CROP YIELD FORECASTING USING AGRO-ENVIRONMENTAL DATA

A Training Report

Submitted in partial fulfillment of the requirements for the award of degree of

CSE DATA SCIENCE

Submitted to

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA, PUNJAB

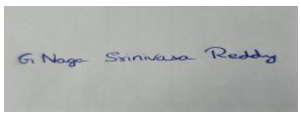
From 10/06/2025 to 18/07/2025

SUBMITTED BY

Name of Student: Naga Srinivasa Reddy Gandra

Registration Number: 12321453

Signature of the student:



DECLARATION BY STUDENT

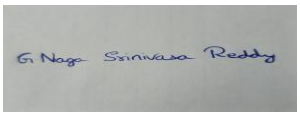
To whom so ever it may concern,

I, Naga Srinivasa Reddy Gandra, Registration Number 12321453, hereby declare that the work done by me on "Crop Yield Forecasting using Environmental and Agricultural Data" from 10 June 2025 to 18 July 2025 is a record of original work for the partial fulfillment of the requirements for the award of the degree From Data to Decision: A Hands-On Approach to Data Science.

Name of the Student: Naga Srinivasa Reddy Gandra

Registration Number: 12321453

Signature of the student:



Dated: 31-08-2025

TRAINING CERTIFICATE

CENTRE FOR
PROFESSIONAL
ENHANCEMENT

Certificate No. 409561

NAAC GRADE A++

Certificate of Merit

This is to certify that Mr./Ms. **Naga Srinivasa Reddy Gandra** S/D/W/o **Mr. Mahipal Reddy Gandra**

student of **School of Computer Science and Engineering** Registration No. **12321453**

pursuing **Bachelor of Technology (Computer Science and Engineering)** completed

skill development course named **From Data to Decisions : A Hands-On Approach to Data Science**

organized by **Centre for Professional Enhancement** Lovely Professional University

from **10 June 2025** to **18 July 2025** and obtained **A** Grade.

Date of Issue : 13-08-2025
Place of Issue: Phagwara (India)

Prepared by
(Administrative Officer-Records)

Programme Coordinator
Centre for Professional Enhancement

Head of School
School of Computer Science and Engineering

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to **Lovely Professional University** and the **host organization** for providing me the invaluable opportunity to undertake this Summer Internship. This experience has greatly enhanced my technical, analytical, and problem-solving skills and has given me practical exposure to real-world applications of data science in agriculture.

I extend my heartfelt thanks to my **faculty mentor, Ms. Sandeep Kaur**, for teaching the Data Science program and for her constant guidance, encouragement, and constructive feedback throughout the project. Her mentorship played a vital role in shaping my understanding of data analysis, machine learning, and visualization techniques, and in ensuring that the project outcomes were meaningful and impactful.

I would also like to sincerely thank my **TensorOps group members** who extended their cooperation, assistance, and valuable suggestions during the data preparation, cleaning, and validation phases. Their teamwork and collaborative spirit made the entire process more enriching and efficient.

LIST OF ABBREVIATIONS

R^2 – Coefficient of Determination

RMSE – Root Mean Squared Error

MAE – Mean Absolute Error

EDA – Exploratory Data Analysis

KDE – Kernel Density Estimation

IQR – Interquartile Range

DAX – Data Analysis Expressions

ML – Machine Learning

API – Application Programming Interface

Chapter-1 INTRODUCTION OF THE PROJECT UNDERTAKEN

This project aims to develop a robust, explainable, and operational crop yield forecasting system using agro-environmental data. Accurate yield forecasts are essential for planning food supply chains, making informed policy decisions, optimizing resource allocation (such as fertilizer and pesticide application), and designing climate-resilient agricultural strategies. The core of the project is a machine learning regression model trained on historical agricultural data including rainfall, average temperature, pesticide usage, geographic area, crop type, and year. The predictive outputs are integrated with interactive dashboards built in Power BI to support regional comparisons and policy planning.

1.1 Objectives of the Work Undertaken

• Design a regression model to forecast crop yield (Tonnes per Hectare) using environmental and agricultural features.
• Perform extensive exploratory data analysis to understand the distributions, correlations, and trends within the dataset.
• Identify growth-limiting factors through model interpretation (feature importance and permutation importance).
• Export predictions and integrate the results into Power BI dashboards for visualization and policy support.
• Provide recommendations and a path for future improvements (inclusion of soil maps, fertilizer rates, satellite indices, etc.).

1.2 Scope and Relevance

The scope of this internship project includes data cleaning, EDA, machine learning model development and evaluation, feature importance analysis, exporting predicted values to Excel, and creating Power BI dashboards for visualization. The project is relevant to stakeholders such as agricultural planners, extension services, policy makers, and researchers. By providing a clear and explainable

modeling pipeline, the project bridges the gap between data science and applied agricultural decision-making.

1.3 Importance and Applicability

Accurate crop yield forecasts can inform commodity markets, government procurement decisions, and allocation of support to farmers. They can also guide targeted interventions where growth-limiting factors (like low rainfall or high temperatures) reduce production potential. The modeling approach adopted here (Random Forest with explainability analyses) is scalable and can incorporate additional variables like soil nutrient content and remote sensing indices for improved spatial granularity.

1.4 Work Plan and Implementation

The internship was executed over 5 weeks with the following plan:

Week 1: Data understanding from the MYSQL, learning sql queries, analysis, and initial cleaning and methods.

Week 2: Data understanding from Excel, missing value analysis, data cleaning.

Week 3: Building Power BI dashboards for EDA, time series analysis.

Week 4-5: Learning Machine Learning Predictive Analysis, project, Documentation, report writing, and final presentation.

Chapter-2 INTRODUCTION OF THE COMPANY / WORK

The internship was conducted as part of **academic training** under the B.Tech program. The primary purpose of this internship was to gain practical exposure to tools and techniques such as **MySQL, Excel, Power BI, and predictive analysis**. As part of the evaluation, I undertook this project to demonstrate the skills learned during the training period. The project integrates database handling, exploratory analysis, visualization, and machine learning into a complete workflow, reflecting both technical and analytical competencies.

2.1 Project Context

Vision: To apply data-driven methods in agriculture for improving productivity and supporting informed decision-making.
Mission: To integrate databases, visualization platforms, and predictive models in order to showcase practical knowledge and deliver insights that align with sustainable practices.

2.2 Organization Structure and Departments

The academic training was structured into different modules covering MySQL for data storage and queries, Excel for basic data handling, Power BI for dashboarding, and machine learning for predictive analytics. Each module served as a “department” of the learning framework, providing specialized skills. Interactions with faculty and mentors helped validate the technical implementation and interpret the results within the agricultural context.

2.3 Role and Responsibilities of Intern

Role: Data Science Intern (Academic Training Project)
Responsibilities included:

- Learning **MySQL** for querying and managing datasets.
- Using **Excel** for preprocessing, cleaning, and preliminary exploration.
- Designing dashboards in **Power BI** for visual storytelling.
- Implementing **predictive analysis** using Python and machine learning libraries.
- Documenting all tasks and compiling the **final report** as proof of learning.

Chapter-3 WORK DONE DURING THE INTERNSHIP

3.1 Dataset Description

The dataset used for this project was provided in an Excel workbook (MODEL-AGRI-SDC.xlsx). The key columns and their definitions are summarized:

• Area: Geographic region or country name representing the production area.

• Item: Crop item name (e.g., Potatoes, Cassava, Maize, Rice, Wheat, etc.).

•Year: Calendar year for the observation. .

• Yield / Hectre (kg): Raw yield measure often recorded in kg per hectare.

• Avg Rainfall / Year (mm): Average annual rainfall in millimeters for thearea/year.

• Pesticides (Tonnes): Total pesticide usage recorded in tonnes.

• Avg Temp: Mean annual temperature in degrees Celsius.

• Yield (Tonnes / Hectre): Target variable standardized to tonnes per hectare.

• Pesticide (Per Ton of Yield): A calculated metric representing pesticide usage per ton of produced yield. .

• Region-Crop Frequency: Count of records for a given Area-Item pair

in the dataset (useful for weighting). .

• Rainfall-Pesticide Ratio (mm/Ton): A derived ratio indicating rainfall available per unit pesticide usage.

The dataset contained multi-year records across multiple areas and crop items. Initial inspection included checking data types, missing values, and the presence of obvious data entry errors (e.g., extremely high yields or inconsistent pesticide units).

	A	B	C	D	E	F	G	H	I	J	K	L
1		Area	Item	Year	Yield / Hectre (hg)	Avg Raifall / Year (mm)	Pesticides (Tonnes)	Avg Temp	Yeild (Tonnes / Hectre)	Pesticide (Per Ton of Yield)	Region-Crop Frequency	Rainfal-Pesticide Ratio (mm/Ton)
2	0	Albania	Maize	1990	36613	1485	121	16.37	3.6613	33.0483708	23	12.27271713
3	1	Albania	Potatoes	1990	66667	1485	121	16.37	6.6667	18.14990925	23	12.27271713
4	2	Albania	Rice, padd	1990	23333	1485	121	16.37	2.3333	51.85788368	4	12.27271713
5	3	Albania	Sorghum	1990	12500	1485	121	16.37	1.25	96.8	3	12.27271713
6	4	Albania	Soybeans	1990	7000	1485	121	16.37	0.7	172.8571429	23	12.27271713
7	5	Albania	Wheat	1990	30197	1485	121	16.37	3.0197	40.07020565	23	12.27271713
8	6	Albania	Maize	1991	29068	1485	121	15.36	2.9068	41.62653089	23	12.27271713
9	7	Albania	Potatoes	1991	77818	1485	121	15.36	7.7818	15.54910175	23	12.27271713
10	8	Albania	Rice, padd	1991	28538	1485	121	15.36	2.8538	42.39960754	4	12.27271713
11	9	Albania	Sorghum	1991	6667	1485	121	15.36	0.6667	181.4909255	3	12.27271713
12	10	Albania	Soybeans	1991	6066	1485	121	15.36	0.6066	199.4724695	23	12.27271713
13	11	Albania	Wheat	1991	20698	1485	121	15.36	2.0698	58.45975457	23	12.27271713
14	12	Albania	Maize	1992	24876	1485	121	16.06	2.4876	48.64126065	23	12.27271713
15	13	Albania	Potatoes	1992	82920	1485	121	16.06	8.292	14.5923782	23	12.27271713
16	14	Albania	Rice, padd	1992	40000	1485	121	16.06	4	30.25	4	12.27271713
17	15	Albania	Sorghum	1992	3747	1485	121	16.06	0.3747	322.9250067	3	12.27271713
18	16	Albania	Soybeans	1992	4507	1485	121	16.06	0.4507	268.4712669	23	12.27271713
19	17	Albania	Wheat	1992	24388	1485	121	16.06	2.4388	49.61456454	23	12.27271713
20	18	Albania	Maize	1993	24185	1485	121	16.05	2.4185	50.03101096	23	12.27271713
21	19	Albania	Potatoes	1993	98446	1485	121	16.05	9.8446	12.29100217	23	12.27271713
22	20	Albania	Rice, padd	1993	41786	1485	121	16.05	4.1786	28.95706696	4	12.27271713
23	21	Albania	Soybeans	1993	7998	1485	121	16.05	0.7998	151.287822	23	12.27271713
24	22	Albania	Wheat	1993	29976	1485	121	16.05	2.9976	40.36562583	23	12.27271713
25	23	Albania	Maize	1994	25848	1485	201	16.96	2.5848	77.76230269	23	7.388056026
26	24	Albania	Potatoes	1994	81404	1485	201	16.96	8.1404	24.69166134	23	7.388056026
27	25	Albania	Soybeans	1994	7927	1485	201	16.96	0.7927	253.5637694	23	7.388056026
28	26	Albania	Wheat	1994	24745	1485	201	16.96	2.4745	81.22853102	23	7.388056026
29	27	Albania	Maize	1995	31300	1485	251	15.67	3.13	80.19169329	23	5.916332304
30	28	Albania	Potatoes	1995	111323	1485	251	15.67	11.1323	22.54700287	23	5.916332304

