Crop Recommendation Using ML

PrasannaKumar Reddy Chirra *Computer Science Dept*  
Eastern Kentucky UniversityRichmond, KY  
[prasannakumar\_chi@mymail.eku.edu](mailto:prasannakumar_chi@mymail.eku.edu)

*Abstract*— Agriculture serves as the backbone of the Indian economy, with over 60% of the country's land dedicated to farming activities. Soil, integral to agriculture, holds paramount importance as its health directly impacts food production. Climatic conditions complement soil health in influencing agricultural yield. This project introduces a sophisticated crop recommendation system designed to assist farmers. It employs machine learning algorithms to analyze diverse parameters, including soil condition, rainfall, temperature, and humidity, to determine optimal crop choices. Tailored to the Indian context, this system offers crop suggestions aligned with specific climatic and soil conditions, enhancing agricultural productivity. Additionally, farmers receive yield predictions to anticipate production levels, facilitating informed decision-making and optimized resource allocation.

Keywords — Farming Innovation, Smart Agriculture, Logistic Regression, Random Forest, Support Vector Machine, Decision Tree Classifier.

# Introduction

India boasts a rich history and ongoing practice of agriculture, with a substantial segment of its population dependent on this sector for livelihood. Agriculture is not only a source of employment for over half of India’s populace but is also instrumental in the nation’s international trade. Commodities like tea, jute, cashew nuts, tobacco, coffee, and spices mark India's agricultural exports on the world map. Land, a treasured asset in India, witnesses considerable investment in irrigation projects, underscoring the sector’s economic significance. Moreover, the intricacies of changing climatic patterns and market dynamics worsen the uncertainty. This intelligent system is not just an answer to crop selection but also a tool for risk mitigation. It integrates real-time data analytics, offering dynamic recommendations that adapt to ongoing changes in weather, soil conditions, and market demand. Furthermore, the system serves as a bridge to reduce the knowledge gap for farmers by providing insights derived from advanced machine learning algorithms. These insights are not only tailored to the unique characteristics of the local ecosystem but are also aligned with global agricultural best practices, ensuring that farmers are equipped with world-class knowledge. By enhancing predictability and reducing uncertainty, the system is envisioned to boost the overall economic stability of the agricultural sector, making a direct positive impact on farmers’ livelihoods.

Additionally, the implementation of this intelligent system is a development towards sustainable farming practices, ensuring that the soil and ecosystems are preserved and nourished, securing the future of agriculture in India for generations to come. Addressing the critical need for informed crop selection, this project introduces an intelligent system designed to offer predictive insights to farmers. The system considers environmental conditions, including temperature and rainfall, alongside soil attributes like pH level, type, and nutrient content.

In essence, the proposed system is more than a crop recommendation tool it’s a comprehensive solution designed to tackle the multifaceted challenges faced by farmers, offering a blend of traditional wisdom and modern technology to usher Indian agriculture into a new era of productivity, sustainability, and prosperity.

# Literature Review

## **Crop Selection Adapting to Climate Change in Africa, Authors: P. Kurukulasuriya, R. Mendelsohn [1]**

The paper analyzes how climate influences crop selection in Africa, utilizing data from over 7,000 farmers. It concludes that crop choice is indeed climate-sensitive, with different crops being favored under varying climatic conditions. As climate change continues, the study anticipates farmers will adapt by shifting towards crops that are more heat and drought tolerant, By Implementing climate-resilient crops and agronomic research can lead to increased agricultural productivity, ensuring food security amidst changing climatic conditions.

**Ensemble An efficient for effective prediction of suitable agricultural crop cultivation Authors: Mahmudul Hasan, Md Abu Marjan, Md Palash Uddin, Masud Ibn Afjal, Seifedine Kardy, Shaoqi Ma, Yunyoung Nam.**

The system collective machine learning model, KRR, applied in Bangladesh, which predicts crop production using historical and environmental data. It aims to address food shortages by optimizing crop selection and production planning. The solution involves gathering, preprocessing, and analyzing data to enable effective crop production forecasting, guiding farmers in selecting the most suitable crops for increased yield.[2]

The implementation of the KRR model leads to data-driven, optimized crop selection, potentially increasing agricultural productivity, ensuring food security, and boosting the economy, especially in developing countries dependent on agriculture.

**Prediction of Potato Crop Yield Using Precision Agriculture Techniques. Al-Gaadi KA, Hassaballa AA, Tola E, Kayad AG, Madugundu R, Alblewi B, Assiri F.**

In this study, the authors employ precision agriculture techniques to predict the yield of potato crops. It serves as an essential resource for understanding how advanced technologies, including machine learning and data analytics, can be applied in real-time to optimize agricultural productivity. The paper provides detailed insights into the methodologies, tools, and techniques for precision farming, focusing on data collection, processing, analysis, and prediction related to potato cultivation[3]. The findings contribute to the broader knowledge of utilizing technology-driven approaches to enhance the accuracy of crop yield predictions, ensuring efficient resource utilization and increased yields.

**Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining Approach Monali Paul, Santosh K. Vishwakarma, Ashok Verma Computer science and Engineering GGITS, Jabalpur**

A system that uses data mining techniques to predict the category of the analyzed soil datasets is presented in this work. Crop yields will be indicated by the predicted category. [4] Naive Bayes and K-Nearest Neighbor methods are used to formalize the crop yield prediction problem as a classification rule.

**Crop Recommendation System for Precision Agriculture S. Pudumalar\*, E. Ramanujam\*, R. Harine Rajashree, C. Kavya, T. Kiruthika, J. Nisha. [5]**

Using Random tree, CHAID, K-Nearest Neighbor, and Naive Bayes as learners, this paper proposes an ensemble model with a majority voting technique to recommend a crop for the site-specific parameters with high accuracy and efficiency. Any of the current designations.

# Dataset Details

In this approach, we utilize a dataset sourced from Kaggle, characterized by its compact size of 150KB and encompassing 2200 data entries[6]. Each row in the dataset encapsulates comprehensive information crucial for crop recommendation, including:

* N - Denotes the Nitrogen content ratio in soil
* P - Indicates the Phosphorous content ratio in soil
* K - Represents the Potassium content ratio in soil
* Temperature - Measured in degree Celsius.
* Humidity - Expressed as relative percentage.
* pH - Specifies the pH value of the soil.
* Rainfall - Quantified in mm.
* Label - Suggests the recommended crop based on the parameters.

We were advantaged to have access to a clean dataset, eliminating the need for an extensive preprocessing phase. This allowed us to channel our efforts into the core objective of training machine learning models. Leveraging algorithms like Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Support Vector Machines, our focus was centered on optimizing accuracy to ensure reliable crop recommendations. The illustrative snippet of the dataset below serves as a testament to the comprehensiveness and structured nature of the data employed in this initiative.

A table with numbers and symbols

Description automatically generated

# Main approach

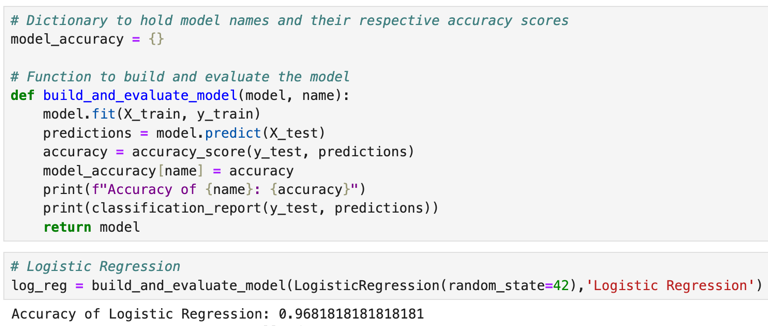
This crop recommendation system is fixed on the implementation of several robust machine learning algorithms, each contributing distinct analytical capabilities to refine the precision and reliability of recommendations. The synergy of these algorithms ensures that diverse aspects of the dataset are accurately evaluated, providing comprehensive insights.

## Logistic Regression:

Logistic regression is a well-established supervised learning classification algorithm, widely used for its efficacy in predicting binary outcomes. By analyzing and interpreting the relationships among one or more independent variables, it predicts a dependent binary variable, making it ideal for scenarios where the outcome is dichotomous.

In this project, every variable, except the 'label', is treated as an independent variable, contributing to the predictive model. The 'label' is used as the dependent variable, embodying the outcome we aim to predict the recommended crop in this context. We partition the dataset into training and testing subsets, allocating 80% for model training and reserving the remaining 20% for model evaluation. This graded sampling ensures this model is both trained and validated on diverse data instances, enhancing its reliability.

After training the logistic regression model on the training subset, we evaluate its performance on the testing data. Key metrics like accuracy and the classification report, encompassing precision, recall, and F1-score, are computed to offer a comprehensive view of the model’s predictive capabilities. This quantitative assessment aids in identifying the model's strengths and areas for improvement, ensuring that it is robust and reliable in real-world scenarios. The following code snippet exemplifies this process, encapsulating the model training, evaluation, and performance metrics computation stages.



## Decission Tree:

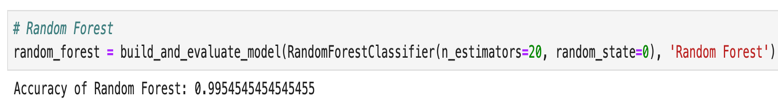
Decision Tree Classifier is a method in supervised learning that makes decisions based on asking a series of questions. For our model, we use the "entropy" criterion to measure the quality of splits, aiming for low entropy indicating organized, homogenous data. The tree is limited to a max depth of 5 to avoid overfitting, ensuring the model generalizes well to unseen data. After training, we evaluate its performance using accuracy and a classification report, offering a concise assessment of its effectiveness in predicting appropriate crops based on the given parameters. Below is the code snippet.



