Efficacy of Machine learning algorithms for the prediction of farm produce in Assam, India

Masters of Technology Phase II
Report

Submitted by

Sushant Swarup

(224154008)

Under the Supervision of

Prof. Ramagopal V.S. Uppaluri

School of Agro and Rural Technology

Indian Institute of Technology, Guwahati

May 2024

Certificate

This is to certify that the work contained in this thesis entitled "**Efficacy of Machine learning algorithms for the prediction of farm produce in Assam, India**" is a bonafide work of **Sushant Swarup** (**Roll No. 224154008**), carried out in the School of Agro and Rural Technology, Indian Institute of Technology, Guwahati under my supervision and it has not been submitted elsewhere for a degree.

Supervisor: **Prof. Ramagopal V.S. Uppaluri Professor**

Professor,

May, 2024 Department of Chemical Engineering,

Guwahati. Indian Institute of Technology Guwahati, Assam.

Acknowledgements

With profound gratitude, I express my heartfelt thanks to the many individuals who have played a pivotal role in supporting and guiding me throughout my time-bound journey towards obtaining my M.Tech degree at IIT Guwahati.

Foremost, I extend my deepest appreciation to my esteemed supervisors **Prof. Ramagopal V. S. Uppaluri** of the Department of Chemical Engineering at IIT Guwahati. His consistent guidance, moral support, patience, motivational encouragement, and insightful knowledge-sharing discussions were instrumental in shaping my M.Tech experience.I am forever grateful for their professional partnership and their empathetic approach to the intricate challenges that often accompany one's academic and personal journey.

I also wish to express my sincere thanks to Ms. Tinka Singh and my fellow batchmates. Collaborating with them and benefiting from their guidance has enriched my understanding of research, projects, and life in general.

Finally, I extend my thanks to the Divine for the blessings that fortified my physical, psychological, intellectual, and emotional resilience throughout the tenure of my M.Tech thesis. These blessings were instrumental in overcoming the myriad challenges and obstacles that presented themselves on the path to realizing this cherished and significant goal in my life. .

Sincerely

Sushant Swarup

Abstract

Crop production prediction heavily relies on machine learning (ML) as a crucial tool for decision-making support, including recommendations on which crops to grow and how to reduce crop loss, among other things. Even though agriculture continues to be the world's most common economic activity, climate change has had a significant adverse effect on it in recent times, which has increased food insecurities. This is due to the fact that extreme weather events brought on by climate change harm the majority of crops and reduce the anticipated level of agricultural production. While it may not be possible to fully prevent such natural phenomena, farmers can significantly benefit from having prior knowledge of future events, as it allows them to make appropriate plans and preparations. In this context, this work uses ML approaches to predict agricultural harvests for the Kamrup(rice)-Jorhat(wheat) district(r and Barpeta(toor) -Golaghat(masoor) in the future, including those for rice, wheat,and other crops . The study uses ML approaches to transmit information about production trends and predict of Crop Production based on weather data. From MERRA-2 and MAFW, weather information and Crop production for the aforementioned crops were acquired. Liner regression(LR), Decision Tree(DT), Random Forest(RF), KNN, Ada-boost, SVM, Decision Tree, Naive Bayes, Gradient boosting ,ANN were used to assess the data that had been gathered. Predictors included temperature, wind speed, precipitation, relative humidity, wind direction and area. The findings show that the Ada-boost is performing good with Train and Test R2 scores are respectively (0.97; 0.84), (0.96; 0.89), . Hybrid models are further employed for long- term predictions. The results of this study will significantly increase the reliance on data for decisions relating to climate change and agriculture, particularly in low-to-middle income.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Agriculture is the mainstay for a significant chunk of India's population, especially in rural areas. With 70Spotting and fine-tuning these crucial factors is key to amping up crop production, directly impacting profitability and economic growth (Salpekar, 2019) [1]. Unlike predicting crop yield (CY), crop production is a broader game, involving considerations like climatic conditions, weather quirks, soil quality, fertilizer use, and seed quirks (Xu et al., 2019). Throwing machine learning (ML) algorithms into the mix to predict crop production is a tough but necessary move in precision agriculture. Many models have taken a swing at this, dealing with diverse datasets to cover all bases in the complex realm of crop production [2]. Current models do an okay job at predicting crop production, but the hunger for more accurate forecasts using straightforward ML techniques, especially with bigger datasets and various crops, is real. powering up district-level crop production needs with ML is a smart move. It's not just about crunching numbers; it's about making a real impact on how we do agriculture in the country. ML steps in to help farmers cut losses and join forces with other farming buddies like horticulture and sericulture. Plus, tech tools like artificial intelligence (AI) and ML are in the mix, making things even more interesting. ML techniques aren't just about data – they're about digging up info that regular methods might miss. By predicting future patterns and traits, ML helps spot the big factors in decision-making, giving farmers solid insights and saving them time . For example. They used ML to help farmers get the best out of their crops in specific spots, even predicting which crops would make the most money [3].found that throwing in extra details about crops, like fertilizers and where they're grown, made predictions way more accurate. The Random Forest (RF) algorithm stood out, showing off against old-school methods like linear regression and decision tree (DT), thanks to its knack for handling

a bunch of variables and making spot- on predictions. Jambekar et al. (2018) [4] went deep, testing ML classification tools like J48, Bayesian inference, and simple cart on a big set of data (218 data points). The 10-fold validation gave a thumbs-up to ML techniques in predicting crop production. This shouts out loud that ML is no sidekick; it's a game-changer in the world of predicting crop production prediction. Few authors investigated the performance of ML algorithms such as DT, polynomial regression and RF algorithms to predict. Thereby, the authors concluded that the RF algorithm performed better to predict crop yield. Further, the DT model performed well for alterations in the dataset and concluded to be more effective to predict the CY. Statistical models and, more recently, tools from machine learning (ML) have been used to model crop yield variability using meteorological indices as inputs. Previous studies have shown the importance of growing season-averaged temperature and precipitation in explainingcrop yield variability. Winter Wheat, for example, has been shown to be particularly susceptible to freezing temperatures during Fall and to heat stress during grain filling and stem elongation (Tack et al., 2015) [5]. This vulnerability to extreme temperatures is believed to be thereason behind a decline in wheat yields across Europe (Brisson et al., 2010). As per a differentstudy (Schauberger et al., 2017), each day above 30∘C causes a decline in maize and soybean yields by upto 6 under rainfed conditions. Similarly, the interannual variation in rainfall alsohas a crucial role to play in crop growth. Although a few studies did consider extreme meteorological indices in their analysis, their scope was either restricted to measuring conditional relationship with yields (Troy et al., 2015) or the extreme event types considered were limited (Lobell and Burke, 2010; Lesk et al., 2016). Nonlinear and threshold-type relationships have been shown to exist between yields and meteorological indices (Schlenker and Roberts, 2009; Lobell et al., 2011a; Troy et al., 2015) [6]. However, most of the previous studies have modeled this nonlinearity using regression models with quadratic terms for meanmeteorological indices without appropriate justification. Understanding the exact relationshipbetween meteorological outcomes and yield is essential given that a prior study reported a significant stagnation and declines in yield for major cereal crops on more than a quarter of global croplands.

1.1 Application of Machine Learning in Crop production Prediction

Machine learning (ML) has found numerous applications in crop production prediction, revolutionizing the way agriculture is practiced. Here are some key areas where ML is applied in crop production prediction:

1.1.1 Yield Prediction

ML models can analyze historical data, including weather patterns, soil conditions, and previous crop yields, to predict future crop yields. This information helps farmers plan for optimal harvesting times, estimate potential profits, and allocate resources efficiently.

1.1.2 Disease Detection and Management

ML algorithms can analyze images of crops to identify signs of diseases or pests. Drones and satellite imagery equipped with ML can monitor large agricultural areas, enabling early detection and intervention to prevent the spread of diseases and reduce crop losses.

1.1.3 Weather Forecasting

ML models can analyze weather data to provide accurate short-term and long-term forecasts. This information is crucial for farmers to make informed decisions about planting, irrigation, and harvesting schedules.

1.1.4 Precision Agriculture

ML enables precision farming by analyzing data from sensors, drones, and other sources to optimize the use of resources such as water, fertilizers, and pesticides. This helps reduce costs, minimize environmental impact, and improve overall efficiency in crop production.

1.1.5 Soil Health Monitoring

ML can analyze soil data, including nutrient levels and moisture content, to assess soil health. This information guides farmers in making informed decisions about soil amendments and crop selection, contributing to sustainable and productive agriculture.

1.1.6 Crop Selection and Rotation

ML algorithms can analyze data on different crops' performance under specific conditions, helping farmers choose the most suitable crops for their land. This also aids in crop rotation planning to maintain soil fertility and reduce the risk of pests and diseases.

1.1.7 Market Price Prediction

ML can analyze market trends, demand-suppl y dynamics, and other factors to predict crop prices. Farmers can use this information to make strategic decisions about when to sell their produce, potentially maximizing profits.

1.1.8 Automated Machinery and Harvesting

ML is employed in the development of autonomous agricultural machinery that can perform tasks like planting, weeding, and harvesting. These machines use sensors and ML algorithms to adapt to the environment and optimize their operations.

1.1.9 Decision Support Systems

ML-based decision support systems provide farmers with real-time recommendations based on various data sources. These systems can offer advice on irrigation schedules, pest control strategies, and other critical aspects of crop management.

