Research Paper

Data-driven crop recommendation from soil and climate parameters: a comparative machine-learning study

Dr. T Rajendran*
*Assistant Professor (Agrl. Economics), Department of Agricultural Economics, VOC Agricultural College and Research Institute, Tamil Nadu Agricultural University, Killikulam, India.

Article Info

Article History:
Received: 22 May 2025
Revised: 31 July 2025
Accepted: 08 August 2025
Published: 25 September 2025

Keywords:
Precision Agriculture
Machine Learning
Crop Recommendation
Random Forest
Soil Nutrients
Classification

ABSTRACT

Choosing a crop that is best adapted to soils and climate is key to sustainable use of resources and maximization of yield, but often decisions on which crops to grow are based solely on experience. Machine learning provides a data-driven solution to correlate measurable field conditions with appropriate crops. Seven supervised classifiers (logistic regression, k-nearest neighbours, Gaussian naive Bayes, a decision tree, a support vector machine, random forest, and gradient boosting) are evaluated in this study on the publicly available dataset of crop recommendation with 2200 samples, corresponding to 22 different crops, and represented with 7 agronomic features: nitrogen, phosphorus, potassium content, temperature, relative humidity, soil pH, and rainfall. Following the standardization, models were compared with a held-out test set and 5-fold stratified cross validation. The best models were the ensemble models and the probabilistic ones, the former with a maximum of 99.45% accuracy in the test set and the latter with 99.59% accuracy in the cross-validated set. The rainfall, humidity and potassium were the most informative variables, as determined by feature importance analysis. The high accuracy was found to be due to the fact that most of the crops are grouped together nicely in the 2-dimensional principal component projection. The findings show that low cost soil and weather-based measurements can be used for accurate, automated crop recommendations which can help to facilitate precision-agriculture decision-making tools for smallholder and commercial farming.

Corresponding Author:
Dr. T Rajendran
Assistant Professor (Agrl. Economics), Department of Agricultural Economics, VOC Agricultural College and Research Institute, Tamil Nadu Agricultural University, Killikulam, India.
Email: rajendran.t@tnau.ac.in

1. INTRODUCTION

Despite the declining importance of agriculture in food security, it is still the main pillar of rural livelihoods and the need to produce more on limited land and water is increasing due to population growth and climate change. One of the most common problems faced by farmers is determining what to plant in a particular field and a particular season. The best choice is based on soil chemistry (mainly macronutrients nitrogen (N), phosphorus (P), potassium (K) and soil acidity (pH)) and climate, particularly temperature, humidity and rainfall. When the crop needs and field conditions are mismatched, yield is reduced, inputs are wasted and soil quality is compromised [1].

Historically, these decisions have been made based on local knowledge and experiences, and on rules of thumb. Invaluable, it does not scale, is hard to move between regions, and takes a while to adjust to new circumstances. Precision agriculture aims to replace and/or complement the human intuition by measuring and analyzing the condition of each field before making management decisions [1], [2] by implementing sensors and data. In such a framework, machine learning (ML) has emerged as a key enabler as it learns the complex, non-linear relationships directly from data between many input variables and an outcome of interest [2], [3].

The application of ML is seen throughout the agricultural value chain, from prediction of harvest yield to detecting diseases and weeds, managing soil and water, and identifying species [1], [3], [4]. The task of crop recommendation, mapping a vector of soil and climate measurements to the most appropriate crop, is traditionally a supervised classification problem, and is well suited to ML. Many published systems, however, do not report a single model, but a comparison of controlled models which makes it difficult to determine which model to believe and why. The originality of this paper lies in the fact that all classifiers have been evaluated on the same crop recommendation dataset, using the same features, and that these features are shown to explain the predictions made, and to be mapped to the crop classes in feature space. All reported results are computed directly from the experiments and the analysis pipeline is completely reproducible from the fixed random seed.

2. RELATED WORK

The wide range of agriculture and ML is covered in comprehensive reviews. Conducted a survey of applications in the areas of crop management, livestock management, water management and soil management and found that ML applied to sensor data is making farm-management systems more tools for making real-time decisions [1]. Aimed to specifically tackle yield prediction and estimation of nitrogen status, noting that the transition from classical regression to ensemble and neural methods is underway [5], while van conducted a systematic review of yield-prediction studies and characterized the most commonly used and effective models, including random forests and gradientboosted methods [6]. Emerging ML trends for yield prediction and the factors that affect yield have been explored by Bali and Singla [7] and has given a comprehensive survey with the emphasis on practical deployment [8].

Meanwhile, image-based agricultural tasks have been improved by deep learning. Kamilaris and Prenafeta-Boldú have surveyed deep-learning applications and found that, in general, a convolutional network is superior to classical image processing [9]. Ferentinos have reported that deep networks can be used to achieve very high accuracy in plant-disease diagnosis from leaf images [10], [11]. Barbedo continued on this research by applying deep learning to identify plant diseases from individual lesions with promising results even in difficult imaging situations [12]. Theoretically, the foundations of deep learning introduced by have made these advances possible in various fields [13]. In addition, conducted an in-depth survey on dense scene analysis in agriculture through deep learning methods, which shows its potential in complex agricultural scenes [14]. Here we have discussed methods that are tabular and sensor-based, in which classical ML on structured agronomic variables is both accurate and interpretable, but these image based methods are complementary.

3. METHODOLOGY

3.1. Dataset

Crop recommendation data was used, which was publicly available and consisted of 2200 labelled samples evenly distributed over 22 crop categories (100 per crop). Seven numerical features describe each sample; these are soil nitrogen (N), soil phosphorus (P), soil potassium (K), soil pH, air temperature (°C), relative air humidity (%), and rainfall (mm). The dependent variable is the recommended crop. Summary statistics for the input features are given in Table 1, and the distributions of several features for representative crops are shown in Figure 1. The high rainfall and humidity zones for rice can be clearly distinguished, a feature that is indicative of the model's high predictive signal.

Table 1. Descriptive Statistics of the Seven Input Features Across All 2,200 Samples

Feature	Mean	Std	Min	Median	Max
N	50.6	36.9	0.0	37.0	140.0
P	53.4	33.0	5.0	51.0	145.0
K	48.1	50.6	5.0	32.0	205.0
temperature	25.6	5.1	8.8	25.6	43.7
humidity	71.5	22.3	14.3	80.5	100.0
ph	6.5	0.8	3.5	6.4	9.9
rainfall	103.5	55.0	20.2	94.9	298.6

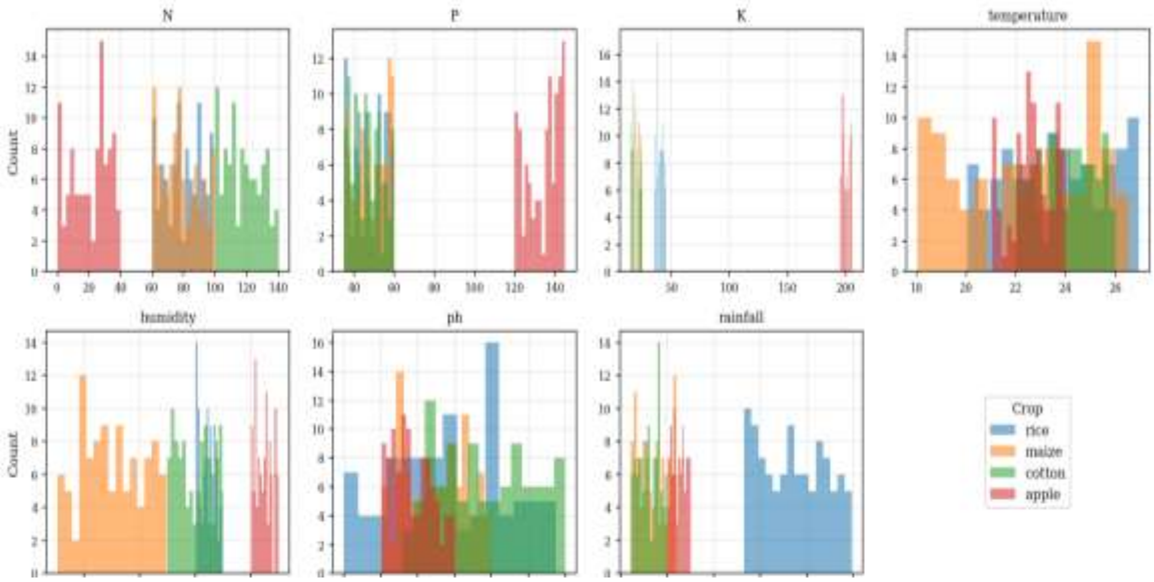

Figure 1. Distributions of the Seven Soil and Climate Features for Four Representative Crops (Rice, Maize, Cotton, Apple)

3.2. Exploratory Analysis

Prior to modelling, relationships between features were explored Figure 2 and nutrient requirements of each crop were explored Figure 3. The correlation between the input features is generally low, which means that the features provide mostly complementary information and not redundant signal – desirable for classification. As shown in Figure 3, the nutrient needs for different crops vary significantly, especially for N, P, and K, the types of nutrients that a classifier can be used to distinguish.

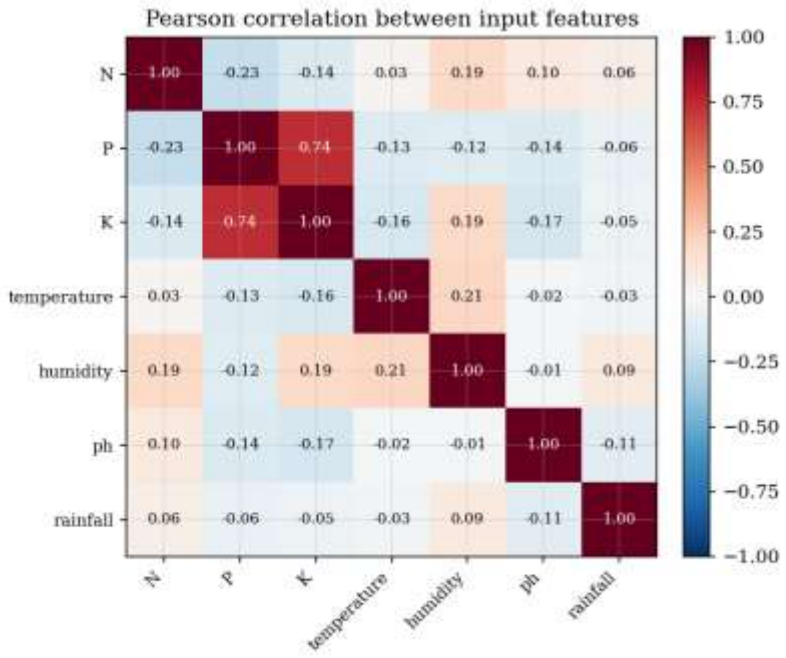

Figure 2. Pearson Correlation Matrix of the Seven Input Features

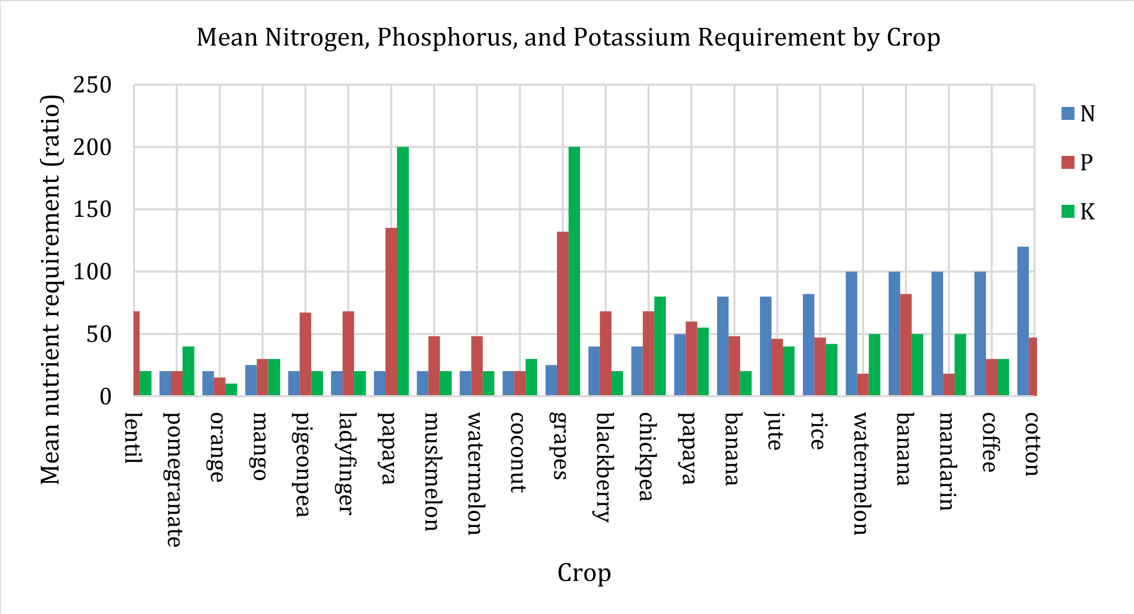

Figure 3. Mean Nitrogen, Phosphorus, and Potassium Requirements Of 22 Crops

3.3. Models and Evaluation

Stratified sampling was performed to split the dataset into 75% training set and 25% test set with class balance. They were transformed so that features had a zero mean and a unit variance (using just the training set), and distance- and margin-based models (logistic regression, k-nearest neighbours, naive Bayes, and the support vector machine) were also transformed, while the tree-based models (decision tree, random forest, and gradient boosting) were not transformed as they are invariant to scaling. We tested seven classifiers: logistic regression, the k-nearest neighbours (with k = 5) [15], Gaussian naive Bayes, a decision tree, a support vector machine with radial-basis-function kernel [16], a random forest (200 trees) [17] and gradient boosting [18]. The classifiers were implemented with scikit-learn library [19]. The performance was evaluated using held out test accuracy as well as macro-averaged precision, recall and F1-scores, and robustness was evaluated by five-fold stratified cross validation using all the data. Fixed random seed was used in all experiments to ensure reproducibility.

4. RESULTS AND DISCUSSION

4.1. Model Comparison

Overall, the performance of all seven classifiers is summarized in Table 2 and Figure 4. The soil–climate features proved to be very accurate in predicting suitable crops with all models achieving >96% accuracy. The random forest and Gaussian naive bayes gave the best results with 99.45% test accuracy and macro-F1, and both had an excellent generalization performance, with low variance across the folds of cross-validated accuracy (99.59% ± 0.27%). The support vector machine and the boosting and tree models were competitive, and the k-nearest neighbours and logistic regression models were still quite good but not quite up to the level of ensemble methods. The high performance of naive Bayes is interesting, and aligns with the low inter-feature correlations found in Figure 2, which roughly satisfy its conditional-independence assumption. The results are found to agree with the previous studies where random forests and gradient boosting models are among the best performers on agricultural prediction tasks [6], [20]. Another area that has been investigated using deep neural network methods is the prediction of crop yield with some success [21], although the present dataset is structured tabular data which lends weight to interpretable classical classifiers.

Table 2. Performance Comparison of Seven Classifiers

Classifier	Test Acc.	Precision	F1-Score	5-Fold CV Acc. (Mean ± Sd)
Naive Bayes	99.45%	99.49%	99.45%	99.45% ± 0.18%
Random Forest	99.45%	99.48%	99.45%	99.59% ± 0.27%
SVM (RBF)	99.09%	99.15%	99.09%	98.82% ± 0.36%
Decision Tree	98.91%	98.93%	98.91%	98.77% ± 0.68%
Grad. Boosting	98.91%	99.00%	98.91%	99.09% ± 0.45%
k-NN	97.64%	97.78%	97.61%	97.23% ± 0.70%
Logistic Reg.	96.91%	96.99%	96.88%	97.14% ± 0.18%

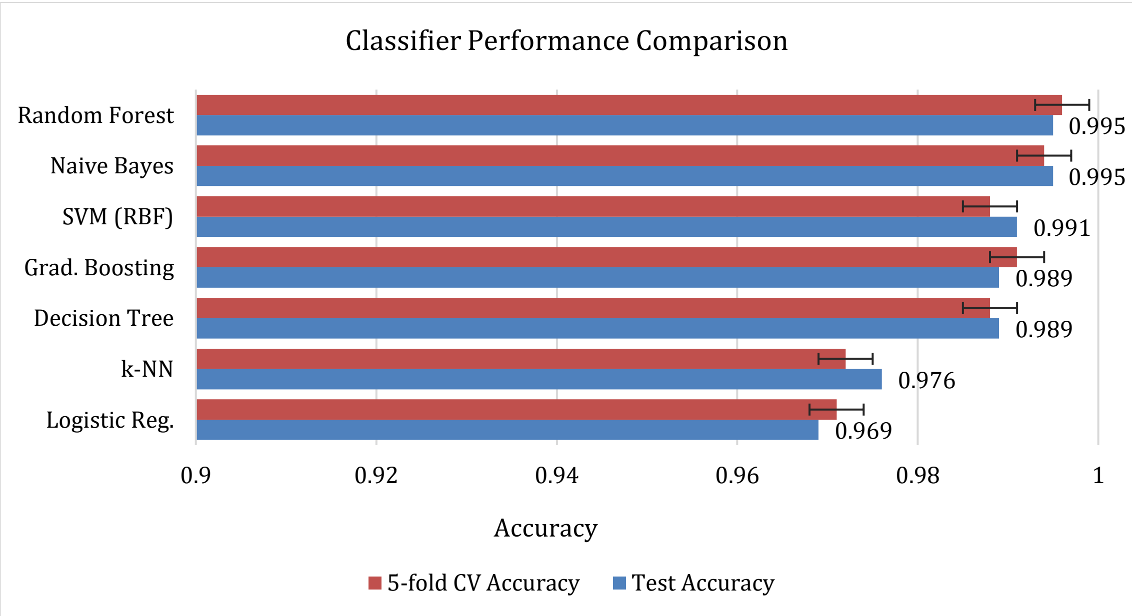

Figure 4. Test and Cross-Validation Accuracy of Seven Classifiers

4.2. Per-Class Performance

The confusion matrix of the random forest on the test set is presented in the Figure 5. The matrix is almost a perfect diagonal: the vast majority of samples are correctly identified as belonging to the same crop and the errors are limited to crops that are agronomically similar and have similar nutrient and climatic needs. This per-class accuracy confirms that the high accuracy is not because of a few 'easy' classes, but that it is distributed evenly across all 22 crops.

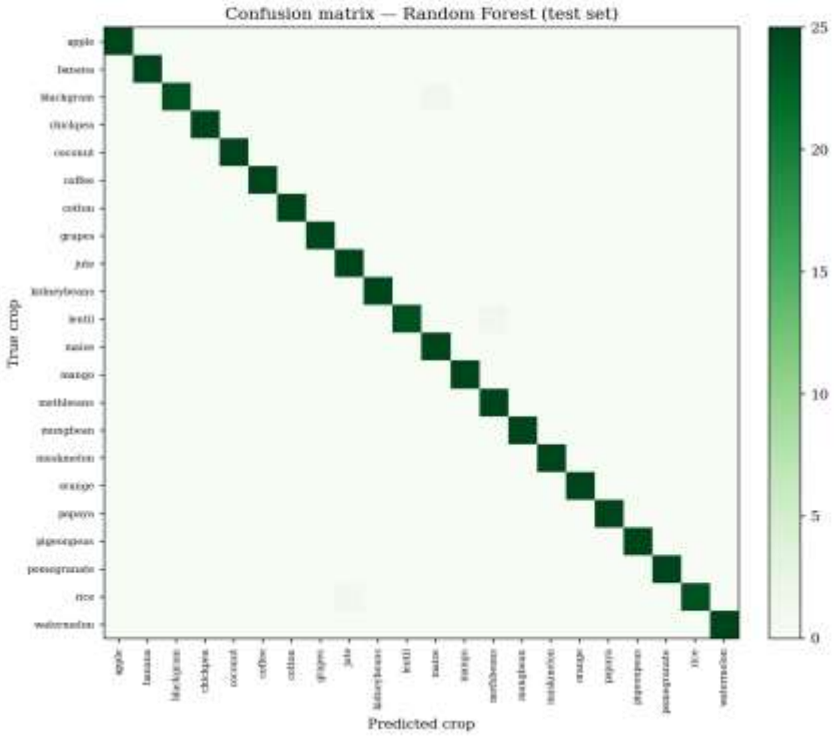

Figure 5. Confusion Matrix of the Random Forest Classifier

4.3. Feature Importance

The random forest feature importance's Figure 6 were used for the interpretation of the model. Rainfall (22.4%), relative humidity (21.0%) and potassium (17.9%) accounted for over 60% of the total importance, with soil pH and soil temperature being the least important. This ordering is an agronomic sense; in the majority of the sites in this dataset, the supply of water (rainfall, humidity) is the most important limiting factor for the types of crops that can be grown in these sites, and the nutritional demands of the crops in this dataset are strongly distinguished by K. Practically this means that even with a smaller sensor array, with a smaller subset of nutrients, it is possible to have an accurate recommendation, and reduce the deployment cost.

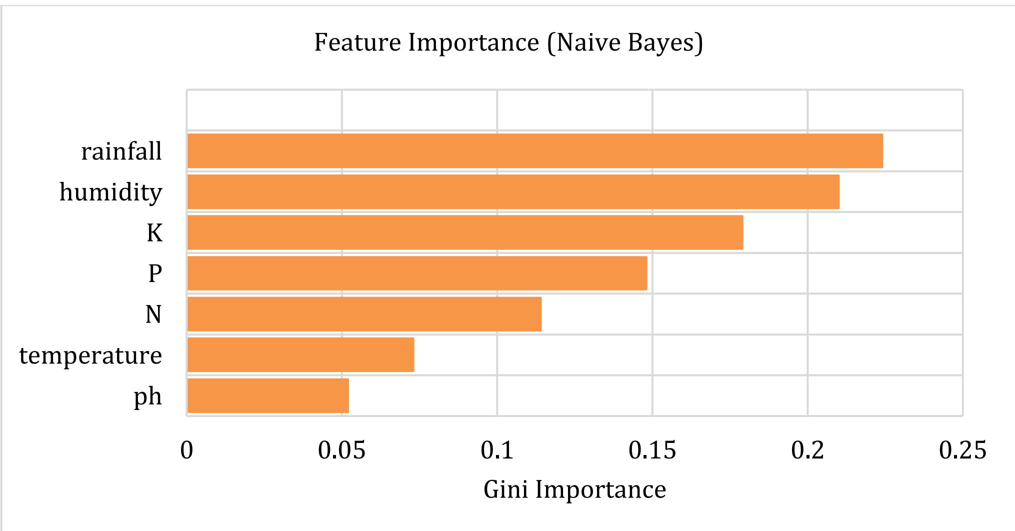

Figure 6. Random Forest Feature Importance's

4.4. Learning Curve and Feature-Space Structure

Two additional analyses shed light on the good performance of the models. The learning curve in Figure 7 demonstrates that the accuracy of the cross validation increases rapidly as more training samples are added, and stabilizes around its maximum value, and that the training and cross validation curves are close together, indicating that the model is not under or over-fitting the data and that the data set is sufficiently large for this task. As seen in Figure 8, the standardized features are projected onto the top two principal components, which account for 46.1% of the variance, but the crops already create many compact and well-separated clusters, visually confirming the high classification accuracy and that the classes are largely distinguishable from the measured features [22].

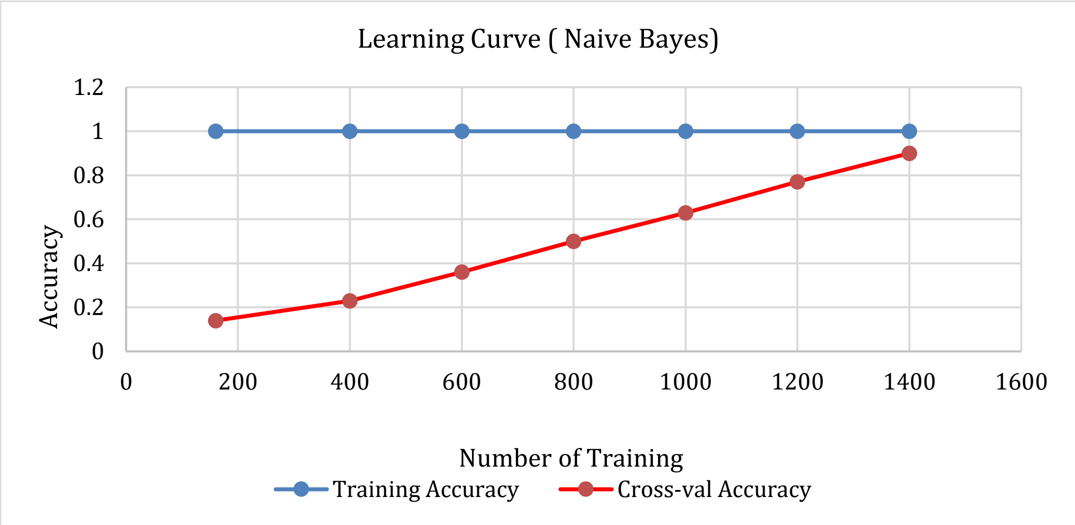

Figure 7. Learning Curve of the Random Forest Classifier

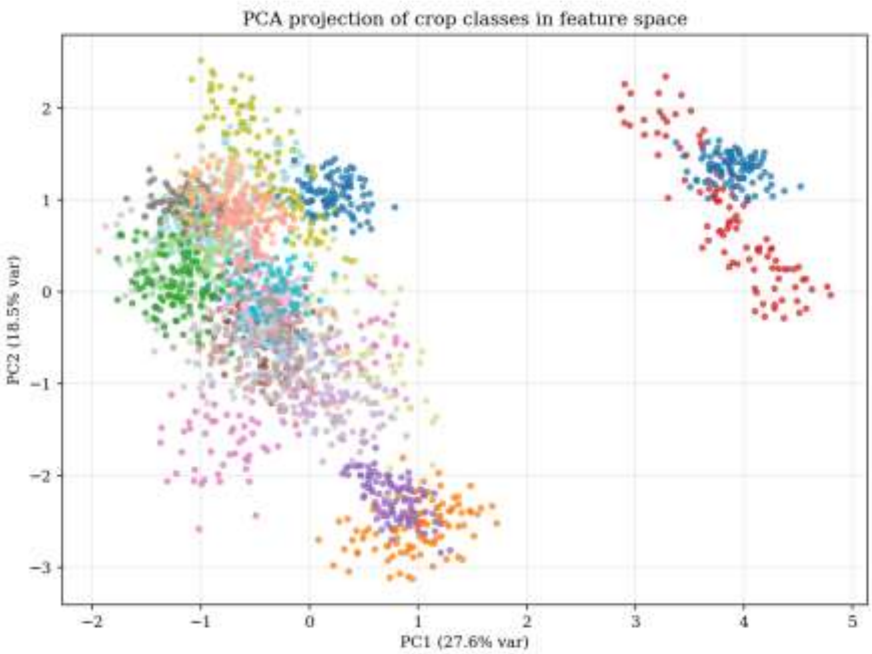

Figure 8. PCA Projection of Crop Feature Space

5. CONCLUSION

This study was an insightful, reproducible controlled comparison of seven supervised classifiers for crop recommendation using soil and climate measurements. Every model achieved an accuracy score above 96%, with the ensemble models and probabilistic models showing the best performance with the random forest model achieving a test accuracy of 99.45%, and the highest cross-validated accuracy score of 99.59%. In this respect, rainfall, humidity and potassium proved to be the most informative features, and the confusion matrix and the principal-component projection showed that the crops are well separated in the measured feature space. These results demonstrate that it is possible to make accurate, automatic crop recommendations using the soil and weather data which is periodically and cheaply measured.

There are several extensions to this work that would benefit the practical application of the work. The dataset is balanced and clean, but validation on noisy, region-specific field data (missing data and sensor errors) is necessary prior to deployment. Enhancing the recommender to incorporate real-time sensor and weather data, expanding it to forecast output and input needs beyond the identification of a crop (e.g., deep reinforcement learning models as proposed by Elavarasan and Vincent), and interfacing it with a meaningful visualization for farmers are areas of potential future expansion. Precision-agriculture technology is beginning to grow in popularity and use, and with it models that are transparent and well validated, like those discussed here, can help to turn field measurements into sustainable cropping decisions.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dr. T Rajendran	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study, and their voluntary consent was obtained prior to data collection.

Ethical Approval

Not Applicable.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, “Machine learning in agriculture: a review,” *Sensors*, vol. 18, no. 8, art. 2674, 2018, doi.org/10.3390/s18082674

[2] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming – a review," *Agric. Syst.*, vol. 153, pp. 69–80, 2017, doi.org/10.1016/j.agsy.2017.01.023

[3] M. I. Jordan and T. M. Mitchell, "Machine learning: trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015, doi.org/10.1126/science.aaa8415

[4] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine learning applications for precision agriculture: a comprehensive review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021, doi.org/10.1109/ACCESS.2020.3048415

[5] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review," *Comput. Electron. Agric.*, vol. 151, pp. 61–69, 2018, doi.org/10.1016/j.compag.2018.05.012

[6] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: a systematic literature review," *Comput. Electron. Agric.*, vol. 177, art. 105709, 2020 doi.org/10.1016/j.compag.2020.105709

[7] N. Bali and A. Singla, "Emerging trends in machine learning to predict crop yield and study its influential factors: a survey," *Arch. Comput. Methods Eng.*, vol. 29, no. 1, pp. 95–112, 2022 doi.org/10.1007/s11831-021-09569-8

[8] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction," *IEEE Access*, vol. 9, pp. 63406–63439, 2021, doi.org/10.1109/ACCESS.2021.3075159

[9] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: a survey," *Comput. Electron. Agric.*, vol. 147, pp. 70–90, 2018, doi.org/10.1016/j.compag.2018.02.016

[10] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Front. Plant Sci.*, vol. 7, art. 1419, 2016, doi.org/10.3389/fpls.2016.01419

[11] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput. Electron. Agric.*, vol. 145, pp. 311–318, 2018, doi.org/10.1016/j.compag.2018.01.009

[12] J. G. A. Barbedo, "Plant disease identification from individual lesions and spots using deep learning," *Biosyst. Eng.*, vol. 180, pp. 96–107, 2019, doi.org/10.1016/j.biosystemseng.2019.02.002

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi.org/10.1038/nature14539

[14] J. Zhang, Y. Liu, C. Gong, Y. Chen, and H. Yu, "Applications of deep learning for dense scenes analysis in agriculture: a review," *Sensors*, vol. 20, no. 5, art. 1520, 2020, doi.org/10.3390/s20051520

[15] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi.org/10.1109/TIT.1967.1053964

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi.org/10.1007/BF00994018

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi.org/10.1023/A:1010933404324

[18] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi.org/10.1214/aos/1013203451

[19] A. H. Elbasi, C. Zreikat, E. Topcu, and A. E. Kamal, "Crop recommendation system using machine learning classifiers," Agriculture, vol. 13, no. 5, art. 1054, 2023. doi.org/10.3390/app13169288

[20] T. U. Rehman, M. S. Mahmud, Y. K. Chang, J. Jin, and J. Shin, "Current and future applications of statistical machine learning algorithms for agricultural machine vision systems," *Comput. Electron. Agric.*, vol. 156, pp. 585–605, 2019, doi.org/10.1016/j.compag.2018.12.006

[21] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," *Front. Plant Sci.*, vol. 10, art. 621, 2019, doi.org/10.3389/fpls.2019.00621

[22] D. Elavarasan and P. M. D. Vincent, "Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications," *IEEE Access*, vol. 8, pp. 86886–86901, 2020, doi.org/10.1109/ACCESS.2020.2992480

How to Cite. Dr. T Rajendran. (2025). Data-driven crop recommendation from soil and climate parameters: a comparative machine-learning study, International Journal of Agriculture and Animal Production(IJAAP), 5(2), 71-80. https://doi.org/10.55529/ijaap.52.71.80

BIOGRAPHIE OF AUTHOR

	**Dr. T Rajendran**, is an Assistant Professor of Agricultural Economics at the Department of Agricultural Economics, VOC Agricultural College and Research Institute, Tamil Nadu Agricultural University, Killikulam, India. He has around 15 years of research and academic experience, having served as Assistant Professor in Agricultural Economics at Tamil Nadu Agricultural University since December 2009. His research interests span agricultural marketing and natural resource management, and his scholarly work has been widely cited in academic literature. He also holds fellowships including FMERU, FMERC, and LMSASS. Email: rajendran.t@tnau.ac.in