INTERNATIONAL E-CONFERENCE ON GLOBAL MULTIDISCIPLINARY RESEARCH AND INNOVATION (ICGMRI - 2026)

PROCEEDINGS

Pencil Bitz

ISBN 978-816800517-4

TABLE OF CONTENTS
ICGMRI - 26

CROP YIELD PREDICTION USING MACHINE LEARNING TECHNIQUES

S.Keerthana
Research Scholar,School of Computing Sciences,
VELS Institute of Technology and Advanced Studies (VISTAS)
Chennai, India
keerthiyugu@gmail.com
Dr.K.Dharmarajan
Professor, School of Computing Sciences,
VELS Institute of Technology and Advanced Studies (VISTAS)
Chennai, India
dharmak07@gmail.com
ORCID : 0000-0001-8493-1301

ABSTRACT
In recent years, the rise of big data has accelerated the adoption of machine learning approaches across diverse domains. Agriculture, in particular, stands to benefit significantly from these advancements, with the potential to enhance crop productivity and quality. This study presents a machine learning–based framework for crop recommendation and yield prediction to support data-driven agricultural decision-making. Exploratory Data Analysis (EDA) is employed to uncover patterns in the dataset, with correlation maps used to visualize feature relationships. The dataset is partitioned into training and testing subsets, and preprocessing includes feature scaling via Standard Scaler and categorical encoding with OneHotEncoder. The proposed architecture integrates Decision Tree Regressor (DTR), Support Vector Regressor (SVR), and Gradient Boosting Regressor as base models, while XGBoost (XGB) and Random Forest (RF) Regressor serve as meta-models within a Stacking Regressor pipeline. Stacking is selected over bagging and boosting due to its ability to combine heterogeneous models in a meta-learning framework. Model performance is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) metrics. Results demonstrate that the stacked ensemble provides an effective and practical tool for enhancing agricultural decision-making through advanced machine learning methodologies.

Keywords: Machine Learning, Decision Tree, OneHotEncoder, Ensemble Learning, Random Forest, XGBoost, Agriculture, Yield Prediction

1. INTRODUCTION:

Agriculture is the most diverse economic sector in terms of population and thus plays a significant role in India's overall socioeconomic structure. Particularly in the era Crop yield prediction is crucial in addressing emerging challenges in food security, particularly in light of current global climate change. Crop yield prediction is a crucial aspect of modern agriculture, helping farmers, policymakers, and agribusinesses make informed decisions about resource allocation, food security, and sustainability. Machine learning (ML) has emerged as a powerful tool to enhance the accuracy and efficiency of these predictions. Farmers can boost production under favorable conditions and decrease production loss under unfavorable conditions by using crop yield estimation. Numerous variables, such as farmer practices, choices, pesticides, fertilizers, weather, and market prices, impact crop yield positive forecasts. It is possible to estimate crop yield using statistical information from yields of previous years, together with climate, area wise output, and rainfall. Recently, machine learning has advanced in various industries, including agriculture. Crop yield prediction using machine learning (ML) applies algorithms to estimate harvest outcomes based on historical data, environmental conditions, and farming practices. Unlike traditional statistical models, ML can capture nonlinear relationships and adapt to diverse datasets, making predictions more accurate and scalable. Yield prediction is an important agricultural problem. In the past farmers used to predict their yield from previous year yield experiences. Thus, for this kind of data analytics in crop prediction, there are different techniques or algorithms, and with the help of those algorithms we can predict crop yield. Random forest algorithm is used. Using all these algorithms and with the help of inter-relation between them, there are growing range of applications and the role of Big data analytics techniques in agriculture. Since the creation of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are concentrated on cultivating artificial products that are hybrid products where there leads to an unhealthy life. Nowadays, modern people don't have awareness about the cultivation of the crops at the right time and at the right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food. By analyzing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India, there are several ways to increase the economic growth in the field of agriculture. There are multiple ways to increase and improve the crop yield and the quality of the crops. The need for technological advancements in the agriculture sector is deemed necessary in order to increase the crop productivity and prevent imminent food crises. Particularly, monitoring the yield of different crops for every season is an important attribute in estimating crop productivity. Analysis of various factors like soil, climate, temperature, etc. can contribute to crop yield prediction.

