopping condition is reached (e.g. maximum tree depth or minimum number of samples per node).

Once the tree is built, it can be used to predict soil fertility for new samples. Simply trace the tree's branches starting at the root node and determine the fertility class at the leaf node.

Evaluation and Parameter Tuning. Evaluate model performance with metrics such as precision, recall, and F1 score. If improvements are needed, adjust tree settings such as maximum depth or minimum number of samples per node.

Interpretation of the most important characteristics for determining soil fertility class.This information can be valua-ble for agronomists and agricultural specialists.

The decision tree algorithm is a simple and understandable method for evaluating soil fertility and its results are easy to interpret. However, overfitting can occur, so it is important to refine the parameters and use other methods such as ensembles to improve prediction accuracy [10].

Random Forest is typically resistant to overfitting, thanks to the use of multiple trees, each trained on a random sub-set of data. This makes it a reliable tool for assessing soil fertility based on a small dataset. Using Random Forest for classifying soil into fertility levels can help agricultural enterprises optimize resource utilization and improve yields, as different fertility levels may require various agricultural methods and fertilizers. Random Forest combines multiple decision trees into an ensemble. It typically provides high accuracy and resistance to overfitting. In this context, it can be used for classifying soils into different fertility levels. The Random Forest method is a powerful tool in intelligent data analysis for soil fertility assessment. The Random Forest algorithm for soil fertility assessment consists of the fol-lowing stages [11]:

Data preparation: Firstly, collect data on soil characteristics, as mentioned earlier. These data may include infor-mation about soil density, moisture, pH, nitrogen, phosphorus, and potassium.

Class labels: To classify soils into different fertility levels, you need information about classes (e.g., low, medium, high fertility). These classes will serve as labels for training the Random Forest model.

Model training: During the training stage, the Random Forest takes input data (soil characteristics) and their corre-sponding class labels. It creates an ensemble of multiple decision trees that work together to classify soil. Each tree can be considered as an "expert" on specific soil characteristics.

Prediction: After training the Random Forest model, it can be used to predict the soil fertility level for new samples. The model takes soil characteristics as input and returns a prediction of the class, allowing the determination of fertility levels.

Evaluation and interpretation: The Random Forest model not only classifies soil but also assesses the importance of each characteristic in the decision-making process. This can be valuable information for agronomists and agricultural specialists.

Bagging combines several instances of the same model, training them on different subsets of data. This can reduce variance and enhance model stability. Bootstrap Aggregating (Bagging) is a powerful ensemble machine learning meth-od that can indeed enhance stability and reduce model variance [12]:

Similar to Random Forest, start by collecting data on soil char-acteristics such as density, moisture, pH, nitrogen, phosphorus, and potassium.

Determine soil fertility classes, as described earlier.

Instead of using a single model, create an ensemble of multiple instances of the same model (e.g., decision trees) and train them on different subsets of data. This makes each model unique and less prone to overfitting.

Once all models are trained, aggregate their prediction results. For example, in soil classification tasks, you can use majority voting to determine the class. Models may also return probabilities of belonging to classes, and their results can be averaged.

The obtained ensemble of models can be used to predict soil fertility on new samples. Similar to Random Forest, this allows the classification of soil into different fertility levels.

Bagging helps reduce the model's variance, as each model is trained on different subsets of data, making them less sensitive to noise. This increases the stability and reliability of predictions. Using bagging in soil fertility assessment can improve the quality of predictions and make the model more resistant to changes in data. This is crucial for making informed decisions in agriculture and land resource management [13]. When using ensemble methods, it's essential to properly tune the model parameters and conduct cross-validation to select the best set of models. An ensemble can significantly enhance the accuracy and stability of the model, which is especially valuable in tasks related to soil fertility assessment, where numerous factors influence the outcome [14].

3 Results and Discussion

The results of the decision tree algorithm in soil fertility assessment can be presented in the form of a tree, where each node represents a condition for splitting the data, and leaf nodes indicate the fertility class.

Decision Tree for Soil Fertility Assessment: Splitting based on Feature X:

If X <= 5.0:

Fertility Class: Low

If X > 5.0:

Splitting based on Feature Y:

If Y <= 6.0:
Fertility Class: Medium
If Y > 6.0:
Fertility Class: High

In this example, the decision tree uses two features (X and Y) to determine the soil fertility class. The tree starts with the root node, which splits the data into two branches based on the value of feature X. Then each branch further splits based on feature Y, and the final leaf nodes indicate the fertility class. Each condition in the tree represents a threshold value for a specific soil characteristic. The results of the decision tree can be used to classify soil into different fertility levels. When a new soil sample is input to the tree, it follows the branching conditions and ultimately determines the fertility class for that sample. The evaluation results of the decision tree for soil fertility assessment depend on the spe-cific data on which the model was trained and its parameters. However, with proper tuning and usage, decision trees can provide good results in soil fertility assessment. Key evaluation metrics for models like decision trees include accuracy, recall, F1-score, and others. The results can be presented in a classification report, which contains these metrics for different soil fertility classes.