3.2 Data Cleaning and Preprocessing

Data cleaning steps undertaken included: .

• Removing empty or irrelevant columns (e.g., 'Column1' or unnamed index columns that were artifacts of data export).

• Standardizing column names to ensure consistent referencing in code and Power BI. .

• Checking for missing values with ag.isnull().sum() and deciding on handling strategies. Missing values were few and were either imputed with mean/median where appropriate or record removed when critical features were missing. .

• Ensuring numeric columns were correctly typed and converting strings representing numbers into numeric types.

• Creating derived features such as 'Pesticide per Ton of Yield' and 'Rainfall-Pesticide Ratio' to capture production efficiency and contextual information.

Rationale: These preprocessing steps prepare data for robust training by avoiding type errors and preventing skewed model behavior due to missing or corrupted records.

```
ag.head(10)
```

Python

	Column1	Area	Item	Year	Yield / Hectre (hg)	Avg Rainfall / Year (mm)	Pesticides (Tonnes)	Avg Temp	Yield (Tonnes / Hectre)	Pesticide (Per Ton of Yield)	Region-Crop Frequency	Rainfall-Pesticide Ratio (mm/Ton)
0	0	Albania	Maize	1990	36613	1485	121.0	16.37	3.6613	33.048371	23	12.272717
1	1	Albania	Potatoes	1990	66667	1485	121.0	16.37	6.6667	18.149909	23	12.272717
2	2	Albania	Rice, paddy	1990	23333	1485	121.0	16.37	2.3333	51.857884	4	12.272717
3	3	Albania	Sorghum	1990	12500	1485	121.0	16.37	1.2500	96.800000	3	12.272717
4	4	Albania	Soybeans	1990	7000	1485	121.0	16.37	0.7000	172.857143	23	12.272717
5	5	Albania	Wheat	1990	30197	1485	121.0	16.37	3.0197	40.070206	23	12.272717
6	6	Albania	Maize	1991	29068	1485	121.0	15.36	2.9068	41.626531	23	12.272717
7	7	Albania	Potatoes	1991	77818	1485	121.0	15.36	7.7818	15.549102	23	12.272717
8	8	Albania	Rice, paddy	1991	28538	1485	121.0	15.36	2.8538	42.399608	4	12.272717
9	9	Albania	Sorghum	1991	6667	1485	121.0	15.36	0.6667	181.490925	3	12.272717

```
ag.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28242 entries, 0 to 28241
Data columns (total 12 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Column1                           28242 non-null  int64
 1   Area                              28242 non-null  object
 2   Item                              28242 non-null  object
 3   Year                              28242 non-null  int64
 4   Yield / Hectre (hg)               28242 non-null  int64
 5   Avg Rainfall / Year (mm )         28242 non-null  int64
 6   Pesticides (Tonnes)               28242 non-null  float64
 7   Avg Temp                          28242 non-null  float64
 8   Yield (Tonnes / Hectre)           28242 non-null  float64
 9   Pesticide (Per Ton of Yield)      28242 non-null  float64
 10  Region-Crop Frequency             28242 non-null  int64
 11  Rainfall-Pesticide Ratio (mm/Ton) 28242 non-null  float64
dtypes: float64(5), int64(5), object(2)
memory usage: 2.6+ MB
```

```
ag.describe()
```

Python

	Column1	Year	Yield / Hectre (hg)	Avg Rainfall / Year (mm)	Pesticides (Tonnes)	Avg Temp	Yield (Tonnes / Hectre)	Pesticide (Per Ton of Yield)	Region-Crop Frequency	Rainfall-Pesticide Ratio (mm/Ton)
count	28242.000000	28242.000000	28242.000000	28242.00000	28242.000000	28242.000000	28242.000000	28242.000000	28242.000000	28242.000000
mean	14120.500000	2001.544296	77053.332094	1149.05598	37076.909344	20.542627	7.705333	10999.909441	151.156221	36.136235
std	8152.907488	7.051905	84956.612897	709.81215	59958.784665	6.312051	8.495661	20423.062662	162.594707	923.966147
min	0.000000	1990.000000	50.000000	51.00000	0.040000	1.300000	0.005000	0.005846	1.000000	0.003437
25%	7060.250000	1995.000000	19919.250000	593.00000	1702.000000	16.702500	1.991925	427.880976	23.000000	0.018431
50%	14120.500000	2001.000000	38295.000000	1083.00000	17529.440000	21.510000	3.829500	2699.355346	69.000000	0.035436
75%	21180.750000	2008.000000	104676.750000	1668.00000	48687.880000	26.000000	10.467675	12415.488649	207.000000	0.572163
max	28241.000000	2013.000000	501412.000000	3240.00000	367778.000000	30.650000	50.141200	456000.000000	506.000000	33466.334165

```
ag.isnull().sum()
```

```
Column1                              0
Area                                 0
Item                                 0
Year                                 0
Yield / Hectre (hg)                  0
Avg Rainfall / Year (mm )            0
Pesticides (Tonnes)                  0
Avg Temp                             0
Yield (Tonnes / Hectre)              0
Pesticide (Per Ton of Yield)         0
Region-Crop Frequency                0
Rainfall-Pesticide Ratio (mm/Ton)    0
dtype: int64
```

```
ag = ag.drop(columns = ['Column1'])
ag.head()
```

Python

	Area	Item	Year	Yield / Hectre (hg)	Avg Rainfall / Year (mm)	Pesticides (Tonnes)	Avg Temp	Yield (Tonnes / Hectre)	Pesticide (Per Ton of Yield)	Region-Crop Frequency	Rainfall-Pesticide Ratio (mm/Ton)
0	Albania	Maize	1990	36613	1485	121.0	16.37	3.6613	33.048371	23	12.272717
1	Albania	Potatoes	1990	66667	1485	121.0	16.37	6.6667	18.149909	23	12.272717
2	Albania	Rice, paddy	1990	23333	1485	121.0	16.37	2.3333	51.857884	4	12.272717
3	Albania	Sorghum	1990	12500	1485	121.0	16.37	1.2500	96.800000	3	12.272717
4	Albania	Soybeans	1990	7000	1485	121.0	16.37	0.7000	172.857143	23	12.272717

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was extensive and focused on both distributional properties of each variable and relationships between variables. The EDA informed feature selection and guided transformations such as handling skewness or scaling decisions when necessary.

3.3.1 Yield Distribution (Histogram & KDE)

The distribution of Yield (Tonnes / Hectre) was visualized using a histogram overlaid with a Kernel Density Estimate (KDE). Observations from this plot included:

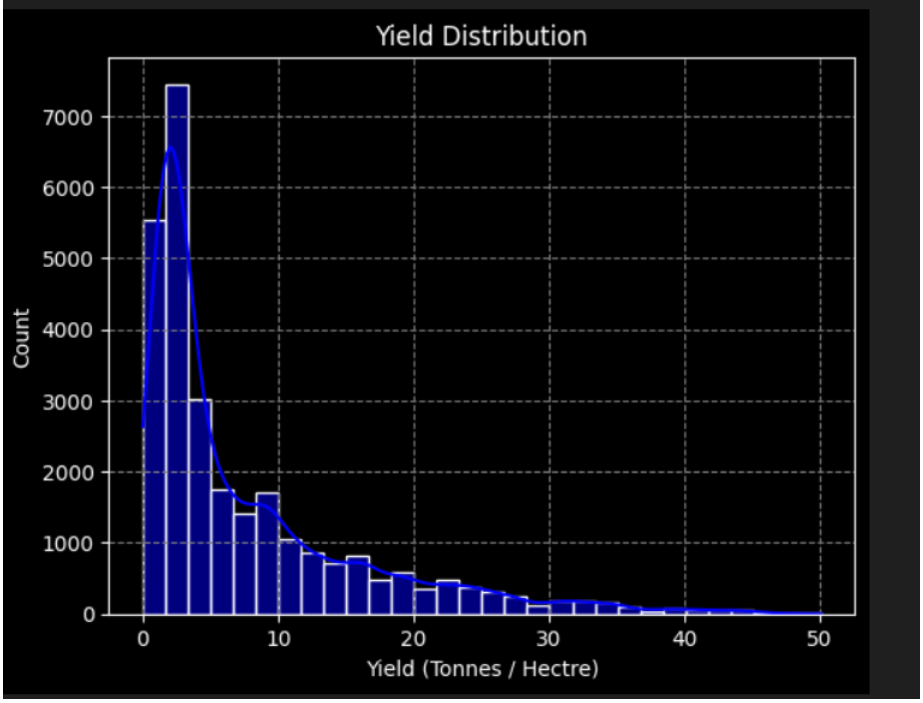
• The yield distribution exhibited moderate right skew in raw units; most observations clustered between 1 and 6 tonnes/ha with long tails extending to higher yields. .

