Research & Review: Machine Learning and Cloud Computing

e-ISSN: 2583-4835
Volume-2, Issue-3 (September-December, 2023)

www.matjournals.com

Evaluation of Soil Quality for Crop Prediction Based on Feature Selection in Machine Learning

B Rasina Begum[1*], S Janarthanan[2], B Naveen[3], P Pandeeswaran[4]
[1]Associate Professor, Department of Computer Science and Engineering, Mohamed Sathak Engineering College, Kilakarai, Tamil Nadu, India
[2-4]Under Graduate, Department of Computer Science and Engineering, Mohamed Sathak Engineering College, Kilakarai, Tamil Nadu, India

[*]Corresponding Author: rasinayousuf@gmail.com

Received Date: May 23,2023 *Published Date:*

ABSTRACT

Farmers can practice an effective understanding of the soil's idiosyncrasies allowing for more crops to be grown with fewer resources. Soil prediction relies heavily on calcium, phosphorus, pH, and soil organic carbon. These traits have a substantial impact on crop productivity. The method employs two independent machine learning models, Using KNN and random forest regression; specific soil parameters can be predicted. Crop productivity is boosted as a result of accurate crop prediction. This is where machine learning in the field of crop prediction comes into play. Crop forecast is influenced by geographic, meteorological, and soil characteristics. An integral aspect of the prediction process used by feature selection techniques is choosing the proper features for the right crop or crops. This paper uses categorization approaches to recommend the appropriate crop or crops for the area and conducts a comparative study of multiple wrapper feature selection methods. According to the experimental findings, the adaptive bagging classifier and recursive feature elimination approach surpass the competition. Based on the Africa soil property prediction dataset, the performance of these models is assessed. Knowing the characteristics of the soil in their particular terrain will be useful to the farmers. This study investigates how effectively various machine learning approaches can predict soil qualities crucial to agriculture using spectroscopic data.

Keywords- Artificial neural network, Convolutional neural network (CNN), KNN regression and random forest regression, Machine learning, Mean absolute error (MAE), Mean squared error (MSE)

INTRODUCTION

India has the second-largest population in the world, 1.27 billion people. It is the seventh-largest country in the world, with 3.288 million square kilometres. For the vast majority of Indians, agriculture is their main source of income. It is the primary source of revenue in India. In rural areas, 82% of farmers are small and marginal, while 70% of inhabitants rely primarily on agriculture for their livelihood. In 2020–21, it was estimated that the world would produce 308.65 million tonnes (MT) of food grains. India is the world's biggest producer, user, and importer of pulses, representing 25% of the world's output, 27% of its consumption, and 14% of its imports [1].

With 190 million cattle and the world's second-largest bovine population, India generated 165 MT of milk annually (2017–18), making it the world's top producer of milk, jute, and pulses. India's agriculture only has 4% water resources and 2.4% arable land, which feeds almost 1.3 billion people, putting a tremendous amount of pressure on the environment to maintain output.

India had enormous advances in agricultural productivity after the 1960s-era "green revolution," which was made feasible by modernity. Farmers now have access to more advanced agricultural technologies, such as enhanced seeds (High Yielding Variety seeds), mechanized farm equipment, chemical fertilizers, irrigation infrastructure, and electrical energy, thanks to technical advancements.

High-yielding wheat and rice varieties

were first introduced in India in the 1960s [2] to enhance food output and lower hunger and poverty. Since the "green revolution," which significantly increased crop production, chemical fertilizers have been overused. However, since the overuse of these chemical fertilizers has had a detrimental effect on soil fertility and agricultural productivity, it has become a worry. Because fertilizer recommendations rarely meet soil requirements, these chemical substances are frequently utilised in excess.

Therefore, the farmer needs correct fertilizer recommendations, and the first step in doing so is to accurately analyze the soil parameters. The Indian Agricultural Research Institute (ICAR) suggests using a balanced, integrated approach to nutrient management that is based on soil tests as a means of reducing the use of chemical fertilizers and avoiding negative environmental consequences, groundwater, and soil health.