1.1.10 Crop Insurance Assessment

ML can be used to assess and predict risks for crop insurance purposes. By analyzing historical data and current conditions, insurers can better estimate potential losses and provide

1.2 Various algorithms used for crop production prediction

Several machine learning algorithms are employed in crop production prediction to analyze and interpret diverse sets of agricultural data. The choice of algorithm depends on the specific task and characteristics of the data. Here are various ML algorithms used for crop production prediction:

1.2.1 Linear Regression

Application:production prediction How: Linear regression models predict crop yield based on various input features such as weather conditions, soil quality, and historical data.

1.2.2 Decision Trees

Application: Crop classification, disease detection How: Decision trees can classify crops based on features and can be used for disease detection by analyzing symptoms and conditions.

1.2.3 Random Forests

Application: Crop classification, yield prediction How: Random forests, an ensemble of decision trees, are effective for improving accuracy in crop classification and yield prediction tasks.

1.2.4 Support Vector Machines (SVM)

Application: Crop classification How:Support Vector Machines (SVMs) are effective in crop production prediction due to their ability to handle high-dimensional data, robustness to overfitting, capability to capture nonlinear relationships using the kernel trick, global optimization for better generalization, handling of imbalanced data, interpretability, and less susceptibility to overfitting with proper regularization.

1.2.5 K-Nearest Neighbors (KNN)

Application: Crop management, precision agriculture How: KNN (K-Nearest Neighbors) algorithm predicts crop production by finding similar historical data points. It looks at factors like weather, soil, and crop type to make predictions. KNN is simple, doesn't make assumptions about data, and can handle various data types. It's useful for dynamic environments like agriculture.

1.2.6 Artificial Neural Networks

Application: Yield prediction, disease detection How: Artificial Neural Networks (ANNs) can aid in crop production prediction by recognizing intricate patterns in diverse data types, capturing nonlinear relationships, modeling temporal and spatial dependencies, scaling to handle large datasets, integrating with IoT and remote sensing technologies, and offering flexibility in model design. Despite requiring ample labeled data and computational resources, ANNs excel in providing accurate predictions for complex agricultural systems.

1.2.7 Naive Bayes

Naive Bayes is a simple and efficient algorithm for crop production prediction. It handles high-dimensional data well, operates on a probabilistic framework, and is scalable. Despite assuming feature independence, it often performs effectively in practice and provides interpretable results. It's especially useful when simplicity and speed are priorities.

1.2.8 Gradient Boosting

Gradient Boosting is a powerful ensemble algorithm for crop production prediction. It combines weak learners sequentially, minimizing errors and capturing complex relationships. It's robust, provides feature importance insights, and popular implementations like XGBoost offer high performance.

1.2.9 Ada-Boost

Ada-Boost is a boosting algorithm that sequentially trains weak learners to focus on difficult cases. It combines their predictions to create a strong model, less prone to over fitting. It's versatile, simple to implement, and effective for crop production prediction.

1.3 Literature Review

In this segment a comparative analysis of diffrent methodologies employed for crop production prediction.Armstrong et al.'s review assesses how temperature, precipitation, and reference crop evapotranspiration impact crop production within a defined range, using Artificial Neural Networks. The study aims to cultivate drought-resistant and heat-tolerant crop varieties for changing climate conditions. Results show a performance metric of RMSE = 590.32 and $R^2 = 0.78$.Kalbande et al. investigate the impact of weather conditions (temperature, precipitation, humidity) on crop growth, utilizing SVR, multi-polynomial regression, and RF regression. Their findings, with SVR outperforming (RMSE=12.52, MAE=56.81, R^2=0.87), inform decisions on crop selection, fertilization, and irrigation.

SL No.	**Objectives**	**Methodology**	**Independent Variable**	**Critical Findings**	**References**
1	To analyze the effects of temperature, precipitation, and reference crop evapotranspiration on crop production within a certain range	(SVM), Naive Bayes, AdaSVM, and AdaNaive	Min-max temperature, precipitation, rainfall, crop evapotranspiration,area	Svm(rice)=90.4 ,Adasvm (cotton)=89.42	Narayanan Balakrishnan

2	To analyze the impact of weather conditions such as temperature, precipitation, and humidity on crop growth and development	SVR, multi polynomial regression, and RF regression	Weather data	SVR outperformed with RMSE=12.52, MAE=56.81, and $R^2 = 0.87$	Kalbande et al
3	To evaluate the individual and combined effects of these factors on plant growth and development	ANN, Naïve Bayes, RF and SVM	Windspeed ,rainfall,soil pH, ,precipitation,temperature, area, soil type, soil pH, pest details, water level, seed type	The accuracy of the model using ANN=87% with RMSE=0.78	Sujatha and Isakki
4	To investigate the relationship between sunlight, rain, frost, and temperature and the yield of wheat crops	SVM, KNN, descision tree	pH value, soil quality, type, rainfall, sunshine hours, humidity, temperature, fertilizers	SVR outperformed with RMSE=0.5 MSE=0.3	Jitendra Kumar Verma

Table 1.1: Summary of Research Objectives and Methodologies (continued)

1.4 Lacunae

1.4.1 Collect and analyze meteorological data for the study area

ML concept has been used in agriculture for several years [7]. Crop yield prediction is one of the challenging problems in precision agriculture, and many models have been proposed and validated so far. This problem requires the use of several datasets since crop yield depends on many different factors such as climate, weather, soil, use of fertilizer, and seed variety (Xu et al., 2019). This indicates that crop yield prediction is not a trivial task; instead, it consists of several complicated steps. Nowadays, crop yield prediction models can estimate the actual yield reasonably, but a better performance in yield prediction is still desirable (Filippi et al., 2019a). The critical gap was that the proper accuracy and performance were not available. The authors have stated that crop yield prediction method also aids in crop information and how to increase yield rate. ANN, DT algorithms, and regression analysis are among the algorithms employed. However, no clear methodology is specified Santra et al., 2016. A study reviewed various applications of ML in agriculture.This method aids in the expansion of the farming sector in countries and the application of more ML applications by using ANN, Bayesian belief networks, DT algorithms, clustering, and regression analysis. The literature gap was that the performance accuracy is lower (Kauri, 2016).

Sujatha Isakki [8], 2016 integration of various data mining methods such as k-means, ID3 algorithms, k-nearest neighbour, and support vector machines for crop yield prediction and risk assessment in agriculture. It does not elaborate on the specific methodologies or tools required to accurately predict pest outbreaks and diseases that impact crop yield .lack of detailed exploration on how these processes can be optimized to enhance forecasting accuracy and decision-making for farmers in India

Jitendra Kumar Verma [9] lack of emphasis on identifying the most effective approach for crop yield forecasting in the Indian agricultural context. does not delve deeply into comparing different machine learning algorithms to determine the most suitable one for crop yield forecasting in Indian agricultural scenarios.

Thomas van Klompenburg [10] The studies mentioned the insufficiency of data as a common problem, emphasizing the importance of using data with more variety for further testing

Narayanan Balakrishnan [11]focuses on predicting crop production using ensemble machine learning models like AdaSVM and AdaNaive, comparing them with traditional methods like SVM and Naive Bayes, The study suggests that further research is needed to explore whether changing the prediction techniques or increasing the dataset size can yield even better results, indicating a gap in exploring other potential methods or scaling up the analysis for more robust prediction.

1.4.2 Pre-process and prepare the data for training and testing the ML models

Djodiltachoumy, 2017 came to the conclusion that this paper's goal is to suggest and put into practice a rule-based system. and project agricultural yield production using historical data. The techniques employed are the clustering method and the k-algorithm. The study's flaw was that it only took into account association criteria and looked at a smaller amount of data. Jain et al. (2017) conducted research that used machine learning (ML) applications in agriculture These two algorithms are used: LR and DT. The lack of precise precision specification in the study is a negative. Rao et al. (2016) used various linear regression techniques to forecast crop yield in order to aid in decision- making. The employed algorithm's susceptibility to noise and overfitting is a drawback. According to Babu and Babu (2016), crop forecast can help farmers find answers and resolve problems with water and fertilizer, which will raise yield output. The agro algorithm is the one being used. Nevertheless, the algorithm's accuracy was poor. According to Ramesh Vishnu (2015), this approach will offer many linear regression techniques that may be used on current data to help with data verification and analysis. 2015 have concluded that the proposed method will provide an agricultural algorithm that will aid in the prediction of suitable crops for the lands. This contributes to crop quality improvement. The agro algorithm is employed. The literature gap is that crop prediction is less accurate. Few authors have presented their method would aid in estimating rainfall and investigating the

causes of low yield. The regression analysis method was used. The gap for the study was the algorithm was is susceptible to over-fitting Raorane and Kulkarni, 2015. Savla et al., 2015 have concluded that this method will aid in the analysis and comprehension of crop yield rates for zones based on attributes. Normalization, Clustering, and Classification are the algorithms employed. The disadvantage is that it only provides a framework. Khairunniza-Bejo et al., 2014 studied how prediction will aid in providing solutions to the few issues that farmers face in obtaining a good yield. ANN algorithms are used. The disadvantage is that the computation time is more.

1.5 Thesis Objective

The primary objective of this research is to help provide essential information that can help in crop yield. The research work was conducted focussing districts of Assam mainly Jorhat-kamrup (rice,wheat) and barpeta-golaghat(toor,masoor). To ascertain the proper methodologies for the yield prediction, particularly the technical methods, the objectives are defined below.

