Data mining for assessing soil fertility

Manzura Inoyatova¹, Davron Ziyadullaev*, Dilnoz Muhamediyeva¹, Sharofiddin Aynaqulov¹, and Sholpan Ziyaeva¹

¹National Research University "Tashkent Institute of Irrigation and Agricultural Mechanization Engineers institute", Tashkent, 100000, Uzbekistan

Abstract. The study is devoted to the use of data mining to assess soil fertility, which is a modern and effective tool in agriculture and ecology. The method includes integrated approaches to data collection, processing and analysis aimed at determining soil fertility, its composition and potential for successful agricultural use. Using a variety of machine learning techniques and statistical models, researchers can predict crop yields, optimize fertilization and soil management strategies, and identify environmental and soil health risks. In particular, the use of the regression method makes it possible to build models that predict the values of fertile soil parameters based on available data. Using machine learning techniques such as Bayes' theorem and support vector machines (SVM), researchers can effectively estimate soil fertility, predict soil characteristics, and optimize agricultural practices. The results of the study demonstrate the high performance of the models in soil sample classification tasks, highlighting their potential for improving soil resource management and increasing crop yields. Such machine learning techniques provide powerful tools for agricultural workers and researchers, facilitating more precise and sustainable agriculture, which is essential for food security and ecosystem resilience.

1 Introduction

In agroecology, sustainable land management, and agriculture, evaluating soil fertility is essential. Crop productivity and agriculture are directly impacted by the composition, quality, and capacity of the soil to supply vital nutrients to plants. As a result, the creation and implementation of cutting-edge techniques for determining and tracking soil fertility have become essential components of farming operations. [1]. Data analytics is becoming more and more important in the assessment of soil fertility in the modern day due to the integration of information technology and data analysis techniques. With this method, one can get around long-standing restrictions and open up new avenues for a more precise, thorough, and dynamic assessment of soil parameters [2]. The primary objectives are methodological considerations, practical implications, and an intellectual study of the data in soil fertility measurement. We will also look at contemporary techniques and technology
that make it possible to forecast and analyze soil conditions and to design farming and soil management plans that take shifting environmental and climatic factors into account [3].

There are two geographical zones in Uzbekistan:

The first geographical zone, which makes up 71.7% of the whole land, is a level, hot, arid region of the desert. This zone, which includes arid regions like the Karakum Desert, is defined by high temperatures and little precipitation. The second geographical zone, which makes up 28.3% of the country's total territory, is made up of mountains and foothills with a generally temperate and humid climate. This zone has a greater amount of precipitation, moderate temperatures, and a wide variety of plants. This region is home to mountain ranges including the Pamir and Tian Shan. These two geographic zones influence the variety of Uzbekistan's flora and fauna as well as the agricultural and industrial advancement of the nation. Assessing the fertility of the soil is a crucial step in figuring out whether it can sustain the growth of healthy plants. Soil fertility can be evaluated using a variety of techniques, such as physical, chemical, and biological elements [4].

Physical properties are an assessment of soil structure, texture (sand, clay and silt content) and density, as well as an assessment of soil moisture, since moisture plays a key role in fertility. Chemical characteristics are determining the level of acidity or alkalinity of the soil pH, analyzing the content of macro- and microelements, assessing the presence of essential nutrients such as nitrogen, phosphorus, potassium, magnesium, iron and others, determining the amount of organic matter in the soil that affects its ability to retain moisture and provide plant nutrition. Biological characteristics: Microbiological analyses: Study of soil-inhabiting microorganisms, such as bacteria and fungi. They play a vital role in the nutrient cycle and decomposition of organic material. Assessment of biological activity: This may include assessing soil respiration activity and enzyme activities. A soil fertility assessment is generally performed to identify measures that can be taken to improve fertility, such as adding fertilizers, incorporating organic substances or adjusting the pH value. For a more precise assessment, we recommend carrying out a soil analysis in the laboratory or consulting specialists in agriculture and agronomy [5].

