A PROJECT REPORT

on

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

**Submitted**

by

|  |  |
| --- | --- |
| 221FA04227  Rohini A | 221FA04456  Mokshagna P |
| 221FA04567 221FA04733  Nandini T Tejashwee N | |

**Under the guidance of**

*Mr. Sourav Mondal*

*Designation*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY**

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“Ensemble machine learning approaches for crop recommendation system using hybrid feature selection techniques”** that is being submitted by 221FA04227 ( Rohini A), 221FA04456(Mokshagna P), 221FA04567( Nandini T), 221FA04733( Tejashwee N)for partial fulfilment of Project is a bonafide work carried out under the supervision of Ms.Ch.PUSHYA, M.Tech., Assistant Professor, Department of CSE.

|  |  |  |
| --- | --- | --- |
| Guide name& Signature |  | Dr.K.V. Krishna Kishore |
| Assistant/Associate/Professor, CSE | HOD,CSE | Dean, SoCI |



**DECLARATION**

We hereby declare that the Field Project entitled **“Ensemble machine learning approaches   
for crop recommendation system using hybrid feature selection techniques”** is being   
submitted by 221FA04227 (Rohini A), 221FA0445 (Mokshagna P), 221FA04567 (Nandini T), 221FA04733 (Tejashwee N)in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Ms.Ch.PUSHYA, M.Tech., Assistant Professor, Department of CSE.

By

**221FA04227 (Rohini A),**

**221FA04456 (Mokshagna P),**

**221FA04567 (Nandini T)  
 221FA04733 (Tejashwee N)**

Date: 21-10-2024

## 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:  
 XG Boost, Gradient Boost, AdaBoost, and Cat Boost, 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.

**TABLE OF CONTENTS**

1. [Introduction 1](#_bookmark0)
   1. What is a Crop Recommendation System and Why is it Important? 2
   2. Challenges Currently Confronted in Crop Recommendation Systems 2
   3. Significance of Accurate Predictions of Weather for Sustainable Agriculture 2
   4. Agricultural Development through Machine Learning 3
   5. Applications of Machine Learning in Crop Recommendation 3
2. Literature Survey 4
   1. [Literature review 5](#_bookmark6)
   2. [Motivation 6](#_bookmark7)
3. [Proposed System](#_bookmark8) 8
   1. [Input dataset 9](#_bookmark9)
      1. Detailed features of dataset 10
   2. [Data Pre-processing 11](#_bookmark10)  
      3.2.1 Data Cleaning 11  
      3.2.2 Normalization 11  
      3.2.3 Correlation 12  
      3.2.4 Train Test splitting 13

3.2.5 Feature selection 14

* 1. [Model Building 16](#_bookmark14)
  2. [Methodology of the system 20](#_bookmark16)
  3. [Model Evaluation 21](#_bookmark17)
  4. [Constraints 24](#_bookmark18)

1. Experimentation and Result Analysis 25
2. Conclusion 30
3. References 34

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1. Correlation Heatmap | 12 |
| Figure 2. Distribution of Target Variables | 13 |
| Figure 3. Chi-Square Selected Features | 15 |
| Figure 4. *Lasso Regression Coefficients* | 16 |
| Figure 5. confusion matrix for XGBoost | 17 |
| Figure 6. confusion matrix for CatBoost | 18 |
| Figure 7. proposed architecture | 20 |
| Figure 8. Train and test | 25 |
| Figure 9. Nutrient requirements for agricultural crops  Figure 10. Prediction pair plot | 26  32 |

**LIST OF TABLES**

Table 1. Performance Table 33

# CHAPTER-1 INTRODUCTION

### INTRODUCTION

The world’s growing population is driving the need for increased food production, but this must be achieved sustain ably 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.

**1.1 What is a Crop Recommendation System and Why is it Important?**Some of the best recommendations in the crop recommendation systems include the use of some algorithms in machine learning that analyse huge datasets of soil properties, nutrient levels, and climatic conditions to predict the best crops for a given region. This will be very important since farmers receive accurate recommendations, they can make decisions based on them. Crop recommendation systems give farm productivity a great advantage, conserve resources, and promote sustainable agriculture by identifying crops which are most likely to thrive in specific environmental conditions. Crop selection aligned with local soil and climate factors greatly enhances the efficiency and yield of farming operations​  
  
**1.2 Challenges Currently Confronted in Crop Recommendation Systems**  
Challenges involve demanding quality, localized data apart from handling complex datasets in regards to the environment. These models will have to take into account issues of soil variability, pH levels, and climate shifts to produce a correct prediction.  
  
**1.3 Significance of Accurate Predictions of Weather for Sustainable Agriculture**Crop recommendation is very fundamental and critical to sustainable agriculture, which provides maximum productivity in land utilization while maintaining minimal environmental degradation. Farmers avoid improper applications of water, fertilizers, and pesticides because they use crops that are suited to the specific soil and climate of the region. This prevents certain environmental impacts, such as soil depletion and pollution. Improper crop choice also eliminates crop losses from farms as the chances of crop failure are minimal and based on inappropriate cultivation conditions. The introduction of correct models through machine learning provides better land utilization with increased food security and encourages environment-friendly farming practices​.  
  
