Performance Analysis and Evaluation of Crop Yield Prediction from Soil using Supervised Machine Learning Techniques

Abstract

This study emphasizes how important soil is in forecasting crop performance. Nutrient analysis can help farmers and soil scientists increase crop output by utilizing cutting-edge techniques. The research focuses on predicting mustard crop production using Kaggle's soil dataset utilizing multiple data prediction techniques. The study evaluates accuracy, recall, precision, and F-score using Decision Trees and SVM, two supervised machine learning techniques. The study finds that the best methods for precisely predicting mustard crop productivity are Logistic Regression, Decision Trees, and SVM.

Keywords: Dataset, Logistic Regression, Decision Tree, SVM

Dr. Jitendra Singh Kushwah
Associate Professor
Institute of Technology and Management,
Gwalior-MP, India
jitendra.singhkushwah@itmgoi.in

Saurabh Shrivastava
Assistant Professor
Institute of Technology and Management,
Gwalior-MP, India
saurabhshrivastava.it@itmgoi.in

Manali Singh
Assistant Professor
Institute of Technology and Management,
Gwalior-MP, India
manali.singh.it@itmgoi.in

Neeraj Gaur
Assistant Professor
Institute of Technology and Management,
Gwalior-MP, India
neerajgaur.it@itmgoi.in

I. Introduction

Self-learning via growth and experience is facilitated by supervised learning. It entails using predetermined models to collect particular inputs and custom algorithms to produce the desired outcomes. Machine learning is used in agriculture to improve crop production and quality by discovering the best conditions for a certain crop using these algorithms.

Crop yield projections are influenced by a number of variables, such as soil type, climate, pests, and terrain [1]. The combination of minerals, water, gases, and living creatures in the soil, each with unique qualities, has a substantial impact on plant growth. For fertility,

productivity, and environmental protection, maintaining soil health is essential. Farmers may choose crops that are best for a certain environment by analyzing the nutrients in the soil. It's essential to identify dietary deficits and keep an eye on changes. For sustained food production and good output, effective management of soil, plants, water, nutrients, and energy is crucial [2].

1.1 Logistic Regression

In supervised learning case studies [12], logistic regression is used to estimate the likelihood of binary (yes/no) occurrences. Using machine learning to anticipate the existence of COVID-19 is an example. Binary logistic regression is frequently used to make yes-or-no decisions. Its practical uses may be found in a variety of fields, including healthcare, where it helps to discriminate between benign and malignant tumors.

- In the financial industry, it finds fraudulent transactions.
- It predicts target audience reactions in marketing.

1.2 Decision Tree

A decision tree [12] creates rules for data classification using characteristics and classes. Decision trees are simple to understand and need little data preparation. They may represent both numerical and categorical data. Intricate trees, however, often lack generalizability and stability as a result of little input changes, resulting in completely different trees.

1.3 Support Vector Machine

Training data are represented by support vector machines as points in space that are divided into groups by a maximum-wide gap. The same region is projected with new samples, which are then categorized based on gap-side location. Utilizing a small portion of the training points allows the decision function to be successful in high-dimensional areas while using little memory.

1.4 Comparison Parameters

Accuracy

Accuracy is defined as the proportion of correctly predicted observations to all observations. Accuracy is the logicalest performance metric.
Accuracy = (True Positive + True Negative) / Total Population
True Positive: The number of correct predictions that the occurrence is positive.
True Negative: The number of correct predictions that the occurrence is negative.