The problem that current Indian agriculture sector is facing is to integrate the technology to predict the desired outputs. In order to perform accurate crop yield prediction various machine learning algorithms are being used. The application of machine learning in agriculture has attracted significant attention in recent years, offering new opportunities to optimize farming practices and enhance food production. Machine learning models have demonstrated notable improvements in predicting insights across diverse crop fields, providing farmers with strategic recommendations that strengthen agricultural decision-making processes. Techniques such as Decision Trees, Support Vector Machines, and ensemble approaches including Random Forest (RF) and XGBoost (XGB) have proven effective in generating highly accurate predictions from complex and multidimensional agricultural datasets. These models not only deliver forecasting results but also provide valuable understanding of the variables influencing crop performance, thereby supporting decisions related to crop selection, input management, and yield optimization. Recent advances in machine learning have further highlighted the ability of modern algorithms to capture non-linearity, spatial variability, and seasonality within agricultural datasets—capabilities that traditional models often fail to exhibit. Consequently, studies have emphasized the advantages of advanced ensemble methods such as Gradient Boosting and Stacking, which integrate multiple algorithms to enhance predictive performance. In particular, stacking has emerged as a robust framework that leverages the strengths of diverse models to deliver more accurate crop recommendations and yield estimations. Data pre processing is a critical step in ensuring that machine learning models perform optimally. In this stage, raw data is transformed or encoded into formats that can be efficiently parsed and analysed by algorithms. Studies estimate that approximately 40–85% of the total time spent on classification tasks is devoted to pre processing. Techniques such as categorical encoding and feature scaling are particularly important for converting agricultural datasets into representations suitable for machine learning models. For datasets containing both numerical and categorical features, recent research highlights the effectiveness of methods such as Standard Scaler and OneHotEncoder. These pre processing techniques enhance model performance by ensuring that all variables are adequately represented and contribute meaningfully to the learning process. In addition, interactive machine learning tools are becoming increasingly relevant, enabling farmers to input specific conditions such as climate, soil type, and crop variety. Such platforms provide personalized recommendations and real-time localized insights, making advanced machine learning technology more accessible at the field level. This paper focuses on improving predictive accuracy through refined pre processing methods, the application of ensemble learning techniques, and the development of interactive features that enhance usability in real-world agricultural scenarios. These advancements underscore the transformative potential of machine learning in optimizing farming practices, increasing crop yields, and supporting sustainable agricultural development. The recent advances in machine learning model have demonstrated the ability to predict and capture the non linearity, spatial variability and seasonality of the agricultural datasets which the traditional models failed to exhibit. Thus, studies have emphasized the benefits of machine learning models such as Gradient Boosting and Stacking which integrate various machine learning algorithms to improve the predictive performance. In particular stacking has become a strong framework that combines the strengths of various algorithms to provide crop recommendations and yield estimations that are more accurate.