- **Development of Crop Production Prediction Model** Develop a machine learning (ML) model to predict crop production, specifically rice and wheat, in the Kamrup District. Use meteorological parameters such as temperature, relative humidity, precipitation, wind speed, and wind direction from the Jorhat District as input features for the prediction model. Additionally, assess the feasibility of adapting the model to predict crop production in other regions. Develop the model also for toor and masoor crops in the Barpeta and Golaghat district.
- **Comparison and Evaluation of Models** Compare and evaluate different ML models for diverse scenarios. Identify the best-performing model for predicting crop production in the both districts based on the provided meteorological parameters from Jorhat -kamrup and Barpeta- Golaghat.

Chapter 2

Materials and Methods

2.1 Study area

The research study was carried out with the Kamrup district and Jorhat district Assam and in the second case it is Barpeta and Golaghat. Kamrup district occupies an area of 4,345 square km (1,678 sq mi) (Srivastava et al., 2009). Fig. 1 shows the study area selected for this research. It is bounded with the North latitude of 25°94 and 26°20 and East of 91°15 and 91°28. The total number of revenue villages in the districts is 991. This was also partial availability of data. The principal crops grown in the region namely rice, cotton, wheat, maize and mustard, have been considered in this study.

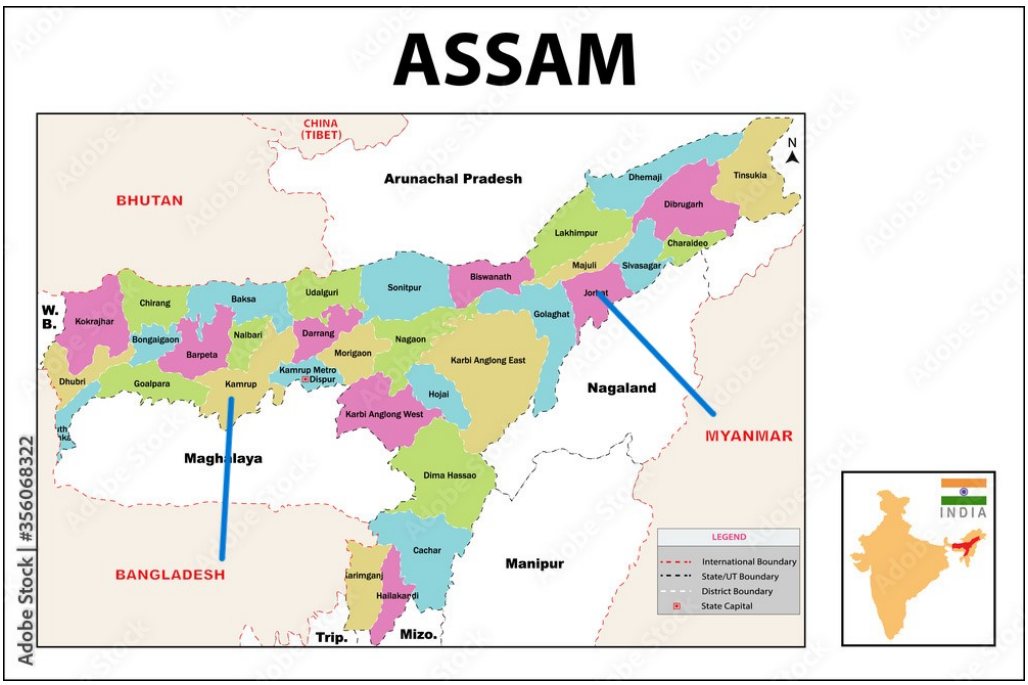


Figure 2.1: Study area selected for the research

2.2 Methodology

The overall modelling framework being implemented in the conducted investigation has been depicted in Fig. 2. This involves various phases such as data gathering, pre-processing, modelling, and analysis. Multiple stages of data pre-processing were required for the preparation of raw secondary data and its transformation into suitable variables for modelling and analysis. These include secondary data extraction, loading of data into appropriate data structures, secondary transformation of data, removal of outlier using filtering processes, and integration into consolidated datasets. The agricultural data was first used to predict CP, and was then sent for the pre-processing to remove any outlier data. The pre-processed data was eventually subjected to a feature extraction process that included considered parameters such as temperature, area, precipitation, humidity etc. The article targeted Two alternate ML algorithms namely DT,LR, and for Two crops (rice, wheat,) to predict the CP in Kamrup district and Jorhat district Assam, India. Thereby, the objective function has been represented as:

$$CP = f(T, RH, P, WD, WS, A) + \epsilon \tag{2.1}$$

where CY, WD, WS, A, T, RH, P refers to Crop Production, Wind Direction, Wind Speed, area, precipitation, Relative Humidity and Pressure respectively. The modelling was performed with reference to the study area Kamrup district (rural) based on 6 cash crops(six). In this study, we have considered the rice crop for the mentioned district. The data analysis and prediction modelling were achieved with Python 3.8. Data loading automation was ensured through Python scripts, pre-processing, and integration. A i7-4790 CPU with 3.60 GHz processor configuration was deployed. Thereby, Matplotlib library, NumPy, Scikit-learn and Pandas packages were used to assist the modelling efforts.

2.3 Model Development for Crop Production Prediction

The development of a model to predict crop production involved a systematic approach and several key procedures. Firstly, I collected a comprehensive dataset comprising historical crop production records along with corresponding environmental, agricultural, and

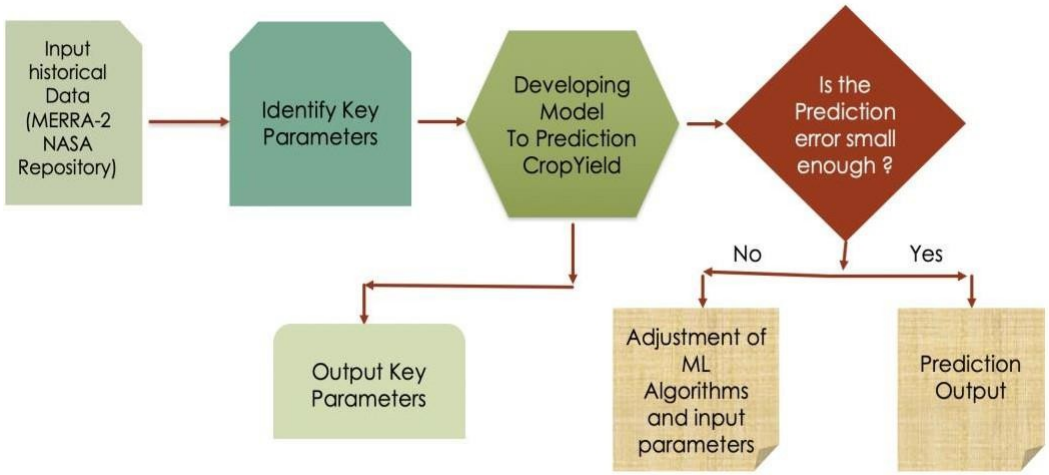


Figure 2.2: Overall methodology using ML algorithms

socioeconomic variables of first district . Next, I conducted data preprocessing steps, including data cleaning, normalization, and feature selection, to ensure the quality and relevance of the input data. Then, I employed various machine learning techniques such as regression analysis, decision trees, to build predictive models. These models were trained using a portion of the dataset and validated using cross-validation techniques to evaluate their performance. Iterative processes of model training and testing on other district data and evaluation were performed to optimize the model's accuracy and generalizability. Additionally, I incorporated domain knowledge and feature engineering techniques to enhance The model's predictive capabilities. Once the model was trained and I validated it on the same distruct „I used it to make crop production predictions for future seasons based on input variables. The model's performance was further assessed by comparing the predicted data with actual yield data for validation. This iterative process of model development and validation ensured the accuracy and reliability of the predictive model for crop production estimation.

2.4 Dataset and Data Pre-processing

The data size contained approximately initally 330 row and where after augmentation it is around 1080 points records of the meteorological parameters (MERRA-2) data for 1997-2023 and cultivation land area (MAFW). These ML models were applied to main cash crops in rural Assam. With the identification of the major factors influencing agricultural

production, new policy actions may be developed to assist the small farmers. Furthermore, ML approach may help future researchers not only to investigate but also to predict the impact of the important environmental variables on crop amount. MERRA-2 data were integrated with the data collected from the Ministry of Agriculture Farmers Welfare (MAFW) and the data was compiled for the year range of 1997-2023. A random train-test split ratio of 70/30 was relevant for the modelling effort. Thereby, a data size of approximately 1080 samples was achieved as training and testing datasets respectively. These datasets were deployed for model development and validation. Table 1 lists the deployed criteria in the conducted research work.

