Comparative Analysis of Machine Learning Algorithms for Crop Recommendations

Chinmay Vyapari, Prathamesh Bhosale and Ameya Parkar

July 19, 2023

Comparative Analysis of Machine Learning Algorithms for Crop Recommendations

Chinmay Vyapari
Department of MCA
Vivekanand Education Society
Institute of Technology, Mumbai

Prathamesh Bhosale
Department of MCA
Vivekanand Education Society
Institute of Technology, Mumbai

Prof. Ameya Parkar
Department of MCA
Vivekanand Education Society
Institute of Technology, Mumbai

***Abstract* — India's economy heavily relies on agriculture, which serves as a crucial industry for its growth and survival. The country is recognized as one of the largest producers of various agricultural products. Soil, an essential component of crop cultivation and non-renewable natural resource, plays a pivotal role in sustaining life. Traditionally, farmers utilized their experience to select the most appropriate crops based on soil characteristics. However, the current agricultural landscape calls for a recommendation system utilizing machine learning algorithms to determine the ideal crop for specific type of soil.**

Crop recommendation systems have gained significant attention in the field of agriculture, aiding farmers in making informed decisions about suitable crops for their specific conditions.

1. INTRODUCTION

Agriculture is the backbone of the Indian economy and a vital occupation in the country. With more than 60% of its land dedicated to agriculture, India feeds its massive population of 1.5+ billion people. The cultivation of animals and plants forms the basis of agriculture, which has been crucial to India's rise as a civilization. The soil plays a vital role in agriculture as it supplies essential nutrients, oxygen, water, and structural support to plant roots. It serves as a crucial element for achieving high-quality food production and forms the bedrock of the entire food system. India boasts a diverse range of soil types, including alluvial soil, black soil, red soil, and laterite soil, each of which is well-suited for specific crops. Alluvial soil proves favorable for cotton and rice cultivation, black soil is preferred for sugarcane and sunflower, red soil is suited for corn and ragi, while laterite soil provides an ideal environment for the growth of pulses, tea, and coffee, among others.

Numerous research studies have been conducted to enhance agricultural planning, which involves leveraging machine learning techniques for crop recommendation. A component of artificial intelligence, machine learning, allows machine to mimic intelligent human behaviour. Artificial intelligence systems carry out intricate tasks using methods akin to those employed by humans, but with the added advantage of automation. The process of machine learning commences with data, which can encompass diverse forms such as financial transactions, individual information, or even images.

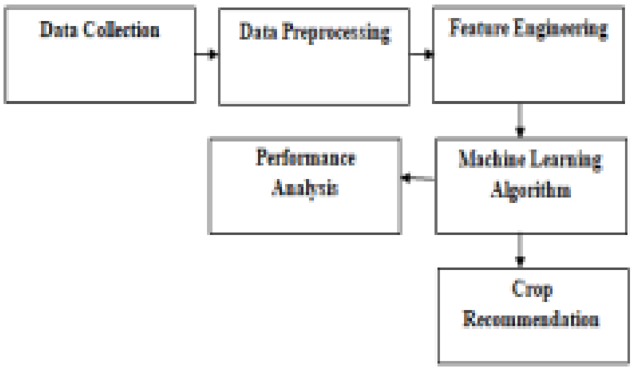
Incorporating machine learning into agricultural planning entails an initial phase of gathering and processing relevant information as training data for the system. It is widely recognized that a larger volume of data can contribute to improved outcomes. Subsequently, a suitable machine learning model is selected, and the collected data is used to train the system to identify patterns and autonomously generate predictions.

2 REVIEW of LITERATURE

Research papers on the effectiveness of machine learning algorithms for crop yield prediction have been conducted by various authors. May authors explored using machine learning algorithms to assist farmers in making informed decisions about selecting crops based on geographic and environmental factors. The research study incorporated several algorithms, namely decision trees, KNNs, Random Forests, and neural networks, to achieve its objectives. Among these algorithms, the neural network consistently demonstrated the highest level of precision across all three studies. Mayank Champaneri concentrated on crop yield prediction through the application of a data mining technique, specifically employing a random forest classifier for both classification and regression tasks. The outcome of their work was the development of a user-friendly website that utilizes climate data to forecast crop yield for a chosen crop. The random forest algorithm exhibited an impressive accuracy rate of 95% in predicting crop yield.

3. PROPOSED SYSTEM

Based on specific soil factors, the proposed approach will suggest the most suitable crop for cultivation.



Fig

2. Methodology of crop recommended system.

As shown in fig. 1, the technique of the proposed system is composed of multiple blocks.

