Ensemble machine learning approaches for crop recommendation system using hybrid feature selection techniques

Rohini A
Dept. of CSE
Vignan's Foundation for Science, Technology and Research
Vadlamudi, India
rohiniayinaparthi@gmail.com

Mokshagna P
Dept. of CSE
Vignan's Foundation for Science, Technology and Research
Vadlamudi, India
moksha2502@gmail.com

Viveka Nandini T
Dept. of CSE
Vignan's Foundation for Science, Technology and Research
Vadlamudi, India
tavitinandini@gmail.com

Tejashwee Nishant
Dept. of CSE
Vignan's Foundation for Science, Technology and Research
Vadlamudi, India
tejnishdhiran@gmail.com

***Abstract*—This paper explores the concept of machine learning algorithms in agriculture, particularly in the optimization of crop yield through the analysis of soil nutrient levels and climatic variables. This study aims to develop a crop recommendation system based on NPK (Nitrogen, Phosphorus, Potassium) content and evolving climatic conditions. Five advanced machine learning boosting algorithms: XGBoost, Gradient Boost, AdaBoost, and CatBoost, along with a hybrid model, are used to evaluate their effectiveness. The dataset is of agricultural parameters, including NPK values, temperature, pH, rainfall, and humidity, while the yield data has 11 major agricultural crops and 10 horticultural crops. This approach demonstrates the potential to provide a user-friendly interface, enhancing decision-making in crop selection and promoting efficient agricultural practices.**

I. Introduction

The world's growing population is driving the need for increased food production, but this must be achieved sustainably to protect the environment. Optimizing crop selection for specific regions is crucial to achieving this goal, as it can significantly impact agricultural productivity and environmental sustainability. However, this is a complex task due to the numerous factors involved, including soil properties, rainfall patterns, and climate conditions.

Machine learning offers a powerful solution by enabling the analysis of large datasets to uncover hidden patterns and correlations between these factors and crop growth. By leveraging machine learning techniques, it is possible to develop a reliable model that can predict the most suitable crops for specific regions, supporting informed decision-making in agriculture and contributing to more sustainable farming practices.

This study aims to explore the potential of machine learning in agriculture, focusing on the development of a data-driven model that can accurately predict crop suitability for different regions. By achieving this goal, we can increase land productivity, reduce environmental degradation, and support the development of sustainable agricultural practices.

II. Literature Review

A notable study published in Heliyon in 2024 evaluated the performance of multiple algorithms, including Support Vector Machines (SVM), Random Forests, K-Nearest Neighbours (KNN), Decision Trees, and XGBoost, achieving accuracy rates of 99.09% and 99.3% [1]. This study highlights the importance of considering multiple algorithms and techniques in predicting crop yields, as well as the need for robust models that can handle complex datasets.

Similarly, a conference presentation in IEEE Xplore in 2021 reported a 95% accuracy rate using SVM, Artificial Neural Networks (ANN), and Random Forests, emphasizing the significance of regional specificity and high-quality data in achieving accurate predictions [2]. This study underscores the need for localized models that take into account regional factors, such as climate, soil type, and weather patterns, and highlights the importance of data quality in machine learning applications.

Other studies have also reported high accuracy rates using various machine learning algorithms. A study in the IJSRCSEIT journal achieved accuracy levels between 90% and 99.31% using Decision Trees and Naive Bayes, while acknowledging the impact of data quality and environmental variability on predictive performance [3]. This study highlights the importance of considering the quality of the data used in training machine learning models, as well as the need to account for environmental variability in predicting crop yields.

In contrast, a study in the IJRASET journal in 2023 assessed KNN and ANN techniques, resulting in a lower accuracy of 65.05%, largely due to the complexity of the datasets and

a lack of automation for integrating real-time environmental data [4]. This study highlights the challenges of working with complex datasets and the need for more robust models that can handle environmental variability, as well as the importance of automation in integrating real-time data.