CP is often affected with climatic factors such as P, T, RH, etc. In a prior-art as well, it was opined to be a critical factor to influence agricultural production (Gandhi et al., 2016). However, climatic factors such as P and T do have seasonal alterations. Thereby, they are critical to influence cropping systems, patterns as and duration of the growth season (Singh et al., 2014). The size of the yields of cultivated crop in various ecological zones of the world is highly complex. Besides these, the impact of climate on the frequency of rain and crop physiological growth are often considered. A crucial understanding into metrological factors such as T, P, solar radiation, CO2 concentration, etc. and their influence on CY will assist in the improvement decision making process and associated adequacy of the adopted steps for maximum CY (Miniappan et al., 2014)

2.5 Machine Learning Algorithms

2.5.1 Linear Regression

Linear regression is a statistical method that is used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The simplest form of linear regression is simple linear regression, which involves only one independent variable. y is the dependent variable, Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. [12] Since linear regression shows the lin-

ear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

$$y = a_0 + a_1 x + \epsilon \tag{2.2}$$

Y= Dependent Variable (Target Variable) X= Independent Variable (predictor Variable) a_0= intercept of the line (Gives an additional degree of freedom) a_1 = Linear regression coefficient (scale factor to each input value). = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Simple Linear Regression

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

2.6 Decision tree

Decision tree is a supervised learning algorithm that is used for classification and regression modeling. Regression is a method used for predictive modeling, so these trees are used to either classify data or predict what will come next. Decision trees look like flowcharts, starting at the root node with a specific question of data, that leads to branches that hold potential answers. The branches then lead to decision (internal) nodes, which ask more questions that lead to more outcomes. This goes on until the data reaches what's called a terminal (or "leaf") node and ends.

In machine learning, there are four main methods of training algorithms: supervised, un-

supervised, reinforcement learning, and semi-supervised learning. A decision tree helps us visualize how a supervised learning algorithm leads to specific outcomes.

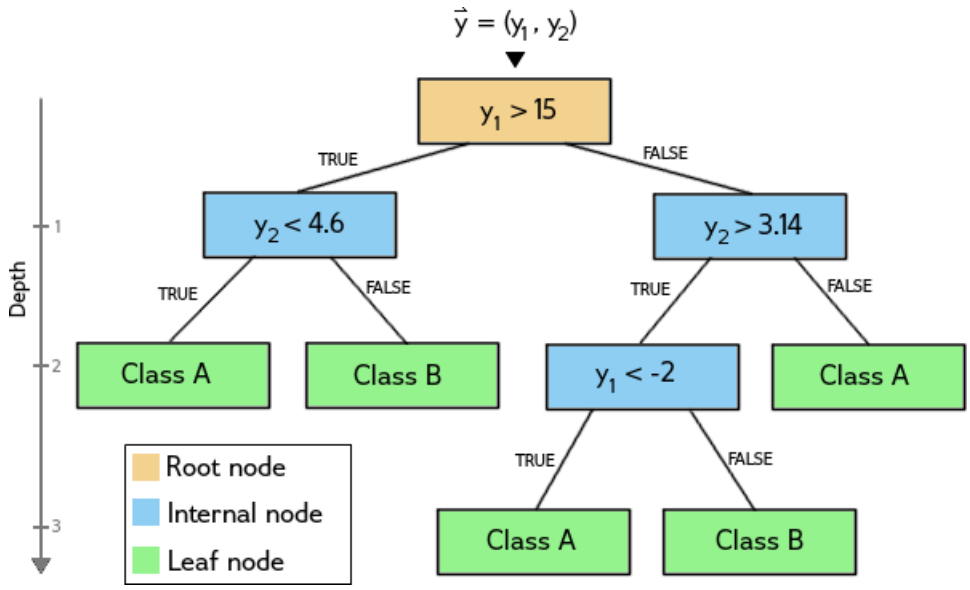


Figure 2.3: Pictorial representation of DT

2.7 Random Forest

Random Forest is a powerful algorithm for crop production prediction:

Ensemble of Decision Trees: It builds multiple decision trees and combines their predictions to make more accurate and robust predictions. Random Feature Selection: Random Forest randomly selects a subset of features for each tree, reducing correlation between trees and improving generalization. Bootstrap Aggregation: it uses bootstrap sampling to create diverse datasets for training each tree, enhancing the model's ability to handle variability in crop production data. Handles Nonlinear Relationships: Random Forest can capture complex relationships between input variables and crop yield, making it suitable for predicting yield based on diverse factors. Reduced Overfitting: By averaging predictions from multiple trees, Random Forest mitigates overfitting compared to individual decision trees, leading to more reliable predictions. Feature Importance: It provides insights into feature importance, helping identify the most influential factors affecting crop production.

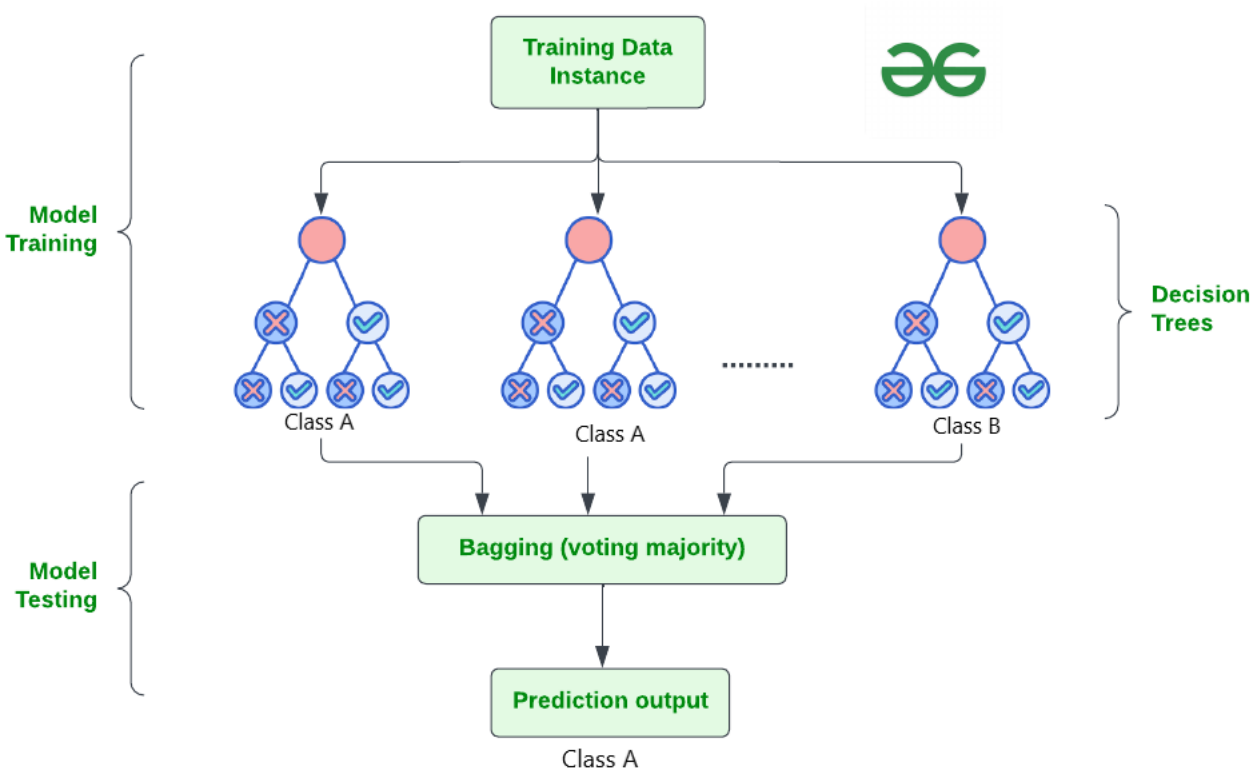


Figure 2.4: Pictorial representation of Random-Forest

2.8 Ada-Boost

Ada Boost (Adaptive Boosting) is a machine learning algorithm used for classification tasks. It belongs to the family of ensemble learning methods, where multiple weak learners are combined to create a strong learner. The idea behind Ada Boost is to iterative train a sequence of weak classifiers, with each subsequent classifier focusing more on the examples that were misclassified by the previous ones. Here's how it generally works: Initialization: Initially, each training example is given an equal weight. Iterative Training: Ada Boost trains a series of weak classifiers, typically decision trees with only a few levels (stumps), in a sequential manner. Weighted Training: At each iteration, Ada-Boost adjusts the weights of the training examples based on whether they were classified correctly or incorrectly by the weak classifier. Classifier Weight: Each weak classifier is assigned a weight based on its accuracy. The more accurate the classifier, the higher its weight. Final Combination: Finally, all weak classifiers are combined into a single strong classifier using a weighted sum or voting mechanism.

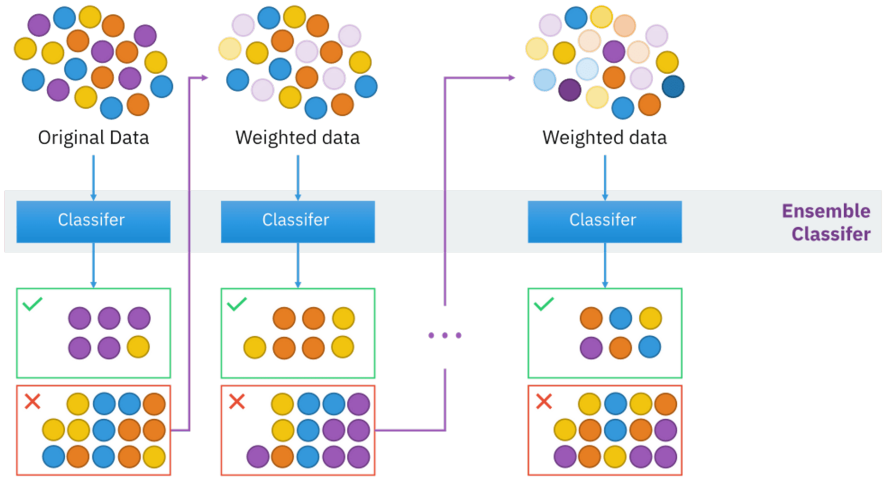


Figure 2.5: Pictorial representation of Ada-Boost

2.9 Gradient Boosting

Gradient Boosting is an ensemble learning technique that sequentially combines weak models, typically decision trees, to improve accuracy. It starts with an initial model and then trains subsequent models to correct errors made by the existing ensemble. This process minimizes a chosen loss function using gradient descent optimization, where each new model adjusts predictions based on the gradient of the loss function with respect to the previous ensemble's predictions. A "shrinkage" parameter is applied to control the contribution of each model, preventing overfitting. The process continues until a stopping criterion, such as reaching a maximum number of models or achieving satisfactory performance, is met. Popular implementations include Gradient Boosting Machines (GBM), XGBoost, LightGBM, and CatBoost, each offering optimizations and extensions to improve efficiency and performance in various applications such as regression and classification tasks.

