AN ENSEMBLE ALGORITHM FOR CROP YIELD PREDICTION

Pavithra S R[1], Priya Darshini G[2], Maheswari.M[3], Malathi.A[4]

[1] *Student, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India*
[2] *Student, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India*
[3] *Assistant professor, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India*
[4] *Assistant professor, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India*

ABSTRACT

Agriculture has a significant part in the Indian economy and jobs. The most common problem faced by Indian farmers is that they do not select crops based on soil requirements, resulting in major productivity concerns. Machine learning algorithms will assist farmers in determining which crop to plant in order to provide the most yield by taking into account parameters such as temperature, rainfall, location, and so on. Precision agriculture can help overcome this challenge. This method considers three elements: soil characteristics, soil kinds, and crop yield data gathering. Based on these parameters, the farmer is recommended a crop to produce. This system includes a model for estimating crop yield that is exact and accurate. This system includes a model that is exact and accurate in predicting crop production and providing the end-user with the appropriate fertilizer ratio recommendations based on atmospheric and soil factors of the land, hence increasing crop yield and farmer revenue. As a result, the suggested system's pre-processing methods are based on feature scaling techniques including Min-Max Scaling, Standardization, and Normalization, which help to extract data on soil quality and weather-related information as an input. The soil's quality, such as nitrogen, phosphorus, potassium, pH, Rainfall, temperature, and humidity are all weather-related variables that can be used to predict a better harvest. We are using datasets from the Kaggle platform for our project. Using the ensemble modeling approach, we were able to do this. Different machine learning techniques, such as the Random Forest algorithm, Adaboost Classifier, Gradient Boosting Classifier, and K-Nearest Neighbors, were compared.

Keyword: - *Precision Agriculture, Feature Scaling, Random Forest algorithm, Adaboost Classifier, Gradient Boosting Classifier, Crop yield and K-Nearest Neighbor.*

1. INTRODUCTION

The Indian economy is based on agriculture. Weather conditions mostly determine agricultural yield in India. Rice cultivation is primarily reliant on rain. Farmers should be able to receive more money from the same piece of land with less labor if productivity is enhanced. Precision agriculture makes it possible. Precision farming, as the name suggests, entails using precise and appropriate inputs such as fertilizers and soil. However, due to globalization, the agricultural trend has radically changed in recent years, and several variables have had an impact on the health of India's agriculture. Many innovative technologies have been developed to help agriculture regain its health. Precision agriculture is applied at the right time to the plant for increased output and yields Precision agriculture systems aren't all created equal. It's a type of site-specific farming technology. Data analysis is the process of examining data sets in order to draw conclusions about the information they contain, sometimes with the aid of specialized tools and software (DA).

Previously, yield prediction was based on the farmer's knowledge of a specific field and crop. Farmers, on the other hand, are compelled to produce an increasing number of crops as the weather fluctuates from day to day. As a result, the proposed system incorporates data on soil quality as well as weather-related data. The quality of the soil, including nitrogen, phosphorus, potassium, and pH. Weather data includes things like rainfall, temperature, and humidity. With the current technique, producers don't have to worry about crop forecasting and instead focus on yield prediction. Pesticides, environmental, and meteorological elements relevant to the crop will not be taken into account using existing approaches unless the correct crop is anticipated.

