Weather-Based Crop Yield Prediction

Dhyey Mehta[1], Juyon Lee[2]

[1]High School Student, Department of Computer Science, S H Mutha School
[2]Post Graduate Student, Department of Computer Science, Oxford University

ABSTRACT

Agriculture plays a critical role in India's economy, providing livelihoods to millions of people and contributing significantly to the nation's GDP (Desai). In fact, agriculture is able to support 45% of India's employed labor force (Damodaran). However, Indian farmers face numerous challenges, including unpredictable weather which leads to poor yields, often leading to financial instability, exacerbating poverty and rural distress ("Agriculture in India"). Predicting crop yields using machine learning models offers a promising solution to this problem. The model this paper proposes leverages meteorological data (temperature, rainfall, etc.) as well as farming practice data (use of pesticide, fertilizer etc.) to help farmers predict their yield. The model presented in this paper ultimately had a mean squared error of 4.16 and a correlation value of 0.761 while predicting yields.

Keywords: Crop Yield Prediction, Machine Learning Model, Environmental Science

1. INTRODUCTION

In this research paper we aimed to answer the following research question: "How can machine learning be leveraged to predict crop yields based on environmental factors and farming practices?"

This question is of paramount significance due to India's heavy dependence on agriculture. In addition to helping farmers' financial security with accurate yield prediction, it can also help them make informed decisions about which crops to grow, when to plant them, and how to manage them effectively. This, clearly, has the potential to significantly improve agricultural productivity, enhance farmers' livelihoods, and contribute to the overall economic development of the country.

2. BACKGROUND

In order to contextualize our work within the pre-existing field, as well as do necessary background research a plethora of past papers were reviewed. The following is a summary of three such papers.

A. The research paper written by Kalimuthu involved a machine learning model, specifically Naive Bayes algorithm, for crop yield prediction to assist beginner farmers. It involved the development of a mobile application for user-friendly access to the prediction system. It also incorporated collection of seed data and relevant parameters for training the prediction model (Naive Bayes). Scope limitations: Only 3 Parameters: Temperature, Humidity and Moisture were considered (Kalimuthu).

B. The research paper by Kale emphasizes the successful development of a crop yield prediction model using an Artificial Neural Network (ANN) with a focus on Maharashtra, a state within India. The research highlights the importance of technology in agriculture and the potential benefits of predictive modeling in crop selection and yield optimization. However, some limitations of the study, such as the reliance on historical data, the need for further validation and refinement of the model, and focuses on a single region of India (Kale).

C. The research paper by Kantanantha applied methods to corn yield and price forecasting in Hancock County, Illinois. Importantly, the paper suggests that the developed methods can be applied to other locations in the US and to different crop types. It demonstrates accuracy (MSE = 234.90 and an R-Squared = 0.8830) in predicting crop yield and price, which can ultimately contribute to better decision-making and planning for farmers. Scope Limitation: This was applied only on corn crops for Hancock County in Illinois (Kantanantha).

3. DATASET

We used an agricultural dataset focusing on crop production in India from 1997 to 2019. The crop information was sourced from Kaggle ("Agricultural Crop Yield in Indian States Dataset") and the weather data was sourced from NASA Power ("Data Access Viewer") and data from the Indian Meteorological Department ("IMD -Data Supply Portal"). The dataset includes following features : crop type, year, season, state, area, production, mean temperature, rainfall, relative humidity, fertilizer usage, pesticide usage, and yield.

The entire dataset has 19,517 entries and 13 columns. The dataset is typically split into training and testing datasets for model development and evaluation, respectively. In this project, it is 80% training and 20% testing.

Before training the model, several preprocessing steps have been carried out. One major such step was the use of One-Hot encodings. As many of the input variables focused on crop production, environmental factors and agricultural practices and were primarily categorical variables, we had to create one-hot encodings. This converts categorical input variables into a numerical input that can be directly given to the models. One-Hot encodings were used on Crop, Season and State.

In our dataset, Yield is defined as Production / Area, so we had to drop Production from our Training and Testing datasets.

Feature significance: Each feature in the dataset plays a significant role in predicting crop yields based on environmental factors.