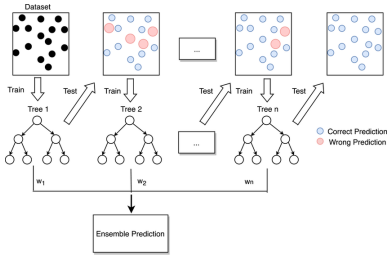


Figure 2.6: Pictorial representation of Gradient Boosting

2.10 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by biological neural networks. They consist of interconnected nodes arranged in layers: an input layer, one or more hidden layers, and an output layer. Each connection between nodes has an associated weight representing its strength. ANNs learn from data through a process called training, where they adjust the weights of connections to minimize the difference between actual and predicted outputs. This is typically done using optimization algorithms like gradient descent and back propagation, which propagate errors backward through the network to update weights. They are powerful for tasks like classification, regression, pattern recognition, and function approximation, especially when dealing with complex, high-dimensional data. Deep Learning, a subset of ANNs with multiple hidden layers, has revolutionized fields such as image and speech recognition, natural language processing, and autonomous vehicles. Popular frameworks for building ANNs include TensorFlow, Keras, PyTorch, and Caffe, offering high-level APIs for easy implementation and deployment of complex neural network architectures.

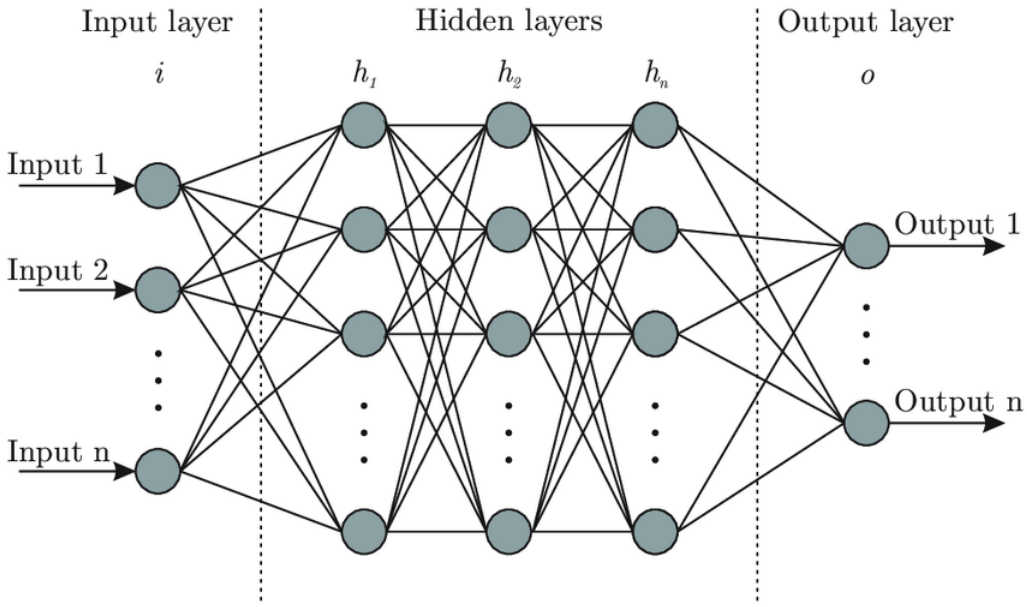


Figure 2.7: Pictorial representation of Artificial Neural Networks

2.11 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a versatile machine learning algorithm for classification and regression tasks. It assigns labels or predicts values based on the majority class or average of the k nearest data points in the feature space. It's easy to understand and implement but sensitive to the choice of k and distance metric. KNN doesn't require explicit training but can be computationally expensive, especially with large datasets, as it calculates distances to all points. Despite its simplicity, it's widely used in recommendation systems, anomaly detection, and pattern recognition, often serving as a baseline model for more complex algorithms.

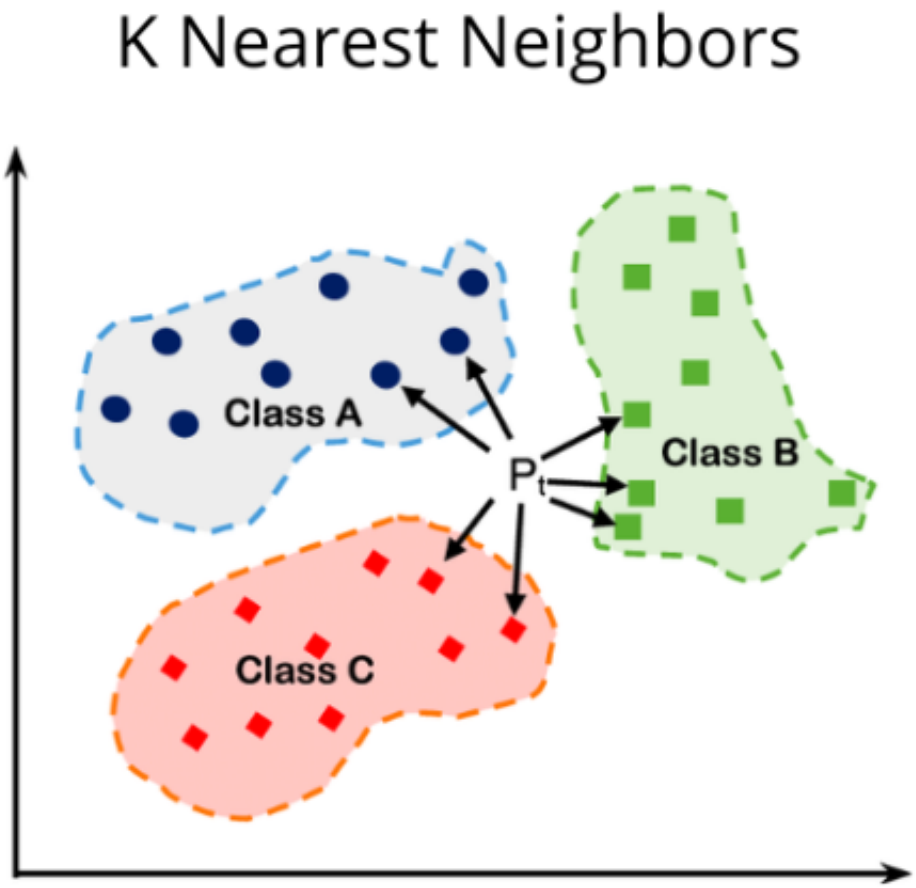


Figure 2.8: Pictorial representation of K-Nearest Neighbors

Chapter 3

Results and Discussion

3.1 Data Features and Correlation Exploration

In the Python programming environment, several Python packages were adopted to facilitate easy implementation of the available predetermined functions. The built-in libraries used were Numpy, Scikit-learn, Panda. Matplotlib and Statsmodels. The datasets deployed (Table 3.1) involved 330 data points and corresponded to MERRA-2 and MAFW data during pre- processing. After pre-processing, the CY data, MERRA-2 and MAFW data augemented and downsized to 1000 datapoints. As a result, the training and testing datasets consisted of around 800 and 200 samples, respectively. To address important problems and answer critical questions, it was necessary to downsize the data points for the purpose of data analysis since high-quality datasets are crucial. Henceforth, data that is inaccurate, repetitive, outdated, or in an unsuitable format will not fulfill its intended function. From now on, cleaning and preparing data is a compulsory step to improve the quality of the data. In addition, reducing the number of data points helps to maximize the utilization of available information, leading to the creation of precise and detailed outcomes at a granular level. Thereby, such an approach allows data transformation into action specific insights. The main objective of the correlation analysis was to establish and prioritize meteorological parameters and yield based on their higher correlation indices. Subsequently, weak or non- correlated parameters were isolated. Such an approach will be beneficial for the concise treatment of the best information towards modelling efforts. A correlation matrix was utilized to evaluate the correlation coefficient between each pair of variables, thus achieving the desired outcome. Fig. 3.1 summarizes the heat map

corresponding to the Pearson correlations of the properties among CY, area and climatic parameters. By looking at the first column in the figure, it can be affirmed that the area had the strongest correlation with a coefficient of 0.82 when compared to all other parameters. Also, an analysis revealed a positive correlation between T (0.80) and P (0.65), leading to the conclusion that the CY rate is significantly impacted by these parameters. The positive but weak correlation of RH-CY (0.26) and WS-CY (0.40), have been consistent with the generic inferences. Also, lower correlation coefficients for the pairs WD-CY (0.17), affirmed weakest correlation. Hence, targeting upon the subduing of the associated multi-collinearity issues, the model pairs RH-WS (-0.048) A-WD (-0.103) and P-WD (0.038) were omitted during the modelling effort. Scattered plots of T (showing the seasonal aspects), A, RH, P, WS and WD with respect to CY rate further illustrate such pairing effect (Supplementary S1 (a-f)). In this figure, a, b and c refers to the temperature trend for summer, winter and autumn seasons respectively.