2. LITERATURE SURVEY

A novel system known as extensible Crop Yield prediction framework is built for precision agriculture using data mining techniques. In this paper there is an investigation of requirement for crop yield prediction and different systems have been utilized and finally it results in a framework which is flexible for prediction accuracy [1]. From many base papers we came to know that Neural Networks, Decision Tree are the most used algorithms for these models. Decision tree uses parameters like maximum depth and n-estimators, so that by adjusting those parameters, we can get better results. After research, we have concluded that ensemble of Decision tree regressor and AdaBoost regressor gave major accuracy. Crop yield prediction subsumes prediction of the yield of the crop from formerly data. Ultimately, this strategy gives us a recommendation of which crop should be cultivated based on the weather conditions of the field location [2]. Variables such as ph, humidity, rainfall, temperature, and so on are included in the dataset. During training, a huge number of decision trees are formed, and the outcome or output is split into classes based on the number of classes. A Decision Tree Classifier is also used in this study to compare the two and select the right choice. Supervised Learning techniques were used to predict the outcome. For training the model, Random Forest has been compared with Decision tree. Crop production using one of the most used boosting methods was introduced [3]. They evaluated two boosting algorithms: AdaBoost and Gradient Boost. The goal of employing a boosting algorithm is to increase a poor learner's performance so that a better outcome may be produced. The results indicated that the AdaBoost Regressor with Decision Tree has 94.67 percent accuracy compared to 94.9 percent for the AdaBoost Regressor with Random Forest. A study of the literature on machine learning models for predicting agricultural productivity using meteorological data was presented [4]. The "Decision Tree" and "KNN" ML classification algorithms were compared in this study. Soil parameters, climatic parameters, and production parameters make up the data set. This study uses machine learning methods to calculate. This paper compared both methods separately, but did not combine them. [5]. According to the report, extension of the search to include additional crop yield-related parameters. Rainfall, temperature, and soil fertility were among the issues addressed. When comparing the experimental values and outcomes for the crop paddy dataset, the deep reinforcement learning model is shown to predict the data with a 93.7 percent higher accuracy and precision than the other methods tested. A classifier-based crop recommendation system was introduced [6]. ML models based on data gathered by collar sensors improved livestock productivity by forecasting reproductive patterns, detecting eating problems, and predicting cow behavior. It illustrates how knowledge-based agriculture may boost long-term production and product quality. For weed prediction, Convolution Neural Networks is the best, Random Forest is better for crop production prediction, and the regression method is superior for weather forecast. A detailed examination of the benefits and drawbacks of machine learning-based crop production prediction, as well as appropriate identification of present and future agricultural sector issues was presented [7]. The user can predict the most suitable crop and its estimated yield for a chosen year. This model uses primary classification, techniques like the random forest, k-NN, logistic regression, decision tree, XGBoost, SVM and gradient boosting classifier for determining the most suitable crop and regression algorithms like Linear Regression, k-NN, DBSCAN, Random Forest and ANN algorithm to estimate the yield of the most optimal crop identified earlier [8]. To know the region-specific crop yield analysis and it is processed by implementing by random forest algorithm. In this project have chosen dataset which in .csv format. For the training purpose 80% of data is used and remaining 20% of data is used for testing. After the successful training and testing next step is finding the accuracy of the model. We have achieved a good accuracy which means this model is good for predicting yield [9]. The workflow can be used to run repeatable experiments (e.g. early season or end of season predictions) using standard input data to obtain reproducible results.[10] The results serve as a starting point for further optimizations. In our case studies, we predicted yield at regional level for five crops (soft wheat, spring barley, sunflower, sugar beet, potatoes) .

3. EXISTING SYSTEM

Using Supervised Learning approaches, Machine Learning algorithms can forecast a target/outcome in a variety of forecasting areas. However, because their study does not include any algorithms, it is unable to provide a clear picture of the suggested work's feasibility. The Random Forest method constructs decision trees on distinct data samples, predicts data from each subset, and then gives the system a better answer through voting. The data was trained with Random Forest using the bagging approach. To improve accuracy, randomization must be inserted in a way that minimizes correlation while preserving strength. They can only deliver an accuracy of 95.7 percent.

EXISTING ALGORITHMS	ACCURACY
AdaBoost Regressor with Decision Tree	95.7%
AdaBoost Regressor with Random Forest Classifier	94.9%
Bagging with KNN classifiers	89%
Decision tree with Gradient Boosting	93.0%
Decision tree with Random Forest Regressor	95.0%

Table: Predictions of the proposed Algorithms

4. PROPOSED SYSTEM

We construct crop prediction utilizing an efficient algorithm in the proposed system. The task at hand is to create an effective model for predicting a better crop. We apply machine learning methods in these projects, which are essentially hybrid classification/ensemble models. An ensemble of models derived from logistic regression, lDA, Decision tree, SVM, Random Forest, and KNN are used in our research. This can improve accuracy and provide a more accurate prediction system.