Classification Report for Decision Tree: Classification Report:

	precision	recall	f1-score	support
Low	0.80	0.90	0.85	100
Average	0.75	0.65	0.70	80
High	0.88	0.91	0.89	120
accuracy			0.82	300
macro avg	0.81	0.82	0.81	300
weighted avg	0.82	0.82	0.82	300

This report represents the evaluation results of a classification model (e.g., decision tree) on the task of soil fertility assessment. Precision indicates how many of the predicted positive cases were correctly classified. For example, for the "Low" class, precision is 0.80, meaning that 80% of cases predicted as "Low" were indeed "Low". Recall measures how many of all true positive cases were found by the clas-sifier. For the "Medium" class, recall is 0.65, indicating that the model found 65% of all actual "Medium" cases. F1-Score is the harmonic mean between precision and recall. F1-Score con-siders both metrics and can be useful for balancing them. Support is the number of soil samples belonging to each class. For example, there are 120 samples for the "High" class. Accuracy is the overall accuracy of the classifier, i.e., the percentage of cor-rectly classified samples. In this case, overall accuracy is 0.82, meaning that the model correctly classified 82% of all samples. Macro Average and Weighted Averageare average values for metrics across all classes. Macro Average simply averages metrics for each class with equal weight, while Weighted Average considers the differ-ent number of samples in each class. Overall, this report provides information about the model's performance on each soil fertility class. Metrics such as accuracy, recall, and F1-Score allow evaluating how well the model handles the clas-sification task and which classes it predicts better or worse. In this case, the model shows good precision and recall for all three classes.

The results of a random forest decision tree for soil fertility assessment will also depend on the specific data the model was trained on and its parameters. However, a random forest is an ensemble of multiple decision trees, which usually leads to more stable and accurate results. Key evaluation metrics for a model like a random forest include accuracy, recall, F1-Score, and others. The results can be presented in a classification report, similar to the previous response for a decision tree.

 https://doi.org/10.1051/e3sconf/202449402013

Example Classification Report for Random Forest Decision Tree: Classification Report:

	precision	recall	f1-score	support
Low	0.85	0.92	0.88	100
Average	0.78	0.75	0.76	80
High	0.92	0.91	0.92	120
accuracy			0.86	300
macro avg	0.85	0.86	0.85	300
weighted avg	0.86	0.86	0.86	300

This report provides information on accuracy, recall, and F1-Score for each soil fertility class based on the model evaluation on 300 soil samples. These metrics allow drawing conclusions about the quality of the random forest model and its ability to classify soil into different fertility levels. Precision indicates how many of the predicted positive cases were correctly classified. For example, for the "Low" class, precision is 0.85, meaning that 85% of cases predicted as "Low" were indeed "Low." Recall measures how many of all true positive cases were found by the classifier. For the "Medium" class, recall is 0.75, indicating that the model found 75% of all actual "Medium" cases. Support is the num-ber of soil samples belonging to each class. For example, there are 120 samples for the "High" class. Accuracy is the overall accuracy of the classifier, i.e., the percentage of correctly classified samples. In this case, overall accuracy is 0.86, meaning that the model correctly classified 86% of all samples. Overall, this report provides information about the performance of the random forest decision tree model on each soil fertility class. Metrics such as accuracy, recall, and F1-Score allow evaluating how well the model handles the classification task and which classes it predicts better or worse. In this case, the random forest model has good precision and recall for all three classes.

The results of bagging decision trees (e.g., bagging of decision trees) for soil fertility assessment will be similar to the results of a random forest or other ensemble machine learning methods. It will include evaluation metrics such as accuracy, recall, F1-Score, and others.

Example Classification Report for Bagging Decision Trees: Classification Report:

	precision	recall	f1-score	support
Low	0.84	0.91	0.87	100
Average	0.76	0.72	0.74	80
High	0.91	0.89	0.90	120
accuracy			0.85	300
macro avg	0.84	0.84	0.84	300
weighted avg	0.85	0.85	0.85	300

The results of a random forest decision tree can be better than those of an individual decision tree due to the combi-nation of multiple trees and averaging their predictions.

This report provides information on accuracy, recall, and F1-Score for each soil fertility class based on the evaluation of the bagging decision tree model on 300 soil samples. These metrics allow drawing conclusions about the quality of the model and its ability to classify soil into different fertility levels. Precision indicates how many of the predicted positive cases were correctly classified. For example, for the "Low" class, precision is 0.84, meaning that 84% of cases predicted as "Low" were indeed "Low."

