Assessing Environmental Impact: Machine Learning for Crop Yield Prediction

Sowjanya Addu, Srujana Sheelam, Samhitha Mekala, Nazma Sulthana, Lohitha Mekala and Zaid Alsalami

1Department of Computer Science, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India.
2The Islamic University, Najaf, Iraq

Abstract. Most agricultural crops have been badly affected by the effect of global climate change in India. In terms of their output over the past 20 years. It will allow policymakers and farmers to take effective marketing and storage steps to predict crop yields earlier in their harvest. This project will allow farmers to capture the yield of their crops before cultivation in the field of agriculture and thus help them make the necessary decisions. Implementation of such a method with web-based graphic software that is simple to use and the machine learning algorithm can then be distributed. This paper focuses mainly on predicting the yield of the crop by applying various machine-learning techniques. The classifier models used here include KNN, Decision Tree, Random Forest, and Voting Classifier. The prediction made by machine learning algorithms will help the farmers decide which crop to grow to induce the most yield by considering factors like temperature, rainfall, humidity, pH, etc. This bridges the gap between technology and the agriculture sector.

1 Introduction

1.1 Overview:

Crop production may also be an advanced development that is affected by the input parameters of soil and conditions. Process parameters of agriculture vary from area to area and producer to producer. It may also be a discouraging challenge to collect such information in an even larger room. The Indian Meteoric Department, however, tabulates the environmental condition information collected in the Republic of India at each 1 sq.m space in various components of the district. The huge sets of such information are often used to predict their effect on that district or place's main crops. Within the agriculture or related sciences field, there are entirely different foretelling methodologies developed and evaluated by researchers around the globe. Some studies of that type are: In alternative countries, agricultural researchers have shown that attempts to increase crop yield by pesticide state maximization have been carried out. Driven strategies for dangerously high chemical use have been introduced. The association between chemical use and crop yield has been stated in these studies. Agriculture is a partner trade sector that has gained
significantly in recent years from the growth of detector technology, information science, and machine learning (ML) techniques. These innovations return to our society's round-faced environmental and population stresses, everywhere studies suggest a need for a robust increase in international agricultural yield to provide food to an increasing population on a warmer planet. The metric capacity unit makes use of some relatively remote sensing information over the farm for most of the work tired of the world of yield foretelling. Agriculture aims to boost and increase the yield of crops and thus the quality of crops to support human life.

1.2 Problem Statement:
In ancient times, agriculture was a major resource to fulfill the human daily needs to survive and is also considered a major occupation by farmers. So, the farmers should be following the conventional methods to get healthy crop yield and they used to keep their agricultural fields and healthy diversity without mixing of chemicals. However, due to changes in climate conditions, food security is rapidly decreasing day to day.

So, this one should be considered a major challenge or issue in determining the new methodology for precision farming. So, the recent automatic detection or prediction technologies such as Machine Learning are one of the best solutions for crop prediction based on collecting the soil and weather parameters.

1.3 Objectives:
These days, machine learning plays a vital role in every field, so in agriculture using machine learning algorithms, we can improve the efficiency of crop prediction for increasing the crop yield. In machine learning technology, the classification of the model is trained with the preceding dataset by feeding them and tested with sample inputs to calculate the prediction of accuracy. In this system, we are collecting the proper dataset to feed them with several machine learning classifiers such as KNN, DT, and RF algorithms. These classifiers are used in many previous systems to predict the name of crop and crop yield. So, to improve the prediction accuracy we are implementing the ensemble Voting Classifier. Finally, this system performed comparisons between several ML classifiers to obtain the best model for crop yield prediction.

2 Literature Review

Naveen Kumar et al. [1] proposed a crop yield prediction system based on ML techniques. In this research paper, the author has used three datasets, namely Rainfall, Yield, and Soil dataset which belongs to the sugarcane crop. These three datasets are combined and trained with machine learning techniques such as SVM, KNN, and LSSVM to calculate the accuracy of three techniques for determining the estimated crop yield estimation. Finally, the best model will predict the estimated crop yield cost like HIGH, MID, and LOW.

Mithila Sompura and Aakash Parmar [2] were implemented the early rainfall prediction with help of a Decision Tree machine learning classifier. According to the authors, early rainfall prediction is important to reduce the destruction of crops and east to maintain the agricultural farms efficiently. Based on traditional methods cannot help to predict accurate results. So, the machine learning classifiers will provide accurate results. Therefore, this research paper trained the highest predictive accuracy model to achieve long-term rainfall prediction. In this research, the rainfall dataset was trained with IDA and CART decision tree algorithms.
The prediction of crops is an important task in an agricultural area, where the farmers are willing to know which type of crop needs to cultivate. So, Prachi Shah and Vinita Shah [3] introduced the prediction of Groundnut crop yield based on various machine learning algorithms. In this research, the authors used the past eight years of the Groundnut dataset to perform the analysis based on ANN, KNN, and Multiple linear regression of machine learning algorithms. This system predicts the Groundnut crop yield by using the attributes like Soil, pH, and Environmental with help of the best model KNN algorithm which is provided the best results compared to other ML classifiers.

