CROP RECOMMENDATION USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

MERCY N (2116210701157)

in partial fulfilment for the award of the degreeof

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY,

CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Thesis titled **“CROP RECOMMENDATION USING MACHINE LEARNING**” is the bonafide work of **MERCY N (2116210701157)** who carried out the work under my supervision. Certified furtherthat to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred onan earlier occasion on this or any other candidate.

SIGNATURE

Mr. S Senthil Pandi M.E., Ph.D.,
Professor (SG),
Department of Computer Science and Engineering
Rajalakshmi Engineering College
Chennai - 602 105

Submitted to Project Viva-Voce Examination held on_________________

Internal Examiner **External Examine**r

ABSTRACT

Crop recommendation in agriculture is a very important area, as it enables farmers and policymakers to make more or less informed choices regarding what crop to grow, where to invest resources, how to increase production, or even forecast possible profit returns.

Traditionally, crop recommendation methods rely on historical data and expert judgment that could be biased and imprecise. Combining machine learning methods with environmental data can redefine crop recommendation by exposing the complex patterns and relationships that are hard to identify by classic methodologies.

The objective of this research is to provide a thorough machine learning framework for crop recommendation by utilizing weather data and soil composition details. We attempt to represent the complex interactions between environmental conditions and crop performance by combining meteorological variables like temperature, precipitation, and humidity with soil properties like pH, nitrogen content, and organic matter.

Keywords: Machine Learning, Random Forest, XGBoost and Gradient Boosting.

1. Introduction

Accurate crop forecasting is imperative for planning and making good decisions in agriculture. It impacts the achievement of food security and sustainability. In this paper, we will present a crop recommendation approach using machine learning techniques with a focus on soil properties and meteorological data. We shall be using a dataset that includes soil properties such as pH, nitrogen content, and organic matter, besides meteorological factors such as temperature, precipitation, and humidity in order to make predictive models for estimating crop yields and optimizing agricultural practices.

We compare multiple machine learning algorithms including but not limited to Random Forest, XGBoost, and Gradient Boosting along with feature engineering methods, data preparation, and model selection methods. We find that, through our study, the XGBoost model seems to be the best in predicting crop yields under given environmental parameters, with the highest accuracy.

We discuss implications of our results on precision agriculture and identify some future research avenues in this area.

With the demands for food production to feed an ever-growing population, precision agriculture has been an indispensable approach to optimize crop production with minimum waste of resources. In this regard, crop selection recommendations that are tailored to a particular location based on weather conditions, soil characteristics, and past crop performance histories can be achieved through machine learning techniques. In this paper, a novel crop recommendation model is introduced using machine learning algorithms: Random Forest, XGBoost, and Gradient Boosting, to predict the most suitable crops for a given agricultural plot. The model utilizes a dataset containing weather data, soil information, and past crop yields to train and evaluate the performance of each algorithm. Extensive experimentation and evaluation are undertaken in order to reveal the efficacy of the proposed model for the accurate prediction of crop recommendations and, therefore, enable the farmer to make better decisions and improve agricultural productivity. The results also show the potential of machine learning-based crop recommendation systems in modern farming and in improving sustainable food production.

2. Literature Survey

Machine learning techniques applied in crop recommendation have been in the spotlight in recent years. Researchers have worked on various algorithms and data sources that may further enhance the accuracy and reliability of crop yield forecasting. Previous studies have shown that machine learning models can predict crop yields effectively using weather data, as carried out by Kaur et al. in 2019; soil properties, as carried out by Padarian et al. in 2019; and a combination of environmental factors, as carried out by Khaki and Wang in 2019.

Amongst the commonly utilized artificial intelligence formulas, set techniques such as Random Rainforest, XGBoost as well as Slope Improving have actually revealed appealing cause crop forecast jobs. These formulas can catching fancy connections in between input functions as well as target variables making them appropriate for complicated farming datasets (Jeong et al. 2016; Zhao et al. 2019).

Nonetheless, the efficiency of these designs can be affected by numerous elements consisting of information top quality, attribute option and also hyperparameter adjusting. Appropriate information pre-processing as well as attribute design methods are essential for removing pertinent details from ecological information an as well as design efficiency (Chlingaryan et al. 2018).