4.1 Dataset collection:

The major purpose is to locate and acquire a dataset in order to execute crop selection and yield prediction at the district level, which is done for crops. Because India has a diverse climate that varies on a regular basis, data is collected at the district level for several Indian states. The state and district names are used to identify each location in the data. Data was gathered from a variety of sources, including the ICRISAT International Crops Research Institute for the Semi-Arid Tropics, Kaggle, and the Indian National Portal. The dataset contains information on crop production for the last 4-5 years and is organized by state, district, crop name, area, production, rainfall, MSP, and soil deficiencies.

4.2 Data Cleaning:

In any machine learning effort, data cleaning is vital. Data cleaning is performed in this module to prepare data for analysis by eliminating or changing data that is erroneous, incomplete, redundant, or badly formatted. You can use a variety of statistical analysis and data visualization tools to investigate tabular data and identify data cleaning activities you might wish to do. We employ algorithms like MinMaxScaler, StandardScaler, Normalizer, Binarizer, and Label Encoder in this pre-processing.

4.3 Feature Extraction:

This is done to reduce the number of attributes in the dataset, resulting in benefits such as faster training and improved accuracy. When an algorithm's input data is too extensive to analyze and is thought to be redundant, it can be reduced to a smaller collection of features (also named feature vector). Feature selection is the process of identifying a subset of the initial features. The selected features should contain the important information from the input data, allowing the intended task to be completed using this reduced representation rather than the entire initial data.

4.4 Model training:

A training model is a set of data used to train a machine-learning algorithm. It is made up of sample output data as well as the equivalent sets of input data that have an impact on the outcome. The training model is used to process the input data via the algorithm in order to compare the processed output to the sample output. The model is modified based on the correlation's result. "Model fitting" refers to this iterative procedure. The model's precision is based on the accuracy of the training or validation datasets. In machine language, model training is the act of supplying data to an ML algorithm to help find and learn suitable values for all of the variables involved. Machines come in a variety of shapes and sizes. The most prevalent forms of machine learning models are supervised and unsupervised learning. To train the model on the cleaned dataset after dimensionality reduction, we employ supervised classification algorithms like linear regression in this module. We used 80% of the dataset to train the model.

4.5 Testing model:

The trained machine learning model is put to the test with the test dataset in this module. In machine learning, model testing is the process of evaluating the performance of a fully trained model on a testing set. Because a programmer normally inputs the data and desired behavior, and the machine elaborates the logic, the aim of machine learning testing is, first and foremost, to ensure that the learned logic will remain consistent no matter how many times the program is called in the testing.

4.6 Performance Evaluation:

In this session, we use performance evaluation measures such as F1 score, accuracy, and classification error to assess the performance of a trained machine learning model. Accuracy is defined as being very close to a measured value or a standard established. In time series analysis, accuracy refers to the anticipated value being very close to the actual value. True positive cases are designated by TP, true negative cases are marked by TN, false positive cases are denoted by FP, and false negative cases are denoted by FN.

The formula for accuracy is A=(TP+TN)/(TP+FP+FN+TN)

Fig 4.6.1: Confusion Matrix

4.7 Prediction:

When predicting the likelihood of a given result, prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data. For each record in the new data, the algorithm will generate probable values for an unknown variable, allowing the model builder to determine what that value will most likely be. This project has a predicted accuracy of 99.4%.

PROPOSED ALGORITHMS	ACCURACY
Linear Regression	95.5
Linear Discriminant Analysis	96.4
K- Nearest Neighbor	98.1
Classification & Regression Tree	98.8
Support Vector Machine	97.4
AdaBoost	14.7
Gradient Boosting Machine	98.9
Random Forest	99.4

Table 4.7.1: Predictions of the proposed Algorithms

4.8 Decision Tree:

It is a supervised learning method of machine learning with a tree structure that is used for classification and regression. Simple decision rules inferred from data attributes are used to build a model that predicts the value of a target variable. Internal nodes represent data set properties, branches represent decision rules, and each leaf node reflects the outcome in the decision tree structure.