## Random Forest:

Random Forest is an ensemble learning method known for its versatility and ease of use for both classification and regression tasks. It builds multiple decision trees during training and outputs the mode of the classes for classification or mean prediction for regression. This forest of decision trees, each trained on random subsets of the training data, votes to reach a final prediction, enhancing model accuracy and robustness. In this approach, we train the Random Forest model using the training dataset and specify the number of trees in the forest with ‘n\_estimators=20’. This parameter helps in building a robust model without significant computational costs. Post training, the model’s performance is evaluated using accuracy and a classification report to provide detailed insights into its precision, recall, and F1-score.

This collective model, with its multiple decision trees, not only aids in making accurate predictions but also offers insights into the importance of different features, supporting informed decision-making for crop selection based on diverse soil and environmental parameters.



## Support Vector Machine:

SVM is renowned for its efficiency and performance in binary classification tasks. It operates by describing a hyperplane that best segregates the two classes of data. Particularly, SVMs are favored for their effectiveness in high-dimensional spaces and their performance optimization with small sample sizes, a common scenario in text classification and similar tasks. In the context of this project, we deploy the SVM model for crop recommendation. We train the model on a labeled dataset, with the SVM effectively categorizing and predicting the most suitable crops based on the given soil and environmental parameters. We set the gamma parameter to 'auto', which defines the influence of a single training example, ensuring the model is responsive yet not overly sensitive to individual data points.

Post-training, we assess the SVM model's performance by examining its accuracy and detailed classification report. These insights enable us to understand the model's predictive strengths and areas for enhancement, ensuring its applicability and reliability in real-world agricultural scenarios.A math equation on a white background

Description automatically generated

# K Means Clustering

In addition to the machine learning models described, KMeans clustering has also been employed in this study to provide deeper insights into the fundamental groupings within the data. While not necessary for the primary objective of crop recommendation, clustering aids in debut underlying patterns and associations that can be instrumental for enhanced decision-making.

In the provided code snippet, KMeans clustering is applied to the dataset after dropping the 'label' column, thereby focusing on the inherent structures based solely on the features. Four clusters are specified initially, with the ‘k-means++’ initialization method optimizing the placement of centroids to hasten convergence. The clustering model is then fitted to the data, assigning each data point to the closest cluster.

A screenshot of a computer code

Description automatically generated

The Elbow Method is engaged to determine the optimal number of clusters. By plotting the within-cluster sum of squares (WCSS) against a range of cluster numbers, the "elbow" point is identified where the reduction in WCSS begins to diminish, indicating the most appropriate cluster number for the data.

A graph of a number of clusters

Description automatically generated

The clusters are then analyzed to discern the suitable season for each crop, based on conditions derived from temperature, humidity, and rainfall. This auxiliary insight, while not central to the crop recommendation, offers valuable contextual information that can be harnessed for more holistic and adaptive agricultural planning.

The crops within each cluster and their corresponding suitable seasons are displayed, offering a nuanced perspective that transcends mere crop recommendation, and delves into the intricate interplay of environmental factors and their collective impact on agricultural outcomes. This layered approach ensures that recommendations are not only accurate but also enriched with contextual insights for maximized agricultural productivity and sustainability.

# Evaluation Metrics

When the process of evaluating and selecting the optimal machine learning model is critical to ensuring that predictions are both accurate and reliable. Utilizing a diverse array of metrics allows for a comprehensive assessment of a model’s performance, ensuring its applicability to real-world, unseen data. In the realm of machine learning, model evaluation is an indispensable phase where various algorithms are scrutinized to select the most efficacious one for prediction tasks. A suite of metrics facilitates this process, offering insights into aspects like accuracy, precision, recall, and the F1-score.

## Accuracy:

Accuracy is a fundamental metric in the realm of machine learning and is often the initial go-to for evaluating model performance. It is calculated as the ratio of the number of correct predictions to the total number of predictions.

A close-up of a symbol

Description automatically generated

## Precission

When there is a dataset imbalance is prominent, precision becomes an indispensable metric. Precision provides insights into the accuracy of a model in predicting positive classes, specifically focusing on the exactness of the prediction. Precision is calculated as the ratio of True Positives (correctly predicted positive observations) to the total predicted positives, which includes both True Positives and False Positives (negative observations incorrectly predicted as positive).

