College Dublin
Computing • IT • Business

Wheat Production Analysis and Supply Chain Optimization

Student Full Name:	Thant Thiha
Student Number:	2025178
Module Title:	ML, Data Prep, Stats – HDip DAB 2025 Feb
Assessment Title:	Integrated CA2
Assessment Due Date:	21 May 2025
Date of Submission:	20 May 2025

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

List of Tables

List of Figures

Data Analytics for Business:

Wheat Production Analysis and Supply Chain Optimization

Author: Thant Thiha,
2025178@student.cct.ie
Higher Diploma in Data Analytics for Business,
CCT College Dublin

INTRODUCTION

This report is for a multinational wheat production and supply chain company operating in Brazil, India, Australia, the United States, Canada, and Syria. It follows the **PACE Framework** for structured analysis, providing actionable insights to improve wheat production efficiency, optimize resource allocation, and enhance market pricing strategies (Yaragal, 2023). The analysis is organized by the **PACE stages**: Plan, Analyze, Construct, and Execute.
The word count for this report is **2,951** (from Introduction to Recommendations excluding cover, contents and references).

PACE Framework Overview

PACE consists of four stages:

P - Plan: Define the problem, identify data sources and set analysis goals.
A - Analyze: Perform data cleaning, exploration and statistical analysis.
C - Construct: Build models, create visualizations and prepare insights.
E - Execute: Apply findings, deploy models and make data-driven decisions.

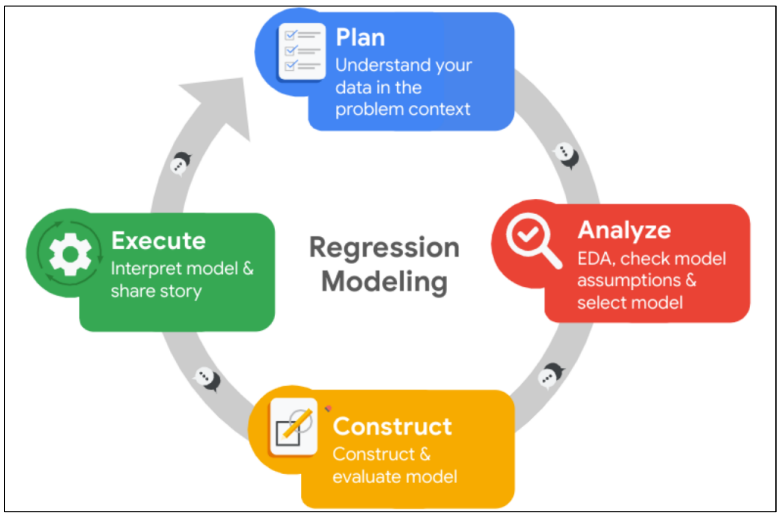

Figure 1: PACE Framework

1. PACE: PLAN STAGE

1.1 Business Scenario: The company seeks to use data-driven insights from the *wheat_production_data.xlsx* dataset to:

- **BO1.** Predict wheat yield(tons/ha) based on environmental and management factors.
- **BO2.** Identify key factors affecting crop performance.
- **BO3**. Optimize pricing strategies to maximize ROI.
- **BO4**. Enhance farmer satisfaction with actionable recommendations.

1.2 About the dataset

The dataset has 900 observations and 9 attributes, covering agricultural production, environmental and economic factors. It includes both categorical and numerical data with minimal missing values.

Column Name	Data Type	Description
Region	Categorical	The region where the wheat is produced such as "North America", "Europe", "Asia", "Africa".
Weather Conditions	Categorical	The weather conditions during the wheat growing season such as "Drought", "Rainy", "Sunny", "Stormy", etc.
Soil Type	Categorical	The type of soil in which the wheat is grown such as "Sandy", "Clay", "Silt", "Loamy", etc.
Fertilizer Usage (kg/ha)	Numeric	The amount of fertilizer used per hectare of land in kilograms per hectare.
Crop Yield (tons/ha)	Numeric	The amount of wheat produced per hectare of land in tons per hectare.
Pest Infestation Level	Numeric	The percentage of pest infestation in the wheat field (0-100%).
Market Price per Ton	Numeric	The price per ton of wheat in the market in USD per ton.
Production Cost ($/ha)	Numeric	The total cost of production per hectare, including labor, equipment, irrigation, etc in USD per hectare.
Farmer Satisfaction Score	Categorical	The satisfaction level of the farmer regarding the wheat production. Values: 1 (Very Unsatisfied) to 5 (Very Satisfied).