  
  
**1.4 Agricultural Development through Machine Learning**Some recent developments on findings in machine learning, such as ensemble models, which comprise ensemble models of such tools as XGBoost, Gradient Boost, AdaBoost, and CatBoost, have been known to increase the effectiveness of crop recommendation systems. These are based on source data that involve the prediction of ideal crops based on certain environmental variables.  
  
**1.5 Applications of Machine Learning in Crop Recommendation**  
Models that can be trained to evaluate and predict which crop is suitable for what condition of the soil are XGBoost, Gradient Boost, and hybrid models. These models usually combine decision trees and gradient boosting to reduce errors and consequently make the recommendations safe for sustainable farming.

# CHAPTER-2 LITERATURE SURVEY

## LITERATURE SURVEY

#### Literature review

A literature survey is a systematic examination of existing research on a particular topic. It serves as the foundation for any scholarly investigation, offering insights into current knowledge, identifying research gaps, and providing context for new studies. By synthesizing and summarizing relevant literature, researchers can formulate precise research questions, build upon existing work, and avoid duplication. In essence, a literature survey is an essential tool for ensuring the validity and relevance of new research within the broader academic landscape.

A literature survey is a systematic examination of existing research on a particular topic. It serves as the foundation for any scholarly investigation, offering insights into current knowledge, identifying research gaps, and providing context for new studies. By synthesizing and summarizing relevant literature, researchers can formulate precise research questions, build upon existing work, and avoid duplication. In essence, a literature survey is an essential tool for ensuring the validity and relevance of new research within the broader academic landscape.

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.

#### Motivation

India's agricultural sector, which provides livelihood for a large portion of the population, faces significant challenges in maximizing crop yield due to the country’s diverse climatic conditions, soil types, and nutrient deficiencies. To sustainably increase agricultural and horticultural productivity, it is essential to optimize resource usage, particularly fertilizers like Nitrogen (N), Phosphorus (P), and Potassium (K), and ensure compatibility with soil health indicators like pH. Additionally, climatic variables such as temperature, rainfall, and humidity further complicate farming decisions.

Despite advancements in farming techniques, many farmers still rely on traditional methods or limited advisory services for crop selection and fertilizer usage, which often leads to suboptimal yields, soil degradation, and resource inefficiency. This issue is particularly prevalent in rural areas where access to real-time agricultural advice is scarce.

Machine learning (ML) presents an opportunity to revolutionize crop farming recommendations by analyzing complex, multivariate data that includes soil properties, climatic conditions, and nutrient availability. By leveraging ML models, personalized recommendations can be provided to farmers on which crops to grow and how to manage nutrient inputs efficiently. These models can predict the best crop choices based on real-time and historical data, tailored to the specific environmental conditions of a region, leading to improved yields, soil health, and sustainable farming practices.

Incorporating NPK, soil pH, and climatic variables in these recommendations is critical to ensure that the advice is holistic and precise. This approach aligns with the broader goal of promoting data-driven, sustainable agriculture in India, thereby improving food security, farmer income, and environmental stewardship.

**Key Points:**

1. **Maximizing Crop Yield**: Farmers need efficient crop recommendations that match the specific nutrient and environmental conditions of their land.
2. **Soil Health & Sustainability**: Addressing soil pH and optimizing NPK usage can prevent long-term soil degradation.
3. **Climate-Smart Agriculture**: Considering climatic variables ensures that recommendations are aligned with environmental changes.
4. **Technology for Rural Empowerment**: Machine learning can democratize access to expert-level agricultural advice in rural areas.
5. **National Significance**: This approach supports India’s agricultural growth while ensuring sustainable practices and food security.

# CHAPTER-3 PROPOSED SYSTEM

### PROPOSED SYSTEM

It recommends crops after detecting the NPK of the soil, pH level and atmospheric conditions like rain. Within this, it uses five more advanced machine learning boosting algorithms (XG Boost, Gradient Boosting, AdaBoost and Cat-boost) to achieve higher prediction accuracy along with the Hybrid algorithm. The data employs agricultural dimensions like NPK values, temperature, pH value and rainfall/humidity/ moisture levels along with the 11 basis for crop rows in agriculture & extra knowledge of more than 10 crops Horticulture done today. It aims to serve as a decision-support system for sowing decisions based on their agro-climatic suitability in order to make farming more productive and environmentally friendly.

**Model Implementation:**

*XG Boost:* A gradient boosting technique that minimizes errors over the learning rate and tree depth-based iterations.

*Gradient Boost:* Building sequential models focused on minimizing the residual errors from previous models.

*AdaBoost:* Pay more attention to the misclassified instances by penalizing them via weight adjustments, so they are corrected in future iterations.

*Cat Boost:* Particularly effective with categorical variables, helps to make the models prediction less biased leading up to better model accuracy and efficiency.

*Hybrid Model:* It is a fusion of all the above algorithms, in order to increase predictive strength by averaging or stacking predictions from them.

#### Input dataset The input dataset used in this system is fetched from Kaggle, covering data points about agricultural factors and crop yield. It includes values for 11 major agricultural crops and 10 horticultural crops. Key variables in this dataset include NPK (Nitrogen, Phosphorus, Potassium), temperature, humidity, pH, and rainfall.