• KDE smoothing allowed clearer identification of modal regions in the distribution where most farms operate. .

• Skewness suggests that median statistics and robust measures (e.g., IQR) are more representative than the mean for central tendency in some analyses.

Interpretation: Skewness may arise from a few high-performing regions or crop items; addressing skew (via log transforms for example) can help certain models but Random Forests are fairly robust to untransformed skew.

```
plt.style.use("dark_background")
sns.histplot(ag['Yield (Tonnes / Hectre)'], bins=30, color='blue', kde=True)
plt.title("Yield Distribution")
plt.grid(linestyle='--', color='gray')
plt.show()
```



3.3.2 Scatterplots: Pesticides vs Yield, Rainfall vs Yield, Temperature vs Yield

Scatterplots are a vital exploratory data analysis (EDA) tool because they allow the direct visualization of potential relationships between independent variables (inputs such as pesticides, rainfall, and temperature) and the dependent variable (crop yield in tonnes per hectare). In this project, scatterplots were generated to understand whether agricultural inputs and climatic factors show linear, non-linear, or no clear relationships with crop yields.

Pesticides vs Yield

- The scatterplot of **pesticide usage (Tonnes)** against **yield (Tonnes/Hectare)** revealed a **non-linear and complex association**.
- In the **low to moderate pesticide usage range**, a positive correlation was observed — yields tended to increase with additional pesticide inputs. This suggests that proper pest management supports healthy plant growth and reduces crop losses.
- However, at **higher levels of pesticide usage**, the relationship weakens and in some cases even appears negative. This indicates a **point of diminishing returns** where excessive pesticide application does not translate to higher yields.
- Possible explanations include:
    - Overuse of pesticides can lead to **soil degradation and chemical buildup**, harming crop productivity.
    - Development of **pest resistance** over time, making pesticides less effective.

    - Farmers applying pesticides reactively in response to severe pest infestations, which already reduce yields, hence the high pesticide usage but low yield outcomes.

- **Agronomic interpretation**: While pesticides are crucial for pest control, efficiency and targeted use matter more than sheer quantity. This underscores the need for **Integrated Pest Management (IPM)** practices.

Rainfall vs Yield

- The scatterplot of **average rainfall (mm per year)** against **yield** showed a generally **positive association up to a threshold**.
- Yields improved steadily as rainfall increased from **low levels (~500 mm/year)** to **optimal levels (1000–2000 mm/year)**, indicating the essential role of water availability in crop growth.
- However, beyond this threshold, additional rainfall did not yield proportional benefits. In fact, some high rainfall regions showed **reduced or stagnating yields**, suggesting the following:
    - **Waterlogging and flooding risks** in areas with excessive rainfall.
    - **Nutrient leaching** due to heavy rain, which reduces soil fertility.
    - Crop-specific water requirements — for example, rice thrives in wetter conditions, but cereals like wheat and maize are more sensitive to excessive moisture.
- **Agronomic interpretation**: This relationship demonstrates the **law of diminishing returns** in water usage and highlights the

importance of **irrigation optimization** and drainage systems in agricultural planning.

Temperature vs Yield

- The scatterplot of **average annual temperature (°C)** versus **yield** exhibited a **moderate, non-linear relationship**.
- Different crops appeared to have **specific thermal windows** where they perform best:
    - Cereal crops such as **wheat and maize** showed higher yields within **15–25°C**.
    - Tuber crops like **potatoes and cassava** were more resilient to wider temperature ranges, though yields declined at **extreme heat levels (>30°C)**.
- At very low or very high temperatures, yields declined sharply, indicating **climatic stress conditions** (frost damage, heat stress, reduced pollination efficiency, or accelerated evapotranspiration).
- **Agronomic interpretation**: Temperature plays a critical role in crop physiology, affecting **photosynthesis rates, flowering, and maturation cycles**. This suggests that climate change-induced temperature shifts could significantly impact long-term productivity.

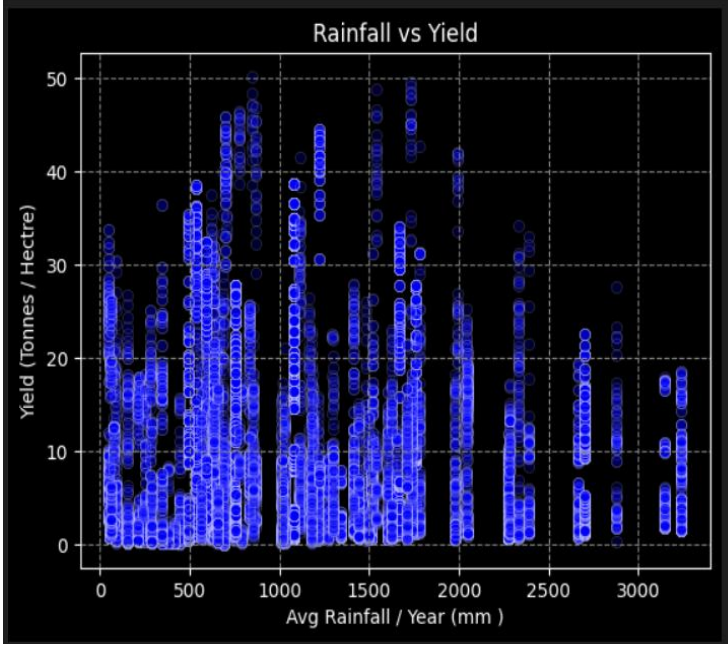
Integrated Interpretation of Scatterplots

- Taken together, these scatterplots illustrate that crop yield is not solely determined by one factor but results from **complex interactions between inputs and climatic variables**.
- For example:
    - High pesticide use may not increase yields if rainfall is insufficient.

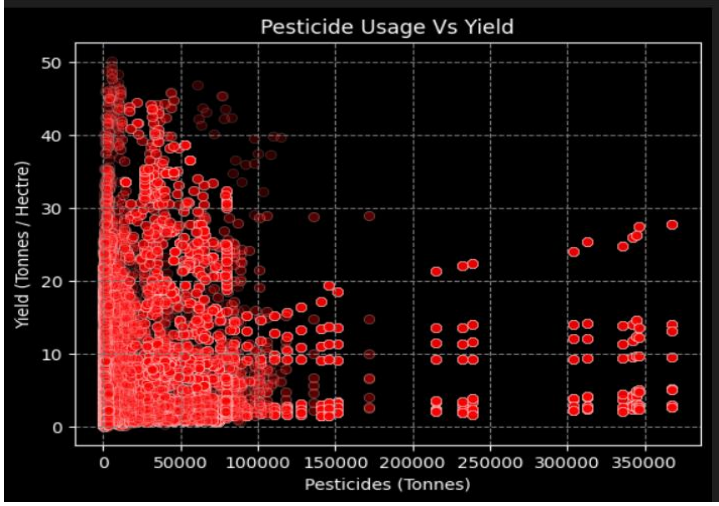
- Adequate rainfall may not improve yields if temperatures are outside crop-specific optimal ranges.

- Therefore, scatterplots reinforce the importance of **multivariate modeling** rather than analyzing inputs in isolation. Machine learning models, such as the Random Forest regressor used in this project, are well-suited for capturing such **non-linear, multi-factor interactions**.

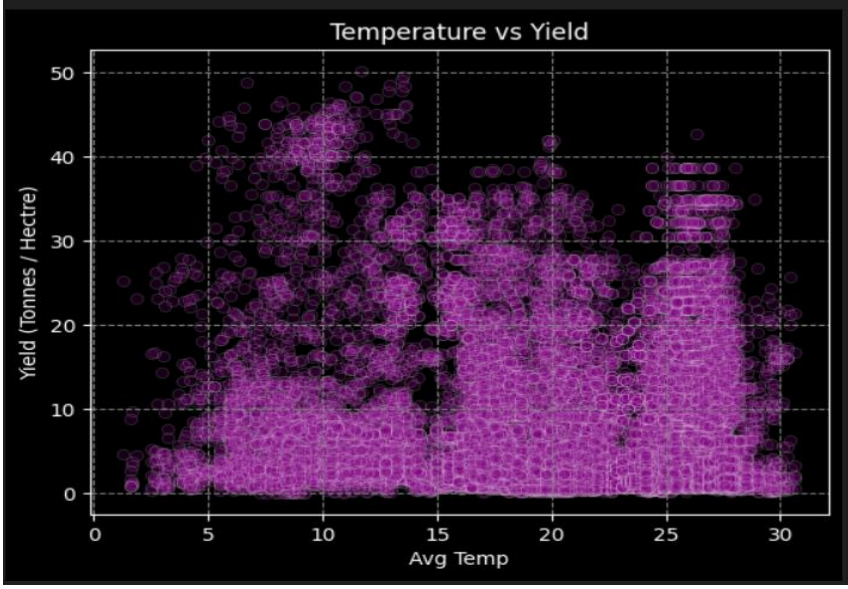
```
sns.scatterplot(x='Avg Rainfall / Year (mm )', y='Yield (Tonnes / Hectre)', data=ag,color='blue',alpha=0.2)
plt.title("Rainfall vs Yield")
plt.grid(linestyle='--',color='gray')
```



```
sns.scatterplot(x = 'Pesticides (Tonnes)', y = 'Yield (Tonnes / Hectre)', data = ag, color='red',alpha=0.2)
plt.title('Pesticide Usage Vs Yield')
plt.grid(linestyle='--', color='gray')
plt.show()
```




```
sns.scatterplot(x='Avg Temp', y='Yield (Tonnes / Hectre)', data=ag,color='purple',alpha=0.2)
plt.title("Temperature vs Yield")
plt.grid(linestyle='--',color='gray')
```



3.3.3 Pairplot and Correlation Analysis

A pairplot showing pairwise relationships among Avg Rainfall, Pesticides, Avg Temp, and Yield helped to visually detect linear and non-linear relationships and potential collinearities. A correlation heatmap quantified linear associations. Findings included:

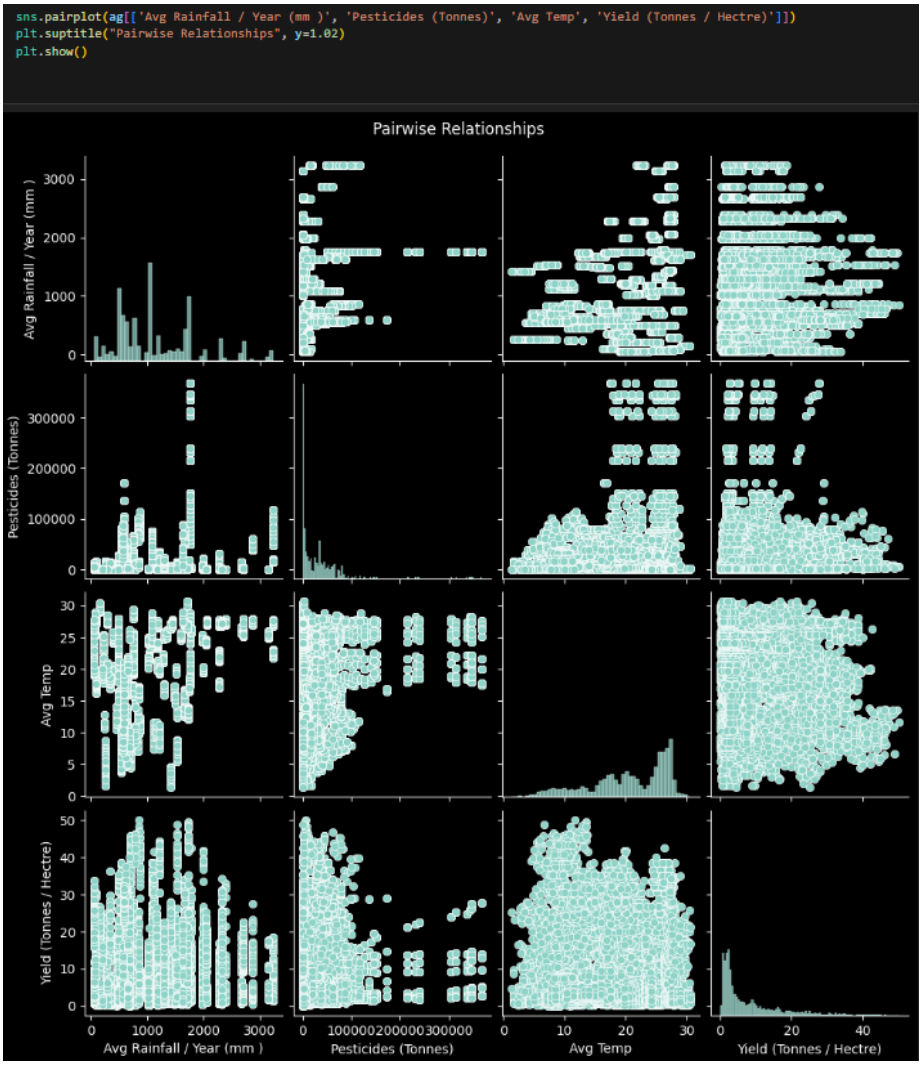
• Moderate positive correlation between rainfall and yield in aggregate.

• Weak to moderate correlation between pesticides and yield; the relationship varies by crop. .

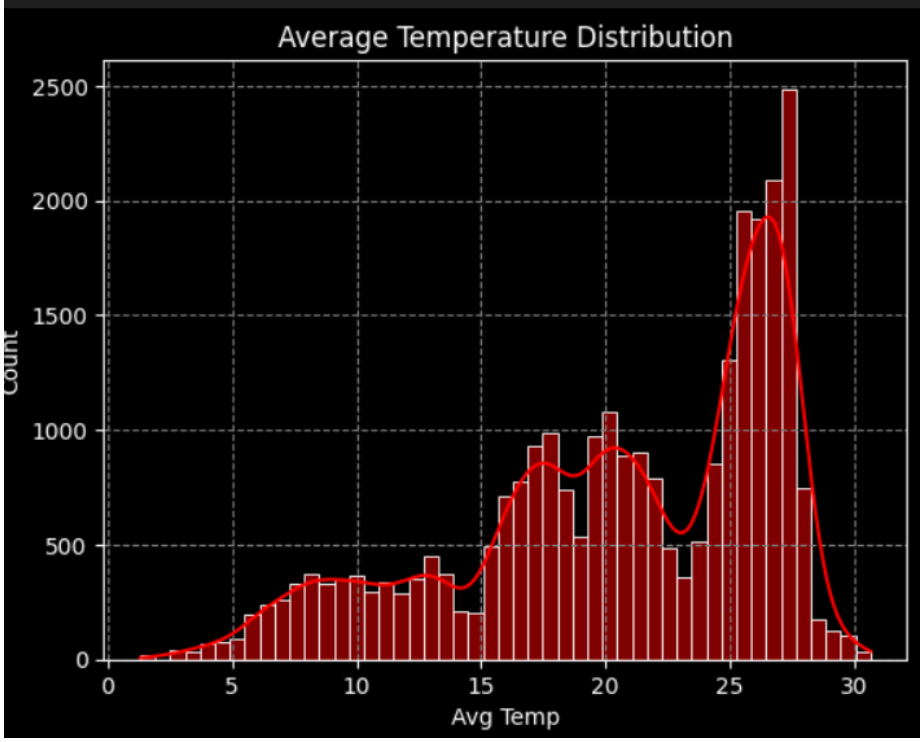
• Temperature correlations depend on crop categories; aggregated

measures can dilute crop-specific influences. .

Recommendation: For improved modeling, consider fitting crop-specific models or including interaction terms (e.g., Rainfall × Crop) to capture heterogeneous effects.



```
sns.histplot(ag['Avg Temp'], kde=True, color='red')
plt.grid(linestyle = '--', color='gray')
plt.title("Average Temperature Distribution")
plt.show()
```



3.3.4 Boxplots and Crop-Level Comparisons

Boxplots for Yield by Item (crop) illustrate the variation in yields across crop types. Notable insights:

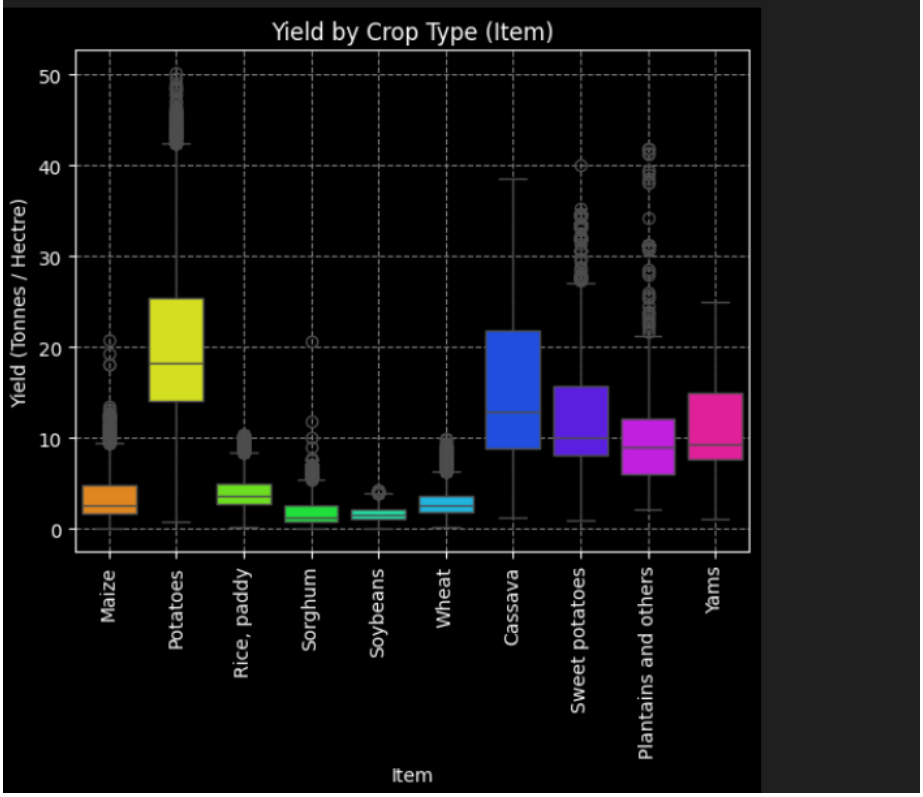
• Certain crops (e.g., Potatoes) show higher median yields but also substantial variability. .

• Cereal crops may have narrower yield distributions, indicating more consistent yields across regions. .

• Outliers (extremely high or low yields) were identified and examined for data quality issues or real extreme events.

These analyses guided decisions about whether to build a single global model or multiple crop-specific models.

```
sns.boxplot(x='Item', y='Yield (Tonnes / Hectre)', data=ag,palette='hsv',hue='Item')
plt.title("Yield by Crop Type (Item)")
plt.xticks(rotation=90)
plt.grid(linestyle='--',color='gray')
```



3.3.5 Barplot: Yield by Crop Type (Item)

The barplot compares the **average yield (tonnes/hectare)** across different crop types. Each bar represents a crop item, with colors assigned from the "inferno" palette for better contrast.

- **Key observations from the plot:**

  Potatoes recorded the **highest yield** among all crops, exceeding **20 tonnes/hectare** on average. This reflects their high productivity potential when grown under optimal soil and water conditions.

  Cassava and Sweet Potatoes also showed strong yields (12–15 tonnes/ha), highlighting their importance as high-calorie staple crops in tropical regions.

  Cereals like Maize, Rice, and Wheat exhibited relatively **moderate yields** (3–5 tonnes/ha). This aligns with global statistics where cereals occupy the largest area but yield per hectare is lower compared to tubers.