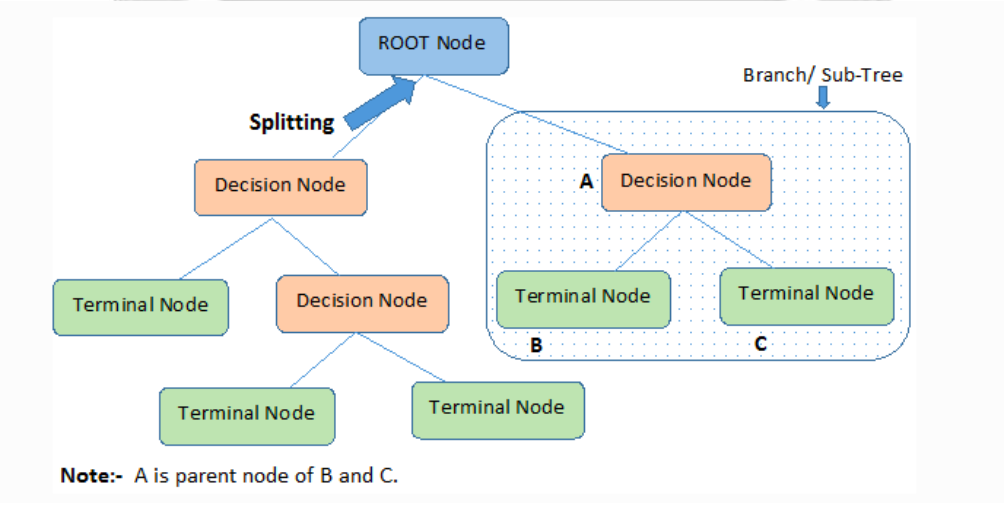

Fig 4.8.1: Decision Tree

Some of the hypotheticals we make when using Decision Tree are as follows
• At first, the entire training set is allowed to be the root.
• It's preferable to have categorical point values. However, they must be discretized before the model can be erected If the values are nonstop.
• Records are recursively distributed grounded on trait values.
• A statistical approach is used to determine the order in which characteristics are placed as the tree's root or an internal node
The Sum of Product(bribe) form is used in Decision Trees. Disjunctive Normal Form is another name for the sum of products(bribe). Every branch from the tree's root to a splint knot with the same class is a confluence(product) of values, while distinct branches terminating in that class constitute a disjunction(sum).
The fundamental problem in enforcing a decision tree is determining which attributes should be considered for the root node and each position. Taking care of this is appertained to as trait selection. At each position, we use several trait selection strategies to discover the trait that can be called the root node.

4.9 Random Forest Algorithm:

Random forest could be a learning algorithm that's supervised. Random Forest may be a classifier that mixes variety of decision trees on different subsets of a dataset and averages the results to extend the dataset's predicted accuracy.
The random forest algorithm has the subsequent steps:
1st step: In Random Forest, n random records are chosen randomly from an information set with k records.
Step 2: for every sample, a personal decision tree is made.
Step 3: Each decision tree produces a result.
Step 4: For classification and regression, the ultimate output is predicated on Majority Voting or Averaging, accordingly.

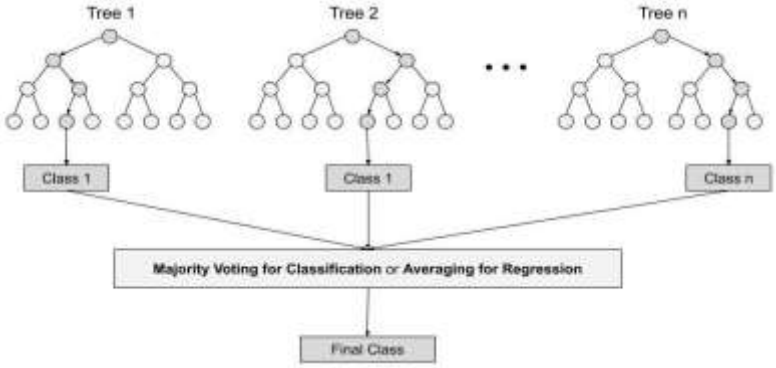

Fig 4.9.1: Random Forest Algorithm.

