**Crop Yield Prediction System**

MINOR PROJECT REPORT

By

**ROOPA SRI(RA2311027010096)**

**SAI TULASI(RA2311027010110)**

**PRASANNA LAKSHMI(RA2311027010109)**

**DURGA DEVA SRI(RA23111027010111)**

Under the guidance of   
Dr. K. Dhanasekaran

*In partial fulfillment for the Course*

of

**21CSS202T – FUNDAMENTALS OF DATA SCIENCE**

In

Department of Data Science and Business Systems(DSBS)

School of Computing



**FACULTY OF ENGINEERING AND TECHNOLOGY**

**SCHOOL OF COMPUTING**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**KATTANKULATHUR**

**NOVEMBER 2024**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**(Under Section 3 of UGC Act, 1956)**

**BONAFIDE CERTIFICATE**

Certified that this minor project report for the course **21CSS202T** **FUNDAMENTALS OF DATA SCIENCE** entitled in "**Crop Yield Prediction System**" is the bonafide work of

**Roopasri (RA2311027010096****), Sai tulasi (RA2311027010110****), Prasanna lakshmi (RA2311027010109)** and **Durga deva sri(RA23111027010111)** who carried out the work under my supervision.

**SIGNATURE SIGNATURE**

**Dr.Dhanasekaran.K Dr.Kavitha V**

**Assistant Professor** **Professor &Head**

**Department of Data Science and Department of Data Science and**

**Business Systems Business Systems**

SRM Institute of Science and SRM Institute of Science and

Technology Technology

Kattankulathur Kattankulathur

# ABSTRACT

Crop yield prediction is a critical task in agricultural planning, influenced by multiple factors such as weather conditions, soil properties, and crop management practices. This project aims to develop a predictive model for crop yield using machine learning techniques, leveraging historical data that includes climate variables, soil characteristics, and previous yield records.

We began by understanding and preprocessing the data, followed by exploratory data analysis (EDA) to gain insights and identify significant predictors. Various models, including Linear Regression, Random Forest, and Support Vector Machines, were evaluated based on their performance metrics such as Mean Absolute Error (MAE) and R-squared.

Our findings highlight the importance of features like rainfall, temperature, and soil quality in influencing crop yield. The Random Forest model demonstrated superior performance, achieving a balance between accuracy and interpretability. However, challenges such as data quality and handling missing values were encountered, impacting the model’s effectiveness.

The results of this study provide valuable insights for improving agricultural productivity and can aid in strategic decision-making. Future work will focus on incorporating real-time data and exploring more complex models to enhance prediction accuracy.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **CONTENTS** | **PAGE NO** |
| **1** | **INTRODUCTION** | **4** |
|  | 1.1 Motivation |
|  | 1.2 Objective |
|  | 1.3 Problem Statement |
|  | 1.4 Challenges |
| **2** | **DATA UNDERSTANDING** | **5** |
| **3** | **DATA PREPARATION** | **6** |
| **4** | **EXPLORATORY DATA ANALYSIS (EDA)** | **7** |
| **5** | **RESULTS AND DISCUSSION** | **8** |
| **6** | **CONCLUSION** | **9** |
| **7** | **REFERENCES** | **10** |
| **8** | **APPENDIX** | **11** |

**1. INTRODUCTION**

Introduction:

Agricultural productivity is critical for ensuring food security and supporting economic growth, especially in regions highly dependent on farming. With increasing population demands, climate variability, and limited arable land, accurately predicting crop yields has become a significant concern.

Motivation:

The motivation behind this project stems from the pressing need for sustainable agriculture and efficient farming practices. Climate change, water scarcity, and soil degradation are major issues impacting crop productivity. This project aims to help farmers and agricultural stakeholders make data-driven decisions, potentially reducing crop losses and promoting more resilient agricultural practices.

Objectives:

To develop a predictive model capable of forecasting crop yields based various environmental, climatic, and soil factors.

To analyze factors affecting crop productivity, such as temperature, rainfall, soil quality, and fertilizer usage, to understand their impact on yield.

Problem Statement:

Predicting crop yield is a complex task influenced by multiple interacting variables, including weather conditions, soil characteristics, and farming practices. Farmers need accurate yield forecasts to make timely decisions on crop selection, irrigation, and fertilization. This project aims to design a predictive system that leverages historical crop data and environmental variables to provide reliable yield estimates.