The use of different algorithms and techniques has also been explored in various studies. For example, a study in the Journal of Agriculture and Food Research combined datasets from Bangladesh and Kaggle, achieving an impressive 97.5% accuracy with the CatBoost algorithm for crop recommendations [5]. This study demonstrates the potential of combining datasets from different sources to improve the accuracy of machine learning models, as well as the importance of considering the interpretability of models in practical applications.

A conference paper in IEEE Xplore in 2022 utilized the Light GBM algorithm, attaining an accuracy of 98% while stressing the need for effective model integration and interpretability [6]. This study highlights the importance of considering the interpretability of machine learning models, as well as the need for effective model integration in practical applications, and demonstrates the potential of using ensemble methods to improve the accuracy and robustness of models.

Another study in IJISAE reported a 95% accuracy with Random Forest and Decision Tree methods, underscoring the necessity for model interpretability and scalability in agricultural applications [7]. This study demonstrates the potential of using ensemble methods, such as Random Forest, to improve the accuracy and robustness of machine learning models, as well as the importance of considering the scalability of models in practical applications.

These studies collectively demonstrate the potential of machine learning techniques in predicting crop yields but also highlight the need for further research to address the challenges and limitations of these techniques. Specifically, there is a need for more robust models that can handle environmental variability and provide accurate predictions across different regions and conditions. Additionally, larger and more diverse datasets are required to improve the accuracy and generalizability of models. Finally, model interpretability and scalability are essential for practical applications, and further research is needed to develop models that can be easily interpreted and scaled up for real-world use.

III. Methodology

The methodology to develop a crop recommendation system using a machine learning approach can be divided into the following steps:

A. About Data

- **Data Sources**: The dataset analyzed in this work was downloaded from the repository of Kaggle [8] which was pooled in terms of time by the Indian Chamber of Food and Agriculture. In total, it has 2100 data points involving 11 crops of agriculture as well as 10 horticulture crops in regards to variables for the supply of NPK fertilizer, soil pH, and climatic aspects like rain, temperature and humidity. Crop yield data constitutes 11 major agricultural crops and 10 horticultural crops.
- **Data Fields**: The following fields are contained in the dataset.
  N: Amount of nitrogen contained in the soil.
  P: It is the amount of phosphorus present in the soil.
  K: Represents the amount of potassium in the soil.
  temperature: Measured in degrees Celsius.
  humidity: The measure or percentage that exhibits the relative air humidity.
  ph: Measures the soil's alkalinity or acidity.
  rainfall: Measured in millimeters.
  Label: Specifies which crop should be grown given the soil conditions present and the needs of the surrounding environment.

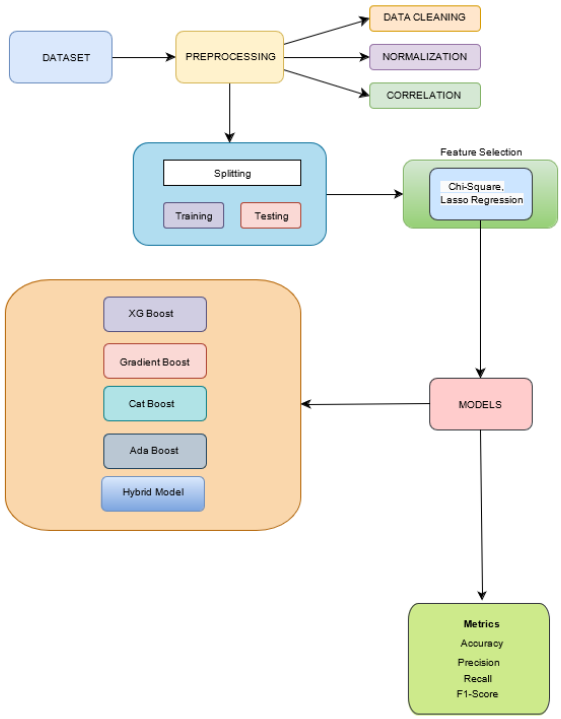

Fig. 1: PROPOSED ARCHITECTURE

B. Data Preprocessing

- **Data Cleaning**: Missing values were treated with techniques like label encoding and feature scaling as well as non-negative transformation. All of these ensure the features are comparable and numerically compatible, hence machine learning algorithms can be trained upon them.
- **Data Normalization**: Features were normalized to allow that variables that have different scales would equally contribute to the performance of the model.
- **Correlation**: To identify the relationship between variables, make enhancing predictive analysis and guide the selection of features, be used to check for multicollinearity, test statistical assumptions, and evaluate the quality