Each soil sample has characteristics. The value describing soil density is usually expressed in g/cm³ or kg/m³. The percentage of soil moisture content determines how wet or dry the soil is. The acidity or alkalinity level of the soil, evaluated on a scale from 0 to 14, where 7 is considered neutral pH. Based on their pH values, soils can be divided into several groups according to their acidity. pH above 7. These soils can be moderately alkaline (pH 7-8) or strongly alkaline (pH above 8). In alkaline soils, some micronutrients may become less accessible to plants. Salty soils have an elevated salt content and can be either acidic or alkaline, depending on the region and soil characteristics [6]. Soil reaction close to neutral pH around 6.5-7.5. This range is considered optimal for most agricultural crops. Classifying soils based on acidity is important for determining the need for pH correction to provide the best conditions for plant growth. This may involve adding lime or ammonium fertilizers to raise pH in acidic soils or adding acid to lower pH in alkaline soils [7].

Soil humus content is typically grouped into the following categories. Poor humus soils contain very low levels of humus, usually less than 1-2% organic matter. They are often referred to as "poor soils" and can be found in regions with limited access to organic materials, such as desert areas or sandy soils. Low humus soils have humus levels below average, typically ranging from 2% to 3%. This is a common type of soil and can be found in various climatic conditions. Moderate humus soils have moderate levels of humus, usually ranging from 3% to 5%. They can support normal plant growth but can be improved with the addition of organic fertilizers. Soils with high humus content contain elevated levels of humus, usually exceeding 5%. They are considered highly fertile and suitable for the successful cultivation of a variety of agricultural crops. These are general
categories, and humus levels can vary depending on the region, climatic conditions, and agricultural practices. The humus content in the soil affects its fertility, ability to retain moisture and nutrients, and its capacity to support healthy plant growth. Managing soil humus content is crucial for agriculture and ecological sustainability [8].

Soils with high nitrification ability and high nitrate content exhibit a high capacity for nitrification and contain significant amounts of nitrates (e.g., over 50 mg/kg). This may be characteristic of fertile soils with intensive agriculture, where nitrogen fertilizers are applied, and nitrification processes occur intensively. Soils with high nitrification ability and moderate nitrate content also have a high nitrification capacity, but the nitrate content is moderate (e.g., from 20 to 50 mg/kg). This may indicate good manageability and efficient use of nitrogen in agriculture. Soils with low nitrification ability and moderate nitrate content have a low nitrification capacity, and the nitrate content is moderate (e.g., from 10 to 20 mg/kg). This may be typical of soils with limited nitrifying bacterial activity. Soils with low nitrification ability and low nitrate content have a low nitrification capacity and contain low levels of nitrates (e.g., less than 10 mg/kg). This may be associated with insufficient nitrifying bacterial activity or low nitrogen content [8-9].

Grouping soils based on nitrate content and their nitrification ability, measured in mg/kg, can help understand nitrogen availability for plants and optimize fertilization. It can also be useful for monitoring and managing nitrates in agriculture and ecosystems [10]. Grouping soils based on the content of available phosphorus (P$_2$O$_5$) and its relationship with the content of phosphoric acid (P$_2$O$_5$) in mg/kg can be conducted as follows [11]. Soils with high available phosphorus and high phosphoric acid content are characterized by high levels of available phosphorus and high phosphoric acid content (e.g., over 50 mg/kg P$_2$O$_5$ and over 50 mg/kg P$_2$O$_5$). This may indicate high phosphorus availability for plants and possibly excess phosphorus fertilization. Soils with high available phosphorus and moderate phosphoric acid content have high levels of available phosphorus but moderate phosphoric acid content (e.g., over 50 mg/kg P$_2$O$_5$ and from 20 to 50 mg/kg P$_2$O$_5$). This may indicate phosphorus availability for plants, but additional fertilization may be needed. Soils with low available phosphorus and high phosphoric acid content have low available phosphorus content but high phosphoric acid content (e.g., less than 10 mg/kg P$_2$O$_5$ and over 50 mg/kg P$_2$O$_5$). This may indicate insufficient phosphorus availability for plants, possibly due to low phosphorus mobility in the soil. Soils with low available phosphorus and moderate phosphoric acid content have low available phosphorus content and moderate phosphoric acid content (e.g., less than 10 mg/kg P$_2$O$_5$ and from 20 to 50 mg/kg P$_2$O$_5$). This may indicate low phosphorus availability for plants and the need for phosphorus fertilization. Grouping soils based on the content of available phosphorus (P$_2$O$_5$) and P$_2$O$_5$ can help assess the availability of phosphorus for plants and develop fertilization strategies to optimize the use of this essential nutrient [12-13].