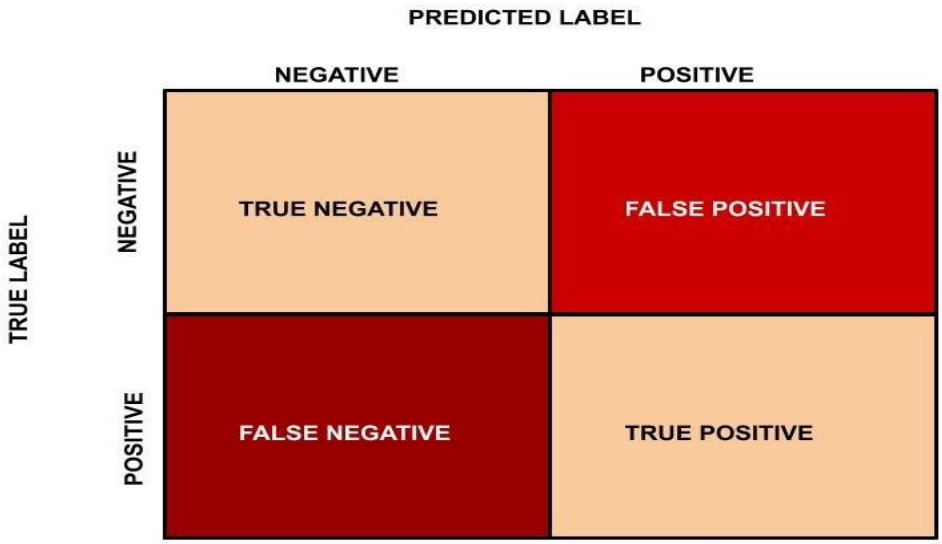

Figure 1: Confusion Matrix

F1-Score

Most people consider the F1-Score, which combines memory and accuracy, to be their harmonic mean. The harmonic mean, which performs better for ratios like recall and accuracy than the traditional arithmetic mean, is used to compute averages. As (2 x Precision x Recall) / (Precision + Recall), it is computed. False positives and false negatives are included in the F1-Score, a weighted average of Precision and Recall in several classification algorithms. The F1-Score frequently performs better than accuracy and is especially useful when class distribution is unequal.

Precision

Precision measures the ratio of positively classified positive samples—including both accurate and inaccurate classifications—to all positively classified positive samples (True Positive). The dependability of a machine learning model is evaluated using the criteria of precision. A classifier that is working well displays a high precision score of 1.
Precision = True Positive/True Positive + False Positive
Precision = TP/TP+FP

Recall

Recall is calculated as the proportion of appropriately labeled Positives in the ratio of correctly recognized Positive samples to all Positive samples. It measures how well the model is able to recognize positive samples, with higher recall suggesting better performance.

Recall = True Positive/True Positive + False Negative
Recall = TP/TP+FN

II. Literature Review

Researchers have lately used machine learning techniques in the agriculture industry. The review that follows summarizes current uses of several machine learning (ML) techniques for predicting agricultural yields using soil data.

[3] states that they suggested using three different classification algorithms—J48, Naive Bayes, and Random Forest—to estimate soil fertility using a soil dataset. The J48 algorithm stood out among them with a remarkable effectiveness of 98.17%.

Two years later, [4] performed a comparison of the Naive Bayes, JRIP, and J48 ML algorithms for predicting soil type. R was used to analyze the 110 soil samples included in the study. Notably, the JRIP algorithm performed better than the competition, with a kappa score of almost 1.0 and a greater accuracy of 98.18%. [5] suggested a model for crop production estimation using data mining techniques in order to increase the value and profitability of agricultural land in the same year. [6] carried out a comparison research using a data analytics tool, contrasting Support Vector Machine with Artificial Neural Network. The two data mining techniques were evaluated using seven hidden node networks.

The same year, using macro- and micronutrient conditions found in the dataset, a variety of machine learning approaches [7] were used to forecast rice crop production categories. KNN, Naive Bayes, and ANN were used to analyze soil data. For soil classification and yield projection based on soil nutrient status, Decision Tree Classifier and Naive Bayes Classifier were shown to be superior models.

A data mining-driven approach for agricultural yield was presented by [8] for the projection of soil dataset categorisation. Data mining technologies and the use of K-Nearest Neighbor and Naive Bayes algorithms made it easier to create and assess system architecture. Additionally, [9] proposed a method for predicting wheat yield using fuzzy c-means clustering and neural networks.

The Random Forest approach was used by [10] to predict agricultural productivity in the state of Tamil Nadu in the same year. The dataset for the study included elements including rainfall, temperature, and agricultural output. R Studio was used to carry out the study.The Random Forest approach was used by [10] to predict agricultural productivity in the state of Tamil Nadu in the same year. The dataset for the study included elements including rainfall, temperature, and agricultural output. R Studio was used to carry out the study.

