Data Mining Classification Algorithms for Analyzing Soil Data

Kazheen Ismael Taher\textsuperscript{1*}, Adnan Mohsin Abdulazeez\textsuperscript{2} and Dilovan Asaad Zebari\textsuperscript{3}

\textsuperscript{1}Akre Technical College of Informatics, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq. \textsuperscript{2}Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq. \textsuperscript{3}Research Center of Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

Authors’ contributions

This work was carried out in collaboration among all authors. Author KIT prepared a detailed review of previous works related to analyzing soil data based on data mining classification algorithms. More so, analysis and discussion of the study have been managed by all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2021/v8i230196

Editor(s):
(1) Dr. G. Sudheer, GVP College of Engineering for Women, India.

Reviewers:
(1) Vikram Bali, Jss Academy of Technical Education, India.
(2) H K Shreedhar, Global Academy of Technology, India.

Complete Peer review History: http://www.sdiarticle4.com/review-history/68035

Received 22 February 2021
Accepted 28 April 2021
Published 04 May 2021

ABSTRACT

Rapid changes are occurring in our global ecosystem, and stresses on human well-being, such as climate regulation and food production, are increasing, soil is a critical component of agriculture. The project aims to use Data Mining (DM) classification techniques to predict soil data. Analysis DM classification strategies such as k-Nearest-Neighbors (k-NN), Random-Forest (RF), Decision-Tree (DT) and Naïve-Bayes (NB) are used to predict soil type. These classifier algorithms are used to extract information from soil data. The main purpose of using these classifiers is to find the optimal machine learning classifier in the soil classification. in this paper we are applying some algorithms of DM and machine learning on the data set that we collected by using Weka program, then we compare the experimental result with other papers that worked like our work. According to the experimental results, the highest accuracy is k-NN has of 84 % when compared to the NB (69.23%), DT and RF (53.85 %). As a result, it outperforms the other classifiers. The findings imply that k-NN could be useful for accurate soil type classification in the agricultural domain.

Keywords: Data mining; soil dataset; classification; Weka.

*Corresponding author: E-mail: kajeen.ismael@gmail.com;
1. INTRODUCTION

Data mining has been used to analyze massive data sets to create classification and patterns. Individuals will reliably anticipate the methods used to evoke substantial information. In the world of agriculture, DM is a complex technology to learn. DM is now used in agriculture for soil classification, wasteland control, and field and pest management [1,2]. In agriculture, we evaluated the affiliation rules of affiliation methods in DM and applied them to soil science to anticipate major relations and provide association rules to various soil types [3]. Agriculture factors such as irrigation, temperature, soil type, pesticides, and fertilizers play a major role in growing output. The primary purpose of agriculture is to grow crops. Crop cultivation is dependent on the quality and nutrients of the soil, and growing land cultivation results in a lack of supplements present in the soil. Soil is very important in crop cultivation. Plants, plants, minerals, and living creatures all benefit from it. Any of this contributes to soil quality management [4,5].

The most challenging aspect of DM is classification. Classification is a machine learning-based DM technique for classifying data items in a dataset into a series of predefined groups. It aids in the discovery of differences between objects and ideas. It also gives you the knowledge you need to analyze comprehensively [6,7]. Classification is a DM technique that analyzes a specified dataset and assigns each instance to a particular class with the least amount of classification error. It is used to derive models from a dataset that correctly describe key data groups. It takes two steps to classify something [8,9,10]. In the first stage, a classification algorithm is applied to the training data set to construct the model, and the derived model is then evaluated against a predefined reference dataset in the second step to assess the model's trained performance and accuracy. As a consequence, classification refers to the method of applying a class label to a dataset that does not yet have a class label [11,12].

Soil is a diverse, nonrenewable, and vital natural resource for agricultural development. It gives plants essential elements such as minerals, water, and air, which assist in their physical production, strong growth, survival, and flourishing. Fertile soil is indeed a good foundation for growing stable and nutritious crops [13]. It performs a variety of productive functions while causing no deterioration or harm to the environment [14]. Soil fertility is specifically affected by its intrinsic physical, physiological, biological, and mineralogical properties [15]. As previously reported, the measurement and assessment of soil properties are normally performed by chemical examination of manually collected soil samples. Since the technique is complex and time-consuming, methods for assessing or estimating some of the properties utilizing previously specified features are needed. The soil must first be divided into distinct homogeneous classes before it can be defined. Without a proper rating, soil analysis is equivalent to conducting field experiments with green plants or laboratory experiments with the bare minimum of soil elements [16]. As a consequence, soil classification has been an essential aspect of soil science.

