Yield Prediction using Soil & Climatic Data using AI

Dr. Nupur Giri[1], Aneesh Kulkarni[2], Tejas Thakur[3], Vedant Wakalkar[4], Vignesh Verma[5]

[1]Dr. Nupur Giri, Dept. of Computer Engineering, V.E.S Institute of Technology, Maharashtra, India
[2]Aneesh Kulkarni, Dept. of Computer Engineering, V.E.S Institute of Technology, Maharashtra, India
[3]Tejas Thakur, Dept. of Computer Engineering, V.E.S Institute of Technology, Maharashtra, India
[4]Vedant Wakalkar, Dept. of Computer Engineering, V.E.S Institute of Technology, Maharashtra, India
[5]Vignesh Verma, Dept. of Computer Engineering, V.E.S Institute of Technology, Maharashtra, India

---

Abstract - *Food production in India is largely dependent on cereal crops including rice, wheat and various pulses. Predicting the crop yield well ahead of its harvest would help the policymakers and farmers for taking appropriate measures in terms of agronomy, crop choice and for proper agricultural planning. Such predictions will also help the associated industries for planning the logistics of their business. The project aims at developing a machine learning model to make such predictions. The model is trained using the dataset which contains rainfall data of the past decade which includes average rainfall (in mm), relative humidity, temperature, and soil data as Features and the crop yield as labels. The data is time-series data. Using appropriate machine learning techniques, the model learns the correlation between the yield and features like Soil Type, Rainfall. The data collected is real-time data of the past several years, provided by online government records for the state of Maharashtra. The predictions can be useful for industries in the agricultural sector and farmers for proper choice of crops etc.*

Keywords: crop, rainfall, soil, yield, data, farmers, predictions.

1. INTRODUCTION

Agriculture is one of the most important industrial sectors in India as well as Maharashtra and the country's economy is highly dependent on it for rural sustainability. It accounts for a large share in GDP (gross domestic product) (16%), and an even larger share in employment (49%). About 60 % of the land in the country is used for agriculture in order to suffice the needs of 1.3 billion people. Due to some factors like climate changes, unpredicted rainfall etc, the level of agriculture has decreased. However, as the conditions change day by day very rapidly, farmers are forced to cultivate more and more crops. Being this as the current situation, many of them don't have enough knowledge about the new crops and are not completely aware of the benefits they get while farming them.

In this project, first, the impact of rainfall and the soil type on the crop yield is analysed by using the data from past several years which consists of climatic and soil data and the yield of crops (kg/ha) of different talukas of Maharashtra.

To know the level of production we perform analytics on the agriculture data and create a machine learning model to predict the yield of crops like Rice, Wheat, Maize, Bajra etc. The proposed system aims at recommending crops which are predicted to have better yield taking into account the current climatic conditions as well as soil conditions of the location of the farm.

2. EXISTING SYSTEM

Currently, there exists no automated system for this purpose which takes into consideration both the climatic and soil factors for prediction of yield. This leaves space for errors which are inevitable. Factors such as uncertain weather and soil are not considered while making choices, which could cost heavily. The proposed system focuses on automating the prediction of crop yields by taking into consideration the climatic conditions and soil data.

3. METHODOLOGY

In order to predict the crop yield, we need to build a model which will take into consideration various factors like soil type, soil quality, climatic conditions, etc. and predict the yield of a particular crop in a particular region.
Following steps are used:

Step 1: Gather Data
Gathering data is the most important step in solving any supervised machine learning problem. We will be gathering the previous year's data of crop yield, climatic conditions and soil types of Maharashtra.

Step 2.1: Exploring Data
Building and training a model is only one part of the workflow. Understanding the key characteristics of the data beforehand, we can build a better model. This could simply mean obtaining a higher accuracy. It could also mean requiring less data for training, or fewer computational resources. Various visualization techniques can be used to detect various correlations and patterns in the training data, which can further help in creating an accurate prediction system.

Step 2.2: Choose a Model
At this point, we have assembled our dataset and gained insights into the key characteristics of our data and choose an appropriate model for accurate classification.

Step 3: Preparing Data
Before our data can be fed to a model, it needs to be transformed into a format the model can understand. For example, conversion of categorical data (nominal or ordinal) /text data into the appropriate numerical format (Like One Hot Encoding), splitting of the dataset in training and validation set etc.

Step 4: Building, Training, and Evaluating the Model
We will work towards building, training and evaluating our model. And then to create and compile our model and train it.

Step 5: Tune Hyperparameters
We have to choose a number of hyperparameters in defining and training the model. We relied on intuition, examples and best practice recommendations. Our first choice of hyperparameter values, however, may not yield the best results. It only gives us a good starting point for training. Every problem is different and tuning these hyperparameters (ex. No. of epochs, batches, hidden layers, loss function, optimizer) will help refine our model to better represent the particularities of the problem at hand.