Table 1: Data Dictionary

1.3 Imports

1.3.1 Import libraries

Libraries such as Pandas, Numpy, Matplotlib, Seaborn and Sci-kit Learn were used. (McKinney, 2010)

1.3.2 Data import and load dataset

The dataset was loaded and the first five rows reviewed.

	0	1	2	3	4
Region	Europe	North America	Asia	North America	North America
Weather Conditions	Cloudy	Cloudy	Rainy	Sunny	Rainy
Soil Type	Silt	Saline	Saline	Loamy	Sandy
Fertilizer Usage (kg/ha)	231.288932	97.797675	134.565776	127.174686	286.758997
Crop Yield (tons/ha)	5.254826	2.897609	4.284138	4.486818	6.522978
Pest Infestation Level (%)	18.267238	13.959887	57.647244	76.188361	78.150565
Market Price per Ton ($)	401.139462	393.143675	450.476991	389.11384	192.736517
Production Cost ($/ha)	1525.682301	1629.678271	1591.178587	1804.424687	784.16515
Farmer Satisfaction Score	5	2	1	3	5

Table 2: First 5 rows of the dataset

1.4 Data Exploration (Initial EDA and data cleaning/validation)

We perform initial EDA to understand more about the variables and clean and validate the data (Ghosh et al., 2018).

1.4.1 Gather descriptive stats about the numerical data

	count	mean	std	min	25%	50%	75%	max
Fertilizer Usage (kg/ha)	898.0	177.316221	72.951801	50.047100	113.642189	181.511382	240.425548	299.928451
Crop Yield (tons/ha)	897.0	5.071745	1.699855	2.012111	3.668775	5.129596	6.522404	7.978390
Pest Infestation Level (%)	898.0	47.744286	29.074580	0.018653	22.619756	46.394047	73.022586	99.762282
Market Price per Ton ($)	897.0	298.814258	113.704917	100.830607	202.037373	300.416095	396.919503	499.619799
Production Cost ($/ha)	899.0	1240.919957	441.624968	500.941275	841.753127	1246.274640	1600.994657	1999.773867
Farmer Satisfaction Score	900.0	2.900000	1.434906	1.000000	2.000000	3.000000	4.000000	5.000000