Figure 1: Correlation Matrix

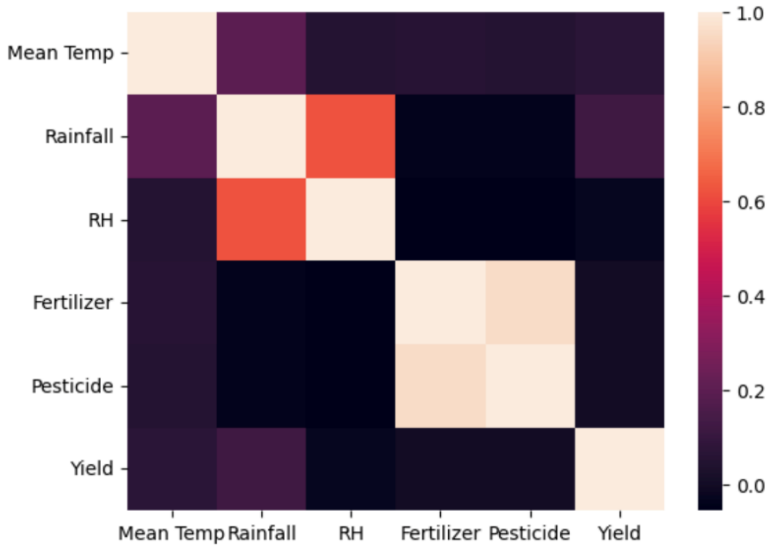


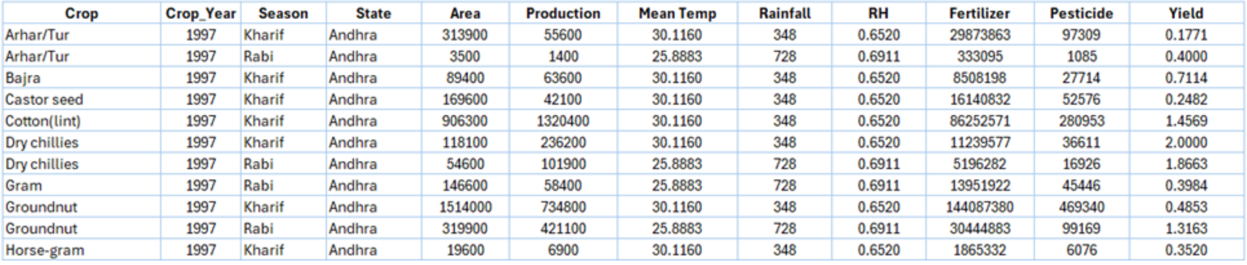
Figure 1 above is the correlation chart, which shows how the different features in the dataset are related to each other. As visible from the table, there is no strong correlation between two features apart from RH (Relative Humidity) and Rainfall. This is important since it abides with the assumptions of machine learning models that the input given to such models must be independent.

- Crop type: Different crops have varying yields and responses to environmental factors.
- Crop year and season: The timing of planting and harvesting can greatly affect crop yields, as different crops thrive under different seasonal conditions.
- State: Environmental conditions, soil types, and farming practices vary across states, influencing crop productivity.
- Environmental factors (mean temperature, rainfall, relative humidity): These variables directly impact crop growth and development, influencing yield potential.
- Agricultural inputs (fertilizer, pesticide usage): Proper - management of inputs is essential for optimizing crop yields while minimizing environmental impact and production costs.

Table-1: Features and its units

Feature	Unit
Area	Hectare
Production	Metric Tonne
Mean Temperature	Degree Celsius
Rainfall	Millimeter
Fertilizer	Kg
Pesticide	Kg
Yield	Metric Tonne/Hectare

Figure-2: First few entries of the dataset



Crop	Crop_Year	Season	State	Area	Production	Mean Temp	Rainfall	RH	Fertilizer	Pesticide	Yield
Arhar/Tur	1997	Kharif	Andhra	313900	55600	30.1160	348	0.6520	29873863	97309	0.1771
Arhar/Tur	1997	Rabi	Andhra	3500	1400	25.8883	728	0.6911	333095	1085	0.4000
Bajra	1997	Kharif	Andhra	89400	63600	30.1160	348	0.6520	8508198	27714	0.7114
Castor seed	1997	Kharif	Andhra	169600	42100	30.1160	348	0.6520	16140832	52576	0.2482
Cotton(lint)	1997	Kharif	Andhra	906300	1320400	30.1160	348	0.6520	86252571	280953	1.4569
Dry chillies	1997	Kharif	Andhra	118100	236200	30.1160	348	0.6520	11239577	36611	2.0000
Dry chillies	1997	Rabi	Andhra	54600	101900	25.8883	728	0.6911	5196282	16926	1.8663
Gram	1997	Rabi	Andhra	146600	58400	25.8883	728	0.6911	13951922	45446	0.3984
Groundnut	1997	Kharif	Andhra	1514000	734800	30.1160	348	0.6520	144087380	469340	0.4853
Groundnut	1997	Rabi	Andhra	319900	421100	25.8883	728	0.6911	30444883	99169	1.3163
Horse-gram	1997	Kharif	Andhra	19600	6900	30.1160	348	0.6520	1865332	6076	0.3520

4. BASELINE MODELS

To begin we ran several baseline regression models to contextualize our results against simpler machine learning models.

