Capstone Project: 01

Crop Yield Prediction Using Machine Learning

IOT Academy, IIT Guwahati

	Prepared By	**Reviewed by**
Name	Ms. Jayashri Pacharane	Mr. Arihant Jain
Designation	Trainee	Manager
Signature (with Date)		

INDEX

1. Title of The Project: Crop Yield Prediction

2. Objective:

Predict the Crop yield (Production per unit area). The objective of crop yield prediction is to accurately estimate the amount of crop produced in a specific area, using a variety of input factors such as weather conditions, crop variety, and farming practices. These predictions are crucial for improving agricultural planning, optimizing resource usage, ensuring food security, and supporting policy decisions. The goal is to predict the future crop yield with high accuracy, allowing farmers, agricultural organizations, and governments to make informed decisions.

3. Brief on The Project:

To develop a machine learning model that can predict the yield of various crops based on historical data, climate conditions, soil health, and other environmental factors. This tool will help farmers, agricultural experts, and policymakers make data-driven decisions to improve crop production, manage resources efficiently, and mitigate risks like drought or pest infestations.
To achieve these goals, a machine learning model is required to predict the yield various crops. Given that there is continuous value to derive. we will utilize a regression model for prediction.

4. Problem Statement:

This dataset encompasses agricultural data for multiple crops cultivated across various states in India from the year **1997 till 2020**. The dataset provides crucial features related to crop yield prediction, including crop types, crop years, cropping seasons, states, areas under cultivation, production quantities, annual rainfall, fertilizer usage, pesticide usage, and calculated yields.

5. Columns Description:

5.1 Crop: The name of the crop cultivated.
5.2 Crop Year: The year in which the crop was grown.
5.3 Season: The specific cropping season (e.g., Kharif, Rabi, Whole Year).
5.4 State: The Indian state where the crop was cultivated.
5.5 Area: The total land area (in hectares) under cultivation for the specific crop**.**
5.6 Production: The quantity of crop production (in metric tons).
5.7 Annual Rainfall: The annual rainfall received in the crop-growing region (in mm).
5.8 Fertilizer: The total amount of fertilizer used for the crop (in kilograms).
5.9 Pesticide: The total amount of pesticide used for the crop (in kilograms).
6.0 Yield: The calculated crop yield (production per unit area).

6. Use Cases:

This comprehensive dataset is valuable for agricultural analysts, researchers, and data scientists interested in **crop yield prediction and agricultural analysis**. It offers insights into the relationship between various **agronomic factors** (e.g., rainfall, fertilizer, pesticide usage) and **crop productivity** across different states and crop types. Researchers can utilize this data to develop robust **machine learning models** for crop yield prediction and identify trends in agricultural production.

7. Project Deliverables:

In this Project, we will focus on prediction of the yield of various crops using Supervised Machine Learning Models. This comprehensive initiative will guide us through each phase of data analysis. we have collected data from website and classified it as a regression problem.

The goal of this project is to predict the crop yield based on various input features using a regression model. The project involves data pre-processing, model development, evaluation, and deployment of a machine learning model that can predict the amount of crop yield given environmental, agricultural, and climatic data.

7.1 Data Collection and Pre-Processing:

The first phase of the project focuses on gathering relevant data from multiple sources, including historical crop yield data, weather patterns (rainfall), and crop varieties. These datasets will be obtained from agricultural departments and weather services. The collected data will be cleaned to address missing values, outliers, and inconsistencies. Data imputation techniques will be used for missing values, and outlier detection methods will be applied to ensure the integrity of the data. To prepare the dataset for model training, the data will be split into a training set (80%) and a testing set (20%)

7.2 Model Development:

The second stage of the project involves the selection and development of regression models. we used Multiple regression techniques will be tested to identify the best fit for the given data. Models such as **Linear Regression**, and **Regularization Techniques like Ridge, Lasso and Elastic net** which is straightforward and interpretable, will be considered for its simplicity and efficiency. More complex models like **Random Forest Regression**, **Decision Tree, Ada Boost, KNN** And **Gradient Boosting Machines (e.g., XG Boost).**

7.3 **Model Evaluation:**

In this phase, the trained models will be evaluated on a testing dataset to assess their performance. Evaluation metrics will include **Mean Squared Error (MSE)** and **R-squared (R^2)** to measure prediction accuracy and the proportion of variance explained by the model. A lower MSE or RMSE indicates a better fit, while a higher R^2 suggests that the model is capturing more of the variability in the

data. Visualizations such as residual plots and error distributions will be used to further assess model performance and identify potential issues, such as underfitting or overfitting.