Objectives

The main objectives of this paper are

- To foresee or identify soil characteristics effectively.
- Putting automated learning techniques to use.
- To improve the classification algorithms' overall performance.
- To assist the farmer in making decisions about the crop's productivity in the future.

RELATED WORK
Literature Survey

Farming is the biggest industry in India produces roughly 14% of India's overall GDP and is important to the socioeconomic development of the nation [3].

Arun Kumar, Balkrishna S. Bhople, and Anil Kumar hypothesised in their study "Indian Agriculture's Prospects: Highly Vulnerable to Massive Unproductivity and Unsustainability" [4] that the country's growth and productivity have been progressively dropping since the green revolution. when inorganic fertilizer consumption increased significantly and peaked at 18.07 million tonnes (mt) of nutrients in 2000. Every year, the soil received roughly the same amounts of nutrients or 16–18 mt. Wheat and rice are the most popular cereal crops in India, and study has shown that their cultivation drew enormous amounts of nutrients from the soil. Recent investigations revealed that nearly 3.7 M hectares of soil had lost much of its organic matter, and there was abundant proof of Land degradation as a result of indiscriminate application of inorganic fertilizers and pesticides. The intensive and continuous application of inorganic fertilizers is a primary source of soil organic matter (SOM) depletion and, as a result, nutrient immobilization. Additionally, general fertilizer recommendations for N, P, and K are followed in India, which rarely correspond to the requirements of the soil for fertility. In various farming systems, secondary and micronutrients are also frequently disregarded. Inorganic fertilizer use is said to have a suppressive effect on SOM mineralization, according to numerous research. The biological relationship between soil carbon and nitrogen is indirect; soil carbon of higher quality and quantity enhances the diversity, abundance, and functions of soil microorganisms, however, there hasn't been enough research done on how long-term fertilization affects carbon mineralization caused by soil microbes.

The author of “Using Soil-Reflected Spectra, A Comparison of Two Regression Methods for Predicting Soil Organic Carbon Content" [5] describes two regression methods for predicting the amount of soil organic carbon in the soil. Sharon Gomes Ribeiro claimed that it is critical to measure the quantity of organic carbon in soil across broad regions to characterize the soil and the repercussions of its treatment. Analytical procedures, on the other hand, can be costly and time-consuming. Reflectance spectroscopy is a standardized process for reviewing the chemical element content of soil. In this work, Hyperspectral remote sensing was used to calculate the amount of soil organic carbon (SOC). The findings came from soil samples obtained in two semi-arid locations in Brazil. The soil samples at wavelengths and spectral reflectance factors were calculated. ranging from 350 to 2500 nm. Normalization, Savitzky-Golay smoothing, and other pre-processing techniques are examples.

According to Biswajit Patra's research, "The soil's mineralogical makeup and the water's C/N level are coastal Odisha, India betel vineyards" [6], betel leaves assist farmers in coastal India as a whole economically. The agricultural soils used in their growth have a considerable impact on the betel leaves' mineralogical makeup. Determining the C/N contents and soil physicochemical characteristics

of vineyards for betel in coastal Odisha was the goal of the current study. In the to investigate their mineralogical composition, samples of the soil and water were taken from P. betel leaf vines in the Odisha districts of Balasore, Ganjam, and Puri. The mineralogy of the soil plays a significant role in understanding the interactions between soil and plants. Numerous techniques are used to determine the soil's mineral and elemental makeup. Several techniques, The elemental and mineralogical composition of soil samples was examined using To investigate their mineralogical composition, samples of the soil and water were taken from P. betel leaf vines in the Odisha districts of Balasore, Ganjam, and Puri. A CHNS analyzer is used to measure the concentrations of hydrogen, nitrogen, carbon, and sulphur. H2O2 (30%) is used to remove organic carbon from bulk soil samples.