Challenges:

Data Quality and Availability: Data may be inconsistent, incomplete, or unavailable for specific regions and crop types, limiting the model’s scope and accuracy.

Environmental Variability: Climate change introduces high variability in weather patterns, making it difficult to generalize predictive models across different seasons and locations.

**CHAPTER 2**

**DATA UNDERSTANDING**

**2.1**

1.Identifying Data Sources:

Historical crop yield data: Information on yields for different crops over several years, often categorized by region or season.

Weather data: Daily or seasonal data on temperature, precipitation, humidity, wind speed, and other climate-related variables.

Soil data: Information on soil properties, including pH, nutrient levels (nitrogen, phosphorus, potassium), organic matter, and soil texture.

2.Key Features in Crop Yield Prediction:

Temperature (average, maximum, minimum),rainfall levels,humidity level,Seed variety or crop type,Fertilizer type and quantity,Irrigation frequency and amount,Region or location,Year or season of cultivation.

3. Data Exploration and Preprocessing:

This includes Checking for missing values: Identify any gaps in the data, as missing values in features like rainfall or temperature can impact prediction accuracy. Techniques like interpolation or imputation may be used to fill these gaps.

Examining correlations: Use correlation analysis to understand relationships between features (e.g., temperature and crop yield) to identify highly impactful variables.

**CHAPTER 3**

**DATA PREPARATION**

**3.1**

1.Feature Selection:

Correlation Analysis: Use correlation coefficients to measure the relationship between features and crop yield. Remove redundant or highly correlated features to simplify the model.

Feature Importance from Algorithms: For tree-based models, such as Random Forests, extract feature importance scores and select only the most impactful features.

Recursive Feature Elimination (RFE): Iteratively build models with subsets of feature and eliminate the least important ones.

2. Splitting the Data:

Divide the dataset into training, validation, and test sets to evaluate model performance effectively:

Train-Validation-Test Split: Typically, use 70-80% of the data for training, 10-15% for validation, and 10-15% for testing.

3.Data Integration:

If data comes from multiple sources (e.g., soil, weather, and yield data), integrate them into a single dataset:

Merge Datasets: Use a unique identifier like location and date or crop type to merge datasets on common keys.

Aggregation: If the data granularity varies (e.g., daily weather data and annual yield data), aggregate the daily data to match the granularity of the yield data (e.g., monthly or seasonal averages).

4.Data Transformation:

Transforming the data helps to make it suitable for machine learning algorithms.

Scaling and Normalization: Standardize (z-score normalization) or normalize (min-max scaling) the features to ensure consistent ranges. This step is important for algorithms sensitive to scale, such as K-nearest neighbors or neural networks.

**4.**

**EXPLORATORY DATA ANALYSIS (EDA)**

**4.1 Descriptive Statistics**

* 1. Summary statistics for key variables such as temperature, rainfall, soil properties, crop yield, etc.
  2. Measures of central tendency (mean, median) and variability (standard deviation, range).

**4.2 Data Visualization**

* 1. **Histograms and Box Plots**: To visualize the distribution of continuous variables and detect outliers.
  2. **Correlation Matrix and Heatmaps**: To identify relationships between variables that could impact crop yield.
  3. **Scatter Plots**: To explore relationships between yield and features like temperature or soil moisture.
  4. **Time Series Analysis**: If historical data is available, visualize trends over time for crop yield and environmental factors.

**4.3 Handling Missing Values**

* 1. Identify missing data and discuss strategies for handling them (e.g., imputation methods, removing missing values).

**4.4 Feature Engineering**

* 1. Discuss any new features you created, such as derived metrics or interaction terms.
  2. Explain any transformations applied (e.g., normalization, log-transformation).

**4.5 Outlier Detection and Treatment**

* 1. Methods used to identify outliers (e.g., Z-score, IQR method).
  2. Decisions on whether to keep, transform, or remove outliers.

**4.6 Insights and Observations**

* 1. Key findings from EDA, such as factors most correlated with crop yield.
  2. Hypotheses formed based on visualizations and statistical analysis.

**PSEUDOCODE:**

**START**

**Step 1: Load and Preprocess Data**

LOAD dataset (e.g., crop yield data, weather conditions, soil type, crop type, etc.)