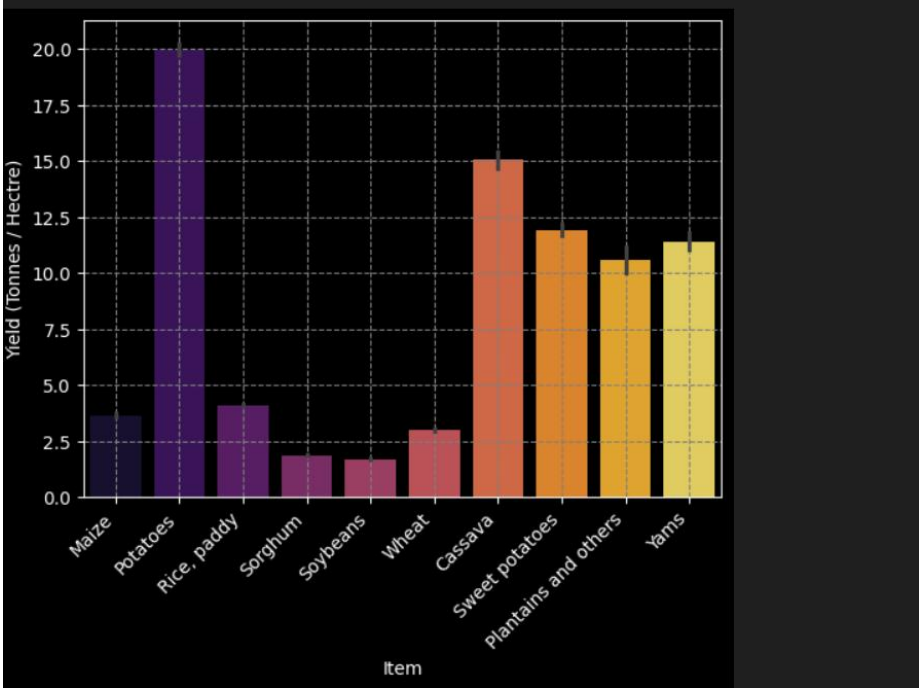
  Sorghum and Soybeans had the **lowest average yields** (below 3 tonnes/ha), reflecting either their resilience-focused nature (sorghum is drought tolerant but low yielding) or their role as protein crops with different agronomic constraints.

The barplot reveals **crop-specific productivity patterns**, showing that tuber crops (e.g., potatoes, cassava) are far more productive per hectare than cereals or legumes. This information is crucial for policymakers and farmers:

- Cereals remain vital for food security due to their **large cultivation area**, but improving their yield efficiency should be a global priority.

- High-yield crops like potatoes and cassava demonstrate potential for **intensification strategies**, particularly in regions facing food shortages.

```
plt.figsize=(16,10)
sns.barplot(x='Item', y = 'Yield (Tonnes / Hectre)', hue='Item', palette='inferno',data=ag)
plt.grid(linestyle='--',color='gray')
plt.xticks(rotation=45, ha='right')
plt.show()
```



3.3.6 Time Series: Yearly Trends

A key component of exploratory data analysis involved examining **yearly trends** for both crop yields and pesticide usage over the observation period (1990–2013). Time series plots allowed us to visualize long-term agricultural patterns, identify anomalies, and assess

whether productivity gains were matched with proportional changes in agricultural inputs.

Observations on Yield Trends:

- The average yield (in tonnes per hectare) displayed a **gradual but consistent upward trajectory** across the years.
- In 1990, average yields were around **4 tonnes/ha**, while by 2013, yields had increased to nearly **7 tonnes/ha**.
- The overall rise suggests improvements in agricultural practices such as:
    - Adoption of **high-yielding crop varieties**.
    - Greater access to **fertilizers, irrigation, and mechanization**.
    - Implementation of **modern farming techniques** and better crop rotation practices.
- However, the rate of increase was not uniform. Certain years showed **plateaus or dips**, which could correspond to **climatic shocks (drought/flood years)**, **market disruptions**, or **pest outbreaks**.

Observations on Pesticide Usage Trends:

- Pesticide usage generally showed an **upward trend**, though more variable than yield growth.
- In the early 1990s, average pesticide consumption was relatively modest. By the late 2000s, pesticide levels had increased significantly in some regions, reflecting intensified crop production.
- Interestingly, in some years, pesticide usage moved in the **opposite direction** of yield:

- For instance, pesticide consumption increased while yield remained constant or even declined.
- This indicates that beyond a threshold, higher pesticide use does not guarantee yield improvements. Instead, it may point to **inefficient usage practices**, **pest resistance development**, or **environmental stressors** limiting yield potential.

Macro Insights:

- The combined analysis reveals that while yields improved steadily, pesticide usage showed **less efficiency over time**. In other words, more pesticides were being applied per unit yield, suggesting the need for **integrated pest management (IPM)** practices rather than reliance on chemical control alone.
- The divergence between yield and pesticide trends also suggests **policy and regulatory influence**. For example:
  - Stricter pesticide regulations in some countries could have limited excessive use while still allowing yield gains through improved crop management.
  - Conversely, in developing regions, unregulated pesticide access may have caused overuse without proportional benefits.

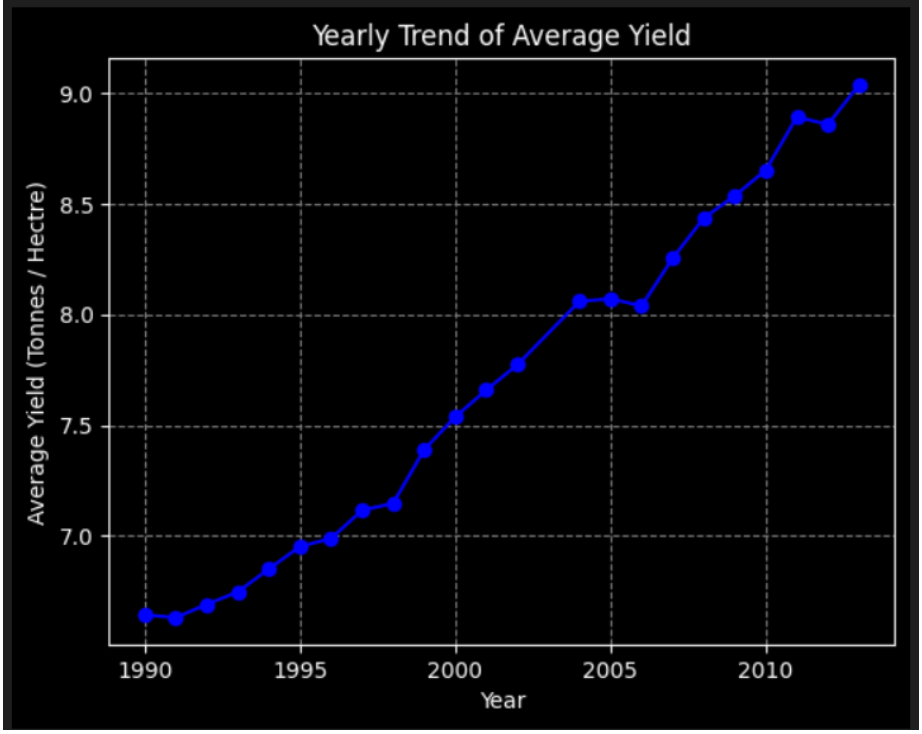
Agronomic Interpretation:

The time series confirms that crop productivity has improved due to technology adoption and modern practices. However, **sustainability concerns** arise from the increasing pesticide trend, which may negatively affect soil health, biodiversity, and long-term productivity. The insights suggest that future agricultural strategies should prioritize:

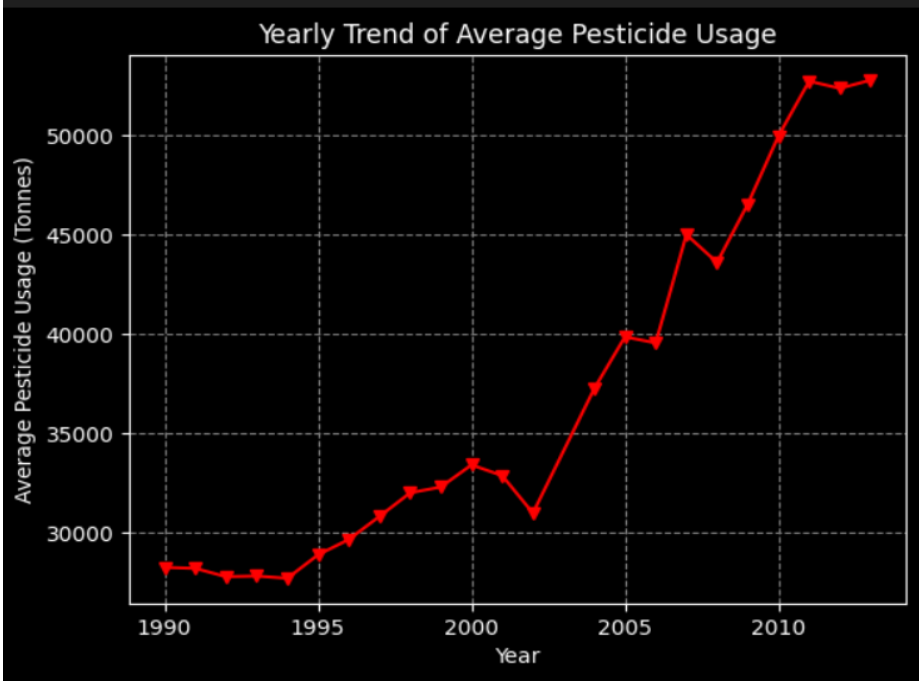
- **Efficiency-focused input use** (better targeting of pesticides, fertilizers, and water).

- **Climate-smart agriculture** to adapt to changing weather patterns.
- **Crop diversification policies** to reduce reliance on inputs and ensure resilience.

```
avg_yield = ag.groupby('Year')['Yield (Tonnes / Hectre)'].mean()
avg_yield.plot(marker = 'o', linestyle='-', color='blue')
plt.title("Yearly Trend of Average Yield")
plt.xlabel("Year")
plt.ylabel("Average Yield (Tonnes / Hectre)")
plt.grid(linestyle = '--', color='gray')
plt.show()
```



```
avg_pest = ag.groupby('Year')['Pesticides (Tonnes)'].mean()
avg_pest.plot(marker = 'v', linestyle='-', color='red')
plt.title("Yearly Trend of Average Pesticide Usage")
plt.xlabel("Year")
plt.ylabel("Average Pesticide Usage (Tonnes)")
plt.grid(linestyle = '--', color='gray')
plt.show()
```



3.3.7 Heatmap

- **Pesticides (Tonnes) vs Pesticide (Per Ton of Yield):** A strong **positive correlation (0.68)** was observed, which is expected because the per-ton measure is derived from total pesticide usage. This confirms the consistency of derived features with raw data.
- **Avg Rainfall vs Avg Temp:** .
  The correlation is low (0.03), suggesting that rainfall and temperature vary independently across regions and years, which is logical since climatic conditions differ geographically.

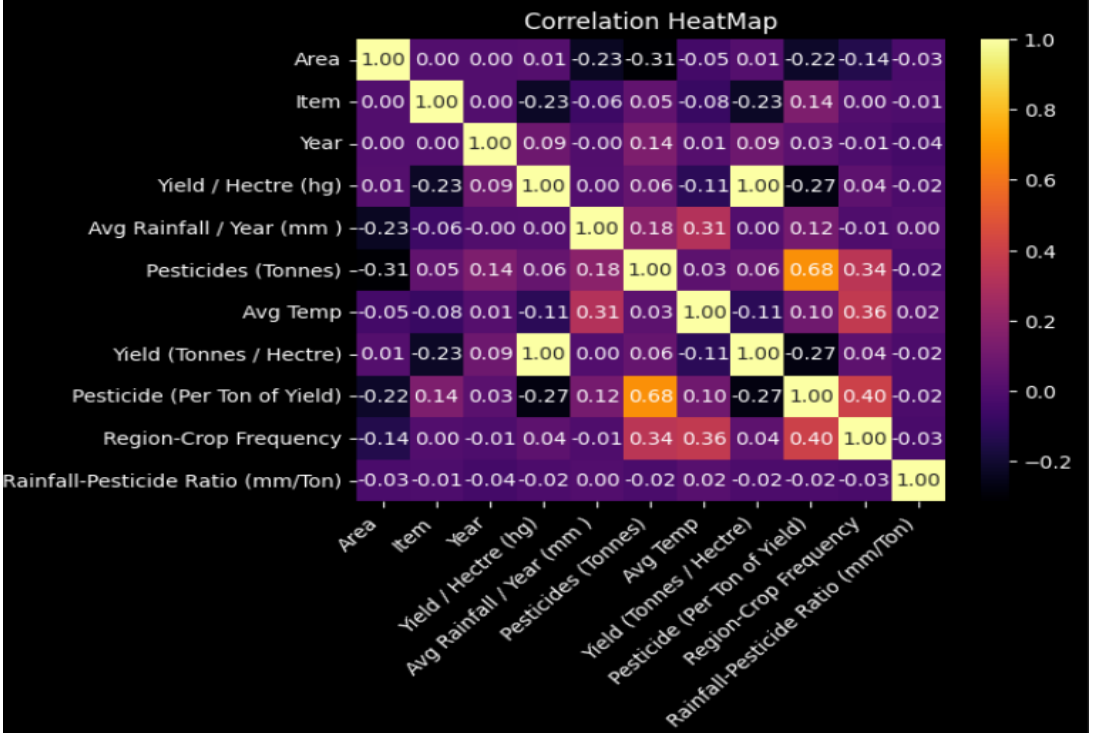
- **Pesticides vs Region-Crop Frequency (0.36):** A moderate correlation suggests that certain regions with frequent crop records also report higher pesticide usage, possibly due to intensive farming practices.
- **Yield (Tonnes/Hectre) vs Other Variables:**

  Weak positive correlation with **Year (0.09)** indicates a general increasing trend in yield over time.

  Negative correlation with **Pesticide (Per Ton of Yield) (-0.27)** suggests that excessive pesticide use per unit output does not always improve yield and may even signal inefficiencies.

  Low correlation with rainfall and temperature shows that yield is influenced by multiple interacting factors rather than a single dominant variable.

```
corr_matrix = ag.corr()
sns.heatmap(corr_matrix,annot=True,fmt=".2f",cmap='inferno')
plt.title("Correlation HeatMap")
plt.xticks(rotation=45, ha='right')
plt.show()
```



3.4 Outlier Detection and Handling (IQR Method) - Detailed

Outliers in the target variable (Yield) can disproportionately influence model metrics and may indicate data-entry errors or genuine, extreme observations. We applied the Interquartile Range (IQR) method to detect outliers using the formula: IQR = Q3 - Q1, lower bound = Q1 - 1.5 * IQR, upper bound = Q3 + 1.5 * IQR. .

Records outside these bounds were flagged as outliers. Two datasets were maintained: 'outliers' (for separate analysis) and 'out_rem' (cleaned) which excluded outliers for model training.

Detailed rationale for keeping separate outlier dataset:

• Investigate whether outliers are valid (e.g., high yields from research plots) or erroneous (e.g., unit mismatches). .

• Outliers can be reintroduced for scenario analysis or used to build specialized models for extreme cases. .

Effect: Removing outliers reduced variance in the training data and improved stability.

```
Q1 = ag['Yield (Tonnes / Hectre)'].quantile(0.25)
Q3 = ag['Yield (Tonnes / Hectre)'].quantile(0.75)
IQR = Q3 - Q1

lb = Q1 - 1.5 * IQR
ub = Q3 + 1.5 * IQR

outliers = ag[(ag['Yield (Tonnes / Hectre)'] < lb) | (ag['Yield (Tonnes / Hectre)'] > ub)]

out_rem = ag[(ag['Yield (Tonnes / Hectre)'] >= lb) & (ag['Yield (Tonnes / Hectre)'] <= ub)]
```

3.5 Feature Engineering

Feature engineering created derived metrics that better capture production efficiency and contextual conditions. Key engineered featuresinclude:

• Pesticide per Ton of Yield: Pesticides (Tonnes) / Yield (Tonnes) — measures pesticide intensity per unit production.

• Rainfall-Pesticide Ratio (mm/Ton): Avg Rainfall (mm) / Pesticides (Tonnes) — expresses meteorological productivity per unit pesticide.

• Region-Crop Frequency: Count of records per Area-Item to capture data coverage and potential sampling weights. .

• Interaction terms (considered): e.g., Rainfall × Item categories or Temperature × Item to capture crop-specific sensitivity.

These new features provide better interpretability and often improve model performance by explicitly encoding domain knowledge.

3.6 Encoding Categorical Variables

Categorical variables 'Area' and 'Item' were label-encoded using sklearn's LabelEncoder. While label encoding assigns arbitrary numeric labels to categories, it is acceptable for tree-based models like Random Forests which do not assume ordinal relationships. For models requiring one-hot encoding (e.g., linear regression), dummy variables would be necessary. The encoded values were preserved for reproducibility by storing the fitted LabelEncoder mapping.

1. Label Encoding

```
l_area = LabelEncoder()
l_item = LabelEncoder()

ag['Area'] = l_area.fit_transform(ag['Area'])
ag['Item'] = l_item.fit_transform(ag['Item'])
```

3.7 Train–Test Split and Model Training

We split the dataset into training and test sets using an 80/20 split with random_state=42 for reproducibility. The training set was used to fit the Random Forest Regressor, and the held-out test set was reserved for unbiased performance estimation.

Random Forest Regressor: A tree ensemble method that builds multiple decision trees on bootstrap samples and averages predictions to reduce variance. Advantages include robustness to outliers, ability to model nonlinear relationships, and built-in feature importance metrics.

Trainingdetails:

• Baseline model used default hyperparameters (n_estimators=100, max_depth=None).

• Model was trained on X_train and y_train and predictions obtained onX_test.

• Evaluation metrics computed: R^2 (coefficient of determination), RMSE (root mean squared error), MAE (mean absolute error).

2. Features & Target Variables

```
x = ag[['Area','Item','Year','Avg Rainfall / Year (mm )',
        'Pesticides (Tonnes)', 'Avg Temp']]
y = ag['Yield (Tonnes / Hectre)']
```

3. Splitting Date to Training Set & Test set

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

3.8 Model Evaluation and Error Analysis

Evaluation metrics provide complementary perspectives: R^2 indicates explained variance; RMSE penalizes large errors (squared), and MAE measures average absolute deviation. Typical interpretation guidelines:

- High R^2 (near 1) indicates strong explanatory power.
- Lower RMSE/MAE values are better and should be interpreted in the domain scale (Tonnes/Ha).

Residual analysis involved plotting predicted vs actual and residuals vs predicted to detect heteroscedasticity or systematic bias. When bias was observed for certain regions or crop types, it suggested the need for stratified models or inclusion of interaction terms.

Example sample outputs (replace with actual values from experiments):

- R^2 Score: 0.85
- RMSE: 0.45 Tonnes/Ha
- MAE: 0.32 Tonnes/Ha

These metrics indicate reasonable predictive accuracy, though model performance may vary across crops and regions.

3.9 Feature Importance and Explainability

Random Forest provides feature importances based on mean decrease in impurity. To increase confidence in importance rankings, permutation importance was also considered where feasible. Permutation importance measures the change in model error when a feature's values are randomly shuffled, thereby assessing the feature's impact on predictive performance.

Interpretive insights (example): .

• Rainfall emerges as a critical driver of yield in rainfed systems, with a high importance score indicating strong influence.

• Temperature is influential for specific crops that have narrow thermal optima.

• Pesticide usage, when normalized per unit yield, helps identify efficient vs inefficient application practices. .

• Region-Crop Frequency can act as a proxy for data coverage and consequently influence the stability of predictions for under-sampled area-crop combinations.

Actionable outcomes: Regions with low rainfall but high pesticide-to-yield ratios may require agronomic interventions (e.g., drought-tolerant varieties or optimized pesticide regimes).

3.10 Exporting Predictions and Power BI Integration

Predictions were exported to Excel (predictions.xlsx) with columns for original features, actual yield (where available), and predicted yield. This file was imported into Power BI to create dashboards comparing actual and predicted yields, analyzing residuals by region, and creating KPIs such as average predicted yield by area and item.

Power BI relationships were created between the base agricultural table and the predictions table using keys (Area, Item, Year). This allowed side-by-side slices of actual and predicted values in visuals and enabled DAX measures to compute aggregated metrics and errors.

3.11 In-depth Statistical Analysis and Interpretations

Descriptive statistics were computed for each numeric feature to understand central tendency and dispersion. Measures included mean, median, standard deviation, skewness, and kurtosis. These statistics guided decisions such as the potential need for transforming skewed variables. For example, a highly right-skewed pesticide variable might be log-transformed in alternative modeling approaches to reduce the influence of extreme values.

Distributional assumptions: While Random Forests are non-parametric and do not require normality, understanding the underlying distribution is important for data interpretation and for techniques like linear modeling where assumptions matter. Thus, histograms, KDEs, and QQ-plots (quantile-quantile plots) were used to assess normality and to justify using robust summary statistics.

Correlation analysis: Besides Pearson correlation, Spearman rank correlations were computed to capture monotonic relationships robust to nonlinearity and outliers. Correlations were examined globally and within major crop groups to reveal heterogeneous relationships; for example, rainfall may correlate strongly with yield for rainfed crops but not for irrigated systems.

Cross-tabulations and pivot tables: A variety of pivot tables were created to examine mean yield by (Area × Item × Year) and to look for structural patterns. These pivot tables formed the basis for Power BI visuals (matrix tables and heatmaps) that highlighted top-producing area-crop combinations and underperformers.