EXISTING SYSTEM

Knowing the properties of the soil helps farmers produce more productively and efficiently in the current system, producing more crops with less usage of resources. In this study, machine learning techniques are used to predict the characteristics of the soil. Sand, soil, calcium, phosphorus, pH, soil organic carbon, and the soil itself are the primary factors in soil prediction. These characteristics have a significant impact on crop productivity. Four well-known machine learning models are used to predict these soil properties: support vector machines, gradient-boosting multiple linear regressions, and random forest regression. The dataset for predicting soil qualities in Africa is used to assess the efficiency of these models. According to the results in comparison to other models, they are outperformed by gradient boosting according to the coefficient of determination. Except for phosphate, gradient boosting can accurately forecast all soil characteristics. The problems with the current system include:

- Inability to handle massive volumes of data; and
- Inaccurate classification results.

SYSTEM ARCHITECTURE

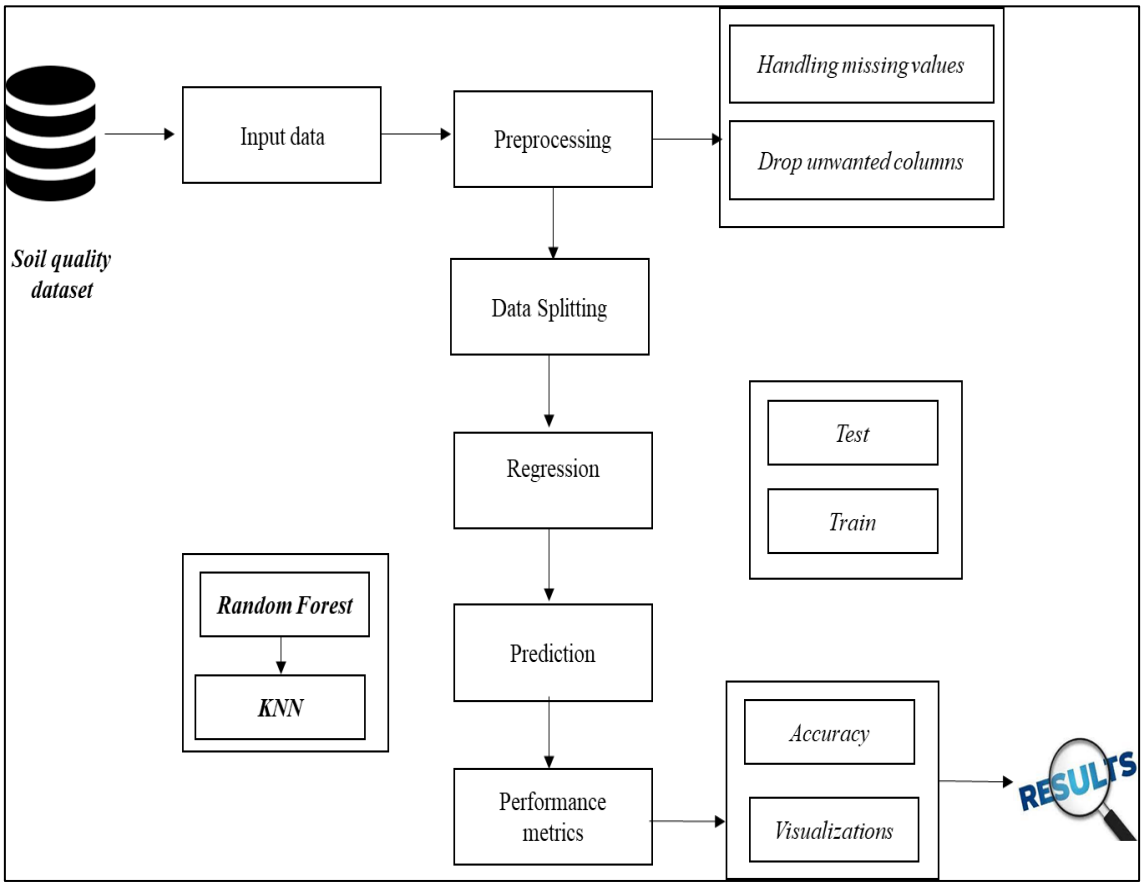

Figure 1: *System architecture.*

The soil properties dataset used in this system is obtained from a dataset repository, such as the UCI repository. After that, the pre-processing stage is carried out. It checks the missing data for incorrect predictions in this stage. The dataset is then divided into test and train, with the test being used for model evaluation and the test being utilized for model prediction. Using ratios, the dataset is divided. Then, it employs KNN regression and random forest regression two examples of machine learning regression algorithms that can be used to predict the qualities of the soil and the quality of the produce. The outcomes of the experiment then demonstrate the accuracy of certain performance measurements, including MAE and MSE. The proposed system makes use of the following modules as shown in Fig. 1.

- Data selection
- Data preprocessing
- Data splitting
- Feature selection
- Classification
- Result Generation

DATA SELECTION

- This technique uses the soil property dataset as its input data, which is obtained from a dataset repository. The data selection approach is used to predict the characteristics of the soil.
- The effectiveness of machine learning models is evaluated using measurements from 1,886 soil samples. The soil was gathered from several places across Africa. A data point has 3,594 different properties in total. PIDN: unique identification for soil samples
- SOC stands for soil organic carbon.
- pH: pH levels
- Ca: Phosphorus extractable from Mehlich-3
- Sand: Sand content • m7497.96 - m599.76: 3,578 mid-infrared absorbance values are available. The "m599.76" column, for example, represents the absorbance at wavenumber 599.76 cm-1.
- Depth: Soil sample depth (Topsoil and Subsoil categories).
- BSA: Long-term average Black Sky Albedo values obtained from MODIS satellite photographs (BSAN = near-infrared, BSAS = shortwave, and BSAV = visible).
