SmartCrop - A Crop Recommendation System

Aastha Dogra
Computer Science
Shiv Nadar University
Delhi NCR, India

Rimjhim Singh
Computer Science
Shiv Nadar University
Delhi NCR, India

Ramya Karna
Computer Science
Shiv Nadar University
Delhi NCR, India

***Abstract*—Crop prediction is a complex procedure that is often determined by several factors such as environment, weather, and the interactions occurring between them. In order to predict yield and recommended crops accurately, we need to achieve a fundamental understanding of the factors in question, that further requires comprehensive datasets and algorithms capable of handling such data and processing it smoothly. We attempt to design a crop prediction system after analyzing data derived from 2 sources and implementing various classifiers and a sequential neural network approach to the dataset.**

I. Introduction

The agriculture sector in India employs nearly half of the workforce in the country. India's production of food grains has been increasing every year, and it is among the top producers of several crops such as wheat, rice, pulses, sugarcane and cotton. However, the agricultural yield (quantity of a crop produced per unit of land) is found to be lower in the case of most crops, as compared to other top producing countries such as China, Brazil and the United States. Although India ranks third in the production of rice, its yield is lower than Brazil, China and the United States. The same trend is observed for pulses, where it is the second highest producer. This issue may be due to the fact that farmers do not have appropriate knowledge about what crops would give maximum yield in a particular area. Hence, we aim to facilitate this process by helping farmers gain insights on what crops are suitable for them to grow at a particular location considering the conditions and environmental factors in that case.

II. Aim

To build a system that can provide Indian farmers with predictive insights, allowing them to make better decisions about which crops to produce considering the soil quality, rainfall. We would also include some additional features like recommendations for crops.

Machine learning is an important decision support tool for crop yield prediction, including supporting decisions on what crops to grow and what to do during the growing season of the crops. Several machine learning algorithms have been applied to support crop yield prediction research. In this project, we tested some machine learning algorithms on a rich dataset and tried to enhance their accuracies by proposing concatenated models.

III. Literature Review

An accurate crop yield production model can help the agricultural sector workers determine what to grow, and when to grow. Elavarasan et al. conducted a survey of publications on machine learning models that are linked to crop yield prediction based on climatic parameters.

In 2015, Somvanshi and Mishra put forward various machine learning approaches for this problem and their following usage in plant biology.

During the course of this project, we came across several research papers on crop prediction. There has been a lot of progress in this field when it comes to crop production recommendation systems, where crop yield was predicted utilizing data analytics and considering a hybrid approach, utilizing the constantly generated agricultural data.

According to Gandhi and Armstrong, who studied the application of data mining in the agricultural sector for various purposes, it was surmised that we need further research to determine how to implement data mining into complex agricultural datasets.

Some of the commonly applied machine learning techniques commonly implemented in this field are multivariate regression, decision trees, associate rule mining and artificial neural networks. One important implication of this topic is that machine learning models treat the output (crop yield) as an implicit function of the input variables, that are typically the genetic and environmental components. This could result in a highly non-linear, complex function.

In a paper by Khaki, they utilize deep neural networks to predict yield, check yield and yield difference of corn hybrids from genotype and environment data. Apart from the numerical and categorical datasets, there is a growing trend in utilizing image datasets as well.

Rusello used convolutional neural networks for crop yield prediction that was based on satellite images in 2018.

IV. Dataset Overview

In our project 2 primary datasets have been used for the predictions and analysis:

A. Dataset 1

This dataset was created by combining 2 datasets that included information about the temperature, humidity, and rainfall requirements for certain crops along with desirable

soil conditions. It had a total of 2201 entries.

CropData

	temperature	humidity	ph	rainfall	label
0	20.87974371	82.00274423	6.502985292	202.9355362	rice
1	21.77046169	80.31964408	7.038096361	226.6555374	rice
2	23.00445915	82.3207629	7.840207144	263.9642476	rice

Fig. 1. Dataset 1

This dataset has been used to run, train and test Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier, Naive Bayes Classifier, Voting Classifier, Gradient Boosting classifier and XGBoost classifier.

B. Dataset 2

This dataset contains more features and records as compared to the first dataset and it contains information specifically about Indian states, districts , seasons and production statistics. It has almost 38,345 entries but we use 9628 for the purpose of our project and analysis. This dataset has been used to train and test the sequential neural network.