4.10 K-Nearest Neighbor:

The supervised machine learning algorithm K-Nearest Neighbor records all instances in n-dimensional space that match to training data points. When it receives an unknown discrete data, it evaluates the closest k number of saved instances (nearest neighbors) and produces Crop Yield Prediction. The most prevalent class is used as the prediction in Supervised Learning Techniques, and for real-valued data, the mean of k nearest neighbors is returned.

5. RESULT AND DISCUSSION:

In a recent study, the effectiveness of carrying it technique was evaluated using Espresso and R-Tool, which are extensively used to apply these carrying it methodologies. When it comes to precision, recognition, and accuracy, the genuine effectiveness carries with it a method through.

Precision Value:

Precision is calculated as the number of true positive predictions (TP) divided by the total number of positive predictions (TP+FP).

Precision value = TP/TP+FP

Recall value:

Recall value is specified to as the relevant datasets that are related to the other request Search.

F measure:

F measure test's accurateness and is define as the weighted harmonic mean of the precision and recall of the analysis Analyze to comparison between K-Nearest Neighbor (Existing System) and Linear Regression (Proposed System) with parameter evaluation.

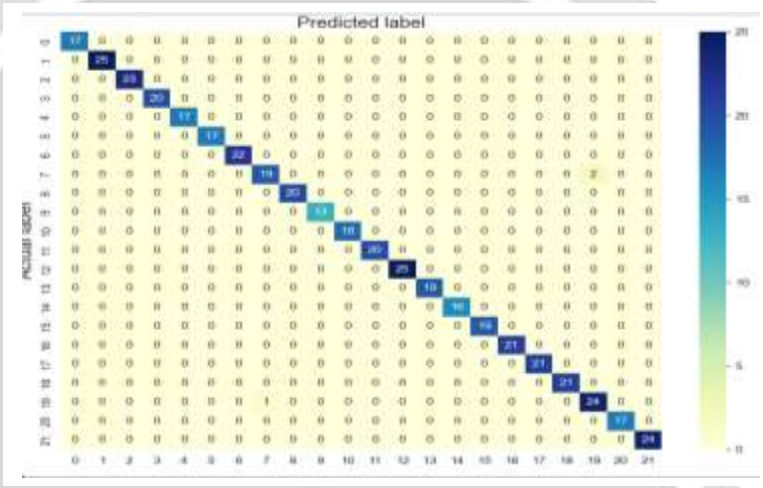

Fig 5.1.1.: Accuracies generated by Confusion matrix.

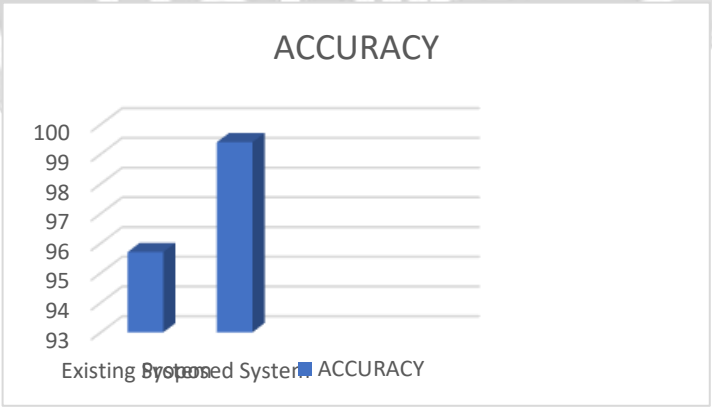

Fig 5.1.2: Comparison of accuracies of Existing and Proposed systems.