- CTI: Compound topographic index derived from elevation data from the Shuttle Radar Topography Mission.
- ELEV: elevation data from the Shuttle Radar Topography Mission.
- EVI: MODIS satellite pictures' average long-term Enhanced Vegetation Index.

DATA PREPROCESSING

Superfluous data are removed from a dataset during pre-processing. Using pre-processing data transformation techniques, the dataset is formatted in a way that is suitable for machine learning. Superfluous data are removed from a dataset during pre-processing. Pre-processing data transformation techniques are used to convert the dataset into a machine learning-friendly format

- Remove any null values, including missing and non-values, from missing data, which are converted to a value of 0 at this stage.
- All anomalies, missing values, and duplicate values are eliminated from the data.
- Encoding Categorical Data: Variables with a limited number of label values are referred to as categorical data.
- Be aware that the bulk of machine learning methods prefer numerical input.

DATA SPLITTING

- The machine learning process needs data for learning to take place.
- In addition to the data required for training, test data are also necessary to evaluate the algorithm's performance and determine its efficacy. In this approach, 70% of the dataset is regarded as training data, and the other 30% is testing data.
- The technique of splitting accessible data into two sections is known as data splitting, and it is often done for cross-validation purposes.
- One set of data is used to build a prediction model, while the other is used to evaluate the effectiveness of the model.
- Data sets are divided into training and testing sets as part of the analysis of data mining models; the bulk of the data in each set is used for training, while the remaining portion is utilised for testing.

FEATURE EXTRACTION

- In machine learning, Principal Component Analysis, a type of unsupervised learning, is used to make dimensions smaller. It is a statistical procedure that converts observations with correlated qualities into a set of linearly uncorrelated data via orthogonal transformation.
- The newly altered features are referred to as Principal Components. One of the tools that is frequently used this one is used for exploratory data analysis and predictive modelling. It is a method for finding important patterns in the given dataset by bringing the variances down.
- When projecting high-dimensional data, PCA often seeks out the surface with the lowest dimensionality. As a high attribute indicates a strong divide between groups, PCA works by taking into account each attribute's variance, which reduces the dimensionality.
- PCA can be applied in practical contexts for tasks including image processing, movie recommendation systems, and optimizing power distribution over a range of communication channels. Because it employs a feature extraction technique, it preserves the useful variables and ignores the irrelevant ones.