8. Resources:

8.1 Data Set Sources:

The Source of the data set is Kaggle.

Dataset link: https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset/data

8.2 Software:

Anaconda with Jupyter Notebook Software is being used to build the Machine Learning Algorithm.

9. Individual Details:

9.1 Name: Jayashri Pacharane
9.2 E-Mail ID: pacharanejayashri@gmail.com
9.3 Contact Details: 8411921362

10. Milestones:

10.1 Define a Problem:

This dataset encompasses agricultural data for multiple crops cultivated across various states in India from the year **1997 till 2020**. The dataset provides crucial features related to crop yield prediction, including crop types, crop years, cropping seasons, states, areas under cultivation, production quantities, annual rainfall, fertilizer usage, pesticide usage, and calculated yields.

10.2 Get the Data:

The Data Set is a CSV file as Attached

crop_yield.csv

OR

Dataset link: https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset/data

10.3 Explore and Pre-Process Data:

This Data Set Contains 19689 rows and 10 columns.

Choosing the Python Platform.
I have Selected Anaconda with Jupyter Notebook as a platform for creating the Supervised Machine Learning Model.

10.4 Create Features:

Following are the Feature for the given Data Set:

10.4.1 **Crop**: The name of the crop cultivated.

10.4.2 **Crop Year**: The year in which the crop was grown.

10.4.3 **Season**: The specific cropping season (e.g., Kharif, Rabi, Whole Year).

10.4.4 **State**: The Indian state where the crop was cultivated.

10.4.5 **Area**: The total land area (in hectares) under cultivation for the specific crop.

10.4.6 **Production**: The quantity of crop production (in metric tons).

10.4.7 **Annual Rainfall**: The annual rainfall received in the crop-growing region (in mm).

10.4.8 **Fertilizer**: The total amount of fertilizer used for the crop (in kilograms).

10.4.9 **Pesticide**: The total amount of pesticide used for the crop (in kilograms).

10.4.10 **Yield**: The calculated crop yield (production per unit area).

11. EDA:

While Performing the EDA, the following are our Observations:

11.1 Understanding the Dataset:

Dataset Features: The dataset may include features such as:

11.1.1 **Numerical features**: Crop Year, Area, Production, Annual rainfall, Fertilizer, Pesticide.
11.1.2 **Categorical features**: Crop, Season, State etc.
11.1.3 **Time-related features**: Year, or season data
11.1.4 **Target Variable**: The target (dependent) variable in this regression problem is typically crop yield

11.2 Data Preprocessing:

11.2.1 **Handling Missing Values**: - Check for missing values in the dataset using .isnull() and handle them appropriately.

11.2.2 Outliers in features like rainfall, production, fertilizer, pesticides, or yield can distort predictions. Use boxplots handle Outliers (either by removal or transformation).

11.2.3 From the Describe Function: Get the mean, median, standard deviation, min, and max values for numerical features (like Annual rainfall, Production, Pesticide, Fertilizer and crop yield) to understand their distributions.

11.2.4 **Correlation Analysis**: Area Shows a Strong Correlation with Fertilizer (0.97) and Pesticide (0.97).

11.2.5 **Correlation Heatmap**: A heatmap to show the correlation between multiple variables at once, particularly the relationship between features and the target variable. Area Shows a Strong Correlation with Fertilizer (0.97) and Pesticide (0.97).

11.2.6 Crop Year: It has almost No Correlation with any other variable.

11.2.7 Production and Yield: Production has a Moderate Positive Correlation with Yield (0.57). Yield shows no significant correlation with other variable except Production.

11.2.8 Fertilizer and Pesticide: A Very Strong Correlation (0.95), implying these two variables tend to move together, possibly due to integrated farming practices.