The yields of the sugarcane crops for the following year were predicted [11] using machine learning methods such KNN, SVM, and Decision Tree. In the research, Python was used, and decision tree forecasts showed the lowest mean square error and 99% accuracy rate[12] estimated soil fertility using machine learning techniques.

III. Methodology and Result Analysis

The objective of the ongoing study is to forecast mustard crop output using soil data and machine learning methods. To do this, experiments were run using the MatLab platform. The study problem was established after gaining a knowledge of the topic, speaking with farmers and soil specialists, and examining the literature. This study's actual data came from Kaggle.

The dataset has 620 occurrences of an output property reflecting soil nutrient status and 11 input factors, including Class Label. The nutrients in the soil are denoted by parameters like Ph (soil pH value), EC (electrical conductivity), OC (organic carbon), N (nitrogen), P (phosphorus), K (potassium), S (sulfur), Cu (copper), Fe (iron), Zn (zinc), and Mn (manganese). Three types of mustard crop production are covered by the output attribute: grapes, mango, mulberry, pomegranate, potato, and ragi. Out of the entire 620 samples, 104 are classified as ragi, 100 as potato, 104 as mango, 104 as grape, 104 as mulberry, and 104 as grape.

The dataset should be split into two halves—training data and test data—to train the model using classification techniques [13]. In this study, Logistic Regression, Decision Tree, and Support Vector Machine classification techniques are explored, taking into account metrics like Accuracy, Precision, Recall, and F1-Score.

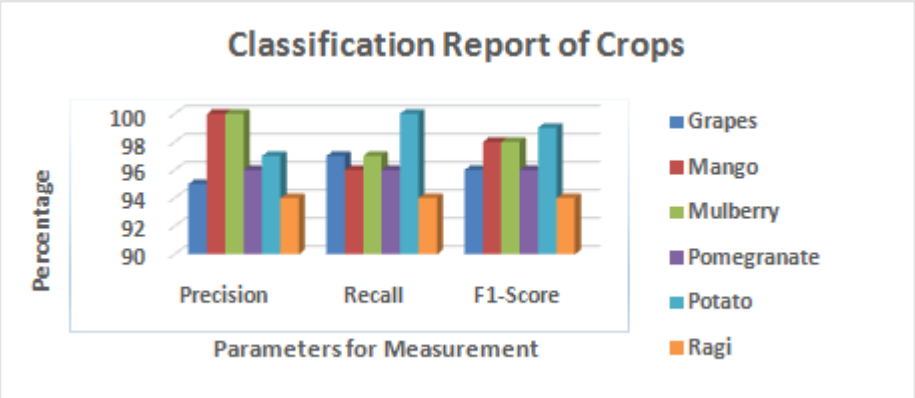

Figure 2: Measurement of Parameters using Logistic Regression

Fig. 2 shows the comparison of crops based on metrics Precision, Recall, and F1-Score. It is shown that the Logistic Regression algorithm finds better results for Mulberry, and Mango crops of Precision.