Table 3: Descriptive Statistics about the numerical data

1.4.2 Gather basic information of the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 900 entries, 0 to 899
Data columns (total 9 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Region                      900 non-null    object
 1   Weather Conditions          900 non-null    object
 2   Soil Type                   900 non-null    object
 3   Fertilizer Usage (kg/ha)    898 non-null    float64
 4   Crop Yield (tons/ha)        897 non-null    float64
 5   Pest Infestation Level (%)  898 non-null    float64
 6   Market Price per Ton ($)    897 non-null    float64
 7   Production Cost ($/ha)      899 non-null    float64
 8   Farmer Satisfaction Score   900 non-null    int64
dtypes: float64(5), int64(1), object(3)
memory usage: 63.4+ KB
```

Table 4: Basic information of the data

1.4.3 Check missing values and Duplicate

10 missing values (<1%) were dropped, leaving 890 complete records. Since they are only less than 1% of total observations and it won't affect the overall distribution of the data (Ghosh et al., 2018).

1.4.4 Validate the data

All variable types and values were validated against data dictionary.

1.4.5 Check outliers

Outliers in numerical variables were checked using boxplots.

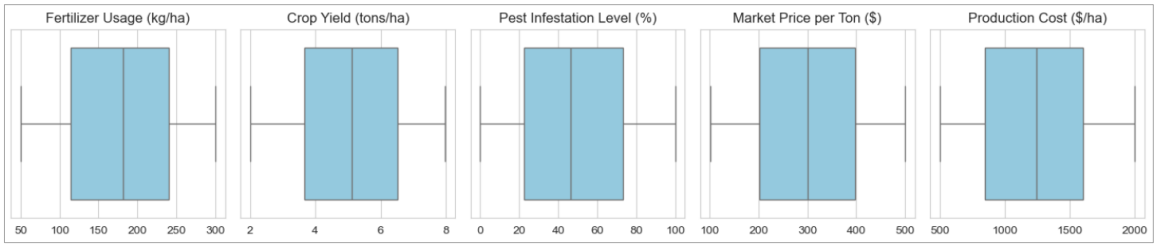

Figure 2: Boxplots of the numerical variables

1.4.6 Rename the columns

Columns were renamed for clarity.

Original Column Name	Renamed Column
Fertilizer Usage (kg/ha)	Fertilizer
Crop Yield (tons/ha)	Yield
Pest Infestation Level (%)	Pest
Market Price per Ton ($)	Price
Production Cost ($/ha)	Cost
Farmer Satisfaction Score	Satisfaction
Region	Region
Weather Conditions	Weather
Soil Type	Soil

Table 5: Renaming of the column names

1.5 Summary of Characterization of the dataset

All categorical and numerical fields fall within expected ranges.

Region: 6 (Europe, North America, Asia, Australia, Africa, South America).

Weather: 5 (Cloudy, Rainy, Sunny, Drought, Stormy).

Soil Type: 5 types (Silt, Saline, Loamy, Sandy, Peaty).

Farmer Satisfaction: 1–5 scale.

Fertilizer: ranges approx. from 50 to 300.

Crop Yield (tons/ha): ranges approx. from 2 to 8.

Pest Infestation Level (%): ranges 0–100%.

Market Price per Ton ($): ranges approx. from 100 to 500.

Production Cost ($/ha): ranges approx. from 500 to 2,000.

Overall, No unexpected values or outliers in base variables.

2. PACE: ANALYZE STAGE

2.1 Feature Engineering

New features were created to better capture efficiency, profitability, and risk. This also enhances model interpretability, drives dimensionality reduction and supports targeted decision-making across production, marketing and extension services (Guyon & Elisseeff, 2003; Kuhn & Johnson, 2013).

- **Fertilizer Efficiency** (*Yield_per_Fertilizer*): Yield per fertilizer applied.
- **Cost Efficiency** (*Cost_per_Ton*): Cost per ton of wheat.
- **Profitability** (*Profit_per_ha*): Net profit per hectare.
- **Return on Investment** (*Return_on_investment*): Profit as a fraction of cost.