Nilesh Mehta et al. [4] was proposed the android based smart agriculture system. In every country, agriculture growth will depend on soil parameters like pH, soil moisture, and weather conditions such as temperature, humidity, rainfall, etc. The rapid growth of machine learning technology is useful to improve crop productivity in various agriculture fields. In this research, the author developed the combinations of web and android based applications based on ML methodology to predict the profitable crop. Here the weather conditions dataset is imported from IMD reposito and Soli parameters dataset are also imported respective repositories.

Finally, the Multiple Linear Regression classifier will be trained with weather and soil datasets and build the model to predict the profitable crop to the farmers for cultivating the proper crop.

The growth of agriculture is playing a vital role in raising the country’s economy, but due to changes of unnatural weather conditions, it will be affected to improve the crop yield. So, Muthaiah and Balamurugan [5] implemented crop yield prediction by using of Random Forest algorithm. Here the Tamil Nādu real-time datasets such as historical crop yield and weather and soil parameters are used to build the model and tested with sample inputs. Majorly this system helps farmers to predict the crop yield before cultivating the specific crop.

Santra et al. [6] proposed a new approach to predict crop production based on several machine learning algorithms such as DT, ANN, KNN, and Regression analysis. Here prediction of early crop production is important to manage the pricing market distribution and exports-imports etc. The prediction of sugarcane crop yield proposed by Ramesh Medar et al. [7] based on supervised ML techniques such as SVM and Naive-Bayes algorithm and Shamim Akhter et al. [8] developed a new approach that is used to predict the weather conditions for remote areas based on random forest machine learning algorithm. In [8] [9] machine learning techniques used for predicting the attack to protect the systems and plant disease detection using ML.

3 Methodology

3.1 System Model:

Figure.1 depicts the proposed system model. From Figure.1, the administrator will perform the data preprocessing process for cleaning the dataset and extracting the features. The admin will calculate the accuracy of all machine learning algorithms and generate the table format with all the algorithm’s performance metrics. This system also generates the system's experimental results with bar charts based on ML performance metrics. Based on the all- ML analysis, the voting classifier has given the best accuracy results so that, the user will enter the current crop field details and weather details such as temperature, humidity, pH, and rainfall then this system will return the prediction results like crop name and crop yield.
3.2 System Modules

**Administration:**
In the following, we are given a brief description of the system’s modules which are operated by the admin.

**Dataset Collection:**
In this system, we are using the prediction of crop yield dataset which was imported from internet resources i.e., Kaggle. This dataset contains four independent attributes such as temperature, humidity, and rainfall, and three dependent attributes or target attributes are crop name, area, and production. Figure 2 depicts the crop yield dataset which contains 7 columns and 3100 records and the target class like crop type has 31 labels.

Fig. 2. Crop Yield Dataset
Data Preprocessing:
The crop dataset will be in the form of raw data which is not understood by ML algorithms. So, in the data preprocessing stage, this system will read the data from the .csv file format and convert it into data frames. Later it will check if the dataset contains any missing values like question marks, special characters, and null values. However, the selected dataset does not contain any missing values.

Train and Test Split:
After the pre-processing stage, the crop dataset will be split into 80 to 20 ratios. By using of \( \text{train}_\text{test}_\text{split}() \) method this system will split the 80 percent training dataset with 2480 records and the 20 percent testing dataset with 620 records.

Training the models:
After splitting the dataset as training and testing, this system will train the ML models with the help of invoking the \( \text{fit}() \) method with input parameters as independent variables \( x_\text{train} \) and target column values \( y_\text{train} \) of the training dataset.

ML Evaluations:
In the ML evaluations, this system will compare the algorithm performances between all ML techniques, so finally, with input parameters of predicted values and actual values then this system will calculate all performance metrics such as accuracy score, precision, recall, and f1score.

User:
In the following, we are given a brief description of the system’s module requirements which are operated by a user.

Registration & Login:
In this system, the user will register by filling out the registration form with mandatory fields. Thereafter, the creation of an account successfully, the user will be authenticated with this system by submitting valid login credentials.

Crop Yield Prediction:
From the ML evaluations, this system will select the best classification model based on algorithm performance metrics. According to this system's experimental results, the voting classifier model was selected as the best model which is predicted with the highest accuracy. So, when the user enters the testing data like temperature, humidity, pH, and rainfall then this system predicts the different crop names, size of area, and productions with various testing data points. Thereafter, this system will return crop yield based on a calculation of area and production.