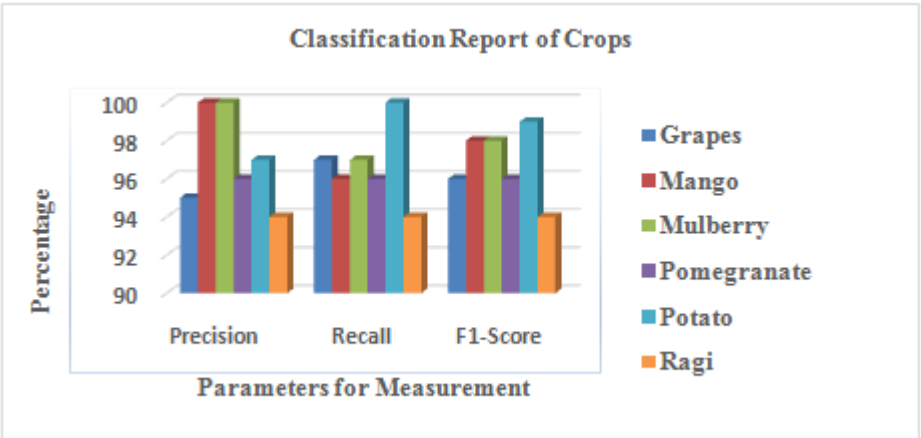

Figure 3: Measurement of Parameters using Decision Tree

Fig. 3 shows the comparison of crops based on metrics Precision, Recall, and F1-Score. It is shown that the Decision Tree algorithm finds the better result for Mulberry, and Mango crops of Precision. Crop Grapes finds the better result of Recall and F1-Score.

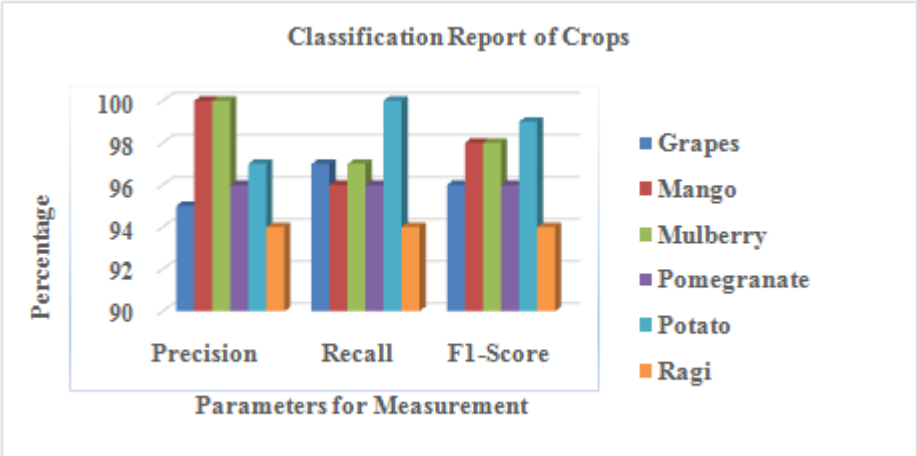

Figure 4: Measurement of Parameters using Support Vector Machine

Fig. 4 shows the comparison of crops based on metrics Precision, Recall, and F1-Score. It is shown that the Support Vector Machine algorithm finds the better result for Mulberry, and Mango crops of Precision. Crop Grapes finds the better result of Recall and F1-Score.

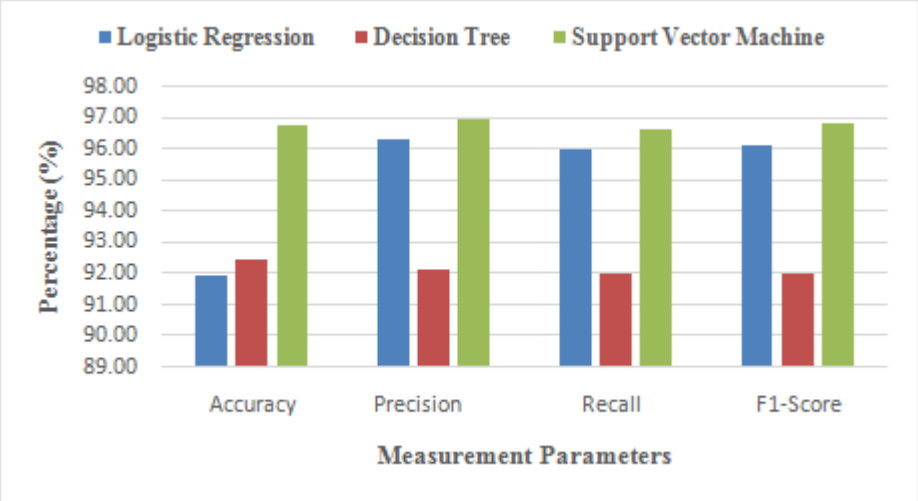

Figure 5: Classification Algorithm Analysis using Metrics

Fig. 5 shows the comparison of classification algorithms based on metrics like Accuracy, Precision, F1-Score, and Recall. It is shown that the Support Vector Machine algorithm finds better Accuracy than other algorithms. It is also shown that SVM finds better Precision, Recall, and F1-Score than other algorithms as per table 1.

Table 1: Analysis of Classification Algorithms

Classification Algorithms	Accuracy	Precision	Recall	F1-Score
Logistic Regression	91.93%	96.33%	96%	96.16%
Decision Tree	92.47%	92.16%	92%	92%
Support Vector Machine	96.77%	97%	96.66%	96.83%

IV. Conclusion

The experimental study emphasizes how useful machine learning techniques like logistic regression, decision trees, and support vector machines are for forecasting mustard crop yield. For predicting mustard crop production, Decision Tree and SVM emerged as the most dependable algorithms. Farmers now have the option to forecast productivity depending on soil parameters thanks to these reliable algorithms. In the future, it appears possible to estimate agricultural yields utilizing substantial soil data in a big data setting. In addition, using crop yield projections to propose fertilizers can help farmers and soil specialists make judgments when faced with a range of crop production estimates.

References

[1] Haneet Kour, Surjeet Singh, Jatinder Manhas, Vaishali Pandith, and Vinod Sharma "Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis" Banaras Hindu University, Varanasi, India, Journal of Scientific Research Institute of Science, Volume 64, Issue 2, 2020.

[2] (2015) Smriti. A Review of the Machine Learning Approach for Soil Property Detection. Science International Journal Online, 4.

[3] V. Bhuyar (2014). Comparative analysis of classification methods used to soil data to forecast fertility rates in the district of Aurangabad. Computer Science International Journal of Emerging Trends & Technology, 3(2), 200–203.

[4] (2016). Rajeshwari, V., and Arunesh, K. Using data mining classification techniques to analyze soil data. 9, 1-4 of the Indian Journal of Science and Technology.

[5] R. Sujatha (2016). A study on the use of classification techniques to forecast crop production. IEEE.

[6] N. Awasthi and A. Bansal (2017). Use of R to Apply Data Mining Classification Techniques to Soil Data. Electronics and Computer Science International Journal, 4, 33–37.

[7] (2017). Singh, V., Sarwar, A., and Sharma, V. Using a machine learning method, the soil is analyzed and crop production (in this case, rice) is predicted. 8(5), 1254–1259, International Journal of Advanced Research in Computer Science.

[8] D. M. Supriya (2017). Using a data mining method, soil behavior is analyzed and crop yield is predicted. 5, 9648–9652, International Journal of Innovative Research in Computer and Communication Engineering.

[9] (2017). Verma, A., Jatain, A., and Bajaj, S. Using Fuzzy C Means Clustering and a Neural Network, predict wheat crop yield. 13(11), 9816-9821, International Journal of Applied Engineering Research.

[10] Muthaiah, U., Priya, P., and Balamurugan, M. (2018). Crop Yield Prediction using Machine Learning Algorithm. Global Journal of Engineering Sciences and Research Technology, 1–7.

[11] The article "A Comprehensive System for Detecting Profound Tiredness for Automobile Drivers Using a CNN" by Jitendra Singh Kushwah et al. Lecture Notes in Electrical Engineering (LNEE, volume 914), print ISBN: 978-981-19-2979-3, online ISBN: 978-981-19-2980-9, Proceedings of ICACIT 2022, Springer, Singapore, 31 August 2022.

[12] Jitendra Singh Kushwah et al. "Comparative study of regressor and classifier with decision tree using modern tools" in First International Conference on Design and Materials (ICDM)-2021 of "Materials Today: Proceedings" ISSN: 2214-7853, Volume 56, Part 6, 2022, Pg. 3571-3576.

[13] Jitendra Singh Kushwah et al. "Analysis and visualization of proxy caching using LRU, AVL tree, and BST with supervised machine learning" in 1st International Conference on Computations in Materials and Applied Engineering-2020 of "Materials Today: Proceedings" ISSN: 2214-7853 DOI: 10.1016/j.matpr.2021.06.224.