State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254	2000
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2	1
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102	321
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176	641

Fig. 2. Dataset 2

V. Proposed Model

The following methodology was followed while making SmartCrop:

Fig. 3. Proposed Workflow

VI. Methodology

A. Data Preprocessing

Dataset 2, containing 38,345 entries was initially utilized for the calculations, with information on various parameters like states, districts, seasons and production statistics. We chose Andhra Pradesh as our primary state and used data from the particular state to analyze the features.

Some of the entries had null values in the production, state and district name categories so those were managed and a new column named yield was constructed, calculated by the given formula: Yield = Production/ Area.

Next, the information in the dataset dates quite far back, so we modified it to only consider crops from after 2004, and the categorical variables like season, crop, states were converted to dummy variables and the original columns were subsequently dropped.

B. Selection of Classifiers

For this project, we have implemented several machine learning algorithms on our datasets. They have been discussed below.

a) Random Forest Classifier: Random Forest developed by Leo Breiman is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process.

Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

b) Decision Tree Classifier: The decision tree classifier (Pang-Ning et al., 2006) creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute. Each leaf represents class labels associated with the instance. Instances in the training set are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path.

c) Support Vector Classifier: A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. Support Vector Classifier is capable of performing binary and multi-class classification on a dataset.

d) Naive Bayes Classifier: Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

e) Voting Classifier: A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output. It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea

is instead of creating separate dedicated models and finding the accuracy for each of them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

f) Gradient Boosting Classifier: Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.

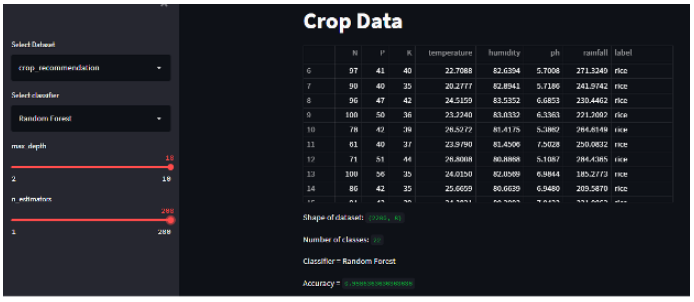
g) XGBoost Classifier: XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way.

h) Neural Network: A deep neural network is a neural network with a certain level of complexity, a neural network with more than two layers. We had imported a sequential neural network from keras for designing our own model. We had used multiple layers of dense fully connected layers so that accuracy is high. We have trained the deep neural network on the 2nd dataset for 50 epochs. For the first Neural Network for the 2nd dataset, it consisted of 1 dense input layer with relu activation function, 3 hidden layers with relu activation function and the output layer was a dense layer with one output node with linear activation function.

The reason for utilizing neural networks for crop prediction is that neural networks can capture the non-linearities that exist in the nature of crop data, and they have the ability to learn these non-linearities from the data without requiring the nonlinear model to be specified before estimation.

C. Front End Visualization

We rendered the data and results of our various machine learning approaches utilizing the Streamlit library. We can use this to analyse and compare the several models and the data in consideration.



VII. Experimentation and Results

A. Scatter Plot

We can use a scatter plot to determine whether or not two variables have a relationship or correlation. In the given image, we have plotted temperature against rainfall.

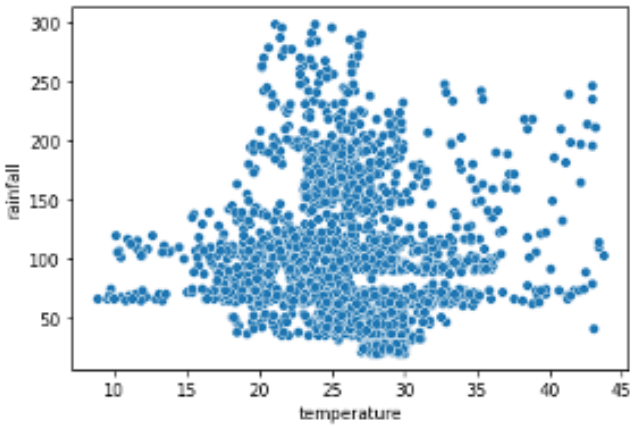


Fig. 4. Scatter plot

It was observed that the maximum number of crops tend to fall in the 25-30 degree range and have their rainfall level at just above 100.

B. Dist Plot

A dist plot plots the univariate distribution of observations. In this case, we plot density and the pH levels of the data and observe the spike in the density when pH is between 6 and 7- towards the neutral side.