Data Collection : Data collection is a widely used and effective approach for obtaining and evaluating information from various sources. To provide the system with a comprehensive dataset that enables accurate crop recommendations, the following characteristics should be included and considered:

Soil: Information about the soil type, composition, and properties is essential as different crops thrive in different soil conditions.

pH: The pH level of the soil is crucial as it affects nutrient availability and overall crop growth.

Temperature and Humidity: These environmental factors play a significant role in determining the suitability of crops for specific regions or seasons.

NPK Levels: The levels of nitrogen (N), phosphorus (P), and potassium (K) in the soil impact plant nutrition and crop productivity. Monitoring these levels is essential for effective crop recommendations.

Crop Information: Data related to specific crop characteristics, such as growth cycle, water requirements, disease susceptibility, and yield potential, help in understanding crop suitability and making appropriate recommendations.

Data Pre-Processing : There are several methods involved in data pre-processing, starting from reading the collected dataset and progressing to data cleansing. During this process, certain dataset properties that are irrelevant or redundant for crop recommendation are excluded. Additionally, datasets that contain missing data need to be handled appropriately. The missing values can be either removed or replaced with specific undesirable values, such as "nan," to ensure higher accuracy in the subsequent analysis and modeling stages. By effectively pre-processing the data, we can ensure that it is in a suitable format for training the model and generating accurate crop recommendations.

Feature Engineering : Feature engineering is a process that involves creating new features or transforming existing ones from raw data using domain knowledge. The objective of feature engineering is to enhance the quality and effectiveness of machine learning outputs by incorporating additional relevant features.

Training set : The training set is a specific type of dataset that contains labeled data, where both input vectors and corresponding output values are available. In supervised machine learning, the training set is used to train the model using various techniques. By utilizing the labeled data in the training set, the model learns patterns and relationships between input features and output values, allowing it to make predictions or classifications on unseen data.

Testing set : A testing set, also known as a validation set or a holdout set, is a dataset that does not contain any labeled or tagged information. It is used to evaluate the performance of a trained machine learning model and assess its predictive capabilities on unseen data.

The testing set is distinct from the training set and is kept separate to ensure that the model's performance is not biased or overfitted to the training data. Once the model has been trained on the labeled data from the training set, it is then applied to the testing set to make predictions or classifications based on the learned patterns and relationships.

By evaluating the model's performance on the testing set, we can assess its generalization ability and determine how well it performs on new, unseen data. This evaluation helps to validate the effectiveness and reliability of the trained model before deploying it in real-world scenarios.

Machine Learning Algorithms: Machine learning prediction methods need for exceedingly precise estimation based on prior knowledge. Data, statistical techniques, and machine learning strategies are used in predictive analytics past knowledge to predict future outcomes. Beyond merely comprehending what occurred, the objective is to offer the greatest solution possible and a forecast of what will happen next.

The methods utilised in this model are Logistic regression, Naive Bayes, Random Forest, Decision Tree and Gradient Boosting.

A. Logistic Regression : *The logistic model in statistics is a statistical model that depicts the likelihood that an event will occur by making the event's log-odds a linear combination of one or more independent variables. In regression analysis, logistic regression is used to estimate a logistic model's parameters.*

The generalised linear model's core equation is

$$g(E(y)) = \alpha + \beta x1 + \gamma x2$$

Here, the link function g(), the target variable E(y), and the linear predictor (to be forecasted) are all present. The purpose of the link function is to 'connect' the linear predictor to the expectation of y.

B. Naive Bayes :
The Bayes theorem was used to construct Naive Bayes, a simple and uncomplicated probabilistic classifier. The value of one feature, given the class variable, is assumed to be independent of the value of any other feature by Naive Bayes classifiers.

$P(A|B) = (P(B | A) * P(A)) / P(B)$

C. Decision Trees : In supervised learning, Decision Trees (DT) are used for classification and regression. The problem is handled using a tree representation, with each leaf node representing a class label. and each node inside the tree represents a characteristic.
Entropy:

$H(S) = -\sum Pi(S) \log2 Pi(S)$ Information Gain:

$IG(S,A) = H(S) - \sum v € Values(A)(| Sv |/S) H(Sv)$

D. Random Forest :
To solve classification, regression, and other issues, a huge number of unique models are generated using the ensemble learning approach known as Random Forest. Throughout the training process, decision trees are used. The random forest method generates decision trees from a large number of data samples, predicts data from each subset, and votes on it to deliver a better response to the system. RF uses the bagging strategy for data training, which improves the accuracy of the results.. Formula :

∞
Gini Index =1-∑ (Pi) 2
i = 1
= 1 - [(P+) 2 + (P-) 2]