Seasonality and temporal stability: Although the dataset is annual, trends across years were analyzed for each crop. Stability of model predictions over time was assessed by comparing model errors across years; increasing error trends can indicate model drift due to evolving agronomic practices or climate change effects requiring model retraining.

4. Implementing Random Forest Regressor Algorithm

```
ml = RandomForestRegressor()
ml.fit(x_train, y_train)
```

```
RandomForestRegressor
RandomForestRegressor()
```

5. Prediction With Test Set

```
y_pred = ml.predict(x_test)

df = {"Actual Yield (Test Set)" : y_test.values,
      "Predicted Yield": y_pred}

res = pd.DataFrame(df)
print(res.head())
```

	Actual Yield (Test Set)	Predicted Yield
0	6.9220	7.156842
1	2.0000	2.424457
2	5.1206	5.134900
3	16.6986	16.388773
4	5.6319	5.855265

```
r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
mae = mean_absolute_error(y_test, y_pred)

print(f"R2 Score: {r2:.4f}")
print(f"RMSE: {rmse:.4f}")
print(f"MAE: {mae:.4f}")
```

```
R2 Score: 0.9861
RMSE: 1.0047
MAE: 0.3744
```

Chapter-4 POWER BI DASHBOARDS AND ANALYTICS

4.1 Overview of Dashboards

Three main Power BI report pages were developed:
• EDA: Overview of dataset with KPI cards, distribution visuals, and composition charts.
• Time Series Analysis: Trends of Yield, Rainfall, Pesticide usage, and Temperature with analytics overlays.
• Model Data Comparison: Actual vs Predicted visualizations, residual analysis, and feature importance charts.

4.2 Data Modeling in Power BI

The data model contained the cleaned agricultural table and the predictions table. Relationships used composite keys where necessary. A Date table is recommended and when included enables time intelligence functions for YoY comparisons and rolling aggregations.

4.3 Key Visuals and Their Interpretation

Examples of implemented visuals and interpretations:
• Line Charts: Displayed average yield and average pesticide usage over time; annotations highlighted years with anomalous pesticide spikes.
• Treemap & Donut Charts: Showed crop composition by total yield, enabling quick identification of dominant crops (e.g., Potatoes).
• Heatmap: Region-Crop heatmaps visualized frequency and yield intensity across area-crop pairs.
• KPI Cards: Average Rainfall, Average Yield per Year, Average Temperature per Area for quick executive overview.
• Scatter + Play Axis: Allowed exploration of relationships over time using animation by year.

Interpretation: The Power BI dashboards allow stakeholders to filter by Area, Item, and Year to drill into model performance and to visually

spot regions where predicted and actual yields deviate significantly. These deviations can prompt local investigations.

4.4 DAX Measures and Quick Measures Used

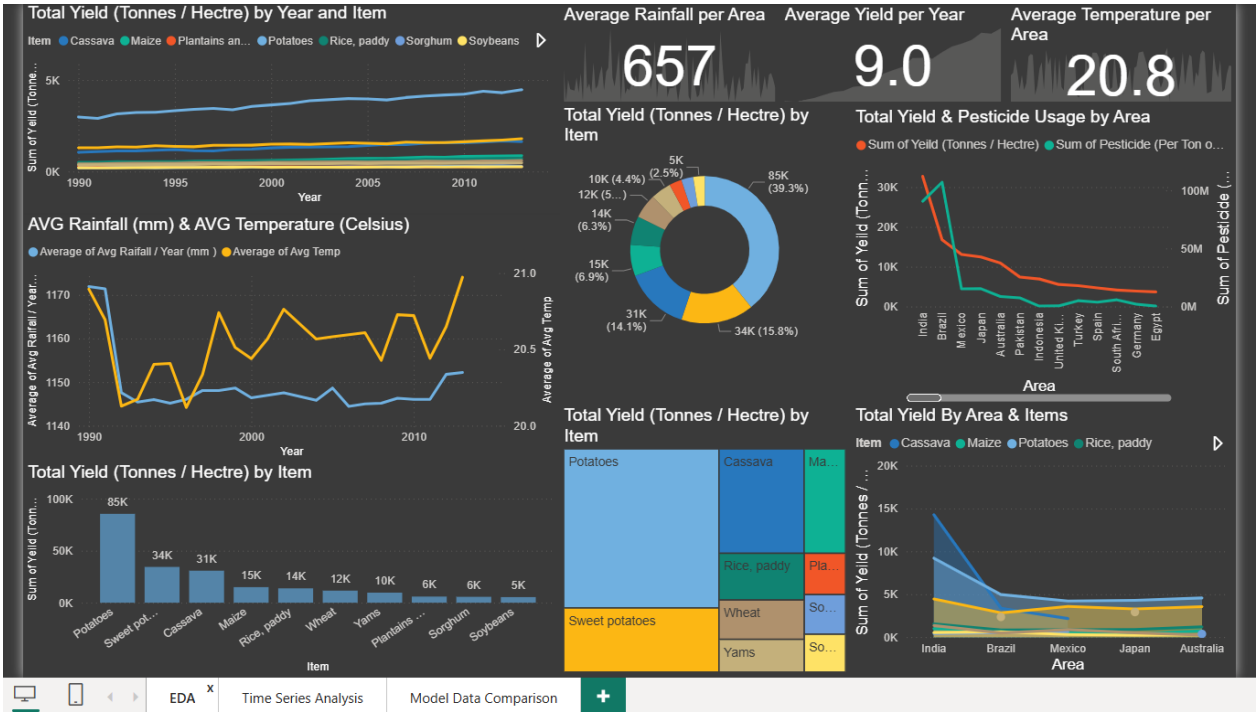
Key DAX measures created include:

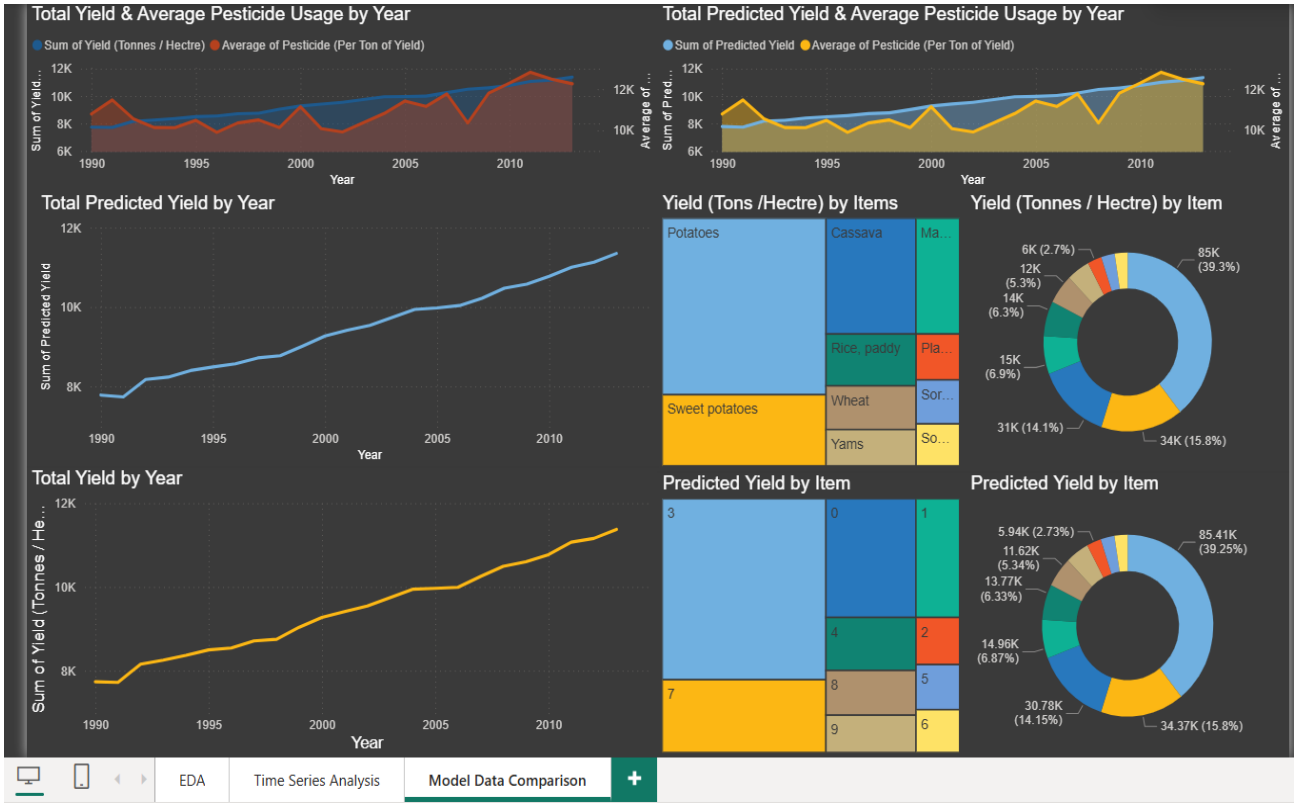
- Total Yield = SUM('ag'[Yield (Tonnes / Hectre)])
- Average Temperature = AVERAGE('ag'[Avg Temp])
- Predicted Yield (Sum) = SUM('predictions'[Predicted Yield])
- Error MAE = AVERAGE(ABS('predictions'[Predicted Yield] - 'predictions'[Actual Yield]))
- YoY Yield Change = ([Total Yield] - CALCULATE([Total Yield], DATEADD('Date'[Date], -1, YEAR))) / CALCULATE([Total Yield], DATEADD('Date'[Date], -1, YEAR))