3.2 Heatmap (Pearson's correlation)

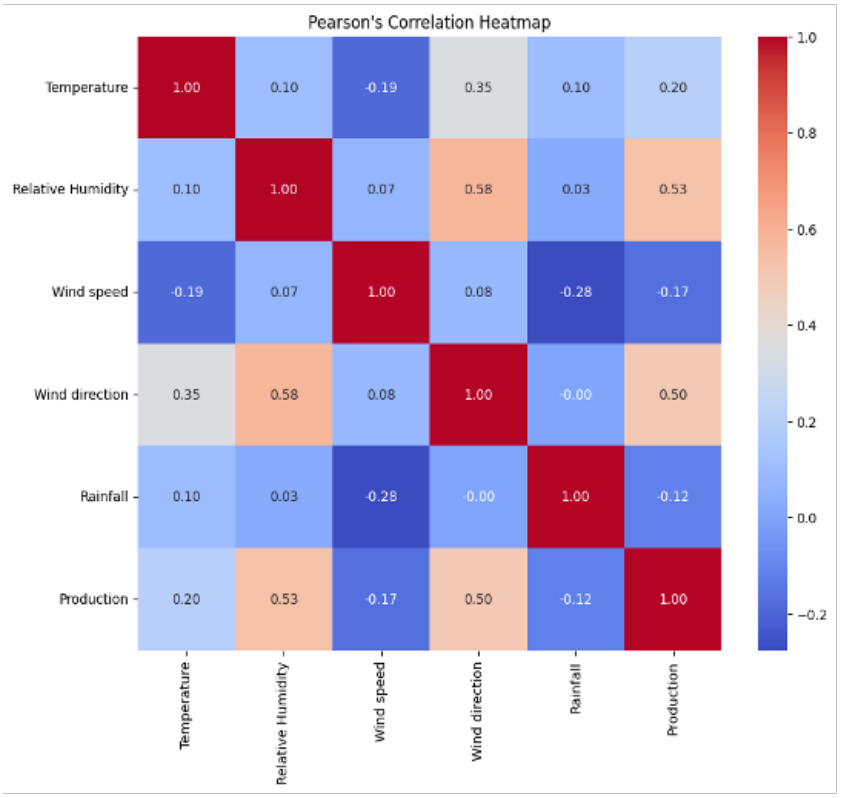


Figure 3.1: Heatmap corresponding to the Pearson correlations for the variables

- There is a weak positive correlation between temperature and relative humidity (0.10).
- There is a moderate positive correlation between relative humidity and wind direction (0.58) and production (0.53).
- There is a weak negative correlation between temperature and wind speed (-0.19).
- There is a weak negative correlation between wind speed and rainfall (-0.28).
- here is a very weak positive correlation between rainfall and production (0.12).

3.3 Modelling and predictive performances of the algorithms

3.3.1 Performance Metrics of Different Algorithms(kamrup(rice) and jorhat(wheat)

Algorithms	**R2 score (Training)**	**R2 score (Testing)**	**RMSE**
Linear Regression	0.84	0.79	0.872
Random forest	0.95	0.73	0.9234
Knn	0.76	0.77	1.7484
Ada boost	0.97	0.84	0.567
SVM	0.71	0.52	1.8475
Decision Tree	0.98	0.71	1.9845
Naive Bayes	0.71	0.52	2.457
Gradient boosting	0.79	0.79	1.1294
ANN	0.81	0.63	1.4683

Table 3.1: Performance Metrics of Different Algorithms

- **Linear Regression** demonstrates solid performance, achieving an R2 score of 0.84 on the training data and 0.79 on the testing data, with an accompanying RMSE of 0.872.
- **Random Forest** emerges as a strong contender, boasting high R2 scores of 0.95 on the training set and 0.73 on the testing set, although with a slightly higher RMSE of 0.9234.
- **K-Nearest Neighbors (KNN)** presents moderate performance, with R2 scores of

0.76 on the training data and 0.77 on the testing data, along with an RMSE of 1.7484, indicating some challenges in prediction accuracy.

- **AdaBoost** stands out as a top performer, exhibiting impressive R2 scores of 0.97 on the training set and 0.84 on the testing set, with a remarkably low RMSE of 0.567, suggesting robust predictive capability.
- Conversely, **Support Vector Machine (SVM)** achieves R2 scores of 0.71 on training and 0.52 on testing, accompanied by an RMSE of 1.8475, while **Decision Tree** presents R2 scores of 0.98 on training and 0.71 on testing, with an RMSE of 1.9845, indicating potential issues with overfitting.
- Both **Naive Bayes** and **Gradient Boosting** demonstrate similar performance, yielding R2 scores of 0.71 on training and 0.52 on testing for Naive Bayes, and 0.79 on both training and testing for Gradient Boosting, with RMSE values of 2.457 and 1.1294 respectively, suggesting room for improvement in predictive accuracy.

3.3.2 Performance Metrics of Different Algorithms of Barpeta-toor and Golaghat-masoor

Algorithms	R2 score (Training)	R2 score (Testing)	RMSE
Linear Regression	0.82	0.79	0.9124
Random forest	0.93	0.81	1.1984
Knn	0.79	0.76	1.9474
Ada boost	0.96	0.89	0.5498
SVM	0.74	0.45	1.7564
Decision Tree	0.96	0.55	2.3753
Naive Bayes	0.68	0.61	2.7895
Gradient boosting	0.88	0.78	1.5673
ANN	0.80	0.66	1.4749

Table 3.2: Performance Metrics of Different Algorithms

- **In evaluating the performance of various algorithms**, several metrics were considered, including R2 scores and Root Mean Square Error (RMSE). **Linear Regression**, a fundamental algorithm in predictive modeling, demonstrated decent performance with an R2 score of 0.82 on the training data and 0.79 on the testing data, accompanied by an RMSE of 0.9124. This suggests that the model explains a signi-

ficant portion of the variance in the data, albeit with some room for improvement in generalization.

- **Random Forest**, a powerful ensemble learning method, exhibited impressive results with an R2 score of 0.93 on the training set, indicating a strong fit, and a slightly lower R2 score of 0.81 on the testing set, suggesting good generalization capability. However, its RMSE of 1.1984 indicates some level of error in prediction.

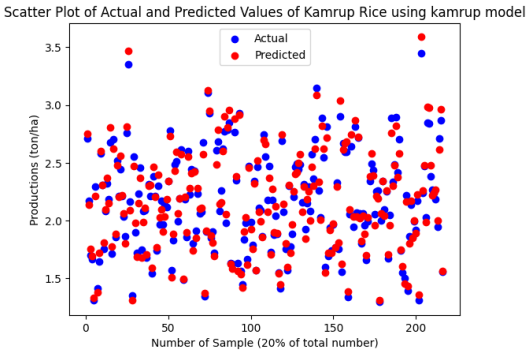
- **K-Nearest Neighbors (KNN)** presented middling performance, with R2 scores of 0.79 and 0.76 on the training and testing data, respectively, and a relatively higher RMSE of 1.9474. This suggests that while KNN captures some patterns in the data, it struggles with generalization, leading to higher prediction errors.

- **AdaBoost** emerged as a top performer among the evaluated algorithms, boasting high R2 scores of 0.96 on the training set and 0.89 on the testing set, along with a remarkably low RMSE of 0.5498. These results indicate robust performance, both in fitting the training data well and in accurately predicting outcomes on unseen data.

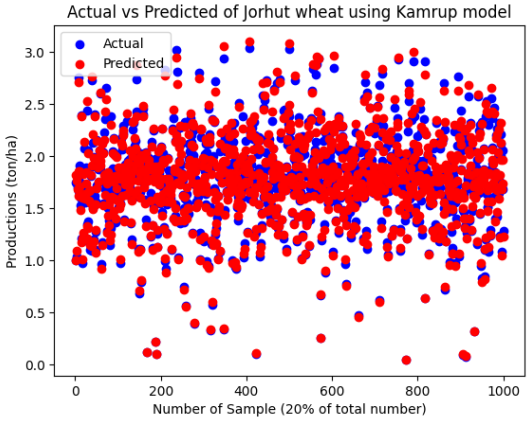
- On the contrary, **Support Vector Machine (SVM)** and **Decision Tree** algorithms displayed less favorable outcomes. SVM demonstrated relatively low R2 scores of 0.74 on training and 0.45 on testing, coupled with an elevated RMSE of 1.7564, suggesting limited explanatory power and poor generalization. Decision Tree, while achieving a high R2 score of 0.96 on training, suffered from a significant drop to 0.55 on testing, indicating overfitting, and yielded a relatively high RMSE of 2.3753.

- **Naive Bayes** and **Gradient Boosting** exhibited moderate performance. Naive Bayes showed R2 scores of 0.68 on training and 0.61 on testing, with an RMSE of 2.7895, implying modest predictive ability. Gradient Boosting, while achieving respectable R2 scores of 0.88 on training and 0.78 on testing, had an RMSE of 1.5673, indicating a reasonable level of predictive accuracy but with room for improvement.