Fig 5.1.3: Home page of the system

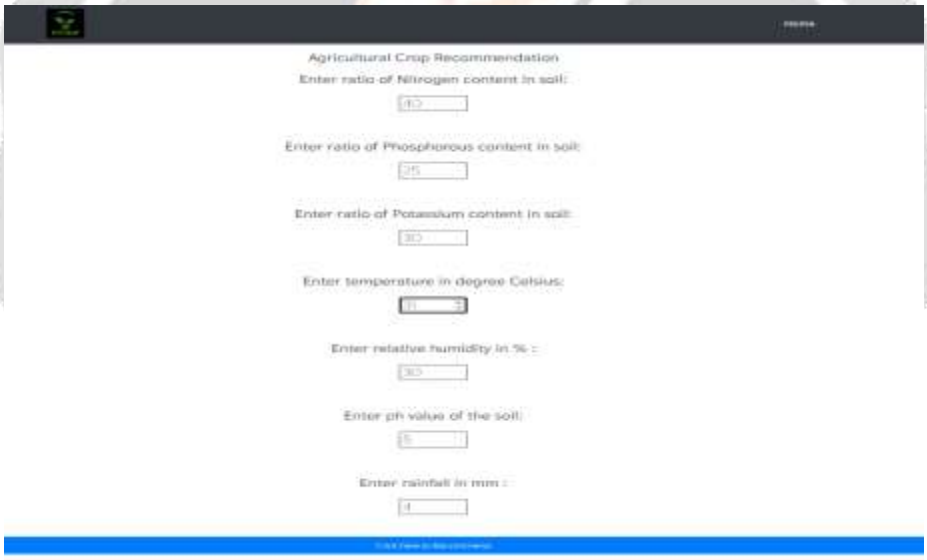

Fig 5.1.4: User interface for data entry.

Fig 5.1.5: Predicted output page

6. CONCLUSION AND FUTURE ENCHANCEMENT

India is a country where agriculture is extremely important. The nation prospers when the farmers prosper. As a result of our work, farmers will be able to sow the appropriate seed based on soil conditions, increasing production and profit. As a result, farmers may plant the appropriate crop, increasing yield and enhancing the nation's overall productivity. Our next work will focus on creating a better data set with a larger number of features, as well as implementing more precise yield prediction. Instead of using solely historical data, we will use real-time weather conditions to generate the most efficient suggestion output.

7. REFERENCES

[1] Aakunuri Manjula, G. Narsimha, "XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction", IEEE Sponsored 9th ISCO, 2015. [6] D. Ramesh, B. Vishnu Vardhan, "Analysis.

[2] Mummaleti Keerthana; K J M Meghana; Siginamsetty Pravallika; Modepalli Kavitha, An Ensemble Algorithm for Crop Yield Prediction, Conference: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)

[3] M. Keerthana, K. J. M. Meghana, S. Pravallika and M. Kavitha, "An Ensemble Algorithm for Crop Yield Prediction," 2021 Third International Conference on IntelligentCommunication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 963-970

[4] Elavarasan, Dhivya, and PM Durairaj Vincent. "Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications." IEEE Access 8 (2020): 86886-86901.

[5] Elisa Kamir, François Waldner, Zvi Hochman, estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods,ISPRS Journal of Photogrammetry and Remote Sensing,Volume 160,2020,Pages 124-135,ISSN, 0924-2716.

[6] Patil, Ajinkya, et al. "Crop Prediction using Machine Learning Algorithms." Kapila Journal of Research 1.1 (2020): 1-8.

[7] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin and N. Khan, "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction," in IEEE Access, vol. 9, pp. 63406-63439, 2021, doi: 10.1109/ACCESS.2021.3075159.

[8] Medar, Ramesh, Vijay S. Rajpurohit, and Shweta Shweta. "Crop yield prediction using machine learning techniques.

[9] PallaviKamath, Pallavi PatilShrilatha, S SushmaSowmyaS Dept. Computer Science and Engineering, Shri Madhwa Vadiraja Institute of Technology and Management (Affiliated to VTU), Udupi 574115, India ,Received 28 June 2021, Accepted 5 July 2021, Available online 13 August 2021, Version of Record 1 November 2021.

[10] Paudel, Dilli; Boogaard, Hendrik; de Wit, Allard; Janssen, Sander; et al. Machine learning for large-scale crop yield forecasting [2021].