Since we have the following models:

a) ***ARIMA*** (hyperparameters p, d, q: Seasonality, Trend and noise)
b) ***Random Forest Regression*** (Number of estimators, max_depth)
We've tuned the hyperparameters using ***Random Search Technique***.

Step 6: Deploying Model
We will train, tune and deploy machine learning models.

4. PROPOSED SYSTEM

The yield prediction will rely majorly on attributes (such as location, soil type, weather, etc) required for maximizing profit related to specific crops. The regression model will take care of these attributes and help predict the crop yield for any crop for the next few years. We have experimented with a few machine learning models like XGBoost Regression model, LGBoost regression model, Random Forest Regression model. The model with the most suitable output was the XGBoost regression model. Thus this model is used for prediction of crop yield.

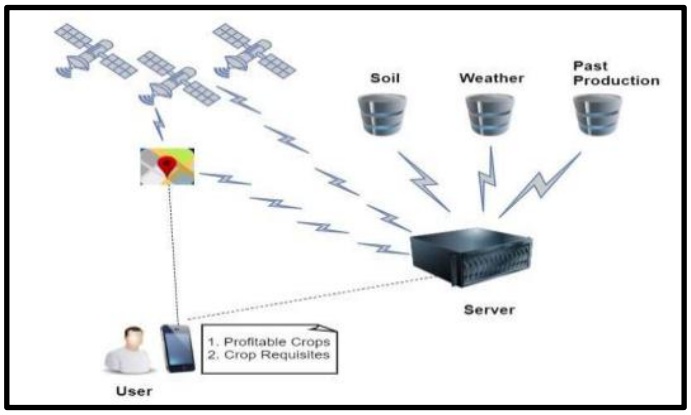

Fig - 1: System Architecture

The system consists of datasets consisting of soil, weather and past production data which is stored in the server. Once a user logs in he/she just shares his/her location along with some basic farm details (area, soil type) the details are then forwarded to the server, where the model is resident, the model fetches the data and outputs its prediction.

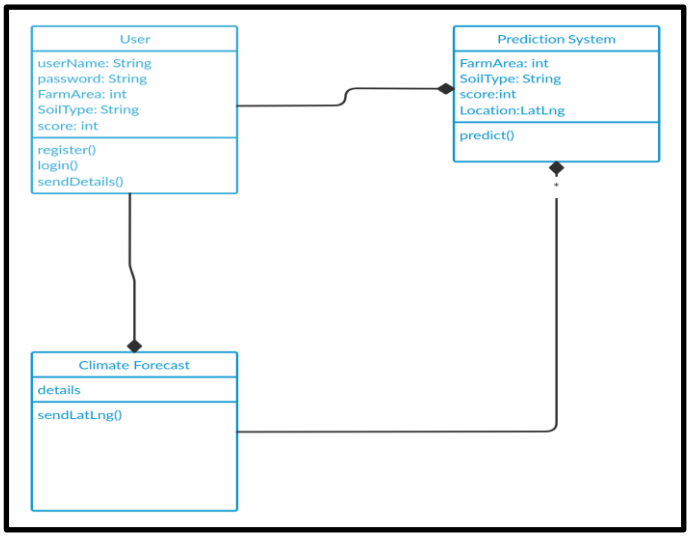

Fig - 2: Modular Diagram

There are 3 modules of functionality:
1. User Module (Basic User Login/ Registration)
2. Climate Forecast Based on User Details
3. Prediction System

The user submits data to the prediction system, the climate forecasting services send the climate forecast based on the user location to the prediction system as well.