CLASSIFICATION

- In this process, it implements approaches for machine learning such as KNN regression and random forest regression [7].
- Unexpectedly in the Forest Using ensemble learning, regression is a form of supervised learning. Ensemble learning combines predictions from several machine learning algorithms to create predictions that are more accurate than those from a single model.
- KNN regression is a non-parametric technique that approximates the relationship between independent variables and continuous output by averaging data from the same neighbourhood.

RESULT GENERATION

The Final Result will be determined based on the overall categorization and projection. Some metrics, such as are used to gauge the efficacy of this advised course of action.

MAE: In statistics, the mean absolute error (MAE) is a metric used to assess a given model's precision [8]. It is calculated as:

MAE = (1/n) * Σ|yi – xi|

Where:

Σ: A Greek symbol that means “sum”

yi: The observed value for the i^{th} observation

xi: The predicted value for the i^{th} observation

n: The total number of observations

MSE: The mean squared error (MSE) is a common way to measure the prediction accuracy of a model. It is calculated as:

MSE = (1/n) * Σ (actual – prediction) ^{2}

Where:

Σ – a fancy symbol that means “sum”

n – sample size

actual – the actual data value

prediction – the predicted data value

CLASSIFICATION

Convolutional Neural Network

Convolutional neural networks (CNNs) are a popular Deep Learning neural network architecture in computer vision [9]. Computer vision, a subfield of artificial intelligence, enables computers to perceive and evaluate visual data or images. In machine learning, artificial neural networks perform amazingly well. Neural networks are utilised in numerous datasets, including those with images, audio, and text. Recurrent neural networks, more precisely an LSTM, are used, for instance, to forecast the order of words. Different neural network types are employed for various purposes. Similar approaches to convolutional neural networks are utilised for picture classification.

Artificial Neural Network

Artificial neural networks are made up of artificial neurons, which are also known as units. These units, which are layered in various layers, comprise a system's entire Artificial Neural Network [10]. The employment of advanced neural networks to uncover the hidden patterns in the dataset will define the number of units in a layer, whether it be a dozen or millions. Input, output, and output layers are

typically combined with hidden layers in artificial neural networks. The neural network receives information from the outside world that it must analyse or learn. This data is changed into useful information for the output layer after going via one or more hidden layers. The output layer represents the Artificial Neural Networks' reaction to the incoming data.

Most neural networks have connections between the units at different layers. The weights in each of these linkages regulate how much one unit influences another. As it advances from one unit to another, the neural network gains more knowledge about the data, finally resulting in an output from the output layer.

PERFORMANCE METRICS

Accuracy

The classifier's accuracy measures how frequently It makes precise forecasts. Accuracy is defined as the ratio of the number of correct forecasts to all predictions.

Precision

Precision explains the proportion of precisely predicted cases that occurred. When False Positives are more problematic than False Negatives, precision is helpful.

F1 Score

It provides a synthesis of the Precision and Recall measurements. When Precision and Recall are the same, it is high.

Recall

The recall is the percentage of real positive cases that our model was correctly able to predict.

TECHNOLOGIES PREFERRED FOR THE PROPOSED WORK

- **Sensors:** soil, Crop, humidity management, Temperature Sensor, Ph Sensor, NPK Sensor
- **Software:** Former Interaction Software, Mobile App
- **Connectivity:** Cellular, LoRa
- **Location:** GPS, Satellite
- **Data storage and Maintenance:** Cloud Computing Vendor
- **Data Analytics:** Standalone analytics programmes, and downstream applications for data pipelines

CONCLUSION AND FUTURE ENHANCEMENT

The dataset of soil properties is used as the input. It is utilised in the preprocessing stage to prevent incorrect prediction. Because this is a time series dataset, machine learning algorithms like KNN regression and random forest regression are used for better performance. The experimental findings demonstrate the high Accuracy of performance indicators like MAE and MSE.