#### Detailed Features of the Dataset

**N (Nitrogen):** It represents the nitrogen present in the soil, gauged in terms of kg/ha.  
**P (Phosphorus):** The variable represents the amounts of phosphorus in the soil, measured also in kg/ha  
**K (Potassium):** The variable represents the amounts of potassium in the soil  
**Temperature:** The variable is measured in degrees Celsius and influences crop growth  
**Humidity:** This variable is expressed as a percentage and influences crop development through moisture levels in the environment  
**pH:** The pH level, together with the measurement, will describe the alkalinity or acidity of the soil, which forms a big component to understand the overall quality of the soil.  
**Rainfall:** In millimetric value, it tells one the amount of water that is usable for crops.  
**Target Variable:** The variable to be predicted in the dataset, which specifies which crop should be grown under those given environmental conditions and soil conditions.

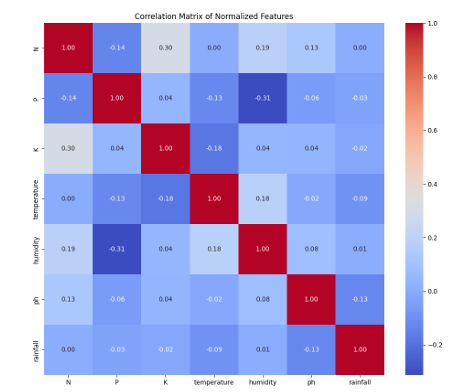
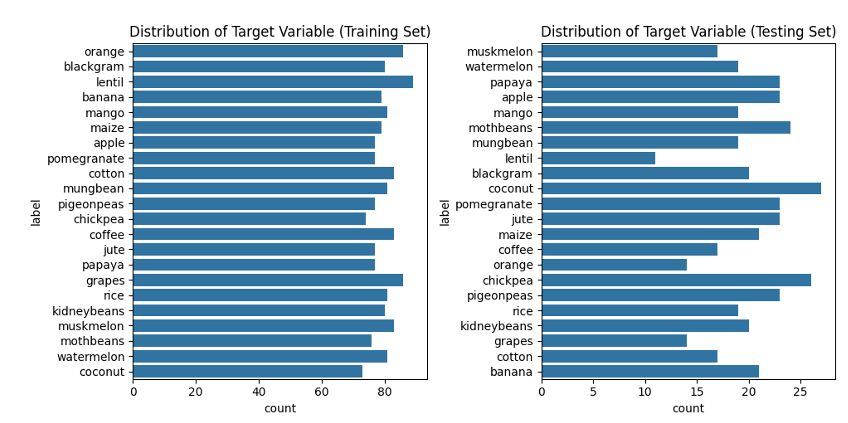
#### Data Pre-processing

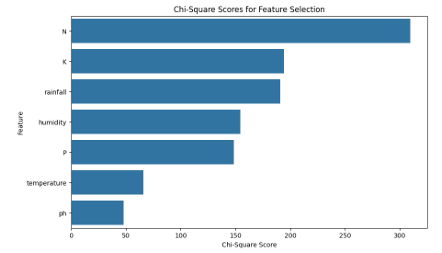
Data pre-processing is the essential process of preparing raw data for analysis and modelling by cleaning, transforming, and structuring it to enhance data quality and utility. It involves tasks like handling missing values, correcting errors, encoding features, and scaling data to ensure it's in an optimal form for further analysis. It encompasses a range of operations and transformations designed to refine raw data, ensuring that it is clean, structured, and amenity subsequent analysis. This process is driven by its manifold significance in data science and analysis.

Through meticulous data cleaning, transformation, feature engineering, dimensionality reduction, outlier handling, scaling, and data splitting, it prepares raw data for more accurate and reliable analysis and modelling. Ultimately, the goal is to obtain more meaningful insights, make informed decisions, and optimize predictive models for a wide range of applications in data science and analysis.

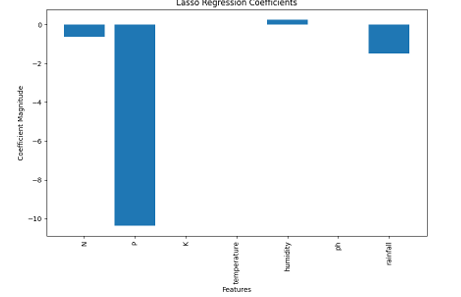
**Discuss on various preprocessing techniques that you have applied for your project**

**3.2.1 Data Cleaning:**It involves the removal of missing values like label encoding, feature scaling and even non-negative transformations to make sure that data is in a form such that it can be used for machine learning algorithms.

**3.2.2 Normalization:**   
the features were normalized to have an equal influence on model development when all variables are in different scales.  
 **3.2.3 Correlation:**  
**1.Axes and Features:**  
The correlation matrix shows all pairs of the variables measured: N (Nitrogen), P (Phosphorus), K (Potassium), temperature, humidity, pH, and rainfall.  
The color for every cell in the matrix is a correlation coefficient, ranging from -1 to 1.  
**2.Color Gradient:**  
Color Scale The color bar on the right illustrates how correlations appear: dark red is used to indicate positive correlations, closer to 1, while dark blue indicates near negative correlations, closer to -1, and lighter hues for weaker or no correlations whatsoever, closer to 0.  
*Example:* Note that we assign deep red colour to represent correlation of 1-as seen in the diagonal cells where each variable is correlated with itself-and use light blue for near-zero correlations.  
**3.Key Takeaways**  
*N and K (0.30):* The concentration of Nitrogen is positively correlated to the level of Potassium at a moderate rate.  
*P and humidity (-0.31):* There is a moderate negative correlation for the level of Phosphorus and humidity.  
*Temperature and other variables:* Apparently, there's little or no relation between temperature and the other variables but somehow it is correlated to the slightly negative side with K, where the correlation stands at -0.18.  
*Humidity and pH (0.08):* Humidity is slightly positively correlated with pH.  
The matrix will provide information to identify which features in a data set are related to each other, which might be valuable to understand dependencies or potential multicollinearity in models.  
  