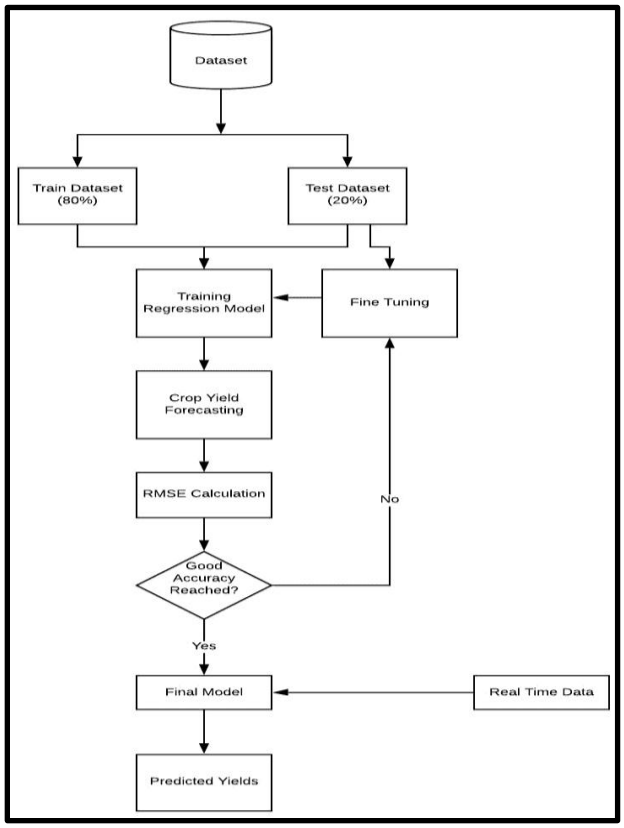

Fig - 3: Flowchart

The flowchart consists of the basic flow of how the model will be trained, what practices will be followed to achieve good accuracy and finally feeding real-time data on the trained model.

4. RESULT AND ANALYSIS

4.1. ACCURACY

```
[ ] import numpy as np
    from sklearn.metrics import mean_squared_error
    from math import sqrt
    rmse = np.sqrt(mean_squared_error(y_test, preds))
    print("RMSE: %f" % (rmse))

    RMSE: 199.822264

[ ] correlation_matrix = np.corrcoef(y_test, preds)
    correlation_xy = correlation_matrix[0,1]
    r_squared = correlation_xy**2
    print(r_squared)

    0.8929197960156591
```

Fig - 4: Loss Metrics on test set

The figure describes the root mean square error and R square error for the prediction model.

4.2. VISUALIZATION

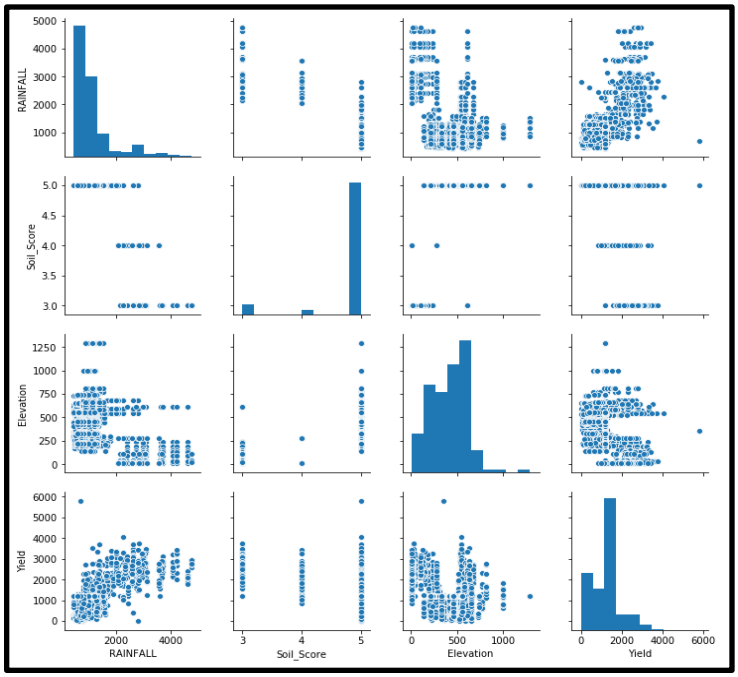

Fig - 5: Scatterplot

The above scatterplot demonstrates the relation between parameters like rainfall, soil score and elevation against the crop yield for a particular crop. It can be deduced that Crop Yield varies linearly with the rainfall of the area. Increase in soil score decreases the rice yield.

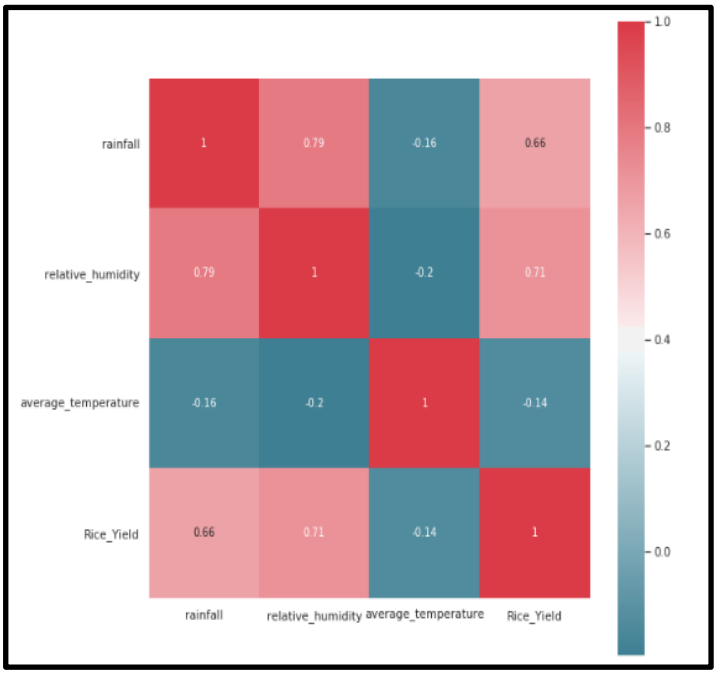

Fig - 6: Correlation (Crop Yield and climate variables)

The graphical illustration demonstrates the correlation between the climatic conditions and the crop yield of Rice. It can be discerned that the correlation between crop yield and relative humidity is the highest. Thus it can be deduced that Crop Yield of the rice crop is largely governed by the values of relative humidity.