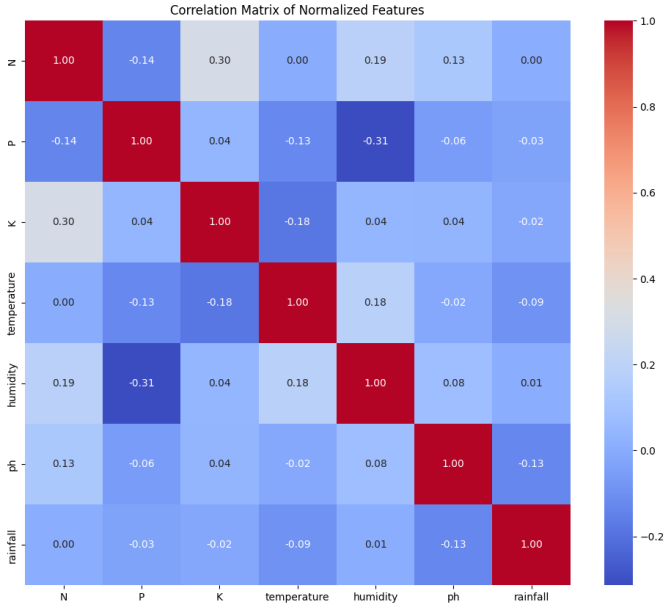

Fig. 2: Correlation Heatmap

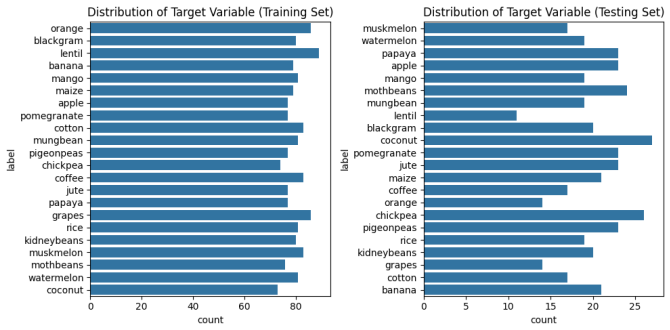

Fig. 3: Distribution of Target Variables

of data while informing further research and enabling visualization. It is one of the basic exploratory tools through which one comes to understand data, eventually leading to better accuracy with models and decisions.

The values in the matrix are between -1 and 1, which correspond to Pearson correlation coefficients of each possible pair of features. These colors and values mean the following:

- 1 (dark red): perfect positive correlation:.
- -1 (dark blue) represents perfect negative correlation.
- 0 indicates no correlation.

- **Data Splitting**: For the evaluation of the performance of the model satisfactorily, the data was partitioned into 80 percent for the training set and the remaining 20 percent for the test set.It creates two count plots showing the distribution of the target variable for training and testing sets that would be useful for class balancing, an important issue in understanding model performance and evaluation.