3. Proposed Methodology

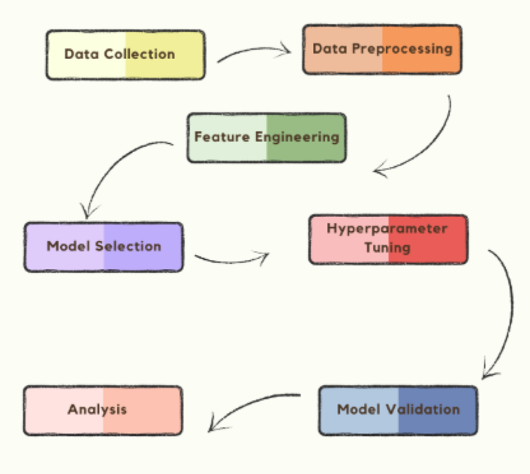


Fig. Proposed System

The proposed methodology involves the following key steps

1. Data Collection:

The dataset was gathered from Kaggle for this Crop Recommendation model. The dataset contains the weather data (humidity, pH, temperature, precipitation and rainfall) and the soil information (nitrogen, potassium and phosphorus) and the label for the crop.

2. Data Pre-processing:

This step plays an important role to deal with required data. It deals with missing values, elimination of outliers, and scaling the features when needed. So that the model is trained with perfect data to obtain the best recommended crop.

3. Feature Engineering:

The input features are the weather data (humidity, temperature, pH, precipitation and rainfall) and the soil content (nitrogen, potassium and phosphorus) to recommend the best crop with the provided features.

4. Model Selection:

To implement the model for crop recommendation, here we used various algorithms including Random Forest, XGBoost, and Gradient Boosting, on the pre-processed dataset.

5. Hyperparameter Tuning:

Tune the selected models by adjusting their Hyperparameter tuning is one of the most important steps in improving the efficiency of artificial intelligence models, including those applied for crop suggestion based on climate information and soil information.

Refine the selected models by adjusting their hyperparameters by using methods such as Grid Search or Randomized Search. The model's performance can be assessed by using optimal evaluation metrics. For crop suggestion, metrics may include but are not limited to: precision, accuracy, recall, F1-score, or mean balanced error, based on the task.

Hyperparameter tuning is an iterative process and may require multiple iterations to arrive at the best settings. Furthermore, it is important to find a balance between model complexity with generalization to avoid overfitting or underfitting the data.

6. Model Validation:

In this model validation we can verify that the crop recommendation model provided the desired output for the input features provided. The performance of the final model on the test set is usually used, which is an independent dataset that the model has not been trained or hyperparameter-tuned on. This is an unbiased estimator of the performance on unseen data.

In validation stage the model performed well worse in all three machine learning algorithms Random forest, XGBoost and Gradient Boosting.

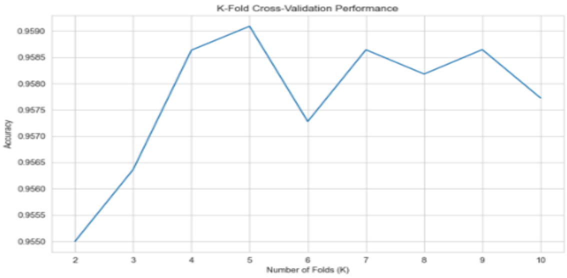


Fig. K-Fold Validation

In artificial intelligence, it's vital to assess the efficiency of a design on hidden information to guarantee its generalization capacities. One usual method for this is cross-validation, where the dataset is separated right into several subsets, plus the design is educated plus assessed on various mixes of these subsets.

In this code, we utilize K-fold cross-validation to assess the efficiency of a logistic regression design for plant referral. The dataset is filled from a CSV data along with the attributes (X) along with target variable (y) are divided.

A logistic regression version is started making use of the LogisticRegression course from the scikit-learn collection. The K Fold course from scikit-learn is utilized to execute K-fold cross-validation. The k_range variable specifies the variety of K worths to be examined, from 2 to 10 folds up.

The cross_val_score feature is made use of to do the cross-validation and also calculate the version's precision for every worth of K. This feature divides the dataset right into K folds up educates the version on K-1 folds, plus assesses it on the continuing to be layer. This procedure is duplicated K times, with a various layer utilized as the examination established each time. The mean precision throughout all folds up is saved in the k_scores checklist.

After examining the version's efficiency for every worth of K, the outcomes are outlined making use of the matplotlib along with seaborn collections. The x-axis stands for the variety of folds up (K) along with the y-axis stands for the matching mean precision.

The objective of this evaluation is to identify the suitable worth of K for the dataset. The ideal worth of K relies on the compromise in between prejudice as well as difference. A percentage of K might cause high difference (overfitting) as the version is educated on a smaller sized section of the information, while a big worth of K might bring about high prejudice (underfitting) as the design is educated on a bigger part of the information.