2. LITERATURE REVIEW:

Increasing the quality and quantity of the agriculture crop has become a crucial task for the food industry in order to ensure food sustainability for the growing population. To get an overview of what has been done on the application of ML in crop yield prediction, we performed a systematic literature review (SLR). Crop yield prediction is an essential task for the decision-makers at national and regional levels (e.g., the EU level) for rapid decision- making. An accurate crop yield prediction model can help farmers to decide on what to grow and when to grow. There are different approaches to crop yield prediction. This review article has investigated what has been done on the use of machine learning in crop yield pre diction in the literature. The literature on crop yield prediction demonstrates the growing role of machine learning in agricultural analytics. Researchers have applied diverse algorithms to forecast yields by leveraging variables such as soil properties, climatic conditions, and crop management practices. The study in [1] developed a hybrid framework combining Artificial Neural Networks (ANN) and Support Vector Machines (SVM) using the Indian Government Agricultural Dataset. This dataset incorporated climate data, soil properties, historical yields, and farming practices. The hybrid model achieved a prediction accuracy of 92%, with an RMSE of 2.0 and MAE of 1.6 tons per hectare. However, challenges included managing the increased computational demands and extensive parameter tuning required for integrating ANN and SVM. Desai et al. [2] conducted a comparative analysis of ensemble algorithms using the Indian Agricultural Research dataset, which included features such as soil properties, historical yields, irrigation, and weather conditions. Among the tested models, stacking achieved the highest accuracy of 93%, with an RMSE of 1.8 and an R^2 value of 0.89. Despite its strong performance, stacking required significant computational resources and expert tuning to optimize model parameters. Kumar et al. [3] applied Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) models on the Kaggle Crop Yield Dataset. RF achieved the best performance, with an MSE of 3 and an accuracy of 85%, effectively handling non-linear relationships. However, limitations included sensitivity to outliers, missing data, and risks of overfitting with more complex algorithms. In [4] Crop production Ensemble Machine Learning model for prediction. International Journal of Computer Science and Software Engineering (IJCSSE). In this paper, AdaNaive and AdaSVM are the proposed ensemble model used to project the crop production over a time period. Implementation done using AdaSVM and AdaNaive. AdaBoost increases efficiency of SVM and Naive Bayes algorithm. In [5]. Machine learning approach for forecasting crop yield based on parameters of climate. The paper provided in International Conference on Computer Communication and Informatics (ICCCI). In the current research a software tool named Crop Advisor has been developed as a user friendly. Rao et.al [6] in their study demonstrates the improvement of crop prediction using soil nutrients and climatic conditions by analysis machine learning models. The comparison of KNN, Decision Trees and Random Forest using Gini and entropy was performed and RF classifiers was found more reliable with an accuracy of 99.32%. Sarowar et.al [7] conducted a systematic survey to identify various attributes and techniques used in machine learning models. It was observed that most used features are temperature, Soil Type and vegetation and most applied machine learning algorithms are Random Forest, Linear regression, Gradient Boosting, Deep learning models. In [8] the authors using the features like soil conditions, crop types, yield records evaluate the predicted model using various algorithms like DT, RForest, and Gradient Boosting (GB). Among the three GB had an accuracy of 92% for yield forecasting and an MAE of 1.4 tons per hectare. The limitations include varied feature importance. In [9]. Machine learning approach for forecasting crop yield based on parameters of climate. The paper provided in International Conference on Computer Communication and Informatics (ICCCI). In the current research a software tool named Crop Advisor has been developed as a user friendly. In[10]. Prediction On Crop Cultivation. International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) Volume 5, Issue 10, October 2016. Presently, soil analysis and interpretation of soil test results is paper based. This in one way or another has contributed to poor interpretation of soil test results which has resulted into poor recommendation of crops, soil amendments and fertilizers to farmers thus leading to poor crop yields, micro-nutrient deficiencies in soil and excessive or less application of fertilizers. Formulae to Match Crops with Soil, Fertilizer Recommendation. In [11]. Analysis of Crop Yield Prediction by making Data Mining Methods. IJRET: The paper provided in International Journal of Research in Engineering and Technology. In this paper the main aim is to create a user friendly interface for farmers, which gives the analysis of rice production based on the available data. For maximizing the crop productivity various Data mining techniques were used to predict the crop yield. Such as K-Means algorithm to forecast the pollution factor in the atmosphere. In [12]. Crop production Ensemble Machine Learning model for prediction. International Journal of Computer Science and Software Engineering (IJCSSE). In this paper, AdaNaive and AdaSVM are the proposed ensemble model used to project the crop production over a time period. Implementation done using AdaSVM and AdaNaive. AdaBoost increases efficiency of SVM and Naive Bayes algorithm. In [13] In this paper, we present a comprehensive review of research dedicated to the application of machine learning in agricultural production systems. Machine learning (ML) has emerged together with big data technologies, techniques, methods and high-performance computing to generate new opportunities to unravel, quantify, and analyse data intensive processes in agricultural operational sectors. By using Support Vector Machines (SVP) the Paper is Implemented. In[14]. Prediction On Crop Cultivation. International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) Volume 5, Issue 10, October 2016. Presently, soil analysis and interpretation of soil test results is paper based. This in one way or another has contributed to poor interpretation of soil test results which has resulted into poor recommendation of crops, soil amendments and fertilizers to farmers thus leading to poor crop yields, micro-nutrient deficiencies in soil and excessive or less application of fertilizers. Formulae to Match Crops with Soil, Fertilizer Recommendation. ML studies consist of different challenges when aiming to build a high-performance predictive model. It is crucial to select the right algorithms to solve the problem at hand, and in addition, the algorithms and the underlying platforms need to be capable of handling the volume of data.