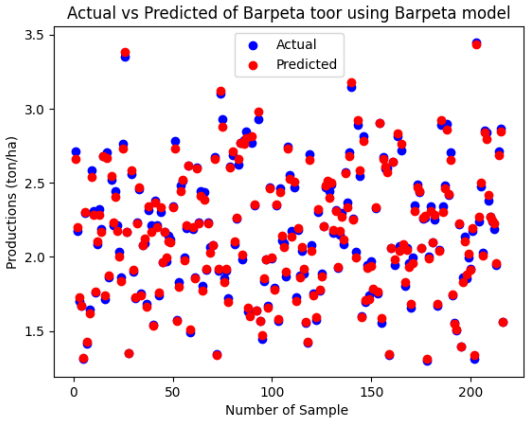
- **scatter plot of actual weather data with respect to production**The scatter plot shows the points are fairly dispersed, without a clear upward or downward trend. There is no evident linear relationship between temperature and production. Visu-



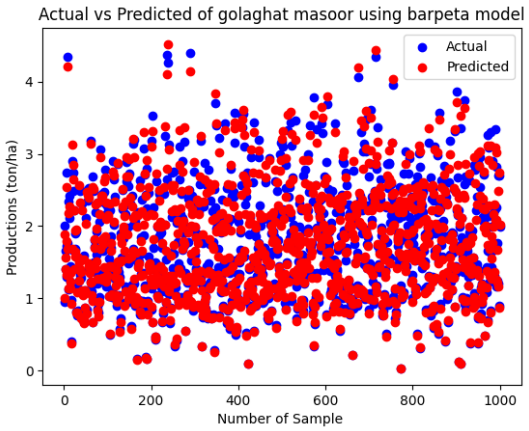
(a) kamrup rice using kamrup model



(b) jorhat wheat using kamrup model



(c) Barpeta toor using Barpeta model



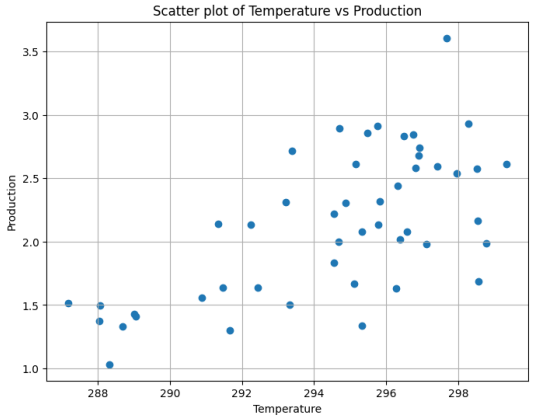
(d) Golaghat masoor using barpeta model

Figure 3.2: scatter plot of Ada- Boost algorithm

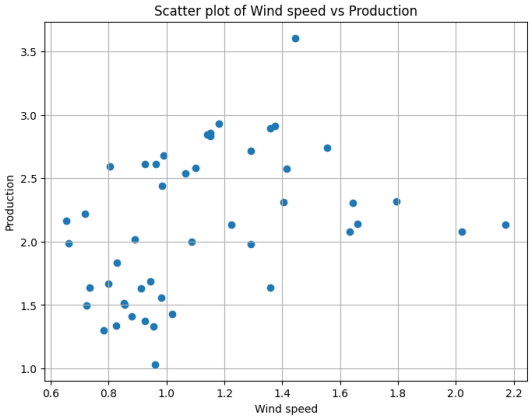
ally, the data points do not align in a way that suggests a strong positive or negative correlation. This indicates that changes in temperature do not consistently correspond to changes in production levels. The data appears to be randomly scattered, suggesting little to no direct correlation between temperature and production.

- **The scatter plot** does not show a strong positive or negative correlation. This indicates that changes in wind speed do not consistently correspond to changes in production levels. The absence of a discernible pattern suggests that wind speed does not have a significant impact on production. To confirm this visual observation, the Pearson correlation coefficient could be calculated. A value close to 0 would support the observation of no significant correlation.

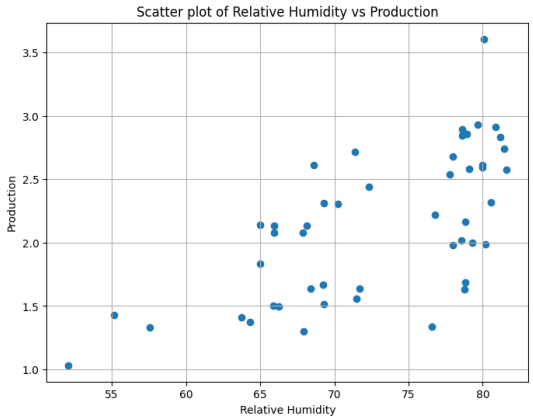
- **the scatter plot analysis** reveals a discernible relationship between relative humid-



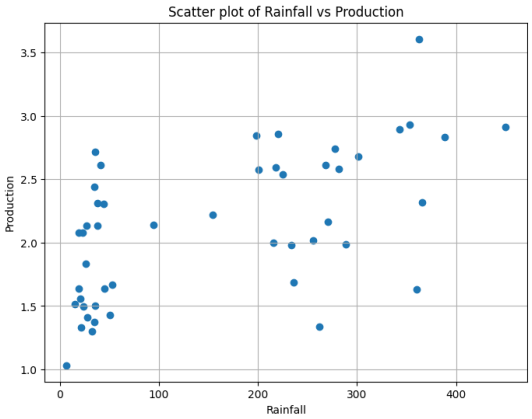
(a) with temperature



(b) with wind speed



(c) with humidity

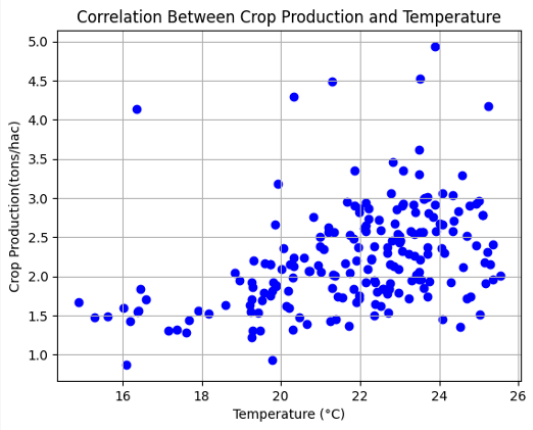


(d) with rainfall

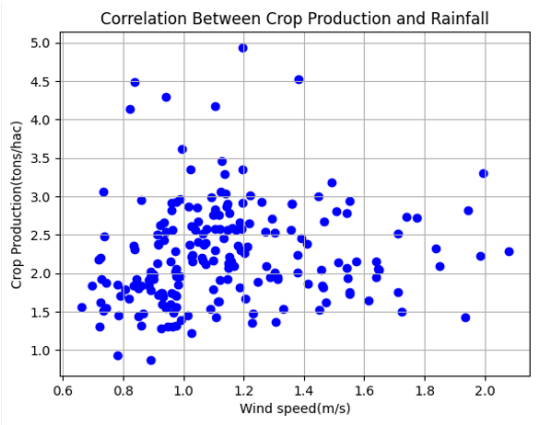
Figure 3.3: scatter plot of actual weather data with respect to production

ity and production data. As relative humidity levels fluctuate, there is a noticeable impact on production output. From the plotted data, it appears that production generally decreases as relative humidity increases. This negative correlation suggests that higher levels of relative humidity may pose challenges or constraints to the production process, potentially affecting efficiency or output. However, it's essential to note that while the scatter plot highlights this relationship, it does not provide insights into causality or the mechanisms underlying the observed correlation. Further analysis, including statistical modeling or controlled experiments, may be necessary to elucidate the precise nature of the relationship and its implications for production management. In conclusion, the scatter plot serves as a valuable tool for visualizing the relationship between relative humidity and production data, providing a

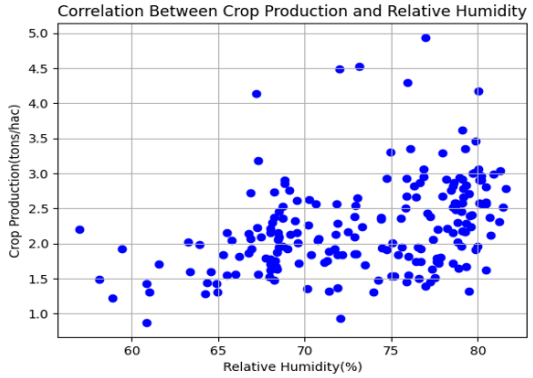
foundation for further investigation and decision-making in optimizing production processes.