The classification method divides the soil data into separate groups based on certain predefined criteria. It also oversees the formal classification of soils based on distinct characteristics, as well as the parameters that describe the choices and options [13,17]. Furthermore, it assists in predicting the action and ability of the land for crop production, soil reduction mitigating environmental degradation, and increasing productivity. The description of soil increases information, comprehension, and coordination [18,19]. The implementation of a classification model that classifies soils based on soil properties as health indicators will increase fertilizer use and farmland reuse for different crop types.

This paper discussed four DM algorithms for the classification of soil data: DT, NB, K-NN, and RF algorithm. The WEKA environment was used to apply these techniques to the collected soil data [20] and the results were compared and analyzed.

2. LITERATURE REVIEW

In recent years, various researchers have used DM techniques in agriculture. The following is a study of the usage of various DM techniques in the field of soil classification analysis over the last few years.

Sorokin et al., 2021 [21] proposed a framework for grouping "black soils" from US, Russian, and international datasets in the space of principal components to better understand if these soils constitute a distinct category. They concluded
that "black soils" are roughly classified as belonging to the Mollisol Order in the US Soil Taxonomy, but that they can also contain some other soils with dark topsoil. They believe that the concept of "black soil" should be wide, with no hard and fast laws. But for a few soils with shallow depths to hardpan or permafrost, the results revealed that the Great Groups of the Mollisol Order in the US Soil Taxonomy had short taxonomic distances within the order. Dark-colored Vertisols and Andisols have been shown to be different from Mollisols and other associated soils found in similar environments, mostly under grasslands. Because of the special properties, potential applications, and maintenance of these soils, they suggest splitting Vertisols and Andisols from the "black soils" cluster. Despite the fact that the properties of the Soil Taxonomy's Vertisols and Mollisols were completely different, the WRB Vertisol Reference group and Russian dark-humus compact soils fit well with the Mollisols cluster, likely due to the different meaning of Vertisols in the studied scheme.

Motia & Reddy, [13] proposed the Ensemble Classifier (EC) outperformed common classifiers such as DT, KNN, and NB in terms of accuracy. For agricultural soils study. Using a publicly accessible agricultural soil dataset, precision of three well-known classification models is compared in this study: k-Nearest-Neighbor (k-NN), Naïve-Bayes (NB), and Decision-Tree (DT). Following the investigation, an Ensemble Classifier (EC) is proposed, which combines the three classifiers previously described. EC has the highest accuracy of 84 percent, as opposed to k-NN (73.56 %), DT (80.84 %), and NB (72.90 %). As a result, it outperforms the other classifiers. The findings suggest that EC may be advantageous for accurately classifying soil types in the agricultural domain.

Bouayad et al. [22] presented a system for soil classification utilizing several cone penetration tests based on the Gaussian mixture (GM) method (CPT). In contrast to hard clustering, the GM model classifies CPT data by treating the probability density function of the measured variables as a mixture of multivariate normal distributions. To determine the optimal number of clusters, a GM model-based expectation maximization (EM) algorithm with a Bayesian information criterion (BIC) is built. Six real CPT data sets from the Dunkerque site in northern France are used. The classification findings are related to the conventional CPT description using the non-normalized soil activity form (SBT) index and the Robertson map. The findings demonstrate that the GM model is capable of reliably identifying soil layers. Additionally, combining all CPTs, rather than considering them individually, can boost soil layer identification by taking into account all relevant site details.

Murugesan & Radha, [23] proposed a novel classification algorithm for effectively classifying soil data that combines attribute category rank with filter-based instance selection. Experiments were conducted using soil data from the Pollachi area in Coimbatore district, Tamil Nadu state, India, which is a common marketplace for a variety of grains, vegetables, and fruits. The proposed model's classification accuracy is also compared to that of other classification models. The proposed model has a higher accuracy rate for soil data, according to the results review. By selecting the instances, they can define the significant attribute category for classifying the soil data using attribute group rank. The proposed model has better classification accuracy than many other current classifiers under review, according to the experimental research. In classifying the soil data of the Pollachi area, the proposed model has 91.2 percent accuracy, 94.4 percent precision, and 94.3 percent recall. The focus of future research will be on analyzing the soil types in and around Coimbatore. Furthermore, in the future, crop prediction for specific soil types, as well as weather and climatic conditions, will be needed, which is critical for increasing agricultural productivity.