Table-3: Results of Baseline Models

Model Name	R-Square value	MSE
Linear Regression	0.841	1,35,774
Decision Tree Regressor	0.966	35,536
Random Forest Regressor	0.969	32,771
Gradient Boosting Regressor	0.891	11,14,867
Lasso CV	0.845	1,63,891

We later come to know that these strong R-Squared values are due to the large outlier values of the yield of the coconut crop.

5. OUTLIER ANALYSIS

We noticed several important patterns within the data and did some outlier analysis. We noticed that the yield for the coconut crop was much higher than other crops. To address the issue with coconut, we decided to create two different regression models: one specifically for the coconut crop and one for the remaining crops without coconut. We then decided to observe the dataset. It was seen that the coconut Crop had a yield substantially higher than the rest of the crops. Hence, we came to the conclusion that Coconut yields were a definite outlier. Coconut had an average yield of about 9,470, compared to the average yield of all other crops, which was approximately 5. On retrospection, this makes sense given the high weight of coconut and that many coconuts grow on a tree which occupies a small area.

There were several yield values in the dataset which were close to 0. This indicated that the crop in those particular conditions produces a very low yield, which isn't particularly ideal. The larger reason we did this is to try and improve our results. Our ultimate aim with the model is to have high accuracy. So, we first decided to run a Classification task, for which we introduced a new row called "Non-Zero Yield" in the dataset. All the yields that were less than 1 were assigned the value of 0 and the rest had the value of 1.

For the Classification Model, the X-Variables are Crop Year, Area, Mean Temperature, Rainfall, RH, Fertilizer, Pesticide along with all the One Hot encoded values of all the Crops, Seasons and States. The Y-Variable is the "Non-Zero Yield" of the Crop.

We ran three different classifiers on this yield / no yield predictor: decision tree classifier, random forest classifier and gradient boosting classifier. The results are summarized below:

Table-4: Results of Classification Models

Model Name	Accuracy Score
Decision Tree Classifier	0.914
Gradient Boosting Classifier	0.843
Random Forest Classifier	0.934

We proceeded with the Random Forest classifier as it yielded the best results. We then returned to predicting the yield which is the final goal of this project, on the non-zero yield values, effectively removing a great number of outliers.

6. FINAL MODEL

Now we return to the original problem of yield prediction. We ran several regression models on the problem: Linear, Decision Tree, Random Forest, Gradient Boosted, and Lasso.

Results of the Various Regression Models before removing any of the Crop Having yield less than 1, i.e. value of Non-Zero Variable = 0

Table-5: Results of Final model – before removing low yield entries

Model Name	R-Square value	MSE
Decision Tree	0.672	51
Random Forest	0.763	37
Gradient Booster	0.700	47

Results of the Various Regression Models after Removing any of the Crop Having yield less than 1, i.e. value of Non-Zero Variable = 0

Table-6: Results of Final model – after removing low yield entries

Model Name	R-Square value	MSE
Decision Tree	0.772	6.130
Random Forest	0.761	6.459
Gradient Booster	0.635	9.837

Here, we see that the R-Squared as well as the MSE values have dropped from the Baseline Models. R-squared dropped because we removed the outlier crop i.e. coconut. MSE dropped because we removed yield/no yield predictor.

Table-7: Comparison of Baseline Model & Final Model

Model Name	MSE – Baseline Model	MSE- Final Model
Decision Tree	35,536	6.130
Random Forest	32,771	6.459
Gradient Booster	11,14,867	9.837

The thing to note is that the R-Squared metric might not be particularly helpful in judging how our Model performs. Since we need to measure how far our Predicted Yield is from the Actual Crop Yield, we came to the conclusion that Mean Squared Error would be a better choice to see how our model performs.

We then decided to perform Hyperparameter Optimization on both Decision Tree Regressor and Random Forest Regressor.

7. RESULTS

Results of Decision Tree and Random Forest Regressor before and after performing the Hyperparameter Optimization are tabulated.

Table-7: Comparison of results before and after Hyperparameter Optimization

Model Name	**MSE – Before Hyperparameter Optimization**	**MSE – Before Hyperparameter Optimization**
Decision Tree	6.130	6.070
Random Forest	6.459	4.157

The reason we feel why the Random Forest Regressor performed better is because the Random Forest Regressor is an ensemble of many Decision Trees. Random forest algorithm avoids and prevents overfitting by using multiple trees. This gives more accurate and precise results.

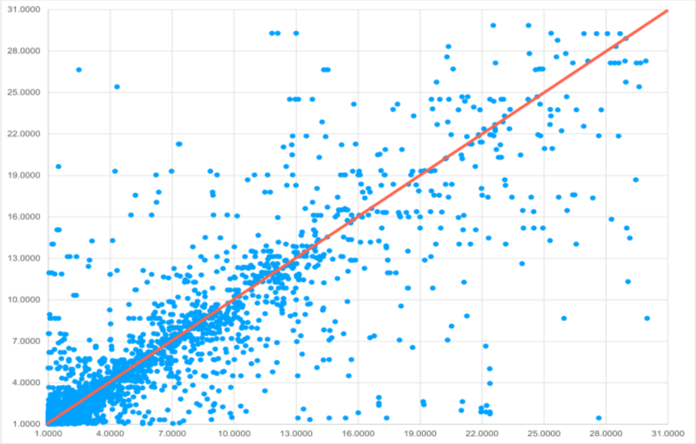
The Parameters for the Hyperparameter Optimization of the Random Forest Regressor are:

- Max_depth: None: This means that nodes are expanded until all leaves are pure or until all leaves contain less than Min_samples_split samples.
- Max_features: Auto: This means that the model will take into consideration all the features which make sense in every tree.
- Min_samples_leaf: 1: This indicates the minimum number of samples required to be at a leaf node.
- Min_samples_split: 2: This indicates the minimum number of samples required to split an internal node.
- N_estimators: 200: This indicates the number of trees to be used in the forest.

We then decided to graph Actual Yield v/s Predicted Yield of all the test values (4672 entries).

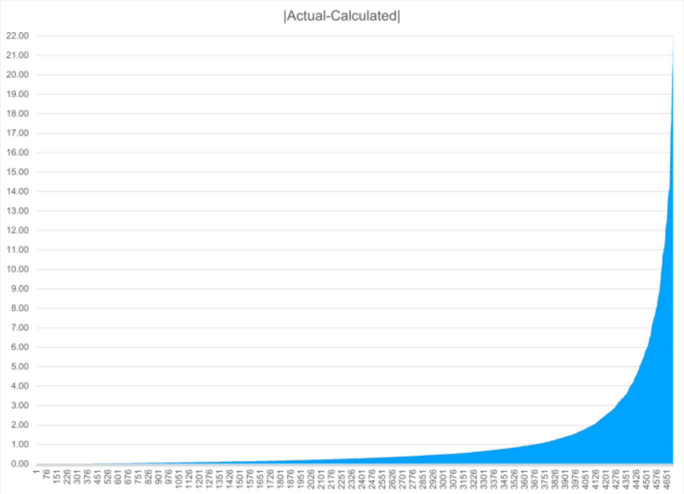
The graph below is of Predicted Yield (y-axis) v/s Actual Yield (x-axis). The Red Line has the equation of y = x. Most of the dots are around the red line, indicating that our model predicts most of the Yields accurately.

Figure-3: Scatter Plot – Predicted Yield vs Actual Yield



For the below graph, we first sorted the |Actual-Predicted| Yield in increasing order. As seen in the graph, the 3601st entry out of 4672 (77th Percentile) has |Actual-Predicted| Yield = 1 and 4201st entry (89.9th Percentile) has |Actual-Predicted| Yield = 2.

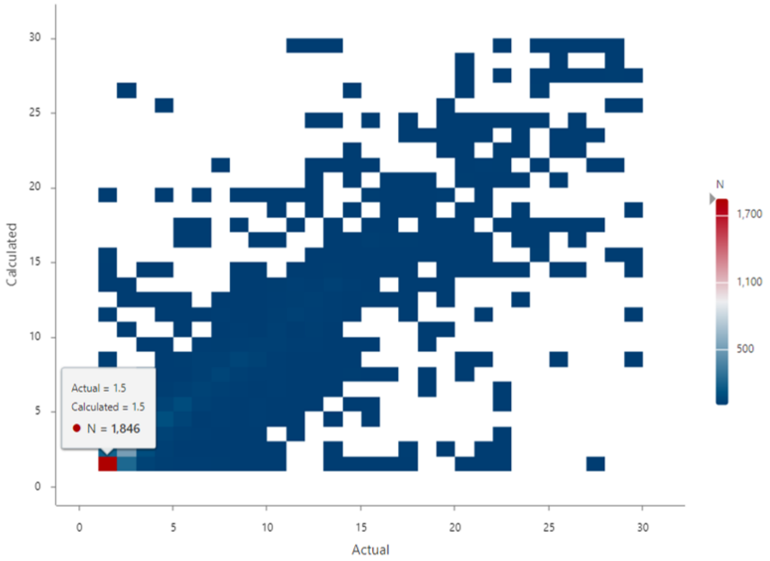
Figure-4: Plot of |Actual-Predicted| Yield



To understand this better, we used a binned scatterplot. It is a variation of scatterplots that can be useful when there are too many data points that are being plotted ("Binned Scatterplots"). It takes all data observations from the original scatterplot and places each one into exactly one group called a bin.

Once every observation is in a bin, each bin will get one point on a scatter plot, reducing the amount of clutter on our plot. For example, in this graph, we see that there are a total of 1846 points for which actual and predicted yield = 1.50. In total, there are a total of 2961 points (out of total 4672 points i.e. 63.3% of points) for which predicted yield = actual yield.

Figure-5: Binned Scatter Plot – Predicted Yield vs Actual Yield



Hence, with the known MSE of 4.16 along with these graphs, we can conclude that this model is able to predict the yield of the crops after analyzing weather conditions with high accuracy.

8. CONCLUSION

In conclusion, for our dataset, we first carried out One-Hot encoding so that categorical variables (Crop, Season, and State) can be converted into a numerical input for our models. We then ran baseline regression models. The best performing model was the Random Forest Regressor (R-Squared Score of 0.969 and an MSE of 32,771). We came to the conclusion that such a high R-Squared score was due to the coconut crop, which had very high yields (average of 9470) compared to other crops in the dataset (average of 5). We also created a new row called "Non-Zero Yield". All the yields that were less than 1 were assigned the value of 0 and the rest had the value of 1. We ran classification models for this new row. The best performing model was Random Forest Classifier (Accuracy Score of 0.934). We then removed all the crops with yields less than 1 in the dataset and again used regression models for yield prediction. The best performing models in terms of MSE were Decision Tree Regressor (MSE = 6.13) and Random Forest Regressor (MSE = 6.459). We then used Hyperparameter Optimization on both these models. Finally, the best performing model was the Random Forest Regressor with R-Squared Score = 0.761 and MSE = 4.157. In addition to this, we believe that if we sort the crops based on their utility eg. cash and food crops (which will be again divided into fruits and vegetables), we will be able to draw some more conclusions, which will ultimately help us to reduce the mean squared error even more. Additionally, we also want to build a model which will be able to predict the prices of these crops accurately. By combining these two models, we will be able to provide farmers with a complete model. Afterwards, we wish to create a website or an application which will be easy to use for the farmers. Using these models, we believe they will be able to make informed decisions which will help them not only to maximize their yields, but also their profitability.

9. REFERENCES

1. "Agricultural Crop Yield in Indian States Dataset." Kaggle, https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset

2. “Agriculture in India.” Wikipedia, https://en.wikipedia.org/wiki/Agriculture_in_India
3. “Binned Scatterplots” LOST, https://lost-stats.github.io/Presentation/Figures/binscatter.html.
4. Damodaran, Harish. “What India's labor force and national income data tell us about jobs shifting from agriculture.” The Indian Express, https://indianexpress.com/article/explained/explained-economics/explained-economics-agriculture-and-employment-8480945/.
5. “Data Access Viewer.” NASA POWER https://power.larc.nasa.gov/data-access-viewer/
6. Desai, Sonalde. “Economy of India.” Wikipedia, https://en.wikipedia.org/wiki/Economy_of_India
7. “IMD - Data Supply Portal.” IMD - Data Supply Portal, https://dsp.imdpune.gov.in/.
8. Kale, Shivani S. “A Machine Learning Approach to Predict Crop Yield and Success Rate.” 2019, https://ieeexplore.ieee.org/abstract/do cument/9105741.
9. Kalimuthu, M. “Crop Prediction using Machine Learning.” IEEE, 2020, https://ieeexplore.ieee.org/abstract/docum ent/9214190.
10. Kantanantha, Nantachai. “Yield and Price Forecasting for Stochastic Crop Decision Planning.” Journal of Agricultural, Biological, and Environmental Statistics, vol. 15, 2010, p. 19, shttps://www.jstor.org/stable/20778474