- **Pest Damage Estimate** (*Pest Damage Estimate*): Estimated tonnage lost to pests.
- **Yield After Pest Impact** (*Yield After Pest*): Net yield after pest losses.
- **Cost-to-Price Ratio** (*Cost-to-Price Ratio*): Production cost relative to revenue.

2.2 Exploratory Data Analysis (EDA) and Statistical Analysis

We perform EDA and statistical analysis to analyze the relationship between variables.

2.2.1 Region vs Key Metrics

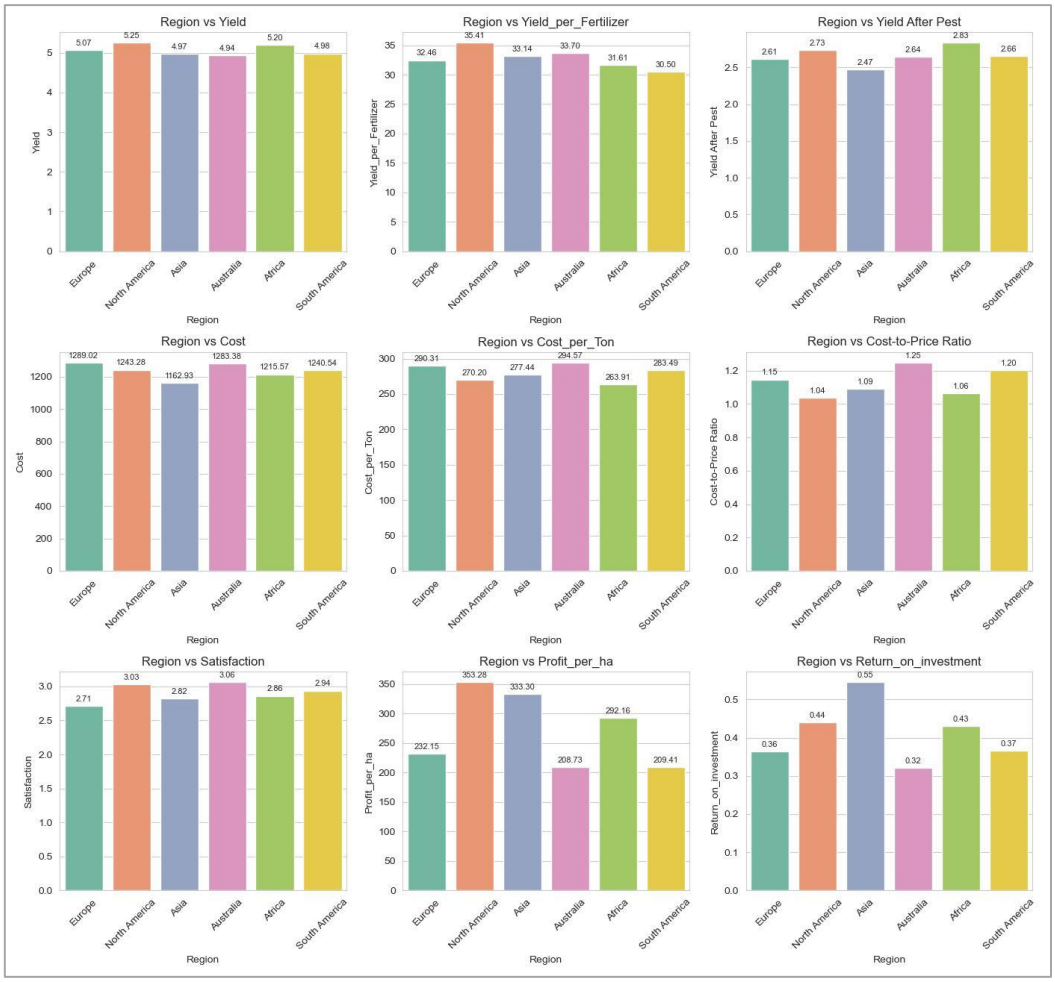

Figure 3: Region vs Key Metrics

Region-Level Insights

- **North America**: Highest yield (5.25 t/ha), fertilizer efficiency, and profit per hectare.
- **Africa:** High net yield after pests, lowest cost per ton, robust ROI.
- **Asia**: Moderate yields, highest ROI, strong profit per hectare.
- **Australia:** Highest farmer satisfaction, but high costs and lowest ROI.
- **Europe:** Moderate yields, highest costs, lowest satisfaction and ROI.

2.2.2 Soil vs Key Metrics

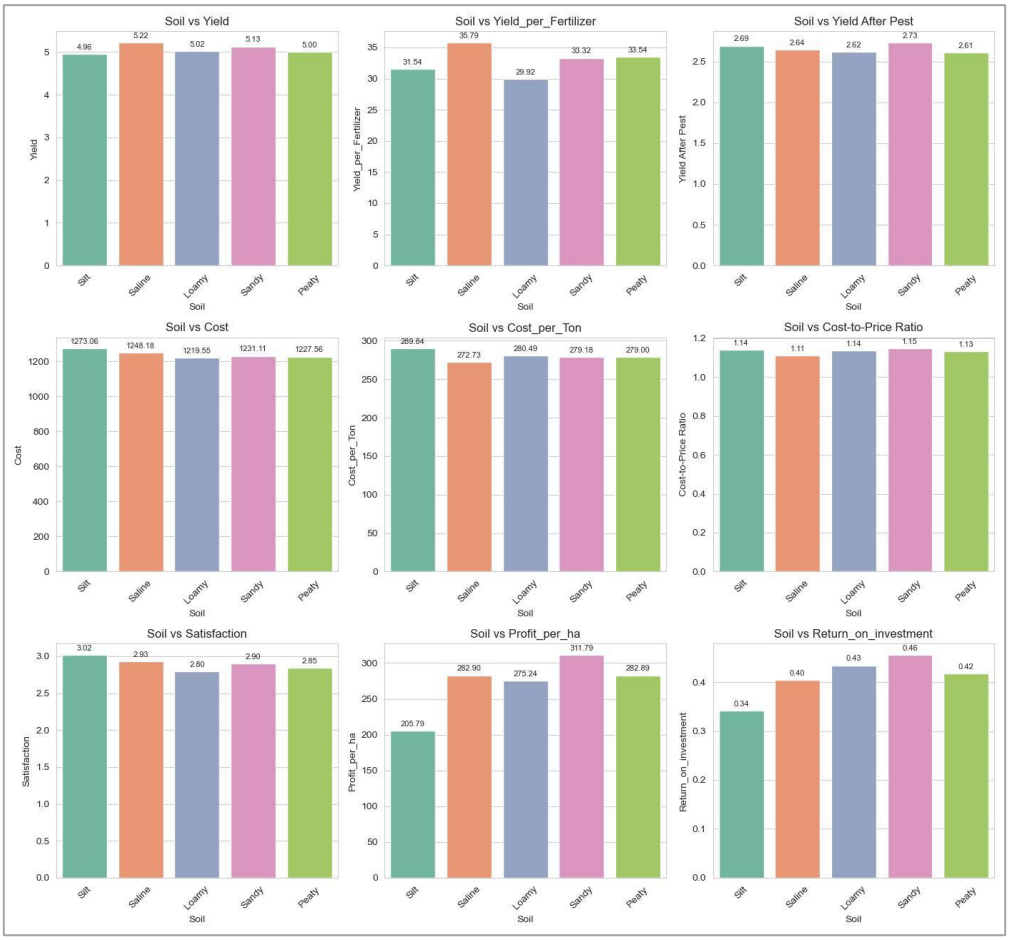

Figure 4: Soil vs Key Metrics

Soil-Type Comparisons

- **Silt:** Slightly below average yield, highest cost per ton, lowest profit and ROI, but highest satisfaction.
- **Saline:** Highest yield and fertilizer efficiency, lowest cost per ton, solid profitability.
- **Loamy:** Average yields, lowest fertilizer efficiency, moderate profit and ROI.
- **Sandy:** High yields and fertilizer efficiency, best profit and ROI.
- **Peaty:** Moderate yields and efficiency, average costs and profits

2.2.3 Weather vs Key Metrics