The use of two different machine learning algorithms is possible in the future. It is also feasible to add to or modify the proposed clustering and classification methods to further improve efficiency. Other clustering algorithm combinations can be utilised to increase the detection accuracy in addition to the tried-and-true combination of data mining methodologies.

REFERENCES

1. J Lim (2021), “The applications of AI in India’s agriculture industry”, [Online] Available at: https://techwireasia.com/2021/09/ai-applications-will-shape-the-future-of-indias-agriculture-industry/ [Accessed on September 2021].
2. A Raeboline Lincy Eliazer Nelson, K Ravichandran and U Antony (2019). The impact of the green revolution on indigenous crops of India, *Journal of Ethnic Foods*, 6, Available at: https://doi.org/10.1186/s42779-019-0011-9.
3. Unacademy, “Role of agriculture in Indian Economy”, [Online] Available at: https://unacademy.com/content/railway-exam/study-material/geography/role-of-agriculture-in-indian-economy/#:~:text=Agriculture%20provided%20the%20first%20motivation%20for%20industrialization%20in,all%20economies%2C%20regardless%20of%20their%20degree%20of%20development.

4. A Kumar, B S. Bhople and A Kumar (2020). Prospective of Indian agriculture: highly vulnerable to huge unproductivity and unsustainability, *Current Science*, 119(7), 1079-1080, Available at: https://currentscience.ac.in/Volumes/119/07/1079.pdf.
5. S Gomes Ribeiro, A dos Santos Teixeira, M Regys Rabelo de Oliveira, et al (2021). Soil organic carbon content prediction using soil-reflected spectra: A comparison of two regression methods, *Remote Sensing*, 13(23), Available at: https://doi.org/10.3390/rs13234752.
6. B Patra, R Pal, R. Paulraj, et al (2020). Mineralogical composition and C/N contents in soil and water among betel vineyards of coastal Odisha, India, *SN Applied Sciences*, 2, Available at: https://doi.org/10.1007/s42452-020-2631-5.
7. D Harshitha Challa (2023), "Crop Prediction Based on Soil Classification using Machine Learning with Classifier Ensembling", [Online] Available at: https://eecs.ku.edu/crop-prediction-based-soil-classification-using-machine-learning-classifier-ensembling#:~:text=A%20comparative%20analysis%20of%20several%20popular%20classification%20algorithms%2C,on%20the%20characteristics%20of%20the%20soil%20and%20environment [Accessed on May 2023].
8. P Schneider and F Xhafa (2022), Anomaly Detection and Complex Event Processing Over IoT Data Streams: With Application to eHealth and Patient Data Monitoring, *1st Edition*. *Academic Press*, Cambridge, Massachusetts, USA. ISBN-10: 0128238186, Available at: https://www.amazon.com/Anomaly-Detection-Complex-Processing-Streams/dp/0128238186.
9. R Yamashita, M Nishio, R Kinh Gian Do and K Togashi (2018). Convolutional neural networks: An overview and application in radiology, *Insights into Imaging*, 9(4), 611-629, Available at: https://doi.org/10.1007/s13244-018-0639-9.
10. N Malik (2005). Artificial neural networks and their applications. *National Conference on 'Unearthing Technological Developments & their Transfer for Serving Masses*. GLA ITM, Available at: https://arxiv.org/ftp/cs/papers/0505/0505019.pdf.

CITE THIS ARTICLE

B Rasina Begum , et al. (2023). Evaluation of Soil Quality for Crop Prediction Based on Feature Selection in Machine Learning, *Research & Review: Machine Learning and Cloud Computing*, 2(3), 1-7.