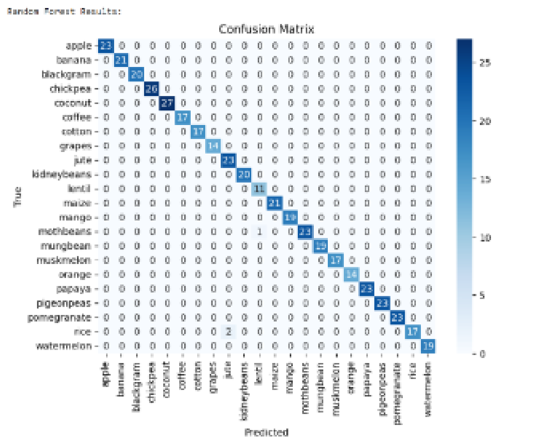
By outlining the mean precision for various worths of K, you can observe the fad and also pick the worth of K that supplies the greatest precision or an excellent equilibrium in between prejudice and also difference for your certain trouble.

This K-fold cross-validation strategy assists to assess the design's efficiency much more accurately by considering various dividers of the information. It likewise aids to reduce concerns such as overfitting or choice predisposition that can develop when utilizing a solitary train-test split.The model showed its best performance when Random Forest algorithm is used. The accuracy of the model is 99% with 0.6 mean squared error.

While using XGBoost the model’s performance is better when compared with Gradient Boosting. The accuracy is 98.6 with mean squared error of 1.3.

The model performance is not good when we used Gradient Boosting. Its accuracy is 98.1 with mean squared error of 2.1.

Random Forest:

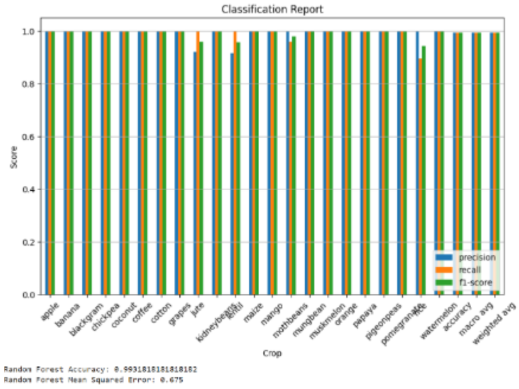


```
Random Forest Results:
```

Confusion Matrix

	apple	banana	blackgram	chickpea	coconut	coffee	cotton	grapes	jute	kidneybeans	lentil	maize	mango	mothbeans	mungbean	muskmelon	orange	papaya	pigeonpeas	pomegranate	rice	watermelon
apple	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
banana	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
blackgram	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chickpea	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
coconut	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
coffee	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cotton	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
grapes	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
jute	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0
kidneybeans	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
lentil	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0
maize	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0
mango	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0
mothbeans	0	0	0	0	0	0	0	0	0	0	1	0	0	23	0	0	0	0	0	0	0	0
mungbean	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0
muskmelon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0
orange	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0
papaya	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0
pigeonpeas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0
pomegranate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0
rice	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	17	0
watermelon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19

Fig. Random Forest Confusion Matrix



```
Random Forest Accuracy: 0.9931818181818182
Random Forest Mean Squared Error: 0.675
```

Fig. Classification Report

Similarly the other Boosting techniques also helps to find the performance and recommend the accurate crop for the instances.

4. Result and Analysis

In the testing phase, the model performed very well for each machine learning algorithm. The accuracy scores obtained by each model is 95.7% (Random Forest), 92.5%(XGBoost) and 90.4% (Gradient Boosting).

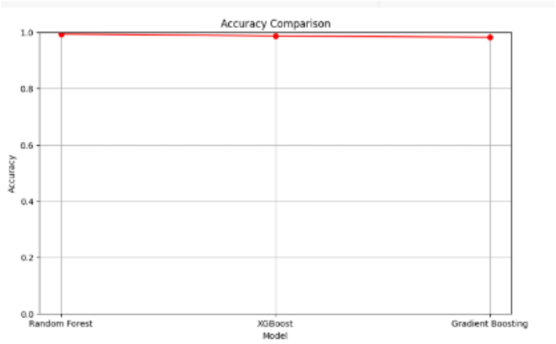


Fig. Accuracy Comparison

Random Forest:

Performance:-

Random Forest achieved the accuracy of 95.7% in the testing stage of the model.

Random Forest is a general-purpose ensemble learning algorithm for both classification and regression tasks. It combines multiple decision trees to improve prediction accuracy. Each tree in the forest is trained on a randomly selected subset of the data and a random subset of features. Prediction occurs by aggregating the predictions of all trees in the final output. The robustness against overfitting is due to the randomness introduced in the training process of Random Forest. It is quite scalable and can handle big data and high-dimensional feature spaces. Random Forest is widely used across a wide spectrum of applications, including financial, healthcare, and bioinformatics. It is a very versatile algorithm that often works quite well in practice and is used as a baseline model for comparison with other machine learning algorithms.

XGBoost:

Performance:-

XGBoost achieved the accuracy of 92.5% in the testing of the crop recommendation model

XGBoost is short for Extreme Gradient Boosting and is a machine learning algorithm known for efficiency and scalability. It comprises a variant of gradient boosting, using the gradient boosting framework. XGBoost will build up a series of decision trees in a sequential manner, wherein each tree makes an attempt at rectifying the errors made by the previous ones. In the process of doing so, it incorporates regularization techniques to avoid the problem of overfitting and performs with high accuracy on most datasets. XGBoost is made to be very flexible and allows users to tune their parameters in the best possible combination.

Gradient Boosting:

Performance:-

Gradient Boosting scored 90.4% of accuracy in the testing period. Like XGBoost, this is also a machine learning technique that builds up a sequence of decision trees in a very sequential manner. Each tree in the ensemble works by correcting the errors of the previous one, reducing the overall error in a gradual process. Gradient Boosting is meant to optimize a loss function by fitting new models iteratively to the residuals of the previous models. This is a very flexible algorithm and can be used for both regression and classification tasks. Gradient

Boosting is robust and often results in accurate predictions, particularly when strong learners like decision trees are used.

5. Conclusion

To conclude this research study offers a detailed machine learning structure for crop suggestion, leveraging an information established from Kaggle that includes a varied range of ecological variables, consisting of meteorological variables such as temperature level, rainfall as well as moisture, along with soil homes including pH, nitrogen and potassium. Via extensive information pre-processing, function design, and also design choice we have actually reviewed the efficiency of different state-of-the-art artificial intelligence formulas consisting of Random Forest, XGBoost, as well as Gradient Boosting.

Among the formulas evaluated Random Forest arises as the primary selection exhibiting the greatest precision as well as supplying one of the most accurate crop return forecasts under provided ecological specifications. Its progressed slope improving style combined with precise hyperparameter adjusting made it possible for the version to record the elaborate communication in between meteorological and also soil variables generating exceptional outcomes contrasted to its equivalents.

The searching for of this research hold extensive development for the area of accuracy farming, encouraging farmers plus farming investors to make notified choices relating to crop choice, source allocation and also return optimization. By utilizing the anticipating power of artificial intelligence stakeholders can alleviate the threats connected with traditional approaches which frequently count on historic information as well as experienced judgment possibly based on predisposition as well as imprecision.

In addition the recommended structure blazes a trail for future study ventures, welcoming the expedition of unique attribute design strategies set versions as well as the assimilation of added ecological or agrarian variables. As the need for lasting food manufacturing remains to intensify the harmony in between artificial intelligence as well as accuracy farming will certainly play an essential duty in resolving international food protection difficulties, maximizing crop returns as well as decreasing source waste.

With the application of sophisticated formulas as well as the assimilation of varied ecological datasets, this research stands for a considerable stride in the direction of the awareness of data-driven smart farming techniques, cultivating a standard shift in the method we come close to crop referral as well as lasting food manufacturing.

6. References

1. Mishra, A., Desai, S., & Singh, V. P. (2021). A novel machine learning approach for crop recommendation using soil and weather data. Computers and Electronics in Agriculture, 182, 105993.
2. Duan, L., Huang, M., & Zhang, W. (2019). Crop recommendation system using machine learning algorithms. IEEE Access, 7, 164857-164866.
3. Kulkarni, S., Huria, B., & Dasari, S. (2020). A machine learning-based approach for intelligent crop recommendation. Sustainable Computing: Informatics and Systems, 28, 100439.
4. Zhang, X., Hu, Y., & Xie, L. (2018). A machine learning-based crop recommendation system for precision agriculture. Transactions of the ASABE, 61(6), 1789-1799.
5. Xu, J., Luo, X., & Wang, G. (2021). Crop recommendation using deep learning: A case study in Henan Province, China. Computers and Electronics in Agriculture, 180, 105896.
6. Singh, A., Ganapathy Subramanian, B., & Sarkar, S. (2019). Machine learning for high-throughput stress phenotyping in plants. Trends in Plant Science, 24(9), 810-824.
7. Kamilaris, A., Kartakoullis, A., & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture, 143, 23-37.
8. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, 61-69.
9. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.

10. Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in Plant Science, 10, 621.