3. PROPOSED MODEL:

In this phase, data are collected from various sources to prepare datasets. There are various online sources like Data.gov.in that provide data. These data can be used for next ten years capturing a wide range of agricultural attributes such as environmental factors, soil nutrients, and crop production metrics. The dataset is designed to support improvements in agricultural practices through both crop recommendation and yield prediction.

Key features include:

- State_Name: Identifies the geographic location where the crop is grown, reflecting the influence of regional and environmental conditions on agricultural outcomes.
- Crop_Type: Specifies the season (e.g., Kharif, Rabi), enabling the alignment of crops with their optimal growing periods.
- Crop: Details the specific crop grown (e.g., cotton, maize), allowing for tailored recommendations based on historical performance and contextual factors.

  Soil Nutrients (N, P, K): Nitrogen, phosphorus, and potassium are critical for crop growth, directly influencing fertilizer recommendations and yield predictions.

  pH: Determines nutrient availability and soil suitability for different crops.

- Rainfall and Temperature: Essential climatic variables that guide crop selection and yield forecasting.
- Area_in_hectares: Represents the cultivated land size, influencing total production and yield estimates.

- Production_in_tons and Yield_ton_per_hec: Serve as key productivity measures and target variables in yield prediction models.
- State_Name, Crop_Type, and Crop: Provide contextual information on geographic location, seasonal crop type (e.g., Kharif, Rabi), and specific crop grown (e.g., cotton, maize), enabling localized and tailored recommendations.

By integrating soil, environmental, and crop-specific data, the dataset supports sustainable farming practices and informed decision-making, ultimately enhancing productivity and resource optimization. By leveraging these attributes, the proposed system aims to enhance predictive accuracy and provide actionable insights for farmers. The integration of environmental, seasonal, and crop-specific variables ensures that recommendations are localized and yield predictions are contextually relevant. This approach not only improves decision-making but also supports sustainable agricultural development by optimizing crop selection and resource management.

3.1 Data Preprocessing

Data preprocessing is a crucial step to ensure that machine learning models perform optimally. Raw agricultural data often contains missing values, categorical variables, and features with varying scales. To address these challenges: Feature Scaling: Standard Scaler is applied to normalize numerical features (e.g., rainfall, temperature, soil nutrients), ensuring uniform contribution to the model.

- Categorical Encoding: OneHotEncoder is used to transform categorical attributes (e.g., State_Name, Crop_Type, Crop) into machine-readable formats.
- Data Cleaning: Outliers and missing values are handled to improve model robustness and accuracy.

The dataset, named Crop_prediction.csv, is loaded into the system for analysis. Missing values are addressed by removing null rows to ensure data integrity. The features in this dataset include 'State_Name', 'Crop_Type', 'N', 'P', 'K', 'pH', 'rainfall', 'temperature', and 'Area_in_hectares'. The target variables are 'Crop' & 'Yield_ton_per_hec'. Categorical factors are transformed through one-hot encoding.

3.2 Feature Engineering:

Here, the features are categorized into numeric and categorial datatypes. The numeric data types include "PH', "rainfall", "temperature and the categorial tyoe type include 'state-name","crop_type". The column consolidates these process making sure that all features get prepared well for modelling.

3.3 Model Training for crop prediction:

For, crop prediction, a sacking regressor is implemented. This method uses SVR, Decision Tree, gradient Boosting Regressor and Random Forest method. The dataset is divided into two sets, One is used for training the model and the other one used for testing phase. Various metrics like MSE and R2 values are used to evaluate model performance. The results are visually represented in with charts. Radient Boosting Regressor is a way of combining multiple simpler models to create one compo model. We combine multiple weak learners(generally decision trees) such that overall loss is minimized. This is why it is also known as an additive model. Trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini is done to minimize the loss. We have used the standard gradient boosting regressor which uses 'least squares regression' as a loss function with 100 boosting stages.

3.4 User input and prediction:

The system has been designed to accept inputs such as state, season along with the values of rainfall, temperature and area. This prediction are processed to data frames to predict the crop yield. Results are shown in the figure. Figure 1 shows the crop yield prediction architecture. It starts with a pre-processed dataset. The trained models are tested using different performance measures. Finally, based on the input the expected yield are predicted with the trained models. Support Vector Regressor: Support Vector Regres sion(SVR) is another supervised machine learning algorithm used for regression. There are multiple kernels such as linear, polynomial and radial basis function kernel (RBF). We utilize RBF kernel