A mathematical equation with a few numbers

Description automatically generated with medium confidence

## Recall

Recall, also known as sensitivity or true positive rate, is a critical metric especially in contexts where missing a positive class is costly. It quantifies the model's ability to identify and correctly classify all relevant instances.

A mathematical equation with black letters

Description automatically generated with medium confidence

## F1-Score

F1 score is a highly valuable metric, especially in contexts where both types of errors false positives and false negatives are significant, and the dataset is imbalanced. It offers a balanced measure by considering both precision and recall computing the model's overall performance.

A close-up of a sign

Description automatically generated

# Results & Analysis

In the search of developing a sophisticated crop recommendation model, various machine learning algorithms were applied and evaluated. We have data visualizations and proceed to employ algorithms including Logistic Regression, Decision Tree, Random Forest, and SVM. The dataset’s features comprised the soil’s N, P, and K values, temperature, humidity, rainfall, and pH values.

A comparative analysis of the applied algorithms unveils Random Forest as the standout performer, boasting an impressive accuracy of 99.5%. Logistic Regression and Decision Tree followed suit, with commendable accuracies of 96.8% and 98.8% respectively. SVM, however, lagged with a 12.7% accuracy, indicating its limited applicability to this specific dataset and problem scenario.

A graph showing different colored squares

Description automatically generated

Random Forest’s superiority wasn’t confined to accuracy alone; it also outperformed in other critical metrics like precision, recall, and F1-score, underscoring its balanced and robust performance. A case in point is when the model, furnished with parameters N=103, P=19, K=31, temperature=23.603016, humidity=61, pH=7, and rainfall=141, adeptly predicted ‘coffee’ as the most suitable crop.

The ensemble nature of Random Forest, characterized by its multiple decision trees and voting mechanism, not only enhances accuracy but also bolsters the model’s resilience against overfitting and noise. This ensures that the model isn’t just accurate but is also reliable and generalizable, offering predictive insights that are both actionable and dependable.

# Future Work

Agriculture stands as the foundation industry globally, essential in supplying food to billions. The evolution from traditional to modern farming, propelled by technological innovations, has reduced manual labor, and amplified both the quality and yield of harvests. Foreseeing broader accessibility, if this system is expanded to a web platform, it could reach millions of farmers nationwide, offering convenience and ease of access. Mobile applications, especially in various native languages, can further enhance accessibility and usability. Collaborating with government entities could significantly enrich the quality and volume of data, honing the accuracy of our predictions. While the current dataset has posed no significant challenges, as it is compact and manageable, integrating a more extensive, real-time dataset could require more involved preprocessing steps, potentially impacting the model’s accuracy. In future iterations, we aim to incorporate additional features, like tailored fertilizer recommendations, elevating the system from offering crop suggestions to providing comprehensive agricultural guidance, ensuring farmers are well-equipped for optimized, informed farming practices.

# Conclusion

In the current landscape, many farmers are yet to integrate technology into their agricultural practices, potentially leading to outdated and less efficient cultivation methods. This project aims to bridge this gap, inspiring farmers to embrace contemporary technological tools, moving beyond traditional approaches. Cultivating a crop involves a nuanced understanding of various factors including soil content, ambient temperature, and soil pH, among others. Technology, especially machine learning algorithms, offers precise insights in these areas with enhanced ease and accuracy. In this project, machine learning has proven to be exceedingly efficient, boasting an accuracy of 99% in predicting optimal crops. The adoption of this technology by farmers is not just a leap towards modernization but a step towards simplified, informed, and optimized farming, ensuring that the agricultural processes are both efficient and yield-maximizing.

REFERENCES

1. Crop Selection Adapting to Climate Change in Africa, Authors: P. Kurukulasuriya, R. Mendelsohn , Kurukulasuriya, Pradeep and Robert O Mendelsohn. “Crop Selection: Adapting to Climate Change in Africa.” Agricultural & Natural Resource Economics (2007): n. pag.
2. Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M. I., Kardy, S., Ma, S., & Nam, Y. (2023). Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. Frontiers in Plant Science, 14, 1234555. https://doi.org/10.3389/fpls.2023.1234555
3. Al-Gaadi KA, Hassaballa AA, Tola E, Kayad AG, Madugundu R, Alblewi B, Assiri F. Prediction of Potato Crop Yield Using Precision Agriculture Techniques. PLoS One. 2016 Sep 9;11(9):e0162219. doi: 10.1371/journal.pone.0162219. PMID: 27611577; PMCID: PMC5017787.
4. Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining Approach Monali Paul, Santosh K. Vishwakarma, Ashok Verma Computer science and Engineering GGITS, Jabalpur
5. Crop Recommendation System for Precision Agriculture S. Pudumalar\*, E. Ramanujam\*, R. Harine Rajashree, C. Kavya, T. Kiruthika, J. Nisha.
6. <https://www.kaggle.com/datasets/>