Pandith et al. [24] suggested five supervised machine learning strategies that were applied to the collected data: Naive-Bayes, k-Nearest-Neighbor (KNN), Multinomial Logistic Regression, Random-Forest, and Artificial-Neural-Network (ANN). Five criteria, namely consistency, memory, precision, specificity, and f-score, were evaluated to determine the success of each technique under review. Experiments have been conducted to determine the most accurate methodology for predicting mustard crop yields. All of the ML methods under investigation may be used to estimate crop yields, according to the findings of the experiments. The highest accuracy was predicted by KNN and random forest (88.67 % and 94.13 %, respectively), whereas the lowest accuracy was predicted by Nave Bayes (72.33 percent). In terms of accuracy, the maximum
value expected by ANN was 99.94 percent, while the lowest value predicted by Logistic regression was 24.17%. But for Naive Bayes, all of the classifiers studied expected recall values of over 90%. It says that Naive Bayes had the maximum false negative rate, whereas Logistic regression had the lowest real negative rate. With specificities of 99.78 percent and 80.72 percent, respectively, and f-scores of 0.9976 and 0.8405, ANN and KNN recorded the highest specificity and f-score.

Ahmed, n.d [25] discussed the DT with Bayesian Model in soil prediction and soil classification. Soil classification has been analyzed using various algorithms such as K-Nearest Neighbor, Support Vector Machine, and DT, as well as a proposed Bayesian approach to DT Algorithm. A comparative analysis of various classification algorithms was presented, along with the proposed algorithm. The Bayesian approach to the DT Algorithm aids in the classification of soil types more accurately than the existing Algorithms KNN, SVM, and DT was chosen for this research paper. Finally, the proposed Bayesian approach to the DT Algorithm outperforms the other three existing algorithms for soil type classification: K-Nearest Neighbor, Support Vector Machine, and DT.

N. Saranya et al.,[7] proposed a method of clustering and predicting the type of crop that can be cultivated in that particular type of soil according to the soil nutrients and micro-nutrients. Machine training algorithms such as k-NN, SVM, Bagged Tree, and logistical regression are used. Several different types of maker training algorithms. Various algorithms are used in machine learning to categorize the soil type. A suitable crop is recommended for a particular soil type. From the test results, SVM was shown to be as accurate as possible. The accuracy of the classification is 96%.

Barman & Choudhury [26] used a linear kernel function and multi-SVM to distinguish soil photographs. The photos were taken with an Android phone camera in the West Guwahati region. Except for loamy fine sand, loamy sand, and silty mud, the three-class classifier and multi-class classifier work well on the actual dataset. Previously, the soil texture was calculated using the conventional hydrometer system and USDA triangle, which is a time-consuming and labor-intensive procedure. For the percentage measurement of sand, silt, and mud, a basic hydrometer test requires at least 24 hours. The suggested scheme, on the other hand, is more precise and requires less time to identify the soil. With the aid of a support vector machine and an android Smartphone, it provides a fast and accurate result for soil classification. The suggested system has an overall precision of 91.37 percent for all soil tests, which is almost identical to the US Department of Agriculture’s soil classification. Jahan [27] proposed three algorithms, including Naive Bayes, zeroR, and stacking, are projected. When compared to the other two classifiers, the Naive Bayes classification algorithm performs better on this dataset and correctly classifies the greatest number of instances. In the soil dataset, 50 instances and 8 attributes were used. They talked about soil in various Indian states, including its properties and fertility. For soil classification, they used three classification algorithms: zeroR, stacking, and naive bayes. For this soil data set, the Naive Bayes classifier performs well. When comparing these three algorithms zeroR, stacking produced the best results.

Arooj et al. [28] presented data mining study possibilities for soil classification utilizing well-known classification algorithms such as J48, OneR, BF Tree, and Naive Bayes. The experiment was carried out on data from the Kasur district of Pakistan. They discovered that the efficacy and reliability of forecasts can be determined by a comparative study of these algorithms with varying levels of precision. However, a greater understanding of soil groups will help farmers maximize production, reduce their reliance on fertilizers, and develop better predictive rules for recommending increased output. the outcomes of different classifiers The important result comes from the Naive Bayes classifier, which has 97.63 % performance, 0.977 precision, and 0.9776 recall. The percentage of correctly identified instances in research data samples is shown by the accuracy scale. In the other hand, the J48 result does not have a significant benefit due to its accuracy of 80.92 %, precision of 0.738, and recall of 0.750. Furthermore, the precision of OneR and BF Tree results was 91.97 % and 77.03 %, respectively. OneR has a precision of 0.846 and a recall of 0.92, while BF Tree has a precision of 0.738 and a recall of 0.750, which is somewhat smaller.
3. CLASSIFICATION OF SOIL

Soil classification is the formal categorization of soils focused on distinguishing characteristics as well as criteria that govern consumption choices. Beginning with the system's framework and advancing to class descriptions and field implementation, soil classification is a complex matter. Soil classification can be viewed from two perspectives: substance and resource [27,4].

The Unified Soil Classification System (USCS) is the most widely used engineering classification system for soils [29]. The USCS classifies soils into three types: coarse-grained soils (such as sands and gravels), fine-grained soils (such as silts and clays), and highly organic soils (referred to as "peat"). For clarity, the USCS splits the three major soil classes into subgroups. Color, in-situ moisture content, in-situ weight, and slightly more detail about the material properties of the soil will be included in a full geotechnical engineering soil specification [30].

4. DATA SOURCE AND PARTICULARS SOIL DATASET

Soil data was collected from Soil Science Department, Ahmadu Bello University, Zaria. The data contains 400 soil samples from the North West zone of Nigeria. The soil extracted data used in the related studies included moisture content, liquid limit, clay content, plastic index, plastic limit, and consistency index [31,20]. This dataset has 13 attributes: CY, SN, SL, PH, CaCl2, OC, N, Ca, P, Mg, K, Na, and EC. Table 1 displays the attribute description, and Table 2 displays the dataset samples with their corresponding percentages of the attributes in Table 1.

| Feature | Particulars                                      |
|---------|--------------------------------------------------|
| CY      | Clay Content of the soil (%)                     |
| SL      | Salinity Of the soil (%)                         |
| SN      | Quantity Of sand of the soil (%)                 |
| PH      | PH value of the soil (ppm)                       |
| CaCl2   | Calcium Chloride content of the soil(ppm)        |
| OC      | Organic Carbon (ppm)                             |
| N       | Nitrogen Content Of the soil (ppm)               |
| P       | Phosphorus Content of the soil (ppm)             |
| Ca      | Calcium Content of the soil (ppm)                |
| Mg      | Magnesium content of the soil (ppm)              |
| K       | Potassium content of the soil (ppm)              |
| Na      | Sodium content of the soil (ppm)                 |
| EC      | Electrical conductivity of the soil (ppm)        |

| sample | CY | SL | SN | PH | CaCl2 | OC | N | P | Ca | Mg | K | Na | EC |
|--------|----|----|----|----|-------|----|---|---|----|----|---|----|----|
| 1      | 9  | 38 | 53 | 6.2| 5.6   | 0.41| 0.07| 2.8| 1.92| 0.4 | 0.19| 1.3 | 4.8 |
| 2      | 9  | 28 | 63 | 6.8| 5.7   | 0.34| 0.07| 3.33| 2.08| 0.4 | 0.14| 0.96 | 4.2 |
| 3      | 17 | 44 | 39 | 6.6| 5.6   | 0.54| 0.14| 2.63| 2.16| 0.46| 0.12| 1.3  | 6.7 |
| 4      | 17 | 40 | 43 | 6.2| 5.5   | 0.6 | 0.07| 2.9 | 2.83| 0.83| 0.09| 0.17 | 5.3 |
| 5      | 15 | 38 | 47 | 6.4| 5.8   | 0.43| 0.07| 5.08| 7.75| 4.4  | 0.19| 0.35 | 14.4|
| 6      | 21 | 42 | 37 | 6.3| 5.4   | 0.34| 0.07| 2.98| 2   | 0.34| 0.87 | 4.6 |
| 7      | 9  | 42 | 49 | 6.5| 5.5   | 0.47| 0.14| 3.68| 1.67| 0.82| 0.05| 0.96 | 4   |
| 8      | 7  | 14 | 79 | 6.7| 5.7   | 0.36| 0.07| 4.03| 2.46| 0.2  | 0.2  | 1.3  | 5.4 |
| 9      | 9  | 20 | 71 | 6.5| 5.8   | 0.41| 0.14| 5.95| 2   | 0.6  | 0.34 | 1.3  | 4.8 |
| 10     | 11 | 46 | 43 | 6.6| 5.9   | 0.73| 0.14| 3.85| 2.78| 0.8  | 0.07| 2.17 | 6.3 |
5. AGRICULTURAL DATA MINING

DM is important for learning about agricultural topics like soil productivity, yield estimation, and soil erosion. Soil prediction is useful for crop management and soil remediation. The aim of classification algorithms is to find rules that divide data into disjoint classes. A classification method produces a series of classification principles that can be used to categorize new data in the future [32]. The sections that follow explain classification algorithms such as the Logistic Regression classifier, the Naive Bayes classifier, the J48 DT classifier, and the K-Nearest Neighbors classifier [33].

Though lazy learning methods are in general are a high demand because of their cognitive strain on the learning mechanisms, the DT, NB, RF, and KNN's strong suit is that it only uses basic computer-based methods with little effort. Classification and regression issues are supported by this tool. When making a forecast, it holds all of the training examples as well as the target and searches through the entire dataset to find k points that are most close to the training point. Thus, there is no other dataset to work with, but the training dataset, which only returns results from queries of the raw dataset. These particular methods would not use any mathematical functions to identify a goal variable that has been pre-defined beforehand. A comparison dataset is used to find soils with matching characteristics; in other words, for soils that have a known counterpart is defined on the elements being queried, the results are scanned. Based on field-case research, it seems that the effectiveness of the procedure relies on the ‘largely’ on the ‘lots of like’ (similar) soils.

5.1 Naive Bayes

The Bayes theorem underpins the Naive Bayes algorithm, which states that every pair of features is autonomous. Naive Bayes classifiers are useful in a variety of real-world applications, including document sorting and spam filtering. This algorithm only needs a small amount of training data to estimate the necessary parameters. When compared to more sophisticated methods, Naive Bayes classifiers are extremely fast. Naive Bayes is notorious for being a poor estimator [33]. A Naive Bayes classifier is a machine learning classifier that belongs to a family of basic probabilistic classification techniques. It is founded on the Bayes theorem with characteristics of freedom. The likelihood of a given instance is used to approximate each class mark. Just a limited amount of training data is required to predict the class mark required for classification [34].

5.2 J48 Decision Tree

A decision tree produces a set of rules that can be used to characterize data given a set of attributes and their groups. Advantages: Decision Tree is easy to comprehend and interpret, needs minimal data processing, and can accommodate both numerical and categorical data. Decision trees may produce dynamic trees that are difficult to generalize, and they can be unreliable since minor changes in the data can result in the generation of an entirely different tree [35,33,36]. J48 is a viaion for (C4.5) in Weka, the J48 algorithm is a classification-decision tree algorithm that is significantly adapted from C4.5. It has the ability to choose the exam that will have the most detail. Ross Quinlan came up with the idea for this algorithm [37]. A mathematical classifier is another name for C4.5. J48 estimates the dependent variable based on the data available. It creates a tree dependent on the training data's attribute values. This categorizes data using the function of data instances that are claimed to have gained knowledge. The pruning principle is used to establish the value of error tolerance [38,39].

5.3 Random Forest

The RF algorithm is a learning algorithm that is supervised. As shown in Fig. 1, RF constructs multiple DTs and merges them to produce a more stable and accurate prediction [40]. While splitting any node, RF looks for the most important parameter among all and then searches for the best among the subset of random features. For the splitting of a node, this algorithm takes only selective features into account [41,42]. The trees can be made more random by using random feature set thresholds rather than searching for the best possible thresholds [43].

5.4 K-Nearest Neighbors

Neighbors-based classification is a form of lazy learning in which it stores instances of the training data rather than trying to construct a general internal model. To define a point, its k closest neighbors must vote by simple majority. This algorithm is simple to use, tolerant of noisy training data, and effective when working with
massive quantities of data. The value of K must be calculated, and the computing expense is large since each instance must be separated from all of the training samples [44].

6. WEKA TOOLS

Weka is a collection of data mining machine learning algorithms. Waikato Environment for Knowledge Learning is the acronym for Waikato Environment for Knowledge Learning. The University of Waikato in New Zealand established it. Pre-processing, regression, sorting, clustering, visualization, and correlation rules are all supported by Weka [8,45]. The Weka workflow are after collections the data in the [20] inserted the excel sheet next change the format to the scv after that opining the Wika program then select the file that chnged format to scv next chose the clasifecation tab after indicate the algorithm that use to analising data fainaly displaed the result.

7. EXPERIMENTS RESULT AND DISCUSSION

Based on the training data collection, the weighted average of the True Positive Rate of the K-NN classifier is 0.848. When the Naive Bayes, DT, and RF TP Rates are 0.692, 0.538, and 0.538, respectively, it suggests a low level. As a result, the data collection was automatically labeled in a higher context by the K-NN classifier. Table 3 shows the detailed accuracy of soil properties.

Table 3. Weighted average detailed accuracy of classifiers

| Classify | TP Rate | FP Rate | Accuracy | Recall | F Measure | MCC | ROC Area | PRC Area |
|----------|---------|---------|----------|--------|-----------|-----|----------|----------|
| NB       | 0.692   | 0.11    | 0.628    | 0.692  | 0.649     | 0.56| 0.767    | 0.714    |
| kNN      | 0.848   | 0.048   | 0.846    | 0.846  | 0.846     | 0.798| 0.862    | 0.773    |
| DT       | 0.538   | 0.104   | 0.615    | 0.538  | 0.573     | 0.457| 0.836    | 0.672    |
| RF       | 0.538   | 0.186   | 0.548    | 0.538  | 0.501     | 0.390| 0.890    | 0.750    |
The classifiers are evaluated comparatively in Table 4. As compared to the other algorithms, k-NN worked best in classification, and the Kappa Statistic value in k-NN algorithm is closest to 1.00.

Fig. 3 shows the amount of Mean Absolute Error classified instances: Here, maximum instances have been classified by RF.

The high prediction precision in the K-NN algorithm is given in Fig. 4. In contrast to K-NN and Naive Bayes, DT and RF algorithms are less reliable.

Study results indicate that the algorithm's performance is not guided by either variable. One of the shortcomings of the method is that large values which fall beyond the optimum ranges, which causes inaccuracy in estimations to exist. Design parameters are set with regard to the size of the reference dataset; however, the optimum design settings rely on the dataset creation. KNN, RF, DT, and NB models are associated with a higher variance of prediction since they use a larger number of input variables.

In the Table 5 Comparison among that algorithm that used in this paper with the algorithm that used in same previse work based on Wika analysis, in proposed work used 10 sample of soil dataset but inthe [24] used 5000 dataset, [26] used 50 image, and experiments on the k-NN accuracy of the proposed work compared to [24] increased 11.05, but compared to [13] 4.06 decreased. Also experiments on the NB accuracy of the proposed work compared to [13,24,28] 5.67, 3.1, 28.4 decreased.

Table 4. Analysis of classifiers in comparison

| Classifier            | NB   | k-NN | DT   | RF   |
|-----------------------|------|------|------|------|
| Correctly-Classified- |      |      |      |      |
| Instances             | 9    | 11   | 7    | 7    |
| Incorrectly-Classified-|      |      |      |      |
| Instances             | 4    | 2    | 6    | 6    |
| Kappa-Statistic       | 0.563| 0.7833| 0.3659| 0.3333|
| Accuracy              | 69.2308% | 84.6154%| 53.8462%| 53.8462%|
| Mean-Absolute-Error   | 0.1515| 0.1085| 0.1795| 0.2718|
Table 5. Comparison proposed Results for Soil Classification with previous studies

| Ref.                      | Data size | Algorithms                        | Accuracy  |
|---------------------------|-----------|-----------------------------------|-----------|
| Pandith et al., 2020 [24] | 5000      | NB                               | 72.33%    |
|                           |           | Multinomial-Logistic-Regression  | 80.24%    |
|                           |           | RF                               | 94.13%    |
|                           |           | k-NN                             | 88.67%    |
|                           |           | Artificial-Neural-Network (ANN)   | 76.86%    |
| Barman &Choudhury, 2020 [26] | 50 images | multi SVM                        | 91.37%    |
| N. Saranya et al., 2020 [7] | -        | K-NN                             | 96%       |
|                           |           | Bagged Tree                      |           |
|                           |           | SVM                              |           |
|                           |           | logistical regression            |           |
| Motia& Reddy, 2021 [13]   | 60        | NB                               | 72.90%    |
|                           |           | k-NN                             | 73.56%    |
|                           |           | DT                               | 80.84%    |
|                           |           | Ensemble-Classifier(EC)          | 84%       |
| Arooj et al., 2018 [28]   | 800       | OneR                             | 91.97%    |
|                           |           | J48                              | 80.92%    |
|                           |           | NB                               | 97.63%    |
|                           |           | BF Tree                          | 77.03%    |
| Murugesan and Radha, 2020 [23] | 3266 samples | Multi classification      | 91.2%     |
| Rahman et al., 2018 [46]  | 438       | Gaussian-SVM                     | 94.95     |
|                           |           | Weighted-k-NN                    | 92.93     |
|                           |           | Bagged-trees                     | 90.91     |
| Proposed Work             | 10 sample | NB                               | 69.23%    |
|                           |           | k-NN                             | 84.61%    |
|                           |           | DT                               | 53.84%    |
|                           |           | RF                               | 53.84%    |

Consequently, class limitations are typically elected subjectively by granting. Because these are not uniform across groups, classes, the default for the project is granting an assumption of ambiguity as to an intermediate. Scale and creating eminence maps for a given category that are uncertain in varying degrees of exactitude. Soil is difficult when it comes to data mining. mechanical recognition of spatial details adds to the difficulty of developing information that is seen through the advent of highly variable geometries and the usefulness of spatial databases. In order to extend the limits on obtainable dirt, this study has to do the following limitations: It's also possible that current traditional soil analysis approaches do not provide accurate classifications of soil activity since the technology of soils is restricted. Then, avoid the prediction errors of soil; hence, classifiers that are not flexible in soil classification should not be used. The process is much more complex to apply because of the additional computing costs involved. When using traditional soil sampling and determination techniques are used, efficiency and precision are decreased. The decrease in productivity in soil data mining techniques led to a substantial reduction in agriculture. There is a big problem with the present soil classification method's usage of soil samples: it creates delays due to the drying phase. Data that is related to the local to the current database or unique to a specific device or other consumer either be expanded in a separate reference database without distorting other databases or having any major effects on them. Additionally, we suggest conducting more research on this technique's potential to predict soil properties.

8. CONCLUSION

The goal of a classification algorithm is to create a method that correctly classifies data using the training data set. The soil is the most important aspect of agriculture. The classification of soils based on the nutrients found in the soil, such as
potassium, nitrogen, sulphur, phosphorus, iron, zinc, manganese, boron, and copper, as well as its physical properties, such as pH, organic carbon, and electric conductivity, is extremely useful for increasing agricultural production. In this paper, comparing of four algorithms such as NB, K-NN, DT, and RF is discussed in this article. In comparison to the other four, the K-NN classification algorithm produces a better result for this dataset, correctly classifying the full number of instances. To forecast soil features, K-NN can be suggested. Also, previous experiments were related to the proposed findings for soil classification.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Cunningham SJ and Holmes G. Developing innovative applications in agriculture using data mining. The Proceedings Of The Southeast Asia Regional Computer Confederation Conference. 1999:25–29.
2. Abdulazeez AM, Sulaiman MA, and Qader D. Evaluating data mining classification methods performance in internet of things applications. 2020;1(2):15.
3. R Zebari, A Abdulazeez, D Zeebaree, D Zebari, and J Saeed. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. JASTT. 2020;1(2):56–70. DOI: 10.38094/jastt1224.
4. Bhargavi P, Jyothish S. Soil classification using data mining techniques: A comparative study. International Journal of Engineering Trends and Technology. 2011;5.
5. Zeebaree DQ, Haron H, Abdulazeez AM, Zebari DA. Machine learning and region growing for breast cancer segmentation. In 2019 International Conference on Advanced Science and Engineering (ICOASE), Zakho - Duhok, Iraq. 2019:88–93. DOI: 10.1109/ICOASE.2019.8723832.
6. Hassan OMS, Abdulazeez AM, TIRYAKI VM. Gait-based human gender classification using lifting 5/3 wavelet and principal component analysis. In 2018 International Conference on Advanced Science and Engineering (ICOASE). 2018:173–178.
7. Saranya N, Mythili A. Classification of soil and crop suggestion using machine learning techniques. Sri Shakthi Institute Of Engineering And Technology, IJERT. 2020; V9(02):IERTV9IS020315. DOI: 10.17577/IJERTV9IS020315.
8. Sadikin M, Alfiandi F. Comparative study of classification method on customer candidate data to predict its potential risk. UJECE. 2018.;8(6):4763. DOI: 10.11591/ijece.v8i6.pp4763-4771.
9. Zeebaree DQ, Haron H, Abdulazeez AM. Gene selection and classification of microarray data using convolutional neural network. In 2018 International Conference on Advanced Science and Engineering (ICOASE), Duhok. 2018;145–150. DOI: 10.1109/ICOASE.2018.8548836.
10. Zeebaree DQ, Haron H, Abdulazeez AM, Zebari DA. Trainable model based on new uniform lbp feature to identify the risk of the breast cancer. In 2019 International Conference on Advanced Science and Engineering (ICOASE), Zakho - Duhok, Iraq. 2019:106–111. DOI: 10.1109/ICOASE.2019.8723827.
11. Nikam SS. A comparative study of classification techniques in data mining algorithms. Oriental Journal of Computer Science & Technology. 2015;8(1):13–19.
12. Kareem FQ, Abdulazeez AM. Ultrasound medical images classification based on deep learning algorithms: A review.
13. Motia S, Sreeji S. Ensemble classifier to support decisions on soil classification. IOP Conf. Ser.: Mater. Sci. Eng. 2021; 1022:012044. DOI: 10.1088/1757-899X/1022/1/012044.
14. El-Ramady HR, et al. Soil quality and plant nutrition. In Sustainable Agriculture Reviews 14, Springer. 2014;345–447.
15. Karlen DL, Ditzler CA, Andrews SS. Soil quality: Why and how?. Geoderma. 2003;114(3–4):145–156.
16. Hartemink AE. The use of soil classification in journal papers between 1975 and 2014. Geoderma Regional. 2015;5:127–139.
17. Brifcani A, Issa A. Intrusion detection and attack classifier based on three techniques: A comparative study. Eng. & Tech. Journal. 2011;29(2):368–412.
18. Kovačević M, Bajat B, Gajic B. Soil type classification and estimation of soil properties using support vector machines. Geoderma. 2010;154(3–4):340–347.
19. Salim NO, Abdulazeez AM. Human diseases detection based on machine learning algorithms: A review. International Journal of Science and Business. 2021;5(2):102–113.
20. Hassan Hayatu I, Mohammed A, Ahmad Isma’eeel B, Yusuf Ali S. K-means clustering algorithm based classification of soil fertility in north west Nigeria. FJS. 2020;4(2):780–787. DOI: 10.33003/fjs-2020-0402-363.
21. Sorokin A, Owens P, Láng V, Jiang Z-D, Michéli E, Krasilnikov P. ‘Black soils’ in the Russian soil classification system, the US soil taxonomy and the WRB: Quantitative correlation and implications for pedodiversity assessment. CATENA. 2021;196:104824. DOI: 10.1016/j.catena.2020.104824.
22. Bouayd D, Baroth J, Dano C. Gaussian mixture model based soil classification using multiple cone penetration tests. IOP Conf. Ser.: Earth Environ. Sci. 2021;696(1):012034. DOI: 10.1088/1755-696X/696/1/012034.
23. Murugesan G, Radha DB. Soil data classification using attribute group rank with filter based instance selection model. 2020;9(06):7.
24. Pandith V, Kour H, Singh S, Manhas J, Sharma V. Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. JSR. 2020;64(02):394–398. DOI: 10.37398/JSR.2020.640254.
25. Ahmed AZ. Application of bayesian approach to decision tree algorithm for classification of soil types. International Journal of Advanced Research in Engineering and Technology (IJARET). 2020;11(8):808-814.
26. Barman U, Choudhury RD. Soil texture classification using multi class support vector machine. Information Processing in Agriculture. 2020;7(2):318–332. DOI: 10.1016/j.ipa.2019.08.001.
27. Jahan R. Applying naive bayes classification technique for classification of improved agricultural land soils. IJRASET. 2018;6(5):189–193. DOI: 10.22214/ijraset.2018.5030.
28. Arooj A, Riaz M, Akram MN. Evaluation of predictive data mining algorithms in soil data classification for optimized crop recommendation. In 2018 International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan. 2018;1–6. DOI: 10.1109/ICACS.2018.8333275.
29. Isbell RF. The Australian Soil Classification., Australian Soil and Land Survey Handbook (CSIRO Publishing: Collingwood, Vic.). 1996;4.
30. Pham BT, Hoang T-A, Nguyen D-M, Bui DT. Prediction of shear strength of soft soil using machine learning methods. Catena. 2018;166:181–191.
31. Patnaik S, Yang X-S, Sethi IK. Eds., Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019. Singapore: Springer Singapore; 2021.
32. Anuradha C, Velmurugan T. A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. Indian Journal of Science and Technology. 2015;8(15):1–12.
33. Rajeswari V, Arunesh K. Analysing soil data using data mining classification techniques. Indian Journal of Science and Technology. 2016;9(19). DOI: 10.17485/ijst/2016/v9i19/93873.
34. Narain B, Kumar S, Patle VK, Chandrakar PK. Study for data mining techniques in classification of agricultural land soils. Journal of Advanced Research in Computer Engineering. 2011;5(1):35–7.
35. Eesa AS, Orman Z, Brifcani AMA. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Systems with Applications. 2015;42(5):2670–2679.
36. Venkatesan E, Velmurugan T. Performance analysis of decision tree algorithms for breast cancer classification. Indian Journal of Science and Technology. 2015;8(29):1–8.
37. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends. 2021;2(01):20–28.
38. Chandrakar PK, Kumar S, Mukherjee D. Applying classification techniques in Data Mining in agricultural land soil. International Journal of Computer Engineering. 2011;2:89–95.
39. Issa AS. A comparative study among several modified intrusion detection system techniques. B. Sc., Computer Science, Duhok University; 2009.

40. Abdulkareem NM, Abdulazeez AM. Machine learning classification based on Random Forest Algorithm: A review. International Journal of Science and Business. 2021;5(2):128–142.

41. Kajol R, Akshay KK. Automated agricultural field analysis and monitoring system using IoT. International Journal of Information Engineering and Electronic Business. 2018;11(2):17.

42. Priya R, Ramesh D, Khosla E. Crop prediction on the region belts of India: A Naive Bayes MapReduce precision agricultural model. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2018;99–104.

43. Keerthan Kumar TG, Shubha C, Sushma SA. Random Forest Algorithm for soil fertility prediction and grading using machine learning. IJITEE. 2019;9(1):1301–1304. DOI: 10.35940/ijitee.L3609.119119.

44. Paul M, Vishwakarma SK, Verma A. Analysis of soil behaviour and prediction of crop yield using data mining approach. In 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India. 2015;766–771. DOI: 10.1109/CICN.2015.156

45. Weka W. 3: Data mining software in Java. University of Waikato, Hamilton, New Zealand (www.cs.waikato.ac.nz/ml/weka). 2011;19:52.

46. Rahman SAZ, Mitra KC, Islam SM. Soil classification using machine learning methods and crop suggestion based on soil series. In 2018 21st International Conference of Computer and Information Technology (ICCIT). 2018;1–4.

© 2021 Taher et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history: The peer review history for this paper can be accessed here: http://www.sdiarticle4.com/review-history/68035