Soils with high exchangeable potassium and high K$_2$O content are characterized by high levels of exchangeable potassium and high potassium oxide content (e.g., over 200 mg/kg K and over 300 mg/kg K$_2$O). This may indicate high potassium availability for plants and possibly excess potassium fertilization. Soils with high exchangeable potassium and moderate K$_2$O content have high levels of exchangeable potassium but moderate potassium oxide content (e.g., over 200 mg/kg K and from 100 to 300 mg/kg K$_2$O). This may indicate potassium availability for plants, but additional fertilization may be needed. Soils with low exchangeable potassium and high K$_2$O content have low exchangeable potassium content but high potassium oxide content (e.g., less than 100 mg/kg K and over 300 mg/kg K$_2$O). This may indicate insufficient potassium availability for plants, possibly due to low potassium mobility in the soil. Soils with low exchangeable potassium and moderate K$_2$O content have low exchangeable potassium content and moderate potassium oxide content (e.g., less than 100 mg/kg K and from 100 to 300 mg/kg K$_2$O). This may indicate low
potassium availability for plants and the need for potassium fertilization. Grouping soils based on the content of exchangeable potassium (K) and K₂O can help determine the availability of potassium for plants and develop fertilization strategies to optimize the use of this important macronutrient.

A soil dataset can be used to develop a machine learning model that predicts the level of soil fertility based on other characteristics [14-17].

2 Materials and methods

Machine learning methods were used to assess soil fertility. The first machine learning method is regression analysis. This equation allows you to predict the value of soil fertility based on the given independent variables.

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \varepsilon, \]  

Where \( Y \) is the dependent variable (soil fertility), \( X_1, X_2, ..., X_i \) are the independent variables (density, moisture, pH, nitrogen, phosphorus, potassium), \( \beta_0 \) is the intercept, \( \beta_1, \beta_2, ..., \beta_i \) are the regression coefficients, and \( \varepsilon \) is the error.

Machine learning methods based on Bayesian machine learning can be used to assess soil fertility, predict soil properties and optimize agriculture. Used Probabilistic approach to modeling and data analysis. This method can be used to classify soil types based on a number of characteristics. For example, soils can be classified as “fertile,” “moderately fertile,” or “infertile” based on their content of nitrogen, phosphorus, potassium, and other indicators. The naive Bayesian classification model is based on the assumption that features are independent, which may be an appropriate assumption for ground data. The Bayesian algorithm for assessing soil fertility can be divided into several phases. In this context, the algorithm is used to classify soil samples based on their properties (e.g. “fertile” or “infertile”). Here is a general Bayesian algorithm for assessing soil fertility:

Step 1: Prepare data on soil samples, including their characteristics such as nitrogen, phosphorus, potassium, pH, organic matter and other parameters.

Step 2: Divide the data into two parts: training set and test set. The training set will be used to train the model, and the test set will be used to evaluate its performance.

Step 3: Train the model to identify statistical relationships between soil characteristics and classes (fertile, infertile, etc.). You can use a simple Bayes classifier or other models based on Bayes' theorem.

Step 4: Estimate the probability for each class based on the trained model. For each soil sample, the model determines the probability of belonging to each class. A score is determined using the Naïve Bayes classification machine learning algorithm which expressed in

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]  

Where: \( P(A) \) is the prior probability that A is correct; \( P(B) \) is the probability of observation B; \( P(B|A) \) is the probability of observing B given that A exists; \( P(A|B) \) is the posterior probability of the class (target) of a given predictor (attribute).

The value of \( P(A|B) \) can be fertile or infertile.

Step 5: Assign a class to the soil sample with the highest probability. For example, if the probability of "fertile" classification is the highest, the soil will be classified as "fertile."
Step 6: Use the test dataset to evaluate the model's performance. Compare the model's predictions with the true classes in the test set to measure accuracy, recall, and other performance metrics.

Step 7: Tuning and Model Improvement. In case of unsatisfactory results, you can tune the model by adjusting parameters and using different features to improve its performance.

Step 8: Model Application. After thorough verification and improvement of the model, you can apply it to classify unknown soil samples. The model will predict their classification based on their characteristics.