3.3 Methodologies

KNN:
The KNN classifier is a supervised machine-learning algorithm. It is mostly used for classification problems. The process of KNN is training the model with independent and dependent variables with the number of \( K \) nearest neighbors (i.e., \( k=3 \)). The test samples are used to predict the results, based on Euclidian distances which return similar results. In the
KNN, to obtain the best accuracy, the $K$ value should be assigned with the highest odd value. In Python, which belongs to the KNN algorithm, we used the \textit{KNeighbors Classifier} module which is imported from \textit{sklearn.neighbors} packages.

**Algorithm Steps:**

1. Calculate the distance between the query instance and all the training samples.
2. Sort the distance and determine the K nearest neighbors based on the distance.
3. Gather the categories of the k nearest neighbors.
4. Return the class with the highest frequency as the predicted class

- **Equation (Euclidean Distance for example):**
  $$\text{Distance}(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

  Where:
  - $p, q$ are two data points.
  - $n$ is the number of dimensions (features).

**Decision Tree:**

The DT is also one of the supervised machine learning algorithms that follow the IF-THEN policies with tree structures. The DT will be constructed with several nodes such as root node, branches, and leaf node. The root need represents attributes or features and the decision rule will be represented by branches, finally, the child or leaf node represents the result or outcome. In the DT, the trees are partitioned based on feature values in a recursive manner which helps in decision making. The predefined class \textit{DecisionTreeClassifier} is used to predict the model which is imported from \textit{sklearn.tree} packages.

**Algorithm Steps:**

1. Select the best attribute using measures like Information Gain or Gini Impurity.
2. Split the dataset into subsets based on the value of the selected attribute.
3. Repeat the process recursively on each subset until one of the stopping conditions is met (e.g., maximum depth reached, minimum samples per leaf).

- **Equation (Information Gain for example):**
  $$IG(D, A) = H(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} \cdot H(D_v)$$

  Where:
  - $IG(D, A)$ is the information gain by splitting dataset $D$ on attribute $A$.
  - $H(D)$ is the entropy of dataset $D$.
  - $D_v$ is the subset of dataset $D$ for which attribute $A$ has value $v$.

**Random Forest:**

The RF classifier was majorly used to resolve the classification problems that belong to the supervised ML algorithm. Here the RF classifier will generate a group of decision trees with the help of a subset of the chosen training set. In the RF, the prediction results will be
generated based on the highest votes of different decision trees. In this system, we are using the Random Forest Classifier class to build the prediction model, this class was imported from sklearn. Ensemble modules.

Algorithm Steps:
1. Randomly select $k$ features from the total features.
2. From the $k$ features, build a decision tree using the best-split point.
3. Repeat steps 1 and 2 to build multiple trees.
4. Aggregate the predictions from each tree (e.g., a majority vote for classification, average for regression).

No Specific equation, but Ensemble Aggregation Mechanisms:
- **Majority voting:** In Classification, the class that gets the most votes from all decision trees is selected as the final prediction.
- **Mean (Regression):** In regression, the average of all the predictions from individual trees is taken as the final Prediction

Voting Classifier:
The voting classifier is an ensemble machine-learning algorithm. It is also called a hybrid classifier which will create the voting classifier object holding the multiple models and it returns the predicted result based on the highest majority of voting. This classifier will support two types of voting such as hard voting and soft voting. In the hard voting, it will collect the outputs of all models and determine the class which is the highest majority of votes. But in soft voting, the average of the highest class is the final prediction class.

- It can be implemented in two ways: hard voting (simple majority voting) and soft voting (weighted average probabilities).
- In hard voting, the class with the majority of votes is chosen as the final prediction.
- In soft voting, the probabilities predicted by each classifier are averaged, and the class with the highest average probability is chosen.
- Mathematically, soft voting can be represented as:

$$P(y|x) = \frac{1}{M} \sum_{i=1}^{M} P_i(y|x)$$

- Where $P(y|x)$ is the probability of class $y$ given input $x$, $M$ is the number of classifiers, and $P_i(y|x)$ is the probability predicted by the $i$th classifier.
Fig. 3. Algorithms Analysis on F1-Score

Fig. 4. Algorithms Analysis on Precision
Fig. 5. Algorithms Analysis on Recall

Fig. 6. Algorithms Analysis on Accuracy
4 Result

In our study, we input factors such as temperature, humidity, rainfall, and soil pH to predict the optimal crop for cultivation in a given area, along with the expected yield per acre. Through the utilization of various machine learning algorithms trained on collected data, we aim to provide farmers with informed recommendations for crop selection, facilitating efficient and productive agricultural practices.

Fig.7. Crop Prediction with Input Data.

Fig.8. Crop Prediction Result

5 Conclusion

This system is all about using advanced computer techniques to predict how well different crops will grow. We've gathered lots of information about past crops, like what they were, what the weather was like, and what the soil was like. We've then used fancy math to figure out which crops are likely to do best in the future. We've combined a bunch of different methods to make our predictions as accurate as possible. By doing this, we're helping farmers decide which crops to plant in their fields for the best harvest. This work is employed to give them the knowledge they need to grow their crops efficiently and effectively.
References