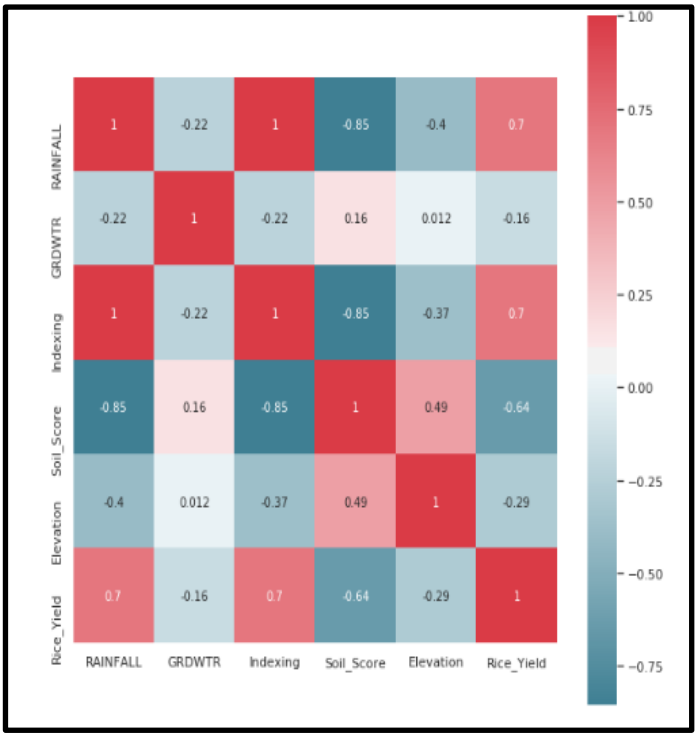

Fig - 7: Correlation Between crop yield and soil variables

The graphical illustration demonstrates the correlation between the soil variables and the crop yield of Rice. It can be discerned that the correlation between crop yield and soil indexing is the highest. Thus it can be deduced that Crop Yield of the rice crop is largely governed by the values of soil indexing.

```
[ ] import numpy as np
    from sklearn.metrics import mean_squared_error
    from math import sqrt
    rmse = np.sqrt(mean_squared_error(y_test, preds))
    print("RMSE: %f" % (rmse))

    RMSE: 199.822264

[ ] correlation_matrix = np.corrcoef(y_test, preds)
    correlation_xy = correlation_matrix[0,1]
    r_squared = correlation_xy**2
    print(r_squared)

    0.8929197960156591
```

Fig - 8: RMSE and R Square for prediction and forecasting.

```
1 get_best_crop(district_name = 'Amravati', year = 2025)

{'Jowar_Yield': 624.906542137872, 'Rice_Yield': 1265.5427730672907}
```

Fig - 9: Output yield based on district_name and year.

Rice Yield for Yavatmal district for next five years is as follows:

	Year	Yield
697	2021	1585.890381
698	2022	1172.127197
699	2023	1323.749634
700	2024	1552.413818
701	2025	1372.265869

Fig - 9: Rice Yield for Yavatmal district

5. FUTURE SCOPE

- Taking inputs like the amount of moisture required in soil and the required soil nutrients, crop recommendations are carried out.
- Consideration of the economic aspect to recommend the farmer's most profitable crop.
- Recommending the farmer best market for selling his cultivated crop.

6. REFERENCES

[1] "Crop Prediction System using Machine Learning" by Prof. D. S. Zingade, Omkar Buchade, Nilesh Mehta, Shubham Ghodekar, Chandan Mehta

[2] "Efficient Crop Yield prediction using Machine Learning Algorithms." by Arun Kumar, Naveen Kumar, Vishal Vats,

[3] "Crop Selection Method to maximize crop yield rate using machine learning technique" by Rakesh Kumar, M.P. Singh, Prabhat Kumar, J.P. Singh.

[4] "Machine learning approach for forecasting crop yield based on climatic parameters" by S. Veenadhari, Bharat Misra, C.D. Singh.

[5] "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach" by Snehal S. Dahikar, Sandeep V. Rod.