This algorithm provides a methodological approach to assess soil fertility using Bayesian methods. This method can be a useful tool in agricultural and environmental research as well as agricultural decision making.

The machine learning algorithm based on Support Vector Machines (SVM) can be applied to assess soil fertility. In this context, SVM can be used for classifying soil samples based on their characteristics, such as classifying soil as "fertile" or "infertile." Here is the SVM algorithm for soil fertility assessment:

Step 1: Data Preparation on soil samples, including their characteristics such as nitrogen, phosphorus, potassium, pH, organic matter, and other parameters.

Step 2: Data Splitting into two parts: the training set and the test set. The training set will be used to train the SVM model, and the test set to evaluate its performance.

Step 3: Kernel Selection for SVM. The kernel determines how data will be transformed into a higher-dimensional space. In this case, you can use a linear, polynomial, or radial-basis function (RBF) kernel depending on the nature of the data.

Step 4: SVM Model Training. The SVM model will seek to find the optimal separating hyperplane that best separates the classes (fertile and infertile soils).

Step 5: Parameter Optimization SVM model such as the regularization parameter C and kernel parameters (e.g., polynomial degree or RBF kernel width). This may require cross-validation to choose the best parameters.

Step 6: Use the trained SVM model to classify soil samples in the test set. The model will determine to which class (fertile or infertile) each sample belongs.

Step 7: Evaluate the model's performance by comparing its predictions with the true classes in the test set. Measure accuracy, recall, F1 score, and other performance metrics.

Step 8: After thorough validation and improvement of the model, apply it to classify unknown soil samples. The model will predict their classification based on their characteristics.

The effectiveness of fertilizers greatly depends on the presence of moisture in the soil. Moisture allows fertilizers to dissolve and become accessible to plants. Water in the soil plays a crucial role in the system of industrial technology for cultivating agricultural crops.

3 Results and Discussion

The results of applying regression methods to assess soil fertility were obtained, which depended on the choice of a specific regression model, data characteristics and modeling quality. Assessing soil fertility using regression involves the task of predicting quantitative values such as nutrient content, pH and other soil parameters. Here are the regression results for assessing soil fertility:

Mean square error (MSE): MSE measures the mean squared difference between model predictions and the true values. A low MSE value indicates good model accuracy.

Mean absolute error (MAE) measures the average absolute difference between predictions and true values. This metric also helps evaluate the accuracy of the model.
Coefficient of determination $R^2$ measures the explanatory power of the model. A value close to 1 indicates a good explanation of the variation in the data by the model. Mean Squared Error: 0.0395463690142137

Linear Regression Equation:

$$Y = 1.31 + 5.14X_1 + -16.53X_2 + 0.03X_3 + 0.24X_4 + -0.04X_5 + 0.04X_6$$  \hspace{1cm} (3)$$

Mean Squared Error (MSE) is a metric that measures the average squared deviation (spread) between the model's predictions and the true values of the data. The MSE value of 0.0395463690142137 indicates that the model closely approximates the data and has a low mean squared error, suggesting good accuracy predictions.

Evaluation of soil fertility using Bayesian methods can yield the following results:

Probabilistic classification: Bayesian methods allow for probabilistic classification of soil samples. The model determines the probabilities of each sample belonging to different classes. Depending on the chosen probabilistic model and settings, high accuracy and recall in classification can be achieved. Accuracy reflects how many of the classified fertile or non-fertile soils actually belong to these classes, and recall measures the model's ability to correctly classify all original samples.

Model accuracy: 0.87

Classification Report:

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0.50</td>
<td>1.00</td>
<td>0.67</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1</td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>0.87</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>macro avg</td>
<td>0.38</td>
<td>0.50</td>
<td>0.42</td>
<td>6</td>
</tr>
<tr>
<td>weighted avg</td>
<td>0.58</td>
<td>0.67</td>
<td>0.61</td>
<td>6</td>
</tr>
</tbody>
</table>

Confusion Matrix:

```
[[3 0 0 0]
 [0 1 0 0]
 [0 1 0 0]
 [0 0 1 0]]
```

The model's accuracy is 0.87, which means the model correctly classified 87% of the samples in your dataset.

precision: This is a metric that measures how many of the predicted positive classes are actually positive. For example, for class "7" the precision is 1.00, which means that all objects predicted as class "7" actually belong to this class.