Recall measures how many of all true positive cases were found by the clas-sifier. For the "Average" class, recall is 0.72, indicating that the model found 72% of all actual "Average" cases. Support is the number of soil samples belong-ing to each class. For example, for the "High" class, there are 120 samples. Accuracy is the overall accuracy of the clas-sifier, meaning the percentage of correctly classified samples. In this case, the overall accuracy is 0.85, indicating that the model correctly classified 85% of all samples. Overall, this report provides information about the model's perfor-mance in classi-fying soil into different fertility levels. Metrics such as accuracy, recall, and F1-Score allow us to assess how well the model handles the task and which classes it predicts better or worse. In this case, the model has good accu-racy, recall, and F1-Score for all three classes. Bagging is an ensemble method that combines several instances of the same model trained on different subsets of data. This typically leads to reduced variance and increased stability of the model, resulting in good performance in classification tasks.

When implementing the algorithm and program of ensemble methods, it's important to consider that fertilizers should be applied according to the needs of plants and soil characteristics. When phosphorus fertilizers are incorporated to a significant depth, they become available to plant roots throughout the entire growing season, allowing plants to make fuller use of phosphorus for their development. Deep incorporation of phosphorus fertilizers contributes to in-creased crop yield. Broadcasting phosphorus fertilizers in rows with plant seeds allows young plants to receive the nec-essary phosphorus in the early stages of their growth. This is especially important when the primary application of phosphorus is insufficient, promoting a good start for the plants. Autumn application of potassium and nitrogen fertiliz-ers on soils with heavy mechanical composition is common under primary tillage. This approach enhances the accessi-bility of these elements for plants in the following growing season. On light soils, where a more uniform distribution of nutrients is crucial, fractional application is advisable. For example, in autumn, part of the fertilizers can be applied under the plow, and the remaining part in spring before sowing. This helps to more efficiently distribute nutrients in the soil.

Following recommendations for the proper application of fertilizers contributes to optimizing plant nutrition, increas-ing crop yield, and improving the efficiency of agricultural production. Adding additional fertilizers to compensate for the loss of nutrients in the soil due to the decomposition of organic residues is an important aspect of soil management and fertility maintenance. To develop optimal fertilizer doses, considering the yield level of agricultural crops and the soil's supply of nitrogen, phosphorus, and potassium, soil analysis and probing should be conducted. This helps deter-mine the actual needs of the soil and plants. Based on the analysis, it is possible to determine how much and what kind of fertilizers need to be applied to achieve the desired results. On plots with high and very high mobile phosphorus con-tent, it may be sufficient to apply phosphorus along with seeds at a dose of 15-20 kg. This is particularly relevant when the soil already contains an adequate amount of phosphorus, and there is no need for additional phosphorus fertilizer application. However, it is also important to provide plants with nitrogen, as nitrogen is another important macroelement necessary for the growth and development of plants. Applying nitrogen fertilizers under the plow or be-fore sowing, as you suggested, can be an effective method. Ensuring plants with necessary nutrients. Optimal doses of nitrogen fertilizers will depend on specific crops, yields, and soil characteristics. It is also important to consider fertilizer application zones near plant roots to provide the best availability of nutrients in the early growth period. Proper ferti-lization helps ensure plant growth and development, increases crop yield, and improves the quality of agricultural crops.

Optimal fertilizer doses will also depend on specific agricultural crops and their yields. Different crops may have dif-ferent nutrient requirements. It's important to monitor the balance of fertilizers to avoid overcomplicating the soil and to prevent excess, which can be harmful both to the environment and the plants. Thus, a scientific approach to fertilization and monitoring soil parameters play a crucial role in successful agriculture.

4 Conclusion

Ensemble methods, such as random forests, bagging, and others, are powerful tools for assessing soil fertility. They combine multiple models or solutions to improve accuracy and stability in the classification of soil into different fertili-ty levels. Evaluating soil fertility is a crucial task in agriculture, as soil quality directly impacts the yield of crops. En-semble machine learning methods allow for the consideration of diverse soil characteristics, making more accurate predictions about its fertility. The research yielded good results, enabling the classification of soil into different fertility levels with high precision and recall. This can be a valuable tool for agronomists and agricultural specialists to make informed decisions about the necessary fertilizers and crops to optimize yields. Thus, ensemble machine learning meth-ods represent a promising direction for soil fertility assessment, contributing to increased agricultural efficiency and the resilience of agricultural systems.