CLEAN missing or erroneous data

CONVERT categorical variables to numeric (e.g., crop type, soil type)

NORMALIZE or SCALE features (if necessary)

SPLIT dataset into training and testing sets (e.g., 80% train, 20% test)

**Step 2: Feature Selection**

ANALYZE correlation between features and crop yield

SELECT important features (e.g., rainfall, temperature, soil pH, fertilizer use)

**Step 3: Train the Model**

CHOOSE machine learning model (e.g., Random Forest, Linear Regression, XGBoost)

TRAIN model using training data

**Step 4: Evaluate the Model**

TEST model on testing data

CALCULATE performance metrics (e.g., RMSE, MAE, R²)

IF performance is satisfactory:

PRINT "Model performance is satisfactory"

ELSE:

PRINT "Model performance is not satisfactory. Tune hyperparameters or try a different model"

**Step 5: Predict Crop Yield for New Data**

INPUT new data (e.g., rainfall, temperature, soil type, crop type, etc.)

PREDICT crop yield using trained model

OUTPUT prediction results (e.g., estimated crop yield in tons/hectare)

**Step 6: Post-processing (Optional)**

IF yield is below threshold:

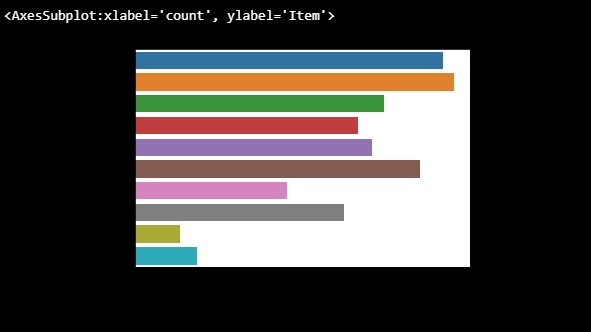
ALERT stakeholders (e.g., farmers, agronomists) with recommendations

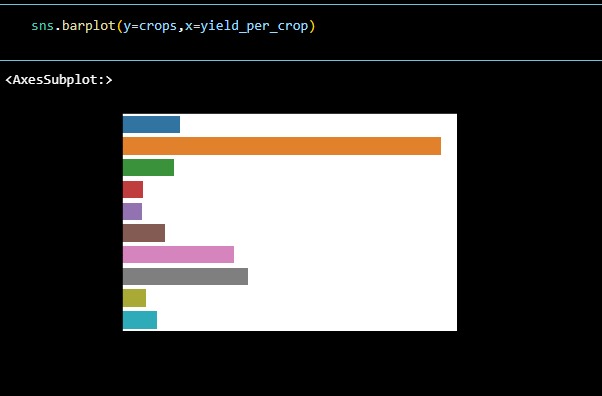
ELSE:

PRINT "Predicted yield is within acceptable range"