This is a metric that measures how many of all true positives were correctly predicted by the model. For class "7" recall is also 1.00, which means that the model correctly identified all positive objects of this class.

f1-score is the harmonic mean of precision and recall. For class "7" it is 1.00.

support: This is the number of samples in each class.

The confusion matrix provides information about how many objects were classified correctly or incorrectly by the model.
Each row represents the true classes, and each column represents the predicted classes. In matrix:

The upper-left element (3) indicates that 3 objects of class "7" were correctly classified as class "7."

The upper-middle element (0) indicates that no objects of class "7" were erroneously classified as class "9."

The middle-left element (0) indicates that no objects of class "9" were erroneously classified as class "7."

The middle element (1) indicates that 1 object of class "9" was correctly classified as class "9."

The rest of the elements in the matrix provide similar information about the classification of other classes. Overall, the model has some accuracy, but there are also false positives and false negatives, especially for classes "10" and "11." Precision, recall, and f1-score metrics help evaluate the balance between precision and recall for each class and can be used for a more detailed assessment of the model's performance.

The results of applying the Support Vector Machine (SVM) algorithm for soil fertility assessment will depend on specific data, model parameters, and the task. However, in general, SVM can provide the following results:

SVM can ensure high accuracy in classifying soil samples into fertile and non-fertile categories. Accuracy can be above 90%, depending on data quality and model parameter choices.

The SVM model generally has a good ability to detect fertile soils (high completeness) and infertile soils (high specificity).

The F-measure, which combines precision and recall, can also be high, indicating a good balance between precision and recall of the classification.

The error matrix allows you to evaluate what classification errors are allowed by the model. It can be used to determine which classes the model is prone to making errors on.

If a classification task has high imbalance between classes, ROC curves and area under the AUC curve can be useful metrics to evaluate performance.

The results can be further improved by optimizing SVM parameters and data preprocessing. It is also important to validate the model on independent data sets to confirm its generalizability.

Model accuracy: 1.00
Classification report:

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

accuracy: 1.00, 6
macro avg: 1.00, 6
weighted avg: 1.00, 6

Confusion Matrix:
[[3 0]
 [0 3]]

The classification report and confusion matrix indicate the high performance of the model in the classification task. Let's look at what the different parts of the report and confusion matrix mean:
The model accuracy is 1.00, which means the model correctly classified all samples in your dataset. This is perfect accuracy, and the model performed the classification task perfectly.

Classification report:
precision: Precision for classes "0" and "1" is also 1.00, which means that the model correctly classified all objects as class "0" and class "1".
recall: Recall for classes "0" and "1" is also 1.00, which means that the model correctly detected all objects belonging to classes "0" and "1".
f1-score: The F1-score for classes "0" and "1" is also 1.00. F1-measure is the harmonic mean of precision and recall, and it also shows excellent classification performance.
Confusion Matrix: The confusion matrix provides information on how many objects were correctly classified by the model. In your matrix:
The upper-left element (3) indicates that 3 objects of class "0" were correctly classified as class "0."
The lower-right element (3) indicates that 3 objects of class "1" were correctly classified as class "1."
All other elements in the matrix are equal to 0, indicating no classification errors.

Fig. 1. The model performed perfectly on this classification task, and its predictions closely matched the ground truth.

4 Conclusion

Machine learning methods provide a powerful tool for assessing soil fertility. They allow you to analyze and use various data such as nutrient content, soil chemical properties, historical yields, and climate factors to predict and classify soil fertility. Machine learning methods can help identify which soil factors and parameters influence soil fertility. The models can predict nutrient levels, pH, humus content and other important soil properties. Machine learning can classify soils based on their fertility. For example, you can create a model that determines whether the soil is fertile, moderately fertile, or infertile. They can help optimize fertilizer use. They can predict how much and what fertilizers should be applied to achieve optimal yields, taking into account many factors including climate conditions, historical data and soil properties. This improves the accuracy of soil fertility estimates. The ability to process large amounts of data allows you to create more accurate and reliable forecasts. Overall, machine learning techniques provide agricultural workers and researchers with a powerful tool to improve land resource management and optimize
agricultural practices. Properly tuned models can significantly increase yields and improve plant resistance to various factors.

References