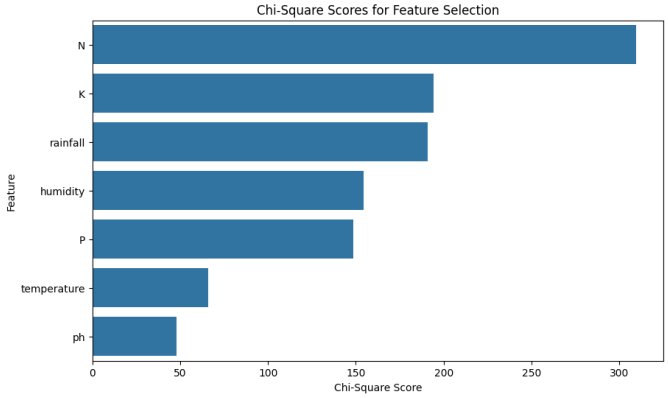

Fig. 4: Chi-Square Selected Features

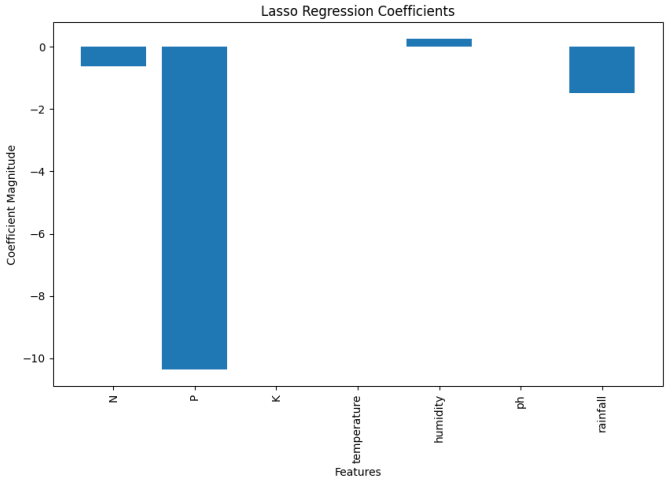

Fig. 5: Lasso Regression Coefficients

- **Feature Selection**: There are two types of features selection techniques; one of them will be chi-square test, which will be adopted in order to select the features with regard to their significance levels in statistics, and the other one lasso regression is going to select features based on the reduction of zeros on irrelevant features.As shown in figures 4 and 5.

C. Model Implementation

The following machine learning models were implemented to predict crop suitability:

- **XGBoost (Extreme Gradient Boosting)**:
  - XGBoost is an ensemble learning technique that combines multiple weak learners (usually decision trees) to form a robust predictive model.
  - It uses an algorithm known as gradient boosting to optimize the model's performance in such a way that it decreases the error from past iterations.
  - The hyperparameters learned consist of the learning rate, the maximum depth of the trees, and the number of estimators whose prediction was maximized.

- **Gradient Boost(GBM)**:
  - Another ensemble technique is Gradient Boosting, building models sequentially in a way that tries to correct errors made by prior models in the sequence.
  - Decision trees are constructed greedily; each new tree reduces the residuals errors that come from all preceding trees.
  - All the key parameters, including learning rate and number of boosting stages, were tuned up to optimize the performance.
- **AdaBoost (Adaptive Boosting)**:
  - AdaBoost combines weak classifiers to create a strong classifier by focusing on the misclassified instances of the dataset.
  - The weights of the misclassified samples are increased and then subsequent classifiers pay more attention to such errors.
  - The number of weak learners and individual learning rates for each were very carefully optimized.
- **CatBoost (Categorical Boosting)**:
  - CatBoost is designed to handle categorical variables natively, which improves its accuracy and efficiency.
  - It relies on ordered boosting that removes prediction bias and treats categorical variables without significant data preprocessing.
  - For the model parameters, depth and learning rate were fine-tuned.
- **Hybrid Model Development**
  - A hybrid model was designed to take the strengths of various algorithms.
  - The hybrid approach combines a weighted average or stacking technique by combining the predictions of the four models, that is, XGBoost, Gradient Boost, AdaBoost, and CatBoost.
  - Hybrid model is used to achieve better prediction accuracy by downplaying individual model weaknesses while building on the potential strengths in an ensemble.

This is a methodology that combines the strengths of each technique in the machine learning methods and addresses the development of a hybrid system for recommending the most suitable crops in terms of their growth on a particular soil under climatic conditions.