11.2.9 **Encoding Categorical Variables**: For categorical variables (like crop type or region), we will use **one-hot encoding** to convert them into numerical form suitable for regression models.

11.2.10 We have use the Square Root Transformation (SQRT) for reducing the Skewness. Then we see the skewness is decreases.

11.2.11 We have Visualizing Trends.

12. Create Model:

12.1 Before creating a model, we split the training and testing data into a ratio of 80:20.

12.2 The Target Variable is Yield.

12.3 Used various Regressor Algorithms like

12.3.1 Linear Regression.
12.3.2 Regularization Techniques Like: Ridge, Lasso, Elastic net.
12.3.3 Decision Tree Regressor.
12.3.4 Random Forest Regressor.
12.3.5 Ada Boost Regressor.
12.3.6 Gradient Boosting Regressor.
12.3.7 XGB Regressor.
12.3.8 K Neighbors Regressor.
12.3.9 Support Vector Regressor.

13. Model Evaluation:

13.1. Performance Metrices are evaluated for each model to determine which model will give better prediction.

13.2 Table of Performance Metrices:

Sr. No	Model Name	Training R-Squared	Testing R-Squared
1	Linear Regression	84%	81%
2	Ridge	84%	81%
3	Lasso	84%	81%
4	Elastic Net	49%	57%
5	Decision Tree	100%	98%
6	**Random Forest**	**98%**	**98%**
7	Ada Boost	95%	92%
8	Gradient Boosting	99%	98%
9	KNN	98%	97%
10	XGBOOST	99%	98%

13.3 Comparing the All Algorithms using “For Loop”.

13.4 In the Performance Metrices, we see that the Random Forest is the best Performance in the both training (98%) and testing in (98%).

13.5 Checked the accuracy score of all the models.

14 Report Writing:

Abstract:

Crop yield prediction plays a vital role in agriculture, enabling farmers, policymakers, and researchers to make informed decisions about production and distribution. Accurate forecasting of crop yields is crucial for ensuring food security and optimizing agricultural practices. This report investigates the application of regression models to predict crop yields based on various factors, such as crop varieties, annual rainfall, production, pesticide, fertilizer, area and the farming practices. Different regression techniques, including linear regression, decision tree regression, and support vector regression, are explored to assess their effectiveness in predicting crop yield and their potential applications in precision agriculture.

14.1 Introduction:

Crop yield prediction is an essential task in agriculture, aimed at estimating the amount of crops that can be harvested from a specific piece of land. Traditionally, crop yield prediction was done through empirical methods, such as farmer experience or field surveys, which were often time-consuming and inaccurate. However, with the rise of data-driven approaches and machine learning, more sophisticated techniques like regression models have become widely used. Regression analysis, a statistical method, helps to model the relationship between a dependent variable (crop yield) and one or more independent variables (e.g., crop year, annual rainfall, area, production, fertilizer, pesticide etc.).

The primary objective of this report is to explore the potential of regression models in predicting crop yields and to determine the accuracy and reliability of these models.

14.2 Problem Statement:

This dataset encompasses agricultural data for multiple crops cultivated across various states in India from the year **1997 till 2020**. The dataset provides crucial features related to crop yield prediction, including crop types, crop years, cropping seasons, states, areas under cultivation, production quantities, annual rainfall, fertilizer usage, pesticide usage, and calculated yields.

This report seeks to answer the following questions:

- 14.2.1 Which regression models can be effectively applied to crop yield prediction?
- 14.2.2 How accurate are these models in estimating crop yield based on available data?
- 14.2.3 What are the key factors influencing crop yield that can be identified through regression analysis?

14.3 Methodology:

To develop an accurate crop yield prediction model, several regression techniques were explored, each of which applies a different approach to model the relationship between inputs and outputs.

14.4 Data Collection:

The data used in this study consists of several features related to crop yield and environmental conditions. The dataset includes historical crop yield data, weather patterns (annual rainfall), Production, area, and farming practices (irrigation, fertilization, pesticide). Data was collected from agricultural research institutions, and on Kaggle link.