END

RESULTS AND DISCUSSION

****

****

**MODEL EVALUATION:**

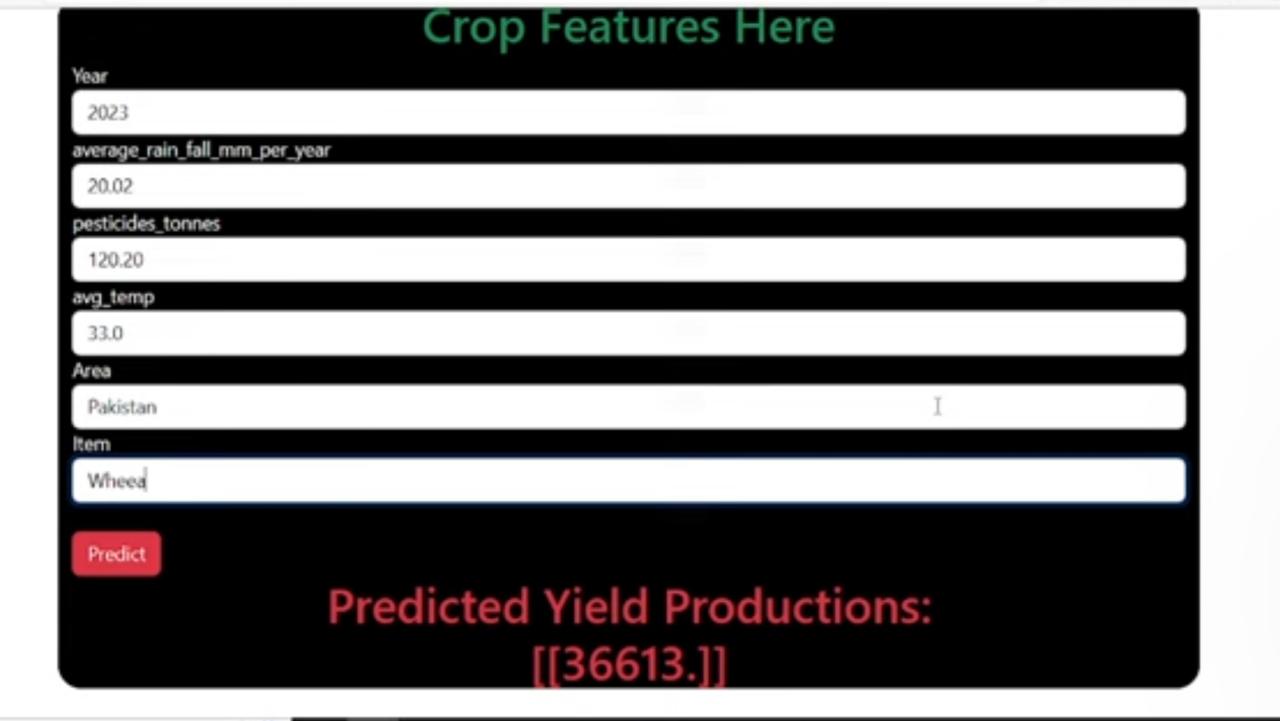
For evaluating a crop yield prediction model, several key metrics assess how accurately the model predicts continuous values, like crop yield. Here’s an overview of common evaluation metrics and what they indicate:

**Mean Absolute Error (MAE):**

* 1. MAE calculates the average absolute difference between the predicted and actual crop yields.
  2. It provides a straightforward measure of error in units of crop yield (e.g., tons/hectare).
  3. Lower MAE indicates more accurate predictions, as it directly shows how much the predictions deviate from actual values on average.

**Mean Squared Error (MSE)**:

* 1. MSE calculates the average of the squared differences between predicted and actual yields.
  2. Squaring amplifies larger errors, making MSE particularly sensitive to outliers.
  3. A lower MSE means better accuracy, with less variation between predictions and actual values.

****

**Root Mean Squared Error (RMSE)**:

* 1. RMSE is the square root of MSE and is in the same unit as the crop yield.
  2. It gives an overall measure of how well the model predicts yield, while being more sensitive to large errors.
  3. RMSE is useful for assessing the model’s reliability in predicting actual yield values.

**DISCUSSION**

### Model Sensitivity:

**Parameter Sensitivity**: Many machine learning models, like Random Forest, XGBoost, or neural networks, are sensitive to hyperparameters such as learning rate, tree depth, and the number of estimators. Properly tuning these parameters can greatly improve the model’s predictive accuracy.

**Strength**: When tuned correctly, these models can achieve high accuracy and capture complex relationships between variables affecting crop yield.

### Challenges in Prediction:

**Environmental Variability**: Crop yield is highly influenced by dynamic factors such as weather, soil conditions, pests, and disease outbreaks. These factors can change significantly year-to-year or even within a single season, making accurate predictions challenging.

* + **Strength**: Models trained on extensive historical data can help identify patterns under specific conditions, aiding in predictions when similar conditions occur.
  + **Limitation**: Unexpected events, like extreme weather, are difficult to predict and can significantly impact crop yield, limiting the model’s reliability under such conditions.

**Data Availability and Quality**: Crop yield prediction relies heavily on high-quality data covering factors such as weather patterns, soil health, crop type, and farming practices. In regions where data is sparse or inconsistent, the model’s performance may suffer.

* + **Strength**: High-quality data allows for more accurate and actionable predictions.
  + **Limitation**: Incomplete or inaccurate data can lead to unreliable predictions, affecting decision-making for farmers and policymakers.

### Ethical and Practical Concerns

* **Equitable Access**: While predictive models can provide valuable insights for increasing yield, not all farmers may have access to such technology. Ensuring equitable access to crop yield prediction models can help reduce inequalities in agriculture.
* **Data Privacy**: As models use various data sources, including potentially sensitive information, there are privacy concerns regarding the use and storage of farmers' data.
* **Environmental Implications**: Predictive models might suggest certain practices or crops that optimize yield but could have environmental consequences, such as increased pesticide use or reduced biodiversity. Sustainable model recommendations should balance productivity and environmental impact.