IV. Results

A. Performance Comparison

- **Model Conclusion:**
  - XGBoost and CatBoost are largely achieving near perfect, so the former two are recommended to achieve the highest accuracy.
  - GBM and the Hybrid Model are acceptable with marginally slightly lower accuracy and excellent results in most instances.
  - AdaBoost was the worst performing model in this experiment. It is probably because it cannot learn the complicated patterns so that it may not be the best algorithm for this specific dataset.
- **Feature Conclusion:**
  - The feature importance plot from the model XGBoost suggests that the most important features are N and K for determining the predictions made by

Model	Accuracy	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
AdaBoost	0.12	0.10	0.12	0.10
GBM	0.98	0.98	0.98	0.98
XGBoost	0.99	0.99	0.99	0.99
CatBoost	0.99	0.99	0.99	0.99
Hybrid	0.98	0.98	0.98	0.98

TABLE I: Performance Comparison of Machine Learning Models

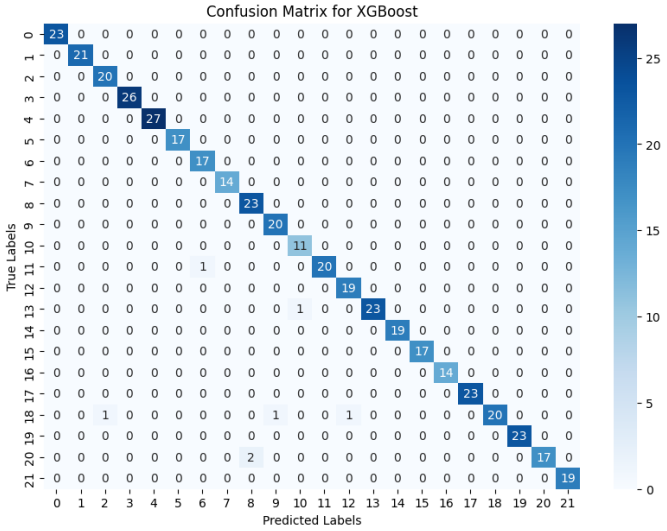

Fig. 6: Confusion Matrix for XGBoost

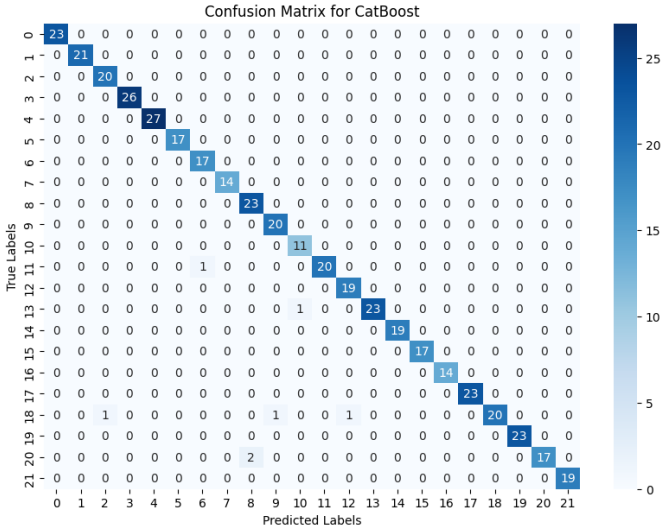

Fig. 7: Confusion Matrix for CatBoost

the model.
  - Of all the features analyzed, pH has the least contribution along with Rainfall and humidity.
  - It shall be a guiding light towards the next phase of further analysis or data collection efforts in the most critical features.

In summary, XGBoost and CatBoost present the best performance and robustness in solving tasks of complex classification. Other GBM and Hybrid models are safe alternatives, and, although AdaBoost might be not so useful in the given problem setting and, therefore, should better be preprocessed or tuned.

B. Performance Measurement

- **Confusion Matrix :** A confusion matrix is visualized for measuring performance to come up with a judgment of the correctness of your model. It will give you a summary of your model's predictions against actual outcomes, enabling you to drill down into how well a model is doing across every class.As shown in figures 6 and 7.

