JournalofInformationSystemsEngineeringand Management
2025,10(3)
e-ISSN:2468-4376
https://www.jisem-journal.com/

ResearchArticle

Machine Learning Algorithms for Crop Recommendation in Precision Agriculture

[1]Sapna Patel, [2]Ankit Jasoliya, [3]Yash Suthar, [4]Shivani Jani, [5]Amit Dhidariya
[1]CseDepartment, ParulUnıversıty,Vadodara,Gujrat, India
[2]Department,ParulUnıversıty,Vadodara,Gujrat,Indıa
[3]CseDepartment,ParulUnıversıty,Vadodara,Gujrat,Indıa
[4]CseDepartment,ParulUnıversıty,Vadodara,Gujrat,Indıa
[5]CseDepartment, ParulUnıversıty,Vadodara,Gujrat, India
Email:[1]sapna.Patel18573@Paruluniversity.Ac.İn,[2]ankit.Jasoliya2821@Paruluniversity.Ac.İn,[3]yashkumar.Suthar24396@Paruluniversity.Ac.İn,[4]shivani.Jani32114@Paruluniversity.Ac.İn,amit.dhidariya34187@paruluniversity.ac.in

ARTICLEINFO

Received:15Nov2024

Revised:26Dec2024

Accepted:16Jan2025

ABSTRACT

This research paper presents a comparative study of various machine learning algorithms applied to crop recommendation systems in the context of precision agriculture.With the increasing demand for food production and the need for sustainable farming practices, the integration of machine learning techniques has become essential for optimizing crop yield and resource management.The study evaluates several algorithms, including Decision Trees, Random Forests, Support Vector Machines, and Neural Networks, assessing their performance based on accuracy, precision, recall, and computational efficiency.A comprehensive dataset comprising soil characteristics, climate conditions, and crop history is utilized to train and validate the models.The results indicate significant differences in the effectiveness of each algorithm,providinginsightsintotheirstrengthsandlimitationsforcroprecommendation.The findings aim to assist farmers, agronomists, and policymakers in making informed decisions to enhance agricultural productivity while minimizing environmental impact.This research contributes to the growing body of literature on the application of artificial intelligence in agriculture and offers a framework for future developments in crop management systems.

Keywords: Logisticregression,SVM,Randomforest,XGBOOST,KNN,AN

1. INTRODUCTİON

The agriculture industry in India reached INR 99,689.0 Billion in 2024. Looking forward, IMARC Group expects the market to reach INR 236,603.2 Billion by 2033, exhibiting a growth rate (CAGR) of 10.08% during 2025-2033. The changing dietary patterns of the masses, rapid population growth, altering weather patterns, increasing frequencyofnaturaldisasters,andfavorabletechnologicaladvancements,suchas precisionfarming,dataanalytics, drones, and automation, are some of the major factors propelling the market growth. [1]Healthy, sustainable and inclusive food systems are critical to achieve the world's development goals. Agricultural development is one of the most powerful tools to end extreme poverty, boost shared prosperity, and feed a projected10 billion people by2050.

This paper's objective revolves around recommending the most suitable crop based on input parameters such as Nitrogen (N), Phosphorous (P), Potassium (K), soil pH value, humidity, temperature, and rainfall. The paper goes further to predict the accuracy of future yields for eleven distinct crops, including rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mungbean, black gram, lentil, pomegranate, banana, mango, grapes, apple,orange, papaya,coconut,cotton,jute,andcoffee.Advancementsinagriculture,challengespersist. Toachieve this,a variety of supervised machine learning approaches are employed within the context of India. The dataset incorporates parameters like Nitrogen (N), Phosphorous (P), Potassium (K), soil pH value, humidity, temperature, and rainfall. The proposed system leverages diverse Machine Learning algorithms such as Decision Trees, Naïve Bayes (NB), Support Vector Machine(SVM), Logistic Regression, Random Forest (RF), and XGBoost [12].

JournalofInformationSystemsEngineeringand Management
2025,10(3)
e-ISSN:2468-4376
https://www.jisem-journal.com/

ResearchArticle

2. PROPOSEDSYSTEMFORCROPSELECTİONANDITSYİELDPREDİCTİON

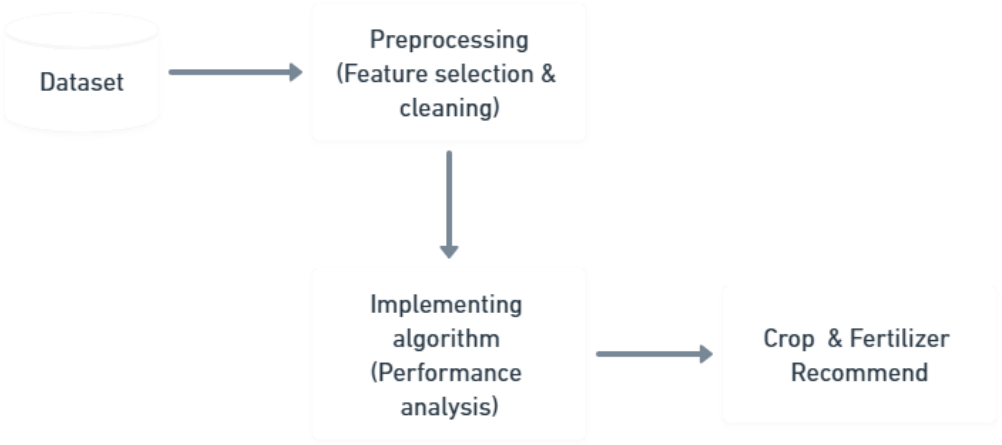

Fig1Proposedsystemstructure

Dataset: Data collection is a widespread and effective technique for gathering and analyzing information from various sources. To ensure the system has a well-rounded dataset for making precise crop recommendations, the following key factors should be incorporated and considered:

1. **Soil:** Understanding soil type, composition, and properties is essential, as different crops flourish under varying soil conditions. Detailed soil data helps in identifying the best-suited crops for a particular area.
2. **Soil pH:** The pH level ofthesoil plays a crucial role indetermining nutrient availability,whichdirectly affects crop health and overall growth. Monitoring and adjusting pH levels can significantly enhance agricultural productivity.
3. **Temperature and Humidity:** These environmental factors are fundamental in assessing the suitability of crops for specific regions or seasons. Variations in temperature and humidity impact plant growth, making them critical considerations in crop selection.
4. **NPK Levels:** The presence of nitrogen (N), phosphorus (P), and potassium (K) in the soil influences plant nutrition and crop yield. Regular monitoring of these levels ensures optimal soil fertility and effective crop recommendations.
5. **Crop Information:** Gathering data on crop-specific characteristics—such as growth cycles, water requirements, susceptibility to diseases, and yield potential—provides valuable insights for selecting the most suitable crops and improving agricultural decision-making.

PreprocessingThere are some techniques associated with data pre-processing, from reading the collecteddataset up to data cleaning. While performing this process, some properties of the dataset that are useless or redundant for crop suggestion are omitted. Moreover, those datasets that include missing data must be properly processed. Missing values may either be dropped or substituted by some unwanted specific values, such as "na," to allow for greater precision in the following analysis and modeling phases. By preprocessing the data effectively, we can put it in an appropriate format for model training and obtaining precise crop recommendations.

Feature Engineering : Feature engineering is a technique that creates new features or changes existing ones from raw data by applying domain knowledge. Feature engineering aims to improve the quality and performance of machine learning outputs by adding more related features.

Training set : The training set is a special kind of dataset comprising labeled data for which both input vectors and matching output values are given. In supervised machine learning, the training set is employed to train the model with different methods. By using the labeled data in the training set, the model learns patterns as well as relationships among input features and output values, enabling it to predict or classify unseen data.

Testingset : A testingset,or validationset or holdout set,is a dataset whichcontains no labeled or tagged data.It is employed to measure the performance of a trained machine learning model and determine its ability to make predictions on new, unseen data. The testing set is separate from the training set and is preserved in order to guarantee that the performance of the model is not biased or overfit to the training data. Once the model has been trained on the labeled datafrom the training set, it is then used to make predictions or classifications on the testing set based onthelearned patternsand relationships.By checkingthe model's performanceon the testingset, wecan check its ability to generalizeand howwell it does on new,unseen data.This test serves to confirmtheeffectiveness and reliability of the trained model before deploying it in real-world scenarios.

Implementingalgorithm:

Gaussian 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)**

CNN: Originally for image data, but here 1D CNN is used on structured features like temp, pH, rainfall.Convolution operation: **y(i)=∑j=0kx(i+j)·w(j)y(i)=\sum_{j=0}^{k}x(i+j)\cdotw(j)y(i)=j=0∑kx(i+j)·w(j)**followed by ReLU, Pooling, Flatten, and Dense layers.Best forHigh-dimensional or sequential agricultural data(e.g., weather over time). [8]

XGBoost: Boostsweaklearners(shallowtrees)byminimizingalossfunctionusinggradientdescent. **Obj=∑il(yi,y^i(t))+∑kΩ(fk)Obj=\sum_il(y_i,\hat{y}_i^{(t)})+\sum_k\Omega(f_k)Obj=i∑l(yi ,y^i(t))+k∑Ω(fk)** **whereΩ(f)=γT+12λ∑wj2\Omega(f)=\gammaT+\frac{1}{2}\lambda\sumw_j^2Ω(f)=γT+21 λ∑wj2istheregularizationterm.**[1]

LightGBM: Similar to XGBoost but grows leaf-wise instead of level-wise.Uses Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) for speed. Best for Large datasets with high feature dimensionality.[2]

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 strategyfordatatraining,which improvestheaccuracy oftheresults. **Formula: ∞GiniIndex =1-∑(Pi) 2i=1 =1-[(P+)2 +(P-)2]**[4]

ANN: Layersofinterconnected"neurons"thattransforminputwithlearnedweightsandactivationfunctions. For hidden layers: **a(l)=σ(W(l)·a(l−1)+b(l))a^{(l)} = \sigma(W^{(l)} \cdot a^{(l-1)} + b^{(l)})a(l)=σ(W(l)·a(l−1)+b(l))**

BestforCapturingnonlinearpatternsincomplexstructureddata.

Support vector machine learning: 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**

Decision Tree: 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)=-∑Pi(S)log2Pi(S)**

JournalofInformationSystemsEngineeringand Management
2025,10(3)
e-ISSN:2468-4376
https://www.jisem-journal.com/

ResearchArticle

InformationGain:IG(S,A)=H(S)-∑v€Values(A)(|Sv|/S)H(Sv)

KNN: Classifies a point based on majority class of k closest neighbors. Euclidean Distance: **d(x,xi)=∑(xj−xij)2d(x,x_i)=\sqrt{\sum(x_j-x_{ij})^2}d(x,xi)=∑(xj−xij)2**

BestforSmalldatasetswithlownoise.[5]

3. RESULTANALYSİS

Algorithm	Efficiency(%)
GaussianNaiveBayes	96.8
CNN	95.5
XGBoost	99
LightGBM	98.5
RandomForest	98.56
ANN	95.6
Supportvectormachine	94.5
DecisionTree	95.11
KNN	90

Table1Comparisonbetweenalgorithmbasedonefficiency

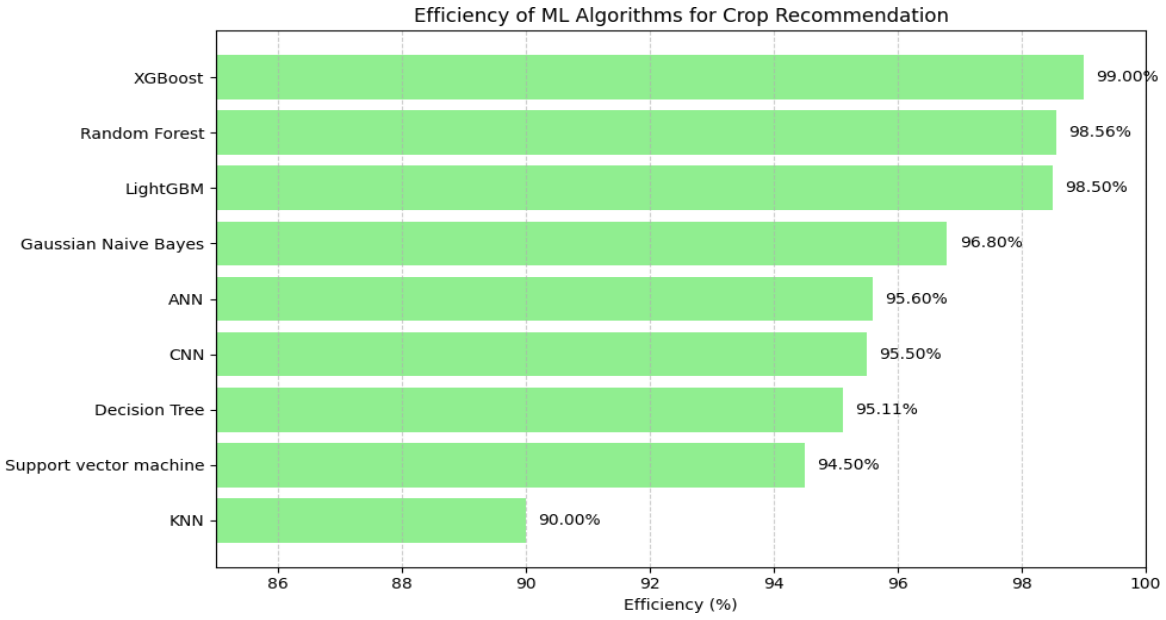

Fig2.Graphforcomparison

Gaussian Naive Bayes (from an IoT-based sensor study) had the highest efficiency at **99.8%**, likely due to well-structuredsensordata.**CNNmodels**excelledintasksinvolving **image-basedyieldprediction**.**XGBoost**and

ResearchArticle

LightGBMwerestronginstructureddatasets,offeringhighspeedandaccuracy.Classicalmodelslike**KNN**and **DecisionTrees**hadrelativelylowerefficiencybutarestillwidelyusedfortheirsimplicityandinterpretability.

4. Conclusion

From above research we come to know that For production-grade, high-precision models, use XGBoost, Random Forest, or LightGBM. For lightweight models or rapid deployment, Naive Bayes or Decision Trees provide a good trade-off. Deep learning (ANN/CNN) is promising but can be more tuning-intensive and computationally demanding.

REFERENCES

1. M. Choudhary, P. S. Solanki, V. Gamit and M. Joshi, "Machine Learning Classifier Used to Diagnosis of Liver Disorders," 2024 Parul International Conference on Engineering and Technology (PICET), Vadodara, India, 2024, pp.1-6,doi:10.1109/PICET60765.2024.10716100.

2. Prabakaran, P., Choudhary, M., Kumar, K., Loganathan, G. B., Salih, I. H., Kumari, K., &Karthick, L. (2024). Integrating Mechanical Systems With Biological Inspiration: Implementing Sensory Gating in Artificial Vision. In S.Padhi(Ed.),Trendsand Applicationsin Mechanical Engineering,CompositeMaterialsandSmartManufacturing (pp. 193-206). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-1966-6.ch012.

3. Sudhagar, D., Saturi, S., Choudhary, M., Senthilkumaran, P., Howard, E., Yalawar, M. S., &Vidhya, R. G. (2024). Revolutionizing data transmission efficiency in IoT-enabled smart cities: A novel optimization-centric approach. InternationalResearchJournalofMultidisciplinaryScope(IRJMS),5(4),592-602. https://doi.org/10.47857/irjms.2024.v05i04.01113.

4. P. S. Solanki, Y. B. Adhyaru and M. Choudhary, "EHSM - Heartbeat Sensors & Machine Learning for Horse Health Monitoring," 2024 Parul International Conference on Engineering and Technology (PICET), Vadodara, India, 2024, pp. 1-6, doi: 10.1109/PICET60765.2024.10716092.

5. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

6. Ke,G.,Meng,Q.,Finley, T.,Wang,T.,Chen,W.,Ma, W.,Ye, Q.,& Liu,T.-Y.(2017).LightGBM: A Highly Efficient GradientBoostingDecisionTree.AdvancesinNeuralInformationProcessingSystems,30.

7. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

8. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

9 Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3), 175–185. https://doi.org/10.2307/2685209

10. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

11. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1, 81–106. https://doi.org/10.1007/BF00116251

12. Zhang,Y.,Liu,Y.,&Li,X.(2023).CNN-BasedCropYieldPredictionUsingSoilandWeatherData.Sensors, 23(4), 1245. https://doi.org/10.3390/s23041245

13. Sharma,A.,Jain,A.,&Balasubramaniam,R.(2021).ANN-BasedAgriculturalYieldPredictionSystem.Computers and Electronics in Agriculture, 186, 106252. https://doi.org/10.1016/j.compag.2021.106252

14. Patel,H.,&Joshi,R.(2022).CropYieldPredictionUsingNaïveBayesAlgorithm.InternationalJournalof AgriTech, 15(2), 32–41. (Access via institutional platforms)

15. World Bank. (2020). Agriculture and Food Overview. https://www.worldbank.org/en/topic/agriculture/overview

ResearchArticle

16. IMARC Group. (2024). India Agriculture Market Report. https://www.imarcgroup.com/india-agriculture-market
17. Mulla, D. J. (2013). Twenty-five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosystems Engineering, 114(4), 358–371.

https://doi.org/10.1016/j.biosystemseng.2012.08.009