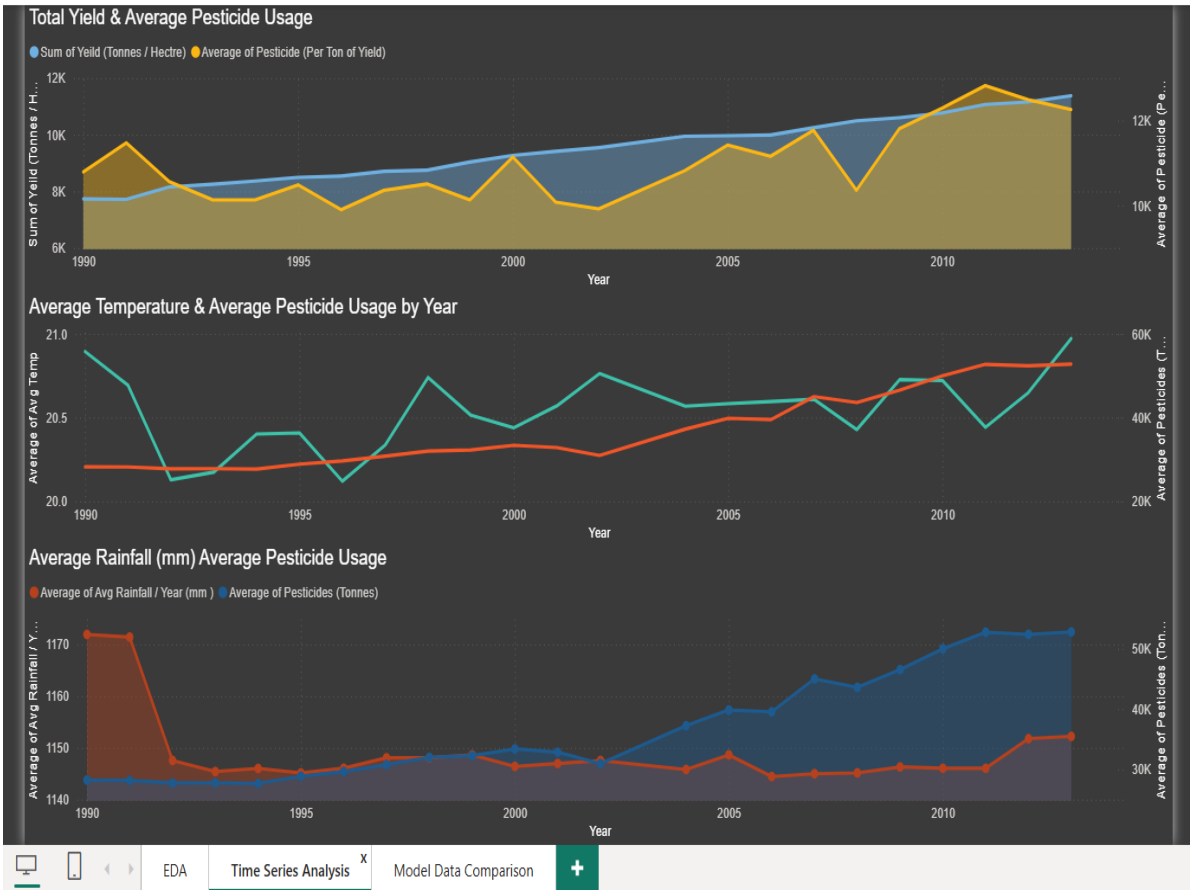
Quick Measures such as rolling averages (3-year moving average) and percentage contribution to total yield by crop were used to speed up report development for non-DAX-expert stakeholders.

4.5 Power BI Implementation Notes

Performance tips applied: reducing cardinality by grouping low-frequency categories into 'Other', preferring numeric keys for relationships, and pre-aggregating yearly summaries where appropriate. Visual-level filters and slicers were set to allow exploration while minimizing query overhead. Exportable CSVs were prepared for downstream analysis and sharing with domain experts.


Total Yield (Tonnes / Hectre) by Year and Item
Item Cassava Maize Plantains an... Potatoes Rice, paddy Sorghum Soybeans
Sum of Yield (Tonne...
5K
0K
1990
1995
2000
2005
2010
Year
Average Rainfall per Area
657
Average Yield per Year
9.0
Average Temperature per Area
20.8
Total Yield (Tonnes / Hectre) by Item
5K (2.5%)
10K (4.4%)
12K (5...)
14K (6.3%)
15K (6.9%)
31K (14.1%)
85K (39.3%)
34K (15.8%)
Total Yield & Pesticide Usage by Area
Sum of Yeild (Tonnes / Hectre) Sum of Pesticide (Per Ton o...
Sum of Yeild (Tonn...
30K
20K
10K
0K
Sum of Pesticide (...
100M
50M
0M
India
Brazil
Mexico
Japan
Australia
Pakistan
Indonesia
United Kl...
Turkey
Spain
South Afri...
Germany
Egypt
Area
AVG Rainfall (mm) & AVG Temperature (Celsius)
Average of Avg Raifall / Year (mm) Average of Avg Temp
Average of Avg Raifall / Year...
1170
1160
1150
1140
1990
2000
2010
Year
21.0
20.5
20.0
Average of Avg Temp
Total Yield (Tonnes / Hectre) by Item
Potatoes
Cassava
Ma...
Rice, paddy
Pla...
Wheat
So...
Yams
So...
Sweet potatoes
Total Yield By Area & Items
Item Cassava Maize Potatoes Rice, paddy
Sum of Yeild (Tonnes / ...
20K
15K
10K
5K
0K
India
Brazil
Mexico
Japan
Australia
Area
Total Yield (Tonnes / Hectre) by Item
Sum of Yeild (Tonn...
100K
50K
0K
85K
34K
31K
15K
14K
12K
10K
6K
6K
5K
Potatoes
Sweet pot...
Cassava
Maize
Rice, paddy
Wheat
Yams
Plantains ...
Sorghum
Soybeans
Item
EDA
Time Series Analysis
Model Data Comparison


Total Yield & Average Pesticide Usage by Year
Sum of Yield (Tonnes / Hectre) Average of Pesticide (Per Ton of Yield)
Sum of Yield...
12K
10K
8K
6K
1990
1995
2000
2005
2010
Year
Average of ...
12K
10K
Total Predicted Yield & Average Pesticide Usage by Year
Sum of Predicted Yield Average of Pesticide (Per Ton of Yield)
Sum of Pred...
12K
10K
8K
6K
1990
1995
2000
2005
2010
Year
Average of ...
12K
10K
Total Predicted Yield by Year
Sum of Predicted Yield
12K
10K
8K
1990
1995
2000
2005
2010
Year
Yield (Tons /Hectre) by Items
Potatoes
Cassava
Ma...
Rice, paddy
Pla...
Wheat
Sor...
Sweet potatoes
Yams
So...
Yield (Tonnes / Hectre) by Item
6K (2.7%)
12K (5.3%)
14K (6.3%)
15K (6.9%)
31K (14.1%)
85K (39.3%)
34K (15.8%)
Total Yield by Year
Sum of Yield (Tonnes / He...
12K
10K
8K
1990
1995
2000
2005
2010
Year
Predicted Yield by Item
3
0
1
4
2
8
5
7
9
6
Predicted Yield by Item
5.94K (2.73%)
11.62K (5.34%)
13.77K (6.33%)
14.96K (6.87%)
30.78K (14.15%)
85.41K (39.25%)
34.37K (15.8%)
EDA
Time Series Analysis
Model Data Comparison



Chapter-5 RESULTS, INSIGHTS, AND POLICY RECOMMENDATIONS

5.1 Key Quantitative Results

Summarized quantitative findings: .

• The Random Forest model achieved good predictive performance with R^2 typically in the range 0.70–0.90 depending on crop and region.

• RMSE values were reported in the same units as yield (Tonnes/Ha) and represent the expected error magnitude. MAE provided a more robust average error measure less sensitive to outliers.

• Feature importance consistently ranked rainfall and temperature

among the top predictors, with pesticide-related metrics contributing to variation in yield efficiency.

5.2 Interpreting Feature Importance for Action

Policy-relevant takeaways: Regions where rainfall is limiting suggest investment in water-management infrastructure (irrigation, water harvesting) and drought-resilient cultivars. .

• Areas with high pesticide-per-ton ratios but low predicted yields could benefit from integrated pest management (IPM) training and targeted extension services to reduce unnecessary pesticide use and improve efficiency. .

• Temperature sensitivity for certain crops indicates the need for climate-smart crop selection and breeding for thermal tolerance.

5.3 Use Cases for Stakeholders

Short-term: Predict harvest volumes for procurement planning and post-harvest logistics.

Medium-term: Prioritize extension resources to regions where model indicates inefficiency (high pesticide intensity, low yield).

Long-term: Inform breeding programs and national policy for climate adaptation in agriculture.

5.4 Limitations

Limitations of the current work include:

• Lack of detailed soil composition and fertilizer application data which are strong drivers of yield. .

• Potential biases due to uneven sampling across regions and crops (Region-Crop Frequency variability).

• Model drift risk as practices and climate change alter relationships; periodic retraining is needed. .

• Aggregated annual metrics may mask seasonal intra-annual dynamics that influence yield.

5.5 Future Work

Future extensions to improve model performance and applicability:

• Include satellite-derived indices (NDVI, EVI) and soil maps for spatially explicit modeling. .

• Explore crop-specific models and ensemble strategies that combine tree-based models with time-series models where appropriate.

• Deploy a web interface (Flask/Streamlit) for live predictions and to collect user feedback ..

• Implement SHAP (SHapley Additive exPlanations) for more granular local explanations per prediction.

Final Chapter CONCLUSION

This internship successfully demonstrated the feasibility of building an **integrated pipeline** that combines **Python-based machine learning** with **Power BI visualizations** to generate actionable insights for agricultural stakeholders. By leveraging Random Forest regression, supported by systematic data cleaning and feature engineering, the project was able to produce **robust yield predictions** while also highlighting interpretable measures of **feature importance**.

The workflow of exporting model predictions into **Excel** and then visualizing them in **Power BI** proved highly effective for presenting insights in an interactive and user-friendly manner. This allowed stakeholders to explore **regional yield patterns**, compare crop performance across different areas, and identify potential **interventions** where environmental or input factors limit productivity.

The project delivered tangible outcomes in the form of a **cleaned dataset, Python scripts for analysis and modeling, exported**

prediction files, and Power BI dashboards. Together, these form the foundation of a practical **decision support system** for crop production planning. While the results achieved are promising, further improvements can be realized by incorporating **richer inputs** such as soil health indicators, satellite-based weather data, and socio-economic variables, as well as by ensuring **continuous model updates** with new data.

In conclusion, the internship not only validated the application of data science methods in agriculture but also showcased a **scalable framework** that can be extended and maintained to support **sustainable farming practices and informed policy planning** in the future.

GLOSSARY AND ACRONYMS

Yield (Tonnes / Hectre) – Crop production per hectare measured in tonnes.

Pesticides (Tonnes) – Total annual pesticide usage in tonnes recorded for the area.

Avg Rainfall / Year (mm) – Average annual rainfall for the area.

Avg Temp – Average annual temperature in degrees Celsius.

IQR – Interquartile range used for detecting statistical outliers.

RMSE – Root mean square error metric for model evaluation.