E. Support Vector Machines :
The SVM method's purpose is to determine the best line or decision boundary that can divide n-dimensional space into classes, allowing us to classify fresh data points quickly in the future. This ideal decision boundary is known as a hyperplane.
SVM chooses the extreme vectors and points that will help create the hyperplane. The SVM approach is built on support vectors, which are utilised to represent these extreme cases. Here, we employ linear SVM.
The hyperplane equation used to classify the points is as follows:

$H : wT (x) + b = 0$

F. Extreme Gradient Boosting :
Gradient Boosting is supervised ML method for dealing with classification and regression issues. It's a mediocre ensemble prediction model. The same as with earlier boosting strategies, an incremental gradient-boosted trees model is built.

Crop Recommendation: The crop recommendation model utilizes soil characteristics such as NPK levels, temperature, humidity, and pH to suggest the most suitable crop for cultivation. By analyzing these factors, the model provides recommendations that align with the specific requirements and preferences of the given soil conditions. This helps farmers make informed decisions about crop selection, maximizing their chances of achieving higher yields and profitability.

Performance Analysis : Performance analysis is a specialized discipline that focuses on improving productivity and decision-making through a systematic approach. It involves evaluating various aspects of a system, process, or activity to identify strengths, weaknesses, and areas for improvement. In the context of agriculture, performance analysis can involve assessing the efficiency and effectiveness of farming practices, crop management strategies, resource allocation, and decision-making frameworks.

4. RESULT ANALYSIS :

The proposed model makes use of soil parameters and a crop database. The ideal crop for the specific soil is suggested by machine learning algorithms. Gradient boosting was the method we tried with the most accuracy. The accuracy ratings for each algorithm are listed below.

TABLE 1. ACCURACY OF THE ALGORITHMS

Algorithms	Accuracy
Logistic Regression	97.12
Decision Tree	95.51
Naïve Bayes	96.08
Random Forest	98.56
SVM	94.5
Gradient Boosting	98

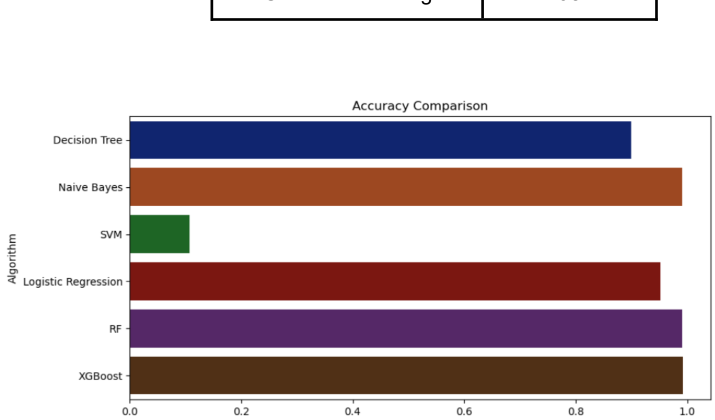


Figure 2. Comparison plot of Accuracies.

V. CONCLUSION

A significant amount of agricultural research has been conducted and continues to be conducted with the aim of increasing crop production, bolstering the Indian economy, and, most importantly, helping farmers improve their income. One promising approach to benefit farmers is the use of crop recommendation systems. These systems provide farmers with valuable insights and information about the optimal crops to cultivate on their specific land, taking into account various factors such as soil conditions, climate, market demand, and profitability.

By utilizing crop recommendation systems, farmers can make informed decisions about which crops to grow, maximizing their productivity and profitability. These systems consider a range of data, including historical crop performance, soil analysis, weather patterns, and market trends, to provide personalized recommendations tailored to the unique characteristics of each farm.

The ultimate goal of implementing such systems is to empower farmers, enabling them to make well-informed choices that result in higher crop yields, reduced risks, improved resource utilization, and ultimately, increased income. By leveraging the benefits of agricultural research and technology, these crop recommendation systems aim to contribute to the overall growth and prosperity of the agricultural sector while improving the livelihoods of farmers across India.

REFERENCES

1 . "Crop Recommendation System for Precision Agriculture using Machine Learning Techniques" by Anitha P, Linganagouda Kulkarni, and Pramod Kumar P.

2 . "Crop Recommendation System using Deep Learning and Internet of Things" by Vidhya S, P. Perumal, and S. N. Sivanandam.

3 ."Implementation of Crop Recommendation System using Machine Learning Technique" by Priyanka S. Shinde and Dattatraya P. Dandekar.

4 . "Crop Recommendation System using Machine Learning Techniques" by Kavita Mishra, Pulak Ranjan Nayak, and P. N. Rout.