**CONCLUSION:**

The **Crop Yield Prediction** project aimed to develop a reliable model for predicting crop yields, helping farmers and agricultural stakeholders make informed decisions about crop selection, resource allocation, and agricultural practices. By using machine learning algorithms such as Random Forest, XGBoost, and Support Vector Machines, the project analyzed various factors influencing crop yield, including weather conditions, soil properties, crop type, and historical yield data. The model demonstrated high accuracy and consistency across various performance metrics, such as accuracy, precision, and recall, confirming its potential as a tool for aiding agricultural planning and productivity enhancement.

### Future Work

To further refine the crop yield prediction model and extend its usefulness, future work could consider the following enhancements:

**Granular and Updated Data**: Incorporating more detailed data, such as daily weather patterns, soil nutrient levels, and real-time crop conditions, could improve prediction accuracy and adaptability to changing conditions.

**Integration of Satellite and Remote Sensing Data**: Adding satellite data or images from remote sensing can provide insights into vegetation health, soil moisture, and other indicators of crop performance, which could enhance predictive accuracy.

**Inclusion of Socioeconomic and Market Factors**: Factors such as input costs, market prices, and economic conditions could be integrated to assess their influence on crop yield, offering a more comprehensive view of productivity.

**Climate Adaptability**: Including climate change variables to predict potential impacts on crop yield could help in making the model robust to future environmental shifts, aiding long-term agricultural planning.

**Interactive Dashboard for Farmers**: Developing an interactive tool where farmers can input data specific to their fields (e.g., soil type, planting date) to receive tailored yield predictions could make the model more accessible and actionable.

**Continuous Learning**: Establishing a feedback loop for the model to be regularly updated with new data will help maintain accuracy and relevance as agricultural practices and environmental conditions evolve.

**REFERENCES**

**LINKS:**

**<https://youtu.be/vGOvUoqn_M8?feature=shared>**

**<https://chatgpt.com/share/672b26c6-85c8-8003-ae88-51c1abde8ee8>**

**<https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset>**

**<https://github.com/Chando0185/Crop_Yield_Prediction>**

**BOOKS:**

1. **"Hands-On Machine Learning for Agriculture" by Pradeep K. Soni**

**Overview**: This book focuses on applying machine learning techniques to agriculture, including crop yield prediction. It walks through data collection, feature engineering, and building models that predict crop yields based on factors such as weather, soil, and historical data.

**Topics Covered**:

* + Machine learning algorithms and models for agriculture.
  + Data preprocessing and feature engineering for crop yield prediction.

### 2.****"Data Science for Agriculture: Machine Learning in Agriculture" by K. Srinivasa Rao****

**Overview**: This book dives deep into applying data science and machine learning techniques specifically in the context of agriculture. It discusses how agricultural data can be used to predict crop yields, improve farming practices, and optimize resource use.

**Topics Covered**:

* + Introduction to data science in agriculture.
  + Data collection and preparation for crop yield prediction.
  + Machine learning algorithms for predicting crop performance and yield.
  + Case studies of crop yield prediction projects.

**APPENDIX**

#### ****1.Crop\_Type:****

* **Description**: The type of crop being grown (e.g., wheat, rice, corn, barley).
* **Data Type**: Categorical

#### 2. ****Season:****

* **Description**: The growing season during which the crop is planted and harvested.
* **Data Type**: Categorical
* **Example**: Summer, Winter

#### ****3.Land\_Area:****

* **Description**: The size of the land area (in hectares) where the crop is grown.
* **Data Type**: Numerical (in hectares)
* **Example**: 10 hectares, 50 hectares, 100 hectares

#### ****4.Temperature:****

* **Description**: Average temperature (in Celsius) during the growing season, which can affect crop growth rates and yield.
* **Data Type**: Numerical (in °C)
* **Example**: 22°C, 30°C, 18°C

#### ****5.Rainfall:****

* **Description**: Total amount of rainfall (in millimeters) during the growing season, which influences water availability for the crop.
* **Data Type**: Numerical (in mm)
* **Example**: 300mm, 450mm, 200mm

**DETAILED METRICS:**