*fig1: Correlation Heatmap* **3.2.4 Train Test splitting:**  
Split the data into 80% for training and 20% for testing.  
  
The distribution of the target variable for a training set (on the left) and a testing set (on the right). The target variable seems to be labels of different crops, and the counts indicate the frequency with which each crop occurs in both sets.  
**1.Training Set (Left):**  
Some crops like rice, coconut, and watermelon have higher frequencies, while some such as pigeonpeas and mungbean have lower frequencies.  
*Distribution:*  
It can be noted that the dataset is a bit balanced but some crops dominate with higher counts, such as coconut, watermelon, and rice.  
**2.Testing Set (Right ):**  
The testing set has the same crops as that of the training set but lower counts all around (which is expected since this would be a smaller subset meant to test this model).  
It so happens that the crop-wise occurrences of coconut, pigeonpeas and pomegranate are higher while that for lentil, orange and kidney beans occurs less often .  
Although the size is small, distribution in testing set goes reasonably well with the training set that is required by accurate model evaluation.  
***Observations:***  
Both the Charts 1 and 2 depict a rather balanced dataset; however, some crops appear more frequently than others.  
The splitting process would probably be random and stratified so that the crops are proportionally represented in both the training set and the test set, and this similar distribution is a good sign.  
This sort of data visualization is beneficial in ensuring the representativeness of a dataset utilized for training and testing and supports the building and evaluation of predictive models.  
  
 *fig 2: Distribution of Target Variables*

**3.2.5 Feature Selection:** We used chi-square tests, along with lasso regression for feature selection which was helpful to remove non informative data and narrowed the key variables into statistically significant one.  
**3.2.5.1 Chi Square Test:**The Chi Square test is typically utilized in feature selection to identify the statistical correlation present for each classification problem between the feature and the target variable.  
Key Features:  
***X-axis (Chi-Square Score):*** This is the Chi-Square score of each feature. A high score implies that it's more relevant in the prediction of the target variable of interest.  
***Y-axis (Feature):*** The following are the features under consideration: Nitrogen (N), Potassium (K), rainfall, humidity, Phosphorus (P), temperature, and pH  
**Observations:***N (Nitrogen):*Has the highest Chi-Square value, higher than 300, meaning that it is the most relevant feature with regard to the target variable.  
  
*K (Potassium) and rainfall:*  
These features also possess high values although they are not quite as high as Nitrogen.  
This means that they are relevant predictors but not so influential as N.  
*Humidity and P (Phosphorus):*  
These features are relatively important but scored only the moderate importance, where humidity was scored higher than P.  
Both of these remain somewhat predictable but are much less impactful than N, K and rainfall.  
  
*Temperature and pH:*These remain to be the poorest-rated Chi-Square features, meaning that in comparison with all other features, they contribute the least to predicting the target variable.  
pH, in particular, remains far apart with a much lower Chi-Square rating than all the other features.  
***Conclusion:***  
The chart conveys the features of importance in the dataset with Chi-Square scores. Those features scored high, such as Nitrogen and Potassium, rainfall, should be considered first when building the model since they are highly associated with the target variable. Conversely, features relatively scored lower, such as temperature, pH, might not significantly add value and thus possibly could be removed from the model without significant sacrifice to its predictive power.  
  
  
 *fig 3: Chi-Square Selected Features*

**3.2.5.2 Lasso regression:**Features in the x-axis are labelled as N, P, K, Temperature, Humidity, pH, and Rainfall. The y-axis represents the magnitude of the coefficients.  
*Feature Importance:*   
The length of each bar represents how important the feature is to predict the target variable. Features with longer bars have a higher impact on the predictions of the model.  
Coefficient Direction Specifies if the given coefficient is positive or negative, which defines the direction of influence of features on the target variable. The larger the feature value, the greater the value of the target variable is provided that coefficient is positive, and vice versa, if the coefficient is negative, then the bigger the value of the feature is, the lower the target variable is.  
*Feature Selection:*   
Lasso regression has its regularization technique that shrinks the coefficients of smaller-important features toward zero to prevent overfitting. You will see in this plot that some coefficients go closer to zero; this might not be very significant in being predictors.  
>> N and P have relatively large magnitudes of coefficients.  
>> K also have an impact, but a much smaller.  
>> Coefficients for temperature, humidity, pH, and rainfall are moderate values. This shows that there is little to no significant influence of these variables towards predicting the target variable.