14.5 Regression Models:

The following regression models were considered for crop yield prediction:

14.5.1 Linear Regression:

A simple yet powerful method that models the relationship between a dependent variable and one or more independent variables using a straight line. It assumes a linear relationship between the input and output variables.

14.5.2 Decision Tree Regression:

A non-linear method that splits the dataset into smaller subsets based on the values of the independent variables. The final prediction is made by averaging the values in each subset.

14.5.3 Support Vector Regression (SVR):

A machine learning model that uses the concept of margins to predict continuous outputs. SVR can handle both linear and non-linear relationships and is robust to overfitting.

14.5.4 Random Forest Regression:

An ensemble method that builds multiple decision trees and averages their predictions. Random forests are often used for their ability to model complex relationships without overfitting.

14.5.5 KNN (K Nearest Neighbors):

KNN is a simple yet effective method for both classification and regression tasks. It has broad applications across various domains such as agriculture (for crop yield prediction), healthcare, finance, and e-commerce. However, its performance can be affected by the choice of **K**, distance metric, and the presence of irrelevant features.

14.5.6 Ada Boost Regression:

In the context of **crop yield prediction**, AdaBoost regression can significantly improve the accuracy of predictions by focusing on the instances where the model makes errors and adjusting accordingly.

14.5.7 Gradient Boosting:

Gradient boosting is an effective and widely-used method for crop yield prediction, thanks to its ability to model complex relationships, handle non-linearity, and focus on the most difficult-to-predict cases. By improving the accuracy of predictions and providing valuable insights into key influencing factors, gradient boosting can help farmers, researchers, and policymakers make more informed decisions. Whether it's predicting the yield of wheat, maize, or any other crop, gradient boosting offers a powerful tool for managing agricultural practices and ensuring food security.

14.5.8 XG Boost:

XG Boost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting that is known for its high performance, efficiency, and scalability. It is a machine learning algorithm specifically designed to speed up the training process, reduce overfitting, and handle large datasets more effectively than traditional gradient boosting methods. XG Boost is widely used in regression, classification, and ranking tasks, making it an excellent choice for complex prediction problems like **crop yield prediction**.

14.6 Model Evaluation:

To evaluate the performance of the regression models, the following metrics were used:

14.6.1 Mean Absolute Error (MAE): Measures the average magnitude of the errors in a set of predictions, without considering their direction.

14.6.2 Root Mean Squared Error (RMSE): Measures the square root of the average squared differences between predicted and actual values.

14.6.3 R-squared (R^2): Represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

14.6.4 Mean Squared Error (MSE): MSE is particularly useful for understanding the **magnitude of prediction errors**. Since MSE squares the errors, it penalizes large errors more heavily than smaller ones.

The dataset was divided into training and testing sets, with a 80-20 split. The models were trained on the training set and evaluated using the testing set.

14.7 Table of Performance of All Algorithms:

Sr. No	Model Name	Training R-Squared	Testing R-Squared
1	Linear Regression	84%	81%
2	Ridge	84%	81%
3	Lasso	84%	81%
4	Elastic Net	49%	57%
5	Decision Tree	100%	98%
6	**Random Forest**	**98%**	**98%**
7	Ada Boost	95%	92%
8	Gradient Boosting	99%	98%
9	KNN	98%	97%
10	XG Boost	99%	98%

The Best Algorithm is random forest, we can see RF performance Balanced both Training and Testing data. We have moving forward with random forest.

15. Conclusion:

The results of this study indicate that regression models, particularly Random Forest Regression, are highly effective in predicting crop yields. These models can assist in improving agricultural productivity by providing farmers and researchers with reliable forecasts based on environmental data. Future work could explore the integration of more advanced machine learning models, real-time weather data, and the impact of climate change on crop yield prediction.

15.1 BEST MODEL: RANDOM FOREST (Training- 98% & Testing- 98%)

Reason: It achieves the highest R Square on testing data (98.81%) with a low RMSE (90.55).

Balanced performance on training and testing data suggests the model generalizes well.

Runner up: XG Boost

Slightly lower R Square and higher RMSE compared to Random Forest but still performs exceptionally well.

Gradient Boosting And KNN: Good Alternative with Slightly worse RMSE Than Random Forest.

END OF THE DOCUMENT