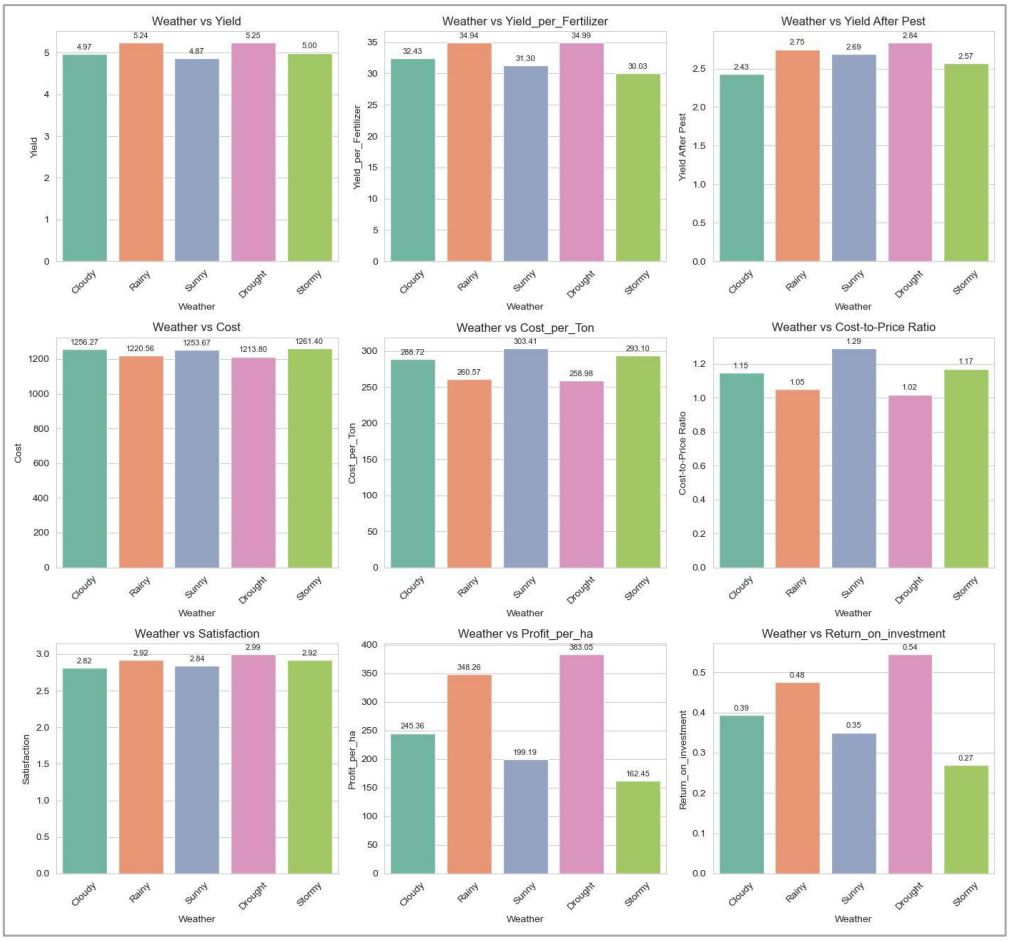

Figure 5: Weather vs Key Metrics

Weather-Condition Effects

- **Drought and Rainy:** Strong yields (>5.2 t/ha), profits($348+/ha), and ROIs (0.48–0.54).
- **Sunny and Stormy:** Lower yields (<5 t/ha), lower efficiency, profits, and ROI.
- Pest-adjusted yields highest in **Drought** and **Rainy** conditions.
- Cost per ton lowest under **Drought** and **Rainy**.

Statistical Techniques for Data Analytics

2.2.3 Descriptive Analysis

This section summarizes key characteristics of the wheat production dataset including central tendency, variability, probability distributions and normality (Field, 2013).

2.2.3a Central Tendency Metrics

Central Tendency Metrics

	mean	median	mode
Fertilizer	177.50	181.69	50.05
Yield	5.07	5.13	4.24
Pest	47.63	46.32	55.28
Price	298.89	300.12	100.83
Cost	1240.55	1246.06	510.56
Satisfaction	2.90	3.00	1.00
Yield_per_Fertilizer	32.79	26.78	6.13
Cost_per_Ton	280.23	242.83	67.02
Profit_per_ha	270.45	145.43	-1685.51
Return_on_investment	0.41	0.14	-0.87
Yield After Pest	2.66	2.39	0.01
Cost-to-Price Ratio	1.13	0.88	0.16

Table 6: Central Tendency Metrics

2.2.3b Variation Metrics

Variation Metrics

	std	var	min	max	IQR
Fertilizer	72.89	5313.54	50.05	299.93	126.71
Yield	1.70	2.87	2.01	7.98	2.85
Pest	29.06	844.62	0.02	99.76	50.40
Price	113.71	12930.72	100.83	499.62	194.71
Cost	440.97	194457.32	500.94	1999.77	753.09
Satisfaction	1.44	2.07	1.00	5.00	2.00
Yield_per_Fertilizer	22.31	497.64	6.13	139.30	22.98
Cost_per_Ton	154.79	23959.34	67.02	922.28	183.57