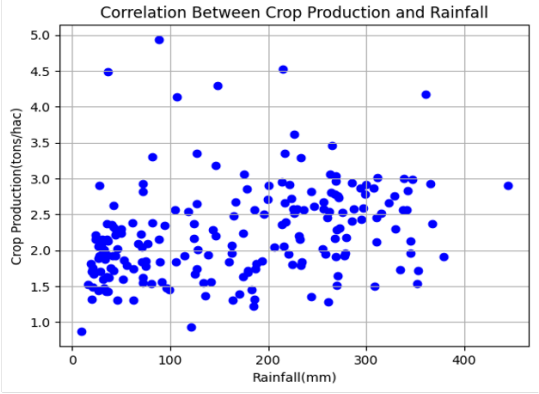
(a) with temperature



(b) with wind speed



(c) with humidity



(d) with rainfall

Figure 3.4: scatter plot of augmented weather data with respect to production

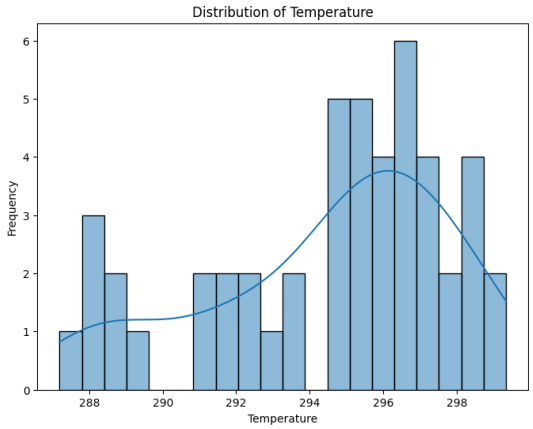
- **Based on the plotted frequency distribution curve**

- The peak of the curve, representing the highest frequency, is around the interval 296-297 kelvin. This indicates that temperatures in this range are the most common within your dataset. The frequency distribution curve suggests that temperatures in the range of 296-297 lelvins are the most common, and the distribution of temperatures is relatively symmetric, with a single dominant mode. This indicates a relatively uniform spread of temperatures without significant skewness.

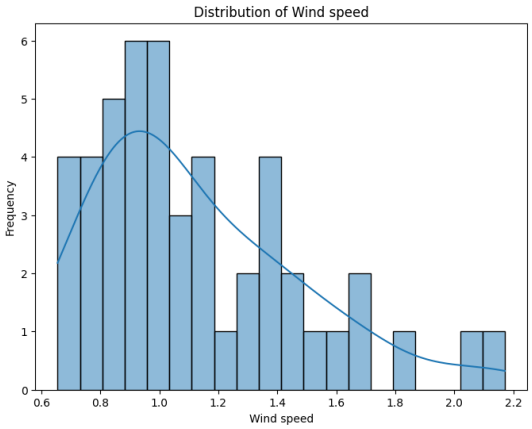
- The peak of the curve, representing the highest frequency, appears to be around the interval 1.0 - 1.5 meters per second. This suggests that wind speeds in this range are the most common within your dataset.he frequency distribution curve suggests

that wind speeds around 1.0 - 1.5 meters per second are the most common, and the distribution of wind speeds is positively skewed, with a single dominant mode. This indicates a predominance of lower wind speeds in my dataset.

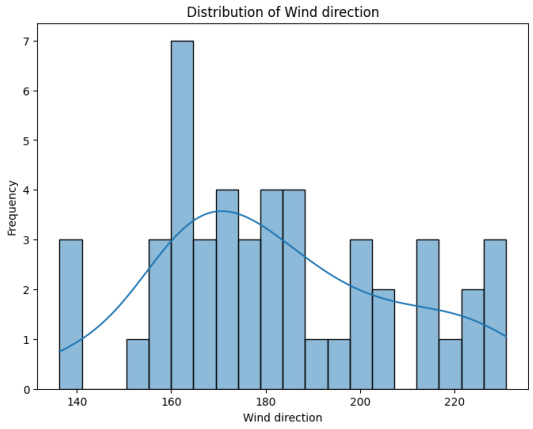
- The peak of the curve, representing the highest frequency, appears to be around the interval 200 - 250 mm. This suggests that rainfall amounts in this range are the most common within your dataset.e frequency distribution curve suggests that rainfall amounts around 200 - 250 mm are the most common, and the distribution of rainfall amounts is positively skewed, with a single dominant mode. This indicates a predominance of lower to moderate rainfall amounts in my dataset, with fewer instances of heavy rainfall.



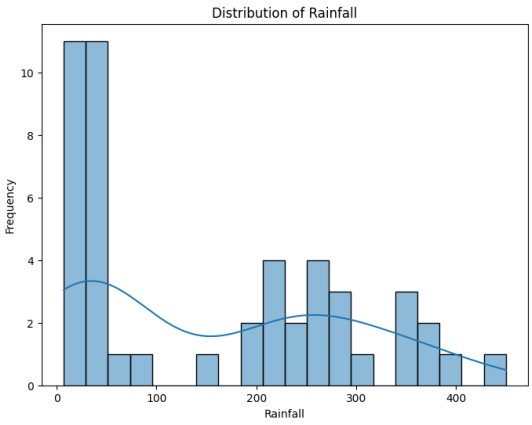
(a) with temperature



(b) with wind speed



(c) with wind direction



(d) with rainfall

Figure 3.5: distribution of augmented weather data

Chapter 4

Conclusions and Future Work

4.0.1 Conclusions

In conclusion, the results obtained from the application of AdaBoost algorithms on our models Kamrup (rice) - Jorhat (wheat) and Barpeta (toor) - Golaghat (masoor) showcase promising performance metrics.

For the Kamrup - Jorhat model, we achieved an impressive R2 score of 0.97 and 0.84, with an RMSE of 0.567. Similarly, the Barpeta - Golaghat model yielded commendable results with R2 scores of 0.96 and 0.89, along with an RMSE of 0.5498. These high R2 scores indicate a strong correlation between the predicted and actual values, while the low RMSE values suggest accurate predictions with minimal error.

Looking ahead, there are several avenues for future work to explore. Firstly, we can investigate additional machine learning algorithms to further optimize the models and potentially improve their predictive capabilities. Additionally, incorporating more features or refining existing ones could enhance the models' performance. Furthermore, exploring different data preprocessing techniques and model tuning parameters may also contribute to refining the predictive accuracy of the models.

Overall, the results obtained demonstrate the potential of our models in predicting crop yields, and further research and refinement can lead to even more robust and accurate predictions, ultimately benefiting agricultural decision-making processes.

4.0.2 Future Work

- **Incorporate More Features or Refine Existing Ones:** Investigate the incorporation of additional features or refine existing ones to enhance model performance. By expanding the feature set, we can capture more nuanced relationships within the data, ultimately improving the accuracy of predictions.
- **Model Tuning Parameters:** Explore different model tuning parameters to further refine predictive accuracy. Fine-tuning parameters such as learning rates, regularization strengths, and model architectures can significantly impact model performance and enhance predictive capabilities.
- **Continued Research and Refinement:** Emphasize the importance of continued research and refinement. Through ongoing exploration and adaptation, we can achieve even more robust and accurate predictions. Continued refinement ensures that our models remain at the forefront of innovation, ultimately benefiting agricultural decision-making processes.
- **Integration of Advanced Technologies:** Investigate the integration of advanced technologies such as Natural Language Processing (NLP) and machine learning to enhance predictive capabilities. Leveraging NLP techniques can enable better analysis of textual data, further enriching our models and improving prediction accuracy.

Bibliography

[1] G. Azzari, M. Jain, and D. B. Lobell, "Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries," *Remote Sensing of Environment*, vol. 202, pp. 129–141, 2017.

[2] B. Bazaya, A. Sen, and V. Srivastava, "Planting methods and nitrogen effects on crop yield and soil quality under direct seeded rice in the indo-gangetic plains of eastern india," *Soil and Tillage Research*, vol. 105, no. 1, pp. 27–32, 2009.

[3] D. Ramesh and B. V. Vardhan, "Analysis of crop yield prediction using data mining techniques," *International Journal of research in engineering and technology*, vol. 4, no. 1, pp. 47–473, 2015.

[4] T. Denœux, O. Kanjanatarakul, and S. Sriboonchitta, "Ek-nnclus: a clustering procedure based on the evidential k-nearest neighbor rule," *Knowledge-Based Systems*, vol. 88, pp. 57–69, 2015.

[5] W. J. Doucette, S. Mendenhall, L. S. McNeill, and J. Heavilin, "The sky is falling ii: Impact of deposition produced during the static testing of solid rocket motors on corn and alfalfa," *Science of the Total Environment*, vol. 482, pp. 36–41, 2014.

[6] M. H. Fatemi and S. Gharaghani, "A novel qsar model for prediction of apoptosis-inducing activity of 4-aryl-4-h-chromenes based on support vector machine," *Bioorganic & medicinal chemistry*, vol. 15, no. 24, pp. 7746–7754, 2007.

[7] A. M. S. Ahamed, N. T. Mahmood, N. Hossain, M. T. Kabir, K. Das, F. Rahman, and R. M. Rahman, "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in bangladesh," in *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial*

Intelligence, Networking and Parallel/Distributed Computing (SNPD). IEEE, 2015, pp. 1–6.

[8] R. Sujatha and P. Isakki, "A study on crop yield forecasting using classification techniques," in *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*. IEEE, 2016, pp. 1–4.

[9] S. K. Sharma, D. P. Sharma, and J. K. Verma, "Study on machine-learning algorithms in crop yield predictions specific to indian agricultural contexts," in *2021 international conference on computational performance evaluation (ComPE)*. IEEE, 2021, pp. 155–166.

[10] T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020.

[11] N. Balakrishnan and G. Muthukumarasamy, "Crop production-ensemble machine learning model for prediction," *International Journal of Computer Science and Software Engineering*, vol. 5, no. 7, p. 148, 2016.

[12] A. Ng, "Cs229 lecture notes," *CS229 Lecture notes*, vol. 1, no. 1, pp. 1–3, 2000.