*Fig 4: Lasso Regression Coefficients*

#### Model Building

>> **Data Collection:**

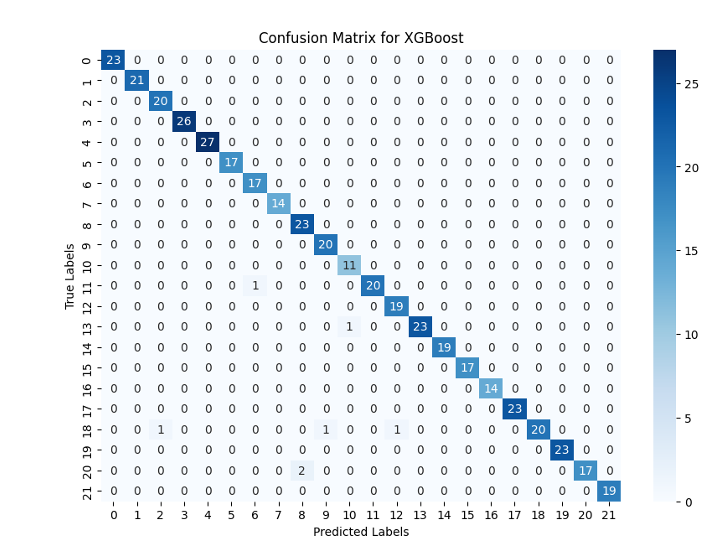
* The data used for building the models was sourced from **Kaggle**, specifically from a dataset curated by the Indian Chamber of Food and Agriculture.
* The dataset contains **2100 data points** covering **11 agricultural crops** and **10 horticultural crops**.
* Key variables include:
  + **N (Nitrogen)**: The nitrogen content in the soil.
  + **P (Phosphorus)**: The phosphorus content.
  + **K (Potassium)**: The potassium content.
  + **pH**: The alkalinity or acidity of the soil.
  + **Temperature**: Measured in degrees Celsius.
  + **Humidity**: The percentage of relative air humidity.
  + **Rainfall**: Measured in milli meters.
  + **Label**: Specifies which crop should be grown based on the conditions.

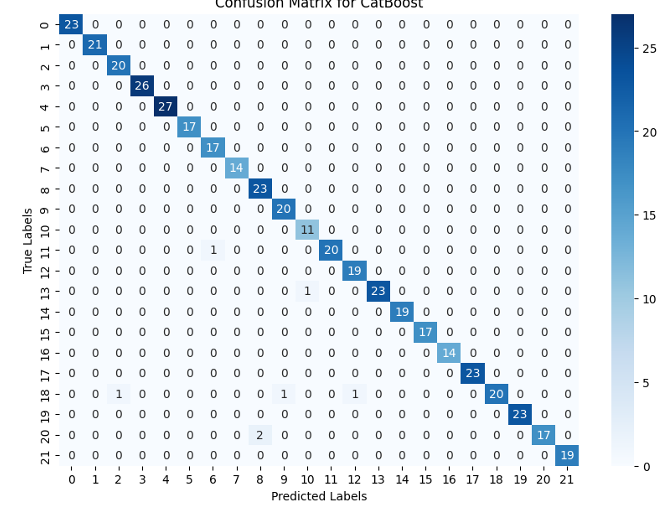
>> **Data Preprocessing:**

* **Data Cleaning:** Missing values were handled through **label encoding**, **feature scaling**, and **non-negative transformation** to make the data usable for machine learning models. These preprocessing techniques ensure that numerical variables are on a similar scale and that the data is free from inconsistencies.
* **Data Normalization:** The features were normalized to allow models to treat all variables equally, regardless of their scale.
* **Data Splitting:** The dataset was split into **80% for training** and **20% for testing**, which is a standard practice to ensure that the models can generalize well on unseen data.
* **Feature Selection:**
  + **Chi-Square Test:** This was used to identify features based on their statistical significance.
  + **Lasso Regression:** This method was applied to reduce the impact of irrelevant features by driving coefficients of irrelevant features to zero.

>> **Model Implementation:** The following machine learning algorithms were implemented in the model-building process:

* **XGBoost (Extreme Gradient Boosting):**
  + **XGBoost** is an ensemble learning method that combines multiple weak learners (decision trees) into a single, strong predictive model.
  + **Gradient boosting** optimizes the model by minimizing the error from previous iterations.
  + Key hyperparameters include **learning rate**, **maximum tree depth**, and the **number of estimators** (trees).

  
 *fig 5: confusion matrix for XGBoost*

* **Gradient Boost:**
  + Similar to XGBoost, Gradient Boosting also combines decision trees in sequence to correct errors from previous models.
  + This method constructs trees sequentially, with each new tree attempting to minimize the residual errors of the prior trees.
  + Key hyperparameters include **learning rate** and **number of boosting stages**.
* **AdaBoost (Adaptive Boosting):**
  + AdaBoost improves model accuracy by focusing on **misclassified instances** in the data.
  + After each iteration, the weights of the misclassified samples are increased so that subsequent classifiers focus on these errors.
  + Important parameters include the **number of weak learners** and **individual learning rates**.
* **CatBoost (Categorical Boosting):**
  + **CatBoost** handles categorical variables natively, improving both model accuracy and efficiency.
  + This model uses an approach called **ordered boosting**, which prevents prediction bias and helps in processing categorical features without heavy preprocessing.
  + Important hyperparameters tuned for this model include **depth** and **learning rate**.  
      
      
     *fig 6: confusion matrix for catBoost*
* **Hybrid Model:**
  + A **hybrid model** was developed by combining the outputs of the above algorithms using techniques like **weighted averaging** or **stacking**.
  + The hybrid approach aims to **minimize the weaknesses** of individual models while **leveraging their strengths** to achieve better predictive performance. This technique boosts the overall accuracy by taking the consensus or the most likely prediction from all models.

>> **Model Tuning:**

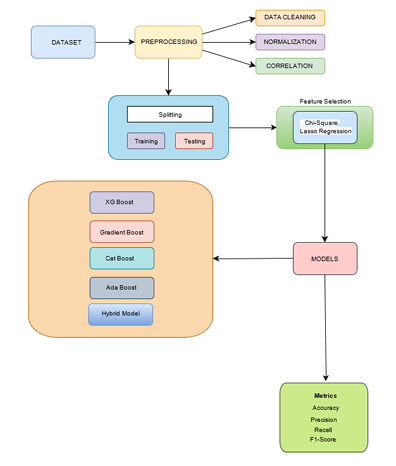
* **Hyperparameter tuning** was performed for each of the models. For example:
  + For **XGBoost**, parameters like **learning rate**, **max depth**, and **number of estimators** were tuned to minimize errors.
  + Similarly, in **AdaBoost**, the number of weak learners was optimized.
* This fine-tuning ensures that the models perform optimally for the crop recommendation task.

>> **Model Integration and Testing:**

* After training, the models were tested on the **20% test set** to evaluate their generalization capability on unseen data.
* The **hybrid model** was found to outperform individual models, as it was able to balance the strengths and weaknesses of the XGBoost, Gradient Boost, AdaBoost, and CatBoost model

#### 3.4 Methodology of the system

Having discussed the foundational elements in the preceding sections, we now venture into the core of our traffic congestion prediction system. In this section, we embark on a journey through the inner workings of our model, unveiling the methodology that drives our system's ability to forecast traffic congestion. Just as a well-orchestrated symphony requires each instrument to play its part harmoniously, our methodology combines data, pre-processing, modelling, and evaluation to create a seamless and efficient prediction system.

**(Discuss on proposed architecture)**  
****

*fig. 7: proposed architecture*

The methodology for developing a crop recommendation system using machine learning involves the following steps:   
A. Data Collection and Preprocessing  
 • Data Sources: The dataset utilized in this research was sourced from the Kaggle repository [39], which was collected over a period of time by the Indian Chamber of Food and Agriculture [40]. It comprises a total of 2100 data points, encompassing 11 agricultural crops and 10 horticulture crops grown under variables for NPK fertilizer, soil pH, and climatic factors such as rainfall, temperature, and humidity. Crop yield data consists of 11 major agricultural crops and 10 horticultural crops.   
• Data Cleaning: Missing values were handled using techniques like label encoding, feature scaling and handling negative values for chi-square test. This help prepare the data for machine learning algorithms by ensuring numerical compatibility and feature comparability.   
• Data Normalization: Features were normalized to ensure that variables on different scales contribute equally to the model’s performance.   
• Data Splitting: The data was split into training (80%) and testing (20%) sets to evaluate model performance accurately.   
B. Model Implementation   
 The following machine learning models were implemented to predict crop suitability:   
• XG Boost (Extreme Gradient Boosting):  
– XG Boost is an ensemble learning technique that combines multiple weak learners (usually decision trees) to form a robust predictive model.  
– It uses gradient boosting to optimize model performance, focusing on reducing the error of previous iterations  
.– Hyperparameters like the learning rate, maximum depth of trees, and the number of estimators were tuned to maximize prediction accuracy.  
 • Gradient Boost:  
– Gradient Boosting is another ensemble technique that builds models sequentially. It aims to correct errors made by previous models in the sequence.  
– Decision trees are constructed in a greedy manner, with each new tree reducing the residuals (errors) from the preceding trees.  
– Key parameters like learning rate and number of boosting stages were adjusted for improved 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 incorrectly predicted samples are increased so that subsequent classifiers pay more attention to these errors.  
– The number of weak learners and their individual learning rates were carefully optimized.  
 • Cat Boost (Categorical Boosting):  
– Cat Boost is designed to handle categorical variables natively, which improves its accuracy and efficiency.  
– It leverages ordered boosting to eliminate prediction bias and deals with categorical data without requiring extensive preprocessing.  
– Model parameters such as depth and learning rate were fine-tuned for the best results.  
 • Hybrid Model Development  
– A hybrid model was developed to leverage the strengths of individual algorithms.  
– The hybrid approach involves combining the predictions of the four models (XG Boost, Gradient Boost, AdaBoost, and Cat Boost) using a weighted average or stacking technique.  
– The goal of the hybrid model is to enhance prediction accuracy by mitigating the weaknesses of individual models and leveraging their collective strengths. This methodology leverages the advantages of each machine learning technique and combines them into a powerful hybrid approach for recommending the most suitable crops based on soil and climatic conditions.

#### Model Evaluation

Model evaluation is a critical aspect of any machine learning project. It involves assessing the performance and accuracy of a trained model on data. This step is essential for several reasons such as:

A competition to build models that predict which crops should be grown in a given region. The hybrid model based on ensemble learning is designed to outperform individual models by combining the strengths of each algorithm. The metrics to measure includes, reduction of error across the model and tuning hyperbolas like learning rate, depth & estimators count for efficient results.  
**>> Performance on Data:**

* After the models (XG Boost, Gradient Boost, AdaBoost, Cat Boost, and the Hybrid Model) are trained on the dataset, they need to be tested on  **data** (20% test set) to evaluate their generalization capabilities. This ensures that the models are not overfitting the training data but can also make accurate predictions on real-world data.

>> **Accuracy Assessment:**

* The system uses various metrics, such as prediction accuracy and error rates, to determine the effectiveness of each model. The **hybrid model** likely improves prediction accuracy by combining the strengths of the other algorithms. The evaluation focuses on how well these models recommend the best crop based on environmental and soil conditions.

>> **Fine-tuning Parameters:**

* During evaluation, **hyperparameters** (such as learning rates, depth of trees, and the number of estimators) are fine-tuned to optimize model performance. This step is essential for improving accuracy and reducing errors. For instance, if XG Boost has a high error rate, tweaking these parameters during evaluation can help minimize it.

>> **Error Reduction:**

* One key objective during model evaluation is to **reduce prediction errors**. This is especially important for agricultural applications where wrong predictions can lead to significant losses. Models like **AdaBoost** focus on minimizing misclassified instances by adjusting weights, and their effectiveness is measured during evaluation.

>> **Selecting the Best Model:**

* The evaluation process helps in identifying the **best-performing model** for practical use. In this case, while individual models (like Cat Boost or Gradient Boost) may perform well, the **hybrid model** is likely chosen for its superior prediction accuracy, which is the result of leveraging the strengths of multiple models.

>> **Addressing Constraints:**

* Evaluation also highlights the **constraints** of the model, such as handling environmental variability, scalability, and computational complexity. If a model performs poorly in certain scenarios (e.g., predicting crops for regions with inconsistent rainfall), the evaluation process helps in identifying these limitations and making necessary improvements.

#### Constraints

In our project, we operate within a framework of specific constraints that shape our approach to designing and developing the traffic congestion prediction system. These constraints ensure that our solution aligns with essential considerations and limitations.

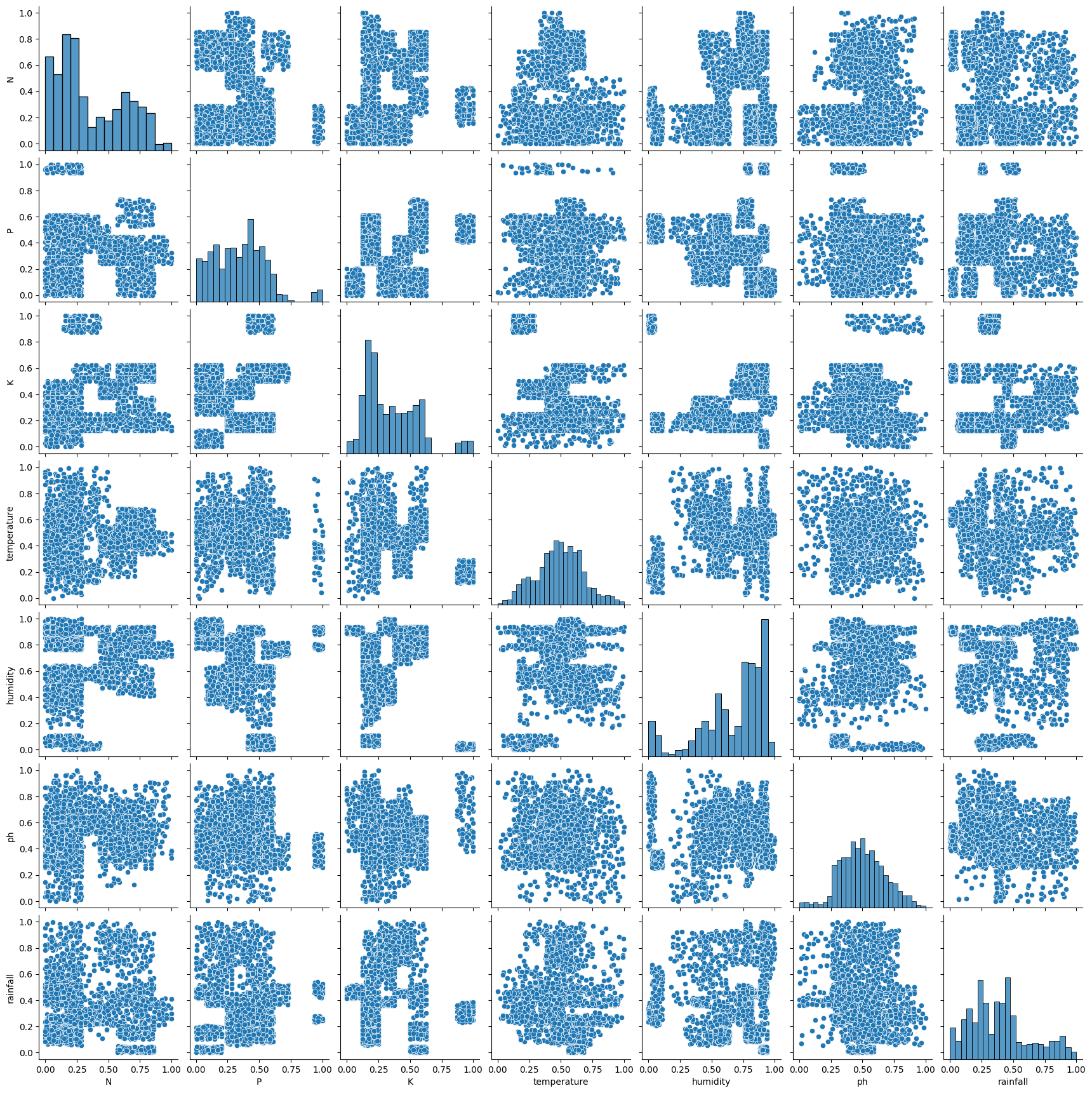
The proposed system faces several constraints:

* **Data Quality:** The accuracy of the recommendation system heavily relies on high-quality and comprehensive datasets. Missing or incorrect data can lead to poor model performance.
* **Environmental Variability:** The system may struggle with real-time environmental changes, as it relies on static datasets rather than dynamic, real-time data.
* **Scalability:** While ensemble methods enhance accuracy, they may also increase computational complexity, making scalability to larger datasets or regions more challenging.
* **Interpretability:** Some machine learning models, especially ensemble techniques, may not be easily interpretable, posing a challenge in practical agricultural decision-making.