Profit_per_ha	887.52	787691.31	-1685.51	2654.88	1296.84
Return_on_investment	0.95	0.90	-0.87	5.15	1.14
Yield After Pest	1.79	3.19	0.01	7.97	2.73
Cost-to-Price Ratio	0.93	0.87	0.16	7.70	0.86

Table 7: Variation Metrics

2.2.3c Normality and Skewness

- Skewness measures the asymmetry of a distribution and kurtosis measures the heaviness of its tails (normal distribution, kurtosis=0) (Kim, 2013).
- Most variables are symmetric with moderate spread.
- Profit per hectare is highly variable due to outliers.
- Since data is normal in null hypothesis, all numeric columns deviate significantly from normality (Shapiro-Wilk $p < 0.05$, reject null hypothesis).
- QQPlots and Histograms are used for graphical visualization.

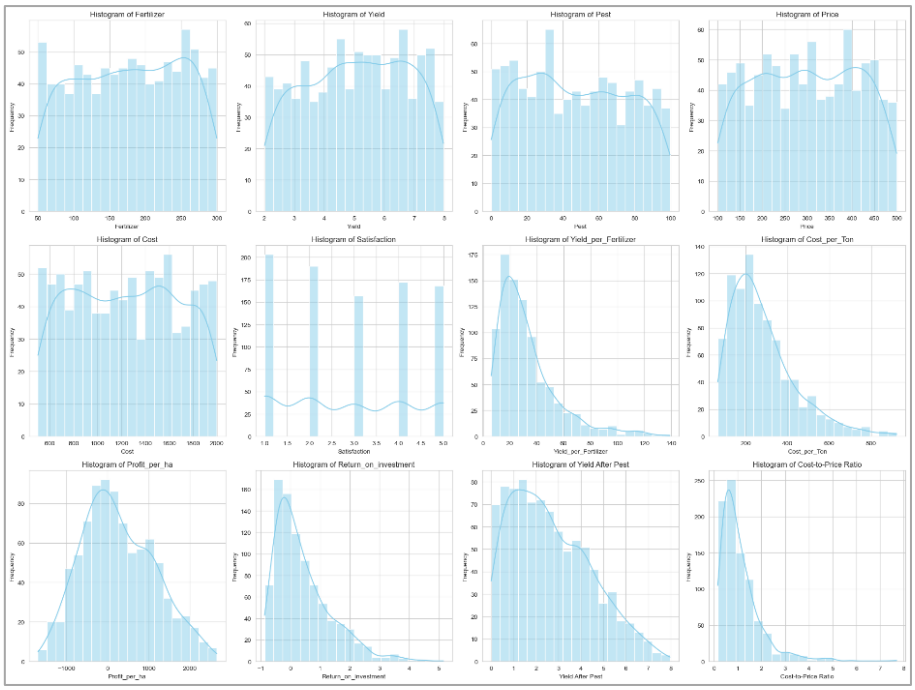

Figure 6: Histograms of key variables

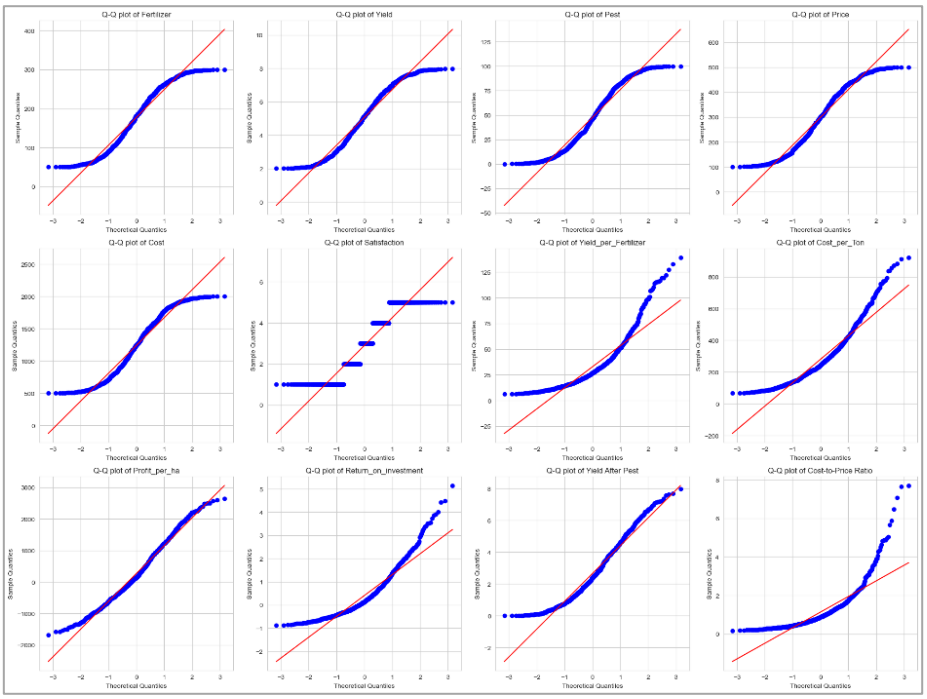

Figure 7: QQ Plots of key variables

Feature	Skewness	Kurtosis	Shapiro-W	Shapiro-p	Normality (p > 0.05)
Fertilizer	-0.0677	-1.2086	0.9527	0.0	No
Yield	-0.0862	-1.1385	0.9590	0.0	No
Pest	0.1009	-1.2114	0.9516	0.0	No
Price	-0.0159	-1.1874	0.9566	0.0	No
Cost	0.0152	-1.2173	0.9518	0.0	No
Satisfaction	0.0941	-1.3429	0.8808	0.0	No
Yield_per_Fertilizer	1.7247	3.4262	0.8416	0.0	No
Cost_per_Ton	1.2541	1.6840	0.9054	0.0	No
Profit_per_ha	0.3419	-0.4666	0.9840	0.0	No
Return_on_investment	1.3569	2.0826	0.8892	0.0	No
Yield After Pest	0.5790	-0.4387	0.9552	0.0	No
Cost-to-Price Ratio	2.7387	11.1917	0.7519	0.0	No

Table 8: Skewness, Kurtosis and Shapiro-Wilk Test Results

2.2.3d Outlier Analