Kong. Res. J. 8 (1): 30-35, 2021
Publisher: Kongunadu Arts and Science College, Coimbatore.
ISSN 2349-2694, All Rights Reserved,
https://www.krjournal.com

DOI:10.26524/krj.2021.5

RESEARCH ARTICLE

PREDICTIVE MODEL CONSTRUCTION FOR PREDICTION OF SOIL FERTILITY USING DECISION TREE MACHINE LEARNING ALGORITHM

Jayalakshmi, R.[1] and Savitha Devi, M.[2,*]
[1]Department of Computer Science, Periyar University, Salem, Tamil Nadu, India
[2]Department of Computer Science, Periyar University Constituent College of Arts and Science, Harur, Dharmapuri, Tamil Nadu, India

ABSTRACT

Agriculture sector is recognized as the backbone of the Indian economy that plays a crucial role in the growth of the nation's economy. It imparts on weather and other environmental aspects. Some of the factors on which agriculture is reliant are Soil, climate, flooding, fertilizers, temperature, precipitation, crops, insecticides, and herb. The soil fertility is dependent on these factors and hence difficult to predict. However, the Agriculture sector in India is facing the severe problem of increasing crop productivity. Farmers lack the essential knowledge of nutrient content of the soil, selection of crop best suited for the soil and they also lack efficient methods for predicting crop well in advance so that appropriate methods have been used to improve crop productivity. This paper presents different Supervised Machine Learning Algorithms such as Decision tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) to predict the fertility of soil based on macro-nutrients and micro-nutrients status found in the dataset. Supervised Machine Learning algorithms are applied on the training dataset and are tested with the test dataset, and the implementation of these algorithms is done using R Tool. The performance analysis of these algorithms is done using different evaluation metrics like mean absolute error, cross-validation, and accuracy. Result analysis shows that the Decision tree is produced the best accuracy of 99% with a very less mean square error (MSE) rate.

Keywords: Agriculture, Machine learning, soil fertility, K-Nearest Neighbour, Support Vector Machine, Decision tree.

1. INTRODUCTION

Data mining could be a fairly immature and interdisciplinary sector of computer science, is that the process that attempts to mining patterns in large data sets. It utilizes methods at the connection of statistics, artificial intelligence, machine learning, and database systems. The data mining task aims to extract information from a knowledge set and transform it into a comprehensible form for further use. Predictive analysis is that the task of gathering information from soil datasets to seek out future outcomes. Machine learning facilitates methods and techniques for accurate diagnosis and analytical facilities within the agricultural domains. Agriculture is one of the crucial industrial sectors in India and also the country's economy relies heavily on that for the sustainability of its rural areas. Because of some factors [1] like climate change, unplanned rainfall, falling water levels, excessive use of pesticides, etc., the extent of agricultural production in India is dilapidated. Most farmers don't achieve expected crop yields for a spread of reasons. To acknowledge production levels, soil fertility is carried out which involves predicting the yield of the crop based on the existing data. Previously, crop yield estimates were supported farmer's specific crops and cultivation experience.

Data mining techniques are useful for predicting the fertility of the soil. Data processing software [2] is an analytical tool that permits users to categorize and assessing identified relationships still as analyzes data at various dimensions. A soil test is an analysis of a soil sample to determine nutrient content, composition, and other characteristics. Tests are usually performed to measure fertility and indicate deficiencies that require being resolved. Soil fertility depends on various factors [3] and depends on:

*Correspondence: Savitha Devi, M., Department of Computer Science, Periyar University Constituent College of Arts and Science, Harur, Dharmapuri, Tamil Nadu, India E. mail: msavithacs@periyaruniversity.ac.in

- Geographical area
- Soil type (saline, alkaline, non-alkaline)
- Weather (Humidity, Temperatures, precipitation)
- Soil composition (pH, N, P, K, EC, OC, Zn, F)

Prediction models are essentially two main categories. 1. Statistical models, which utilize one forecast function that has all samples. 2. Machine learning, emerging technology to explore knowledge that connects input and output variable models. Machine learning has the potential to learn the machine without defined computer programming, so it enhances machine performance by detecting and characterizing the pattern of constraining data. Machine learning can be classified into three types according to the learning methods are supervised learning, unsupervised learning, and reinforcement learning. This kind of algorithm is used to build the most accurate and effective model. It involves the construction of a machine learning predictive model that's supported on labeled samples. The SVM, KNN, and Decision trees are used for soil fertility prediction.

This paper is organized as follows: Section II presents the related work, whereas the proposed method is discussed in Section III. Then, the experimental result analysis of agricultural data is described in Section IV. Finally, the conclusion is given in Section V.

2. RELATED WORK

Machine learning in Agriculture may Novel research field; an excellent deal of labor has been a tired field of Agriculture utilizing Machine learning. Agricultural scientists in Pakistan have demonstrated that endeavors of harvest yield amplification through expert pesticide state strategies have prompted a hazardously high pesticide use. These examinations have revealed a negative relationship between's pesticide use and harvest yield [4]. In their investigation, they have explained that how data mining incorporated farming information including irritation exploring, pesticide utilization and meteorological information help streamline of pesticide use. Topical data identified with agribusiness which has spatial properties was accounted for in one in every of the study [5]. Their research went for perceiving patterns in farming creation with references to the accessibility of information assets. K-means method was applied to hold out gauges of the contamination within the air [6], the k- nearest neighbor become connected for mimicking day by day precipitations and other climate elements [7], and numerous ability changes of the weather situations are dissected utilizing SVM [8]. Statistics mining techniques are often wont to have a glance at soil qualities. For example, the k-means method is employed for segmenting soils in a mixture with GPS-based technology [9]. A decision tree classifier for agriculture information changed into proposed [10]. This new classifier uses new facts expression and may address each entire records and in entire records. Inside the test, a 10-fold cross-validation technique is employed to test the dataset, horse-colic dataset, and soybean dataset. Their results showed the proposed selection tree is capable of classifying all varieties of agriculture records. A yield prediction version became proposed in one in the entire take a glance at [11] which makes use of mining techniques for category and prediction. This model worked on entering parameters crop name, land location, soil type, soil ph, pest information, climate, water stage, seed type, and this model anticipated the plant boom and plant diseases and so enabled to pick the good crop supported climate information and required parameters.

3. PROPOSED METHODOLOGY

In the proposed system, we use supervised learning to create a model, which provides predicted fertility of soil as Ideal or Not Ideal. The proposed system is described in the following stages like Problem study, data collection, dataset description, preprocessing step, parameter study, and applying machine learning techniques as shown in figure 1.

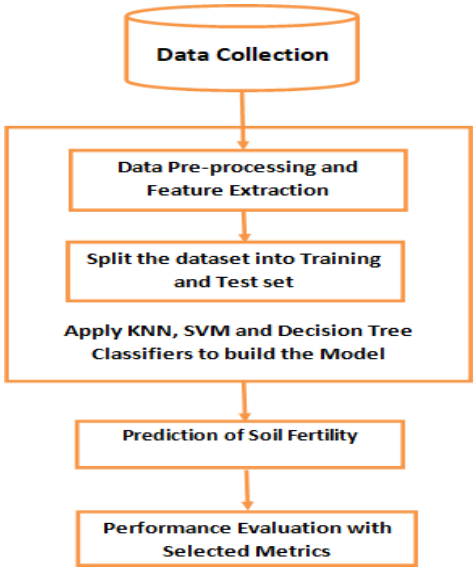

Figure 1. Flow diagram of Proposed Methodology

Problem study

A brief study of problems associated with maximization of the productivity and prediction of soil fertility has been done by hunting the related literature review, the brief discussions with soil analysts and broader view of research problem have been gained.

Data collection

After gaining insight into the research problem the related data has been collected from Soil Testing Laboratory, Melalathur, Vellore District. The dataset consists of 1000 samples. Further data has been divided for training and testing purposes. 700 samples are used for training and 300 data samples are used for testing purposes, Data has been preprocessed and has been transformed into two excel sheets one for training and one for testing purposes.

Preprocessing steps

This step could be very significant in machine learning. Preprocessing consists of inserting the missing values, the suitable data range, and extracting the functionality. Soil attributes like Sample no, Ph, EC, OC, N, P, K, S, Cu, Fe, Zn, Mn are taken as feature variables, and therefore the NULL values additionally as redundant values from the dataset are removed. The type of dataset is critical to the analysis process. During this work, we've used the anyNA method for the treatment of missing values.

Attributes description

The collected dataset consists of soil composition parameters and is one in all subsets for the prediction of yield. The dataset consists of 12 parameters out of Sample no, Ph, EC, OC, N, P, K, S, Cu, Fe, Zn, Mn, out of which 7 (Ph, EC, OC, N, P, K, S) are classified as Macro-Nutrients and remaining 4 parameters (Cu, Fe, Zn, Mn) are Micro-Nutrients and soil were classified into two class labels: Ideal and Not Ideal has been further used for the making decision. Table 1 shows an attribute description.

Table 1. Attributes & its description

Attributes	Description
Sample No	Sample Identification Number
pH	pH value of soil
EC	Electrical conductivity
OC	Organic Carbon
N	Nitrogen
P	Phosphorous
K	Potassium
S	Sulphur
Cu	Copper
Fe	Iron
Zn	Zinc
Mn	Manganese
FI	Class label (Ideal, NotIdeal)

Split the Dataset into Train and Test Set

The dataset is partitioned into training and testing set of input data. The loaded data is split into two sets like training data and test data, with a division ratio of 70% or 30%, such as 0.7 or 0.3. In an exceedingly learning set, a classifier is employed to make the available input data. During this step, create the classifier's support data and preconceptions to approximate and classify the function. During the test phase, test data were set used to test the trained model.

Applying Machine Learning Techniques

We have used three different supervised machine learning algorithms for soil fertility prediction which is given as follows

KNN Algorithm

KNN could be a nonparametric supervised learning technique that uses training sets to segment data points into given categories. In simple classifications, the word collects information from all educational cases and similarities supported the new case. Observe the training for the foremost similar (neighbor) K cases and predict the new instance (x) by summarizing the output variables for these K cases. Classification is the class value mode. A flow diagram of the KNN algorithm is shown in Figure 2.

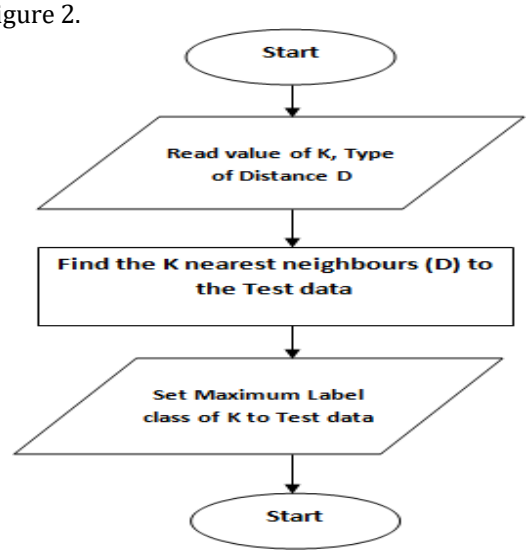

Figure 2. Flow chart for KNN algorithm

Support Vector Machines

Support vector machines (SVM) which is the most powerful and flexible supervised machine learning algorithms used both for classification and regression problems. SVM can extremely popular due to its capability to handle categorical and multiple continuous variables. SVM divides the given data into the decision surface. The decision surface further divides the information into the hyperplane of two classes. Training points define the supporting vector which defines the hyperplane. The hyperplane is generated iteratively by SVM so that the error can be minimized. Probably, a hyperplane with the maximum distance to the closest learning data point typically has better margins and bigger errors due to the larger margins, the generalization of classifiers is weak. The flow chart for SVM is given in figure 3, it shows the steps involved in the SVM algorithm.

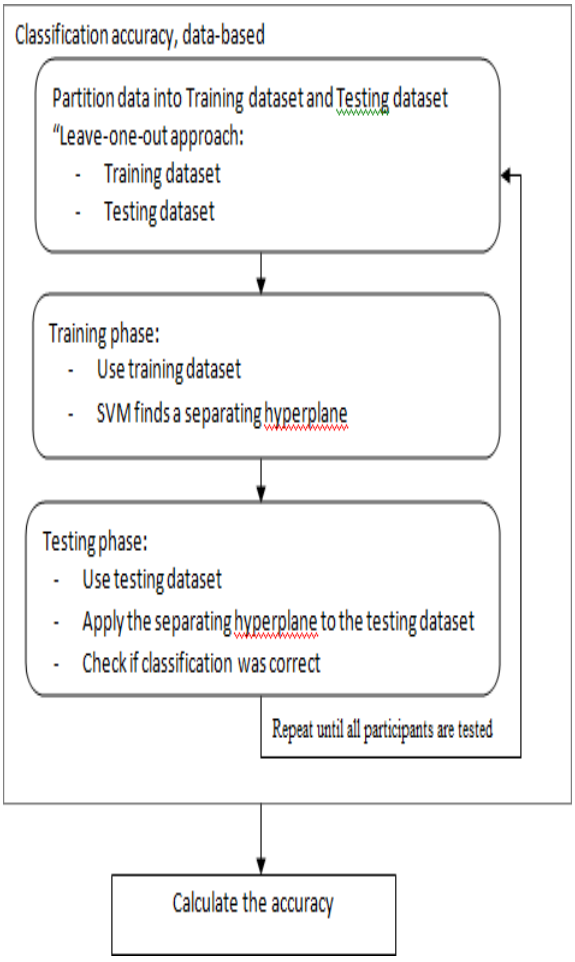

Figure 3. Flow chart for SVM algorithm

Decision tree

A decision tree can be a predictive model which works by checking condition at every level of the tree and proceeds towards the bottom of the tree where various decisions are listed. The condition depends on the appliance and also the outcome may well be in terms of the decision. There are various kinds of Decision tree algorithms like C4.5, CART, and ID3 algorithm. In this work, we used the C5.0 algorithm for model building. Information Gain is the most important measure used to create a decision tree because it was worn to choose the variable that best splits the data at each node of a Decision Tree. The variable with the highest IG is used to split the data at the root node. A C5.0 model works by dividing the sample based on the area that provides the highest information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the task repeats until the subsamples can't be split further. Finally, the lowest-level splits are again reviewed, and those that do not contribute drastically to the value of the model are pruned. The C5.0 node can predict only a categorical target.

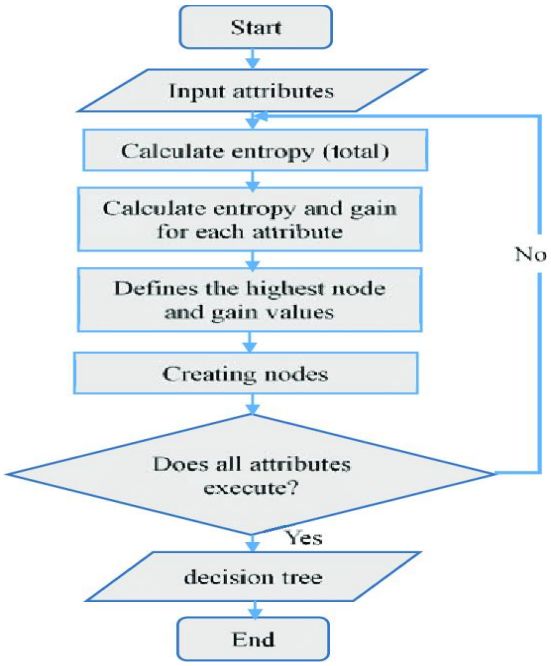

Figure 4. Flowchart of Decision tree

4. RESULTS ANALYSIS AND DISCUSSION

This section shows the result obtained after the implementation of the Machine learning algorithms on the collected dataset. Three Machine learning algorithms KNN, SVM, and Decision Tree have been applied to a trained dataset. The R Tool

version 3.5.3 has been used in this work. The Training data consists of 12 attributes based on the availability of Macro and Micro Nutrients present in the soil. In R Tool the training data with 1000 samples have been used to train the model separately by KNN, SVM, and Decision Tree classifiers. An efficiently trained model has been applied to the testing data set which is different from the training data set. The different evaluation parameters for these algorithms were mean squared error, accuracy, and cross-validation which are used to estimate the efficiency of the method as shown in table II.

Table 2. Comparison table for different parameters

Algorithms	Accuracy	MSE
Decisiontree	99	0.01
KNN	74	0.6897
SVM_linear	78.9	0.6552
SVM_rbf	73	0.559

Machine learning algorithms have been applied individually using the Cross-Validation techniques with 10 folds and the accuracy of prediction has been observed for each of them. In this paper the accuracy for SVM was calculated for two different kernels i.e, SVM _rbf, and SVM_linear among these two RBF kernels was showing more error rate. Among the three algorithms, Decision Tree proves to be a better classifier as compared to SVM and KNN which produced more accuracy with very little MSE. As shown in the graph the accuracy of the decision tree algorithm is more and also it is showing less error rate.

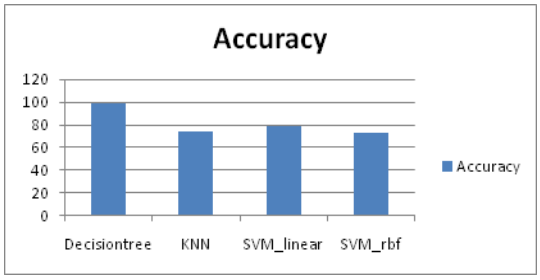

Figure 5. Comparison graph for Result Analysis based on Accuracy

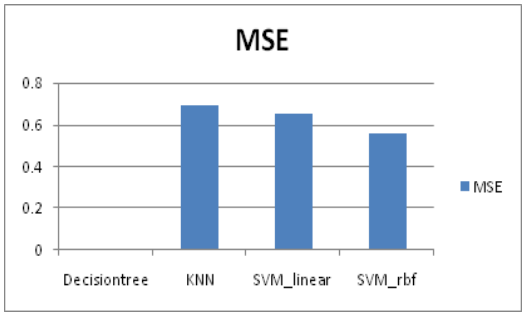

Figure 6. Comparison graph for Result Analysis based on MSE

5. CONCLUSION

Different machine learning algorithms have been implemented on agricultural data to evaluate the best performing method. In this work, we used three different supervised learning algorithms, such as SVM, KNN, and Decision tree. The data set consists of a variety of parameters that are useful for identifying the status of fertility and conducting supervisory training on data sets collected from the agriculture domain to divide information into multiple classes. This paper shows the performance evaluation of three different algorithms like Decision tree, KNN, and SVM. These algorithms were used to train the 0.7 or 70 percent of the input data and are tested with the remaining 0.3 or 30 percent of the test dataset and results of the algorithms were compared based on accuracy and mean square error. Here, the decision tree algorithm is produced the best accuracy of 99%, and also the mean square error for this algorithm is also very less. This article will provide the solution to equip the farmers with the required information that necessary to gain great yield and therefore improving their surplus and consequently will reduce difficulties. In the future, our goal is to analyze an extended soil dataset using artificial Neural Networks in machine learning under different climate conditions for obtaining better prediction with high accuracy.

REFERENCES

1. Arun Kumar, Naveenkumar and Vishal Vats (2018). Efficient crop yield prediction using machine learning algorithms. International Research Journal of Engineering and Technology (IJRET), 5: 3151-3159.
2. Priya, P., Muthaiah, U. and Balamurugan, M. (2018). Predicting yield of the crop using machine learning Algorithm. International Journal of Engineering Sciences & Research Technology, 7: 1-7.

3. Sujatha R. and Isakki, P. (2016). A study on crop yield forecasting using classification techniques. International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), 1-4.
4. Abdullah, A., Brobst, S., Pervaiz, I., Umer, M. and A. Nisar, (2004). Learning dynamics of pesticide abuse through data mining. Proceedings of Australian Workshop on Data mining and Web Intelligence 32: 63-68.
5. Kiran Mai, C., Murali Krishna, I.V. and Venugopal Reddy, A. (2006). Data Mining of Geospatial Database for Agriculture Related Application. Proceedings of Map India, New Delhi, 83-96.
6. Jorquera, H., Perez, R., Cipriano, A. and Acuna, (2001). Short term forecasting of air pollution episodes". In. Zannetti P (eds) Environmental Modeling, pp.221-237.
7. Rajagopalan, B. and Lall, (1999). A k- nearest neighbor daily precipitation and other weather variables". Wat Res Research 35: 3089-3101.
8. Geetha, M.C.S. (2015). Implementation of association rule mining for different soil types in agriculture. International Journal of Advanced Research in Computer and Communication Engineering, 4: 520–522.
9. Verheyen, K., Adrianens, M., Hermy, and Deckers, S. (2001). High-resolution continuous soil classification using morphological soil profile descriptions". Geoderma 101: 31- 48.
10. Jun Wu, Anastasiya Olesnikova, Chi-Hwa Song, and Won Don Lee. (2009). The Development and Application of Decision Tree for Agriculture Data". IITSI, 16-20.
11. Dakshayini Patil and Shirdhonkar, M.S. (2017). Rice Crop Yield Prediction using Data Mining Techniques: An Overview. International Journal of Research in Computer Science and Software Engineering, 7: 427-431.
12. Jayalakshmi, R. and Savithadevi, M. (2019). "Relevance of Machine Learning Algorithms on Soil Fertility Prediction Using R. International Journal of Computational Intelligence and Informatics. 8: 193-199.
13. Jayalakshmi, R. and Savithadevi, M. (2019). Soil Fertility Prediction for Yield Productivity and Identifying the Hidden Factors through Data Mining Techniques. International Journal of Computer Sciences and Engineering, 7: 596-600.

About The License

The text of this article is licensed under a Creative Commons Attribution 4.0 International License