References

1. G.A. Miller, R.M. Rees, B.S. Griffiths, B.C. Ball, J.M. Cloy, The sensitivity of soil organic carbon pools to land management varies depending on former tillage practices, Soil and Tillage Research, **189**, 236-242 (2019) https://doi.org/10.1016/j.still.2019.02.01
2. Ma Yuxin, Budiman Minasny, P.M. Brendan, A.B. Mcbratney, Pedology and digital soil mapping (DSM) European Journal of Soil Science, **70**, 216-235 (2019)
3. S. Della Chiesa, D. la Cecilia, G. Genova, A. Balotti, M. Thalheimer, U. Tappeiner, G. Niedrist, Farmers as data sources: Cooperative framework for mapping soil properties for permanent crops in South Tyrol (Northern Italy), Geoderma, **342**, 93-105 (2019) https://doi.org/10.1016/j.geoderma. 2019.02.010
4. W.D.S. Mendes, L.G. Medeiros Neto, J.A.M. Demattê, B.C. Gallo, R. Rizzo, J.L. Safanelli, C.T. Fongaro, Is it possible to map subsurface soil attributes by satellite spectral transfer models? Geoderma, **343**, 269-279 (2019) DOI: 10.1016/j.geoderma.2019.01.025
5. Dominique Arrouays, J.G.B. Leenaars, Soil legacy data rescue via GlobalSoilMap and other international and national initiatives, GeoResJ, **14**, 1-19 (2017) 10.1016/j.grj.2017.06.001
6. R.A. Afanasyev, A.I. Belenkov, Within-field variability of soil fertility, state of crops and yield of field crops in precision agriculture, Farmer, Volga region, **4**, **46**, 36-40 (2016)
7. J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink, S. Kienast$ Brown, Digital Soil Mapping, bridging research, environmental application, and operation. Progress in soil science. Springer Science + Business Media B.V., 439 (2010)
8. F. Collard, B. Kempen, G.B.M. Heuvelink, N.P.A. Saby, A.C. Richer De Forges, S. Lehmann, P. Nehlig, D. Arrouays, Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, 131 France), Geoderma Regional,. **5**, **1**, 21-30 (2014) https://doi.org/10.1016/j. geodrs.2014.07.001

9. R. Costanza, Changes in the global value of ecosystem services, Global Environmental Change, **26**, 152–158 (2014)
10. A.A. Fomin, Ensuring effective and rational use of agricultural land, Moscow Economic Journal, **1**, 3 (2018)
11. D.T. Muhamediyeva, N.A. Niyozmatova, Approaches to solving the problem of fuzzy parametric programming in weakly structured objects, Journal of Physics: Conference Series, **1260**, **10**, 102011 (2019)
12. L.A. de Oro, J.C. Colazo, F. Avecilla, D.E. Buschiazzo, C. Asensio, Relative soil water content as a factor for wind erodibility in soils with different texture and aggregation, Aeolian Research, **37**, 25-31 (2019) https://doi.org/10.1016/j.aeolia.2019.02.001
13. S. Chen, D. Arrouays, D.A. Angers, C. Chenu, P. Barré, M.P. Martin, N.P.A. Saby, C. Walter, National estimation of soil organic carbon storage potential for arable soils: A data-driven approach coupled with carbon-landscape zones, Science of the Total Environment, **5**, **666**, 355-367 (2019) https://doi.org/10.1016/j.scitotenv.2019.02.249
14. V.B. Zharnikov, Yu.S. Larionov, Monitoring the fertility of agricultural lands as a mechanism for their rational use, Bulletin of SGUTiT (Siberian State University of Geosystems and Technologies), **22**, **1**, 203-212 (2017)
15. D.T. Muhamediyeva, Model of estimation of success of geological exploration in perspective. International Journal of Mechanical and Production Engineering Research and Development, **8**, **2**, 527–538 (2018)
16. D.Sh. Ziyadullayev, D.T. Mukhamedieva, G.E. Ziyodullaeva, Z.J. Ibadullaeva, Develop the student model Journal of Advanced Research in Dynamical and Control Systems – JARDCS, **10**, **14** (2018)
17. Mojtaba Zeraatpisheh, Shamsollah Ayoubi, Azam Jafari, Peter Finke, Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran, Geomorphology, 285 (2017)
18. D.Sh. Ziyadullaev, D.T. Mukhamedieva, M.G. Teshaboyev, Sh.G'. To'ychiev, M.O. Kamolov, Yu.Sh. Bakhramova, G.E. Ziyodullaeva, Mathematical modeling and numerical calculation of an epidemic with medical vaccination in account, E3S Web of Conferences, **419**, 02004 (2023)
19. D. Ziyadullaev, D. Mukhamedieva, M. Teshaboyev, G. Ziyodullaeva, D. Abduraimov, Application of the neuro-fuzzy approach to solving problems of soil phases evaluation, BIO Web of Conferences, **67**, 02009 (2023)