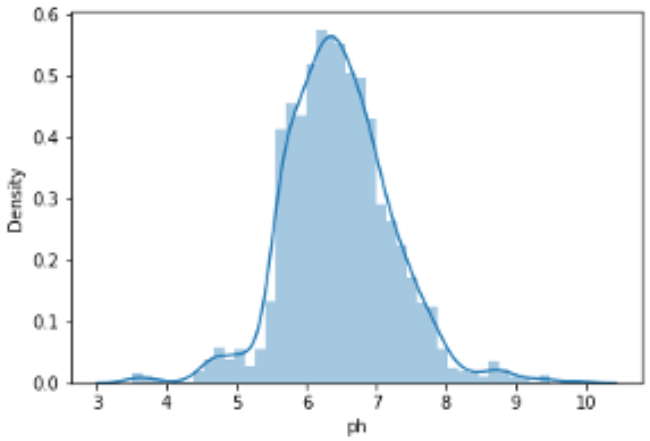


Fig. 5. Dist plot

C. Heat Map

A heat map is a two-dimensional representation of data in which values are represented by colors. [Fig 6]

We observe that P and K- the variables for the elements Phosphorus and Potassium had the highest correlation.

D. Implementation and Results

Using the preliminary insights from the exploratory data analysis, the features were outlined to be N, P, K, temperature, humidity, pH level and rainfall. A 20 percent test size was set aside and we implemented a decision tree classifier.

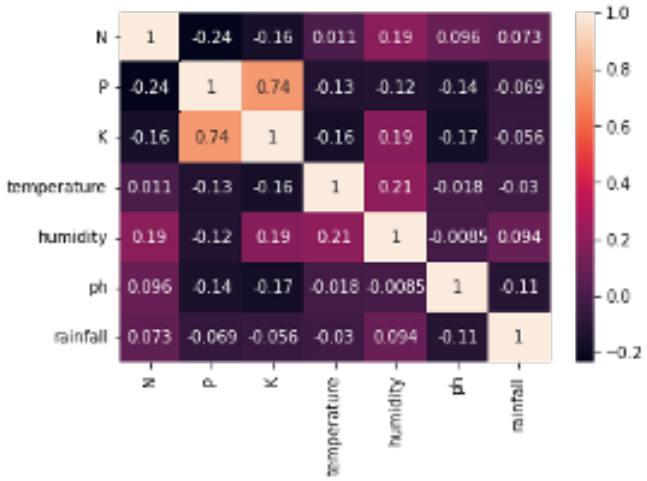
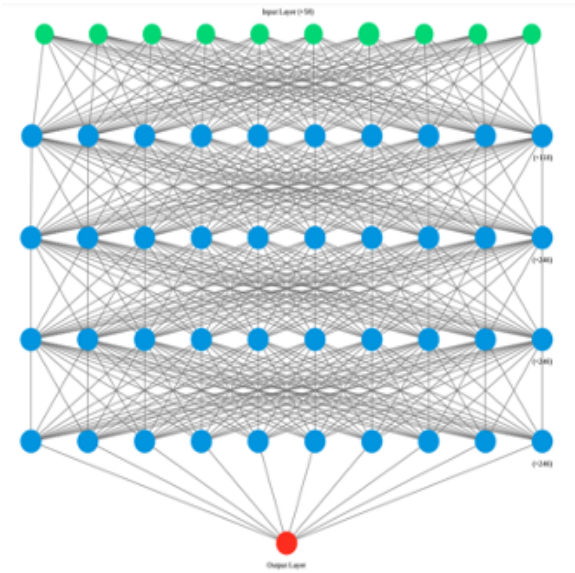


Fig. 6. Heat Map for Dataset 1

The cross validation score was formulated with a parameter value of 5, the highest being approximately 94 percent.
Subsequently, a Support Vector Machine was constructed with the C value taken as 1, as this was observed to give the best performance, where we tried out the classifier with values between 0.01 and 10. The accuracy was 97.72 percent. Once again, a k fold cross validation procedure was carried out in order to prevent overfitting of the data, and to correctly estimate the skill of the model.
Next, a Random Forest Classifier was implemented with the n-estimators considered to be 1, 2, 4, 8, 16, 32, 64, 100 and 200.

Due to the highest accuracy score of 0.995 being assigned



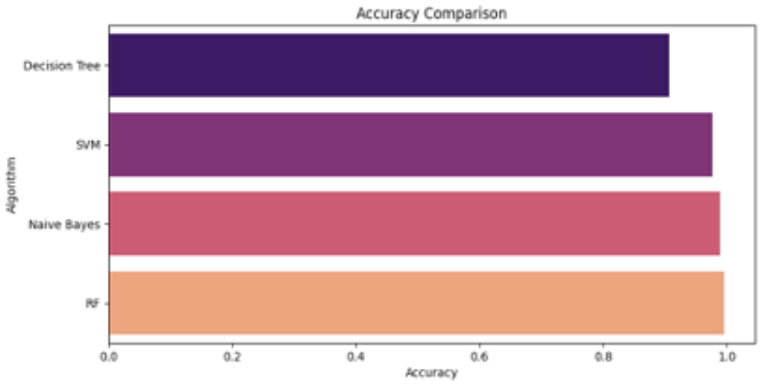
to the value when n-estimators = 16, we select the particular value and utilize it to predict the random forest model.
In case of Deep Neural Network the results have been evaluated using:

- Mean Absolute Error: The magnitude of difference between the prediction of an observation and the true value of that observation. The Mean Absolute Error(MAE) is the average of all absolute errors.
- Mean Squared Error: The mean squared error of an estimator measures the average of the squares of the errors-the average squared difference between the estimated values and the actual value.
- R2 value: It is the proportion of the variance in the dependent variable that is predictable from the independent variable.

Dataset 1					
MODELS	**Precision**	**Recall**	**Accuracy**	**Cross-val score**	**F1 -score**
Decision Tree	0.95	0.91	0.9068	0.926	0.91
SVM	0.98	0.98	0.9772	0.979	0.98
Naive Bayes	0.99	0.99	0.9886	0.993	0.99
Random Forest	1.00	0.99	0.9954	0.993	1.00
Voting Classifier	-	-	0.9818	-	-
Gradient Boosting Classifier	0.98	0.98	0.992	-	0.98
XGBoost	-	-	0.9848	-	-

E. Accuracy Comparision

The following diagram summarizes the accuracies of different models used on dataset 1.
We observe that Random Forest has the highest accuracy as



compared to the other two models, and we can utilize it to predict a particular crop prediction for a set of input values.
We generated a voting classifier utilizing our decision tree model and random forest model in hopes of increasing the accuracy of the decision tree classifier, and we succeeded as the final accuracy for the voting classifier with the weights

designated as [1,1] was 98.18 percent, that was a major improvement from the previous accuracy of the decision tree. Finally, a Gradient Boost algorithm and the XGBoost classifier were implemented. The best learning rate was observed to be 0.25 as this gave a validation accuracy score of 99.2 percent. The XGBoost classifier presented an accuracy of 98.4 percent.

VIII. Conclusions and Limitations

We presented a machine learning approach for crop yield prediction, which helped us make predictions for the most suitable crop in a particular environment, considering the factors using various primary classifiers and more importantly, a neural network. The crop is predicted as an output for the given input parameter. This work has the potential to aid farmers who might have inadequate knowledge in predicting the particular crops for a sustainable outcome.

The sequential neural network was able to learn comparatively more nonlinear and complex relationships between environmental conditions, and interactions from previously existing data and make reasonably accurate predictions for new plants in other locations with known weather conditions. We observe that the performance of the model is heavily influenced by the quality of the predicted data of the weather and temperature parameters, so the importance of weather prediction techniques is understood.

References

[1] F. F. Haque, A. Abdelgawad, V. P. Yanambaka and K. Yelamarthi, "Crop Yield Prediction Using Deep Neural Network," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1-4, doi: 10.1109/WF-IoT48130.2020.9221298.

[2] Russello, H. (2018). Convolutional neural networks for crop yield prediction using satellite images. IBM Center for Advanced Studies.

[3] Gandhi, N., Petkar, O., Armstrong, L. J. (2016, July). Rice crop yield prediction using artificial neural networks. In 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 105-110). IEEE.

[4] D. Elavarasan, P.D. Vincent. Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access, 8 (2020), pp. 86886-8690

[5] S. Khaki, L. Wang. Crop yield prediction using deep neural networks. Front. Plant Sci., 10 (2019), p. 621

[6] P. Somvanshi, B.N. Mishra. Machine learning techniques in plant biology. PlantOmics: The Omics of Plant Science, Springer India, New Delhi (2015), pp. 731-754