This system, by combining multiple machine learning approaches, aims to deliver a robust crop recommendation tool tailored to specific regions and environmental conditions

# CHAPTER-4 EXPERIMENTATION AND RESULT ANALYSIS

**Inference**The joint distributions of six variables: N, K, temperature, humidity, pH, and rainfall. Each subplot in the plot in the main panel is a pairwise comparison of two variables: the joint distribution of the two and marginal histogram of each.



#### *Fig 10: prediction pair plot*

This is a **pairplot**, commonly generated using the Seaborn library in Python. Pairplots are useful for visualizing relationships between multiple variables in a dataset.

**Key Observations:**

1. **Diagonal Plots**:
   * The diagonal contains histograms, showing the distribution of each individual variable (like N, P, K, temperature, humidity, ph, and rainfall).
   * This helps identify skewness, multimodality, or uniformity in the distribution of these variables.
2. **Scatter Plots (Non-Diagonal Cells)**:
   * Each scatter plot shows the relationship between two variables, with one variable on the X-axis and the other on the Y-axis.
   * You can use these scatter plots to detect patterns, correlations, or clusters. For example, if points align along a diagonal trend, there is likely a correlation between the two variables.
3. **Use Case**:
   * Since variables like temperature, humidity, ph, and rainfall are environmental factors, this visualization can be useful for understanding how these factors interact. This is likely from an agriculture or environmental dataset where nutrients (N, P, K) are also measured.
4. **Insights to Look For**:
   * Correlations: Strong or weak relationships between variables, e.g., does rainfall have any linear or non-linear correlation with humidity or ph?
   * Outliers: Are there any isolated points that look out of place?
   * Clusters: Are there visible clusters of points indicating subgroups in the data?

This pairplot is likely meant to analyze potential **dependencies** between variables for tasks like prediction or classification, possibly as part of an **agriculture-related predictive model** (aligned with your interest in rainfall prediction for agriculture). You could further explore correlations by overlaying regression lines or using a heatmap with Pearson correlation coefficients.

#### CHAPTER-5 CONCLUSION Distributions: N and K: These are right-skewed distributions indicating that they have a preponderance of lower values with some extreme outliers. Temperature, Humidity, and pH: These variables are well described as normal, indicating a central tendency with roughly symmetrical spread. Rainfall: It is potentially bimodal distribution, which might indicate that there are two possibly different types or regimes of rainfall. Correlations: N and Temperature: Weak positive correlation where higher N shows a relation to higher temperatures N and Humidity: The correlation seems weak negative where greater humidity will occur when N is smaller. K and Temperature: The correlation showed a strong positive correlation indicating that higher values of K correspond to higher temperatures. K and Humidity: There is a very significant negative correlation, indicating an inverse relationship between K and humidity. Temperature and Rainfall: A weak positive correlation was obtained, suggesting a slight tendency for higher rainfall amounts to be associated with higher temperatures. Humidity and Precipitation: This is weakly negatively correlated, implying low precipitation to tend to be associated at higher humidities. The six variables will further depict their interrelation in a pair plot. A first test for such observed distributions and correlations would alert to further investigating and understanding the underlying patterns or dependencies in the data.

**Model Conclusion**

- **Best Performers**: **XGBoost** and **CatBoost** are the top performers, achieving

  near-perfect results. They should be preferred when aiming for the highest accuracy.

- **Good Alternative**: **GBM** and the **Hybrid Model** also perform well, with slightly lower accuracy but still excellent results.

- **Underperforming**: **AdaBoost** significantly underperformed in this scenario, likely due to its inability to capture complex patterns, which might indicate it is not the best choice for this dataset.

**** *Table 1: Performance table* **Feature Conclusions:**

The feature importance plot from the XGBoost model indicates that N (Nitrogen) and K (Potassium) are the most influential features in determining the model's predictions. Rainfall and humidity also play significant roles, while \*\*pH\*\* has the least impact among the features analysed. This insight can guide further analysis or data collection efforts, focusing on the most critical features.

In summary, XGBoost and CatBoost show superior performance and robustness, making them ideal for complex classification tasks. GBM and Hybrid models are reliable alternatives, while AdaBoost may not be suitable without further tuning or preprocessing.

In 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 learn ing 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/S24058440240114350  
 [2] Gaurav Chauhan, Alka Chaudhary, “Crop Recommendation System using Machine Learning Algorithms,” in Proceedings of the IEEE Conference, 2021. <https://ieeexplore.ieee.org/document/967621>   
[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 Fig. 8: Confusion Matrix for XGBoost 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, Ramakrish nan 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