C. Inference

Finally a code snippet is used to demonstrate how to train a CatBoost classifier for the recommendation of a crop by using the various input agricultural characteristics such as nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, pH, and rainfall. It is first assumed that any input training and test datasets are predefined. The model CatBoost with a number of iterations = 100, learning rate = 0.1 is then created with default parameters as tree depth. The function predict crop accepts input values for agricultural features, creates a DataFrame, normalizes the input using a scaler pre-calculated and used for the training set. It then classifies the crop type by the trained model. Then, using a Label Encoder, it decodes the predicted label back to the original crop name and returns that label.

V. Conclusion

This study has been prepared to show high potential machine learning algorithms with the aim of optimization of agricultural practices through efficient crop recommendation. The boosting algorithms advanced in the analysis presented here are XGBoost, CatBoost, Gradient Boost, and AdaBoost. Here, we have shown how critical factors like soil nutrient levels and climatic conditions can be analyzed in order to predict which crops are suitable for cultivation. These results show that performance metrics for XGBoost and CatBoost are systematically higher while also recommended choices for more accurate crop prediction, and hybrid seems like an acceptable alternative. In addition, analysis of feature importance suggests the prime influence is attributed to nitrogen and potassium levels in suitability for crops, and further recommendations for data collection and research endeavors based on this can be proposed. According to the results, data quality, the interpretability of the models, and region-specific adaptations are important considerations in the use of machine learning for agriculture. In summary, this work outlines user-friendly, data-driven tools for accelerating decision-making in crop selection while enhancing productivity in agriculture at the face of increased global demands for food. Further development of such models, as well as integration with real-time data streams from the environment, may be needed to further enhance their performance and robustness in multiple agro-environments.

References

[1] Alka Chaudhary,Gaurav Chauhan, C. Author, "Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables," *Heliyon*, vol. XX, no. XX, pp. 1-10, 2024.https://www.sciencedirect.com/science/article/pii/S2405844024011435

[2] Gaurav Chauhan, Alka Chaudhary, "Crop Recommendation System using Machine Learning Algorithms," in *Proceedings of the IEEE Conference*, 2021. https://ieeexplore.ieee.org/document/9676210

[3] Dhruvi Gosai, Chintal Raval, Rikin Nayak, Hardik Jayswal, Axat Patel,"Crop Recommendation System using Machine Learningin IJSRCSEIT," *IJSRCSEIT*, vol. XX, no. XX, pp. 100-110, 2020. https://ijsrcseit.com/paper/CSEIT2173129.pdf

[4] Ms. Sarika Gambhir , Manish Sharma , Khushboo Agarwal , Keshav Kumar , Lakshya Kumar , Mayank Chaudhary, "Crop Recommenda tion System Using Machine Learning in IJRASET,"*IJRASET*, vol. XX, no. XX, pp. 200-210, 2023. https://www.ijraset.com/best-journal/crop-recommendation-system-using-machine-learning-879

[5] Pradyot Ranjan Jena a , Purna Chandra Tanti a, , Keshav Lall Mahar jan, "Journal of Agriculture and Food Research," *Journal of Agriculture and Food Research*, vol. XX, no. XX, pp. 300-310, 2023. https://www.sciencedirect.com/journal/journal-of-agriculture-and-food-research/vol/11/suppl/C

[6] R Jaichandran, T. Murali Krishna,Sri Harsha Arigela, Ramakrishnan Raman, N Dharani, Ashok Kumar, "Light GBM Algorithm based Crop Recommendation by Weather Detection and Acquired Soil Nutrients," in *Proceedings of the IEEE Conference*, 2022. https://ieeexplore.ieee.org/document/10047765

[7] Benny Antony, "Prediction of the production of crops with respect to rainfall IJISAE," *IJISAE*, vol. XX, no. XX, pp. 400-410, 2023. https://www.sciencedirect.com/science/article/abs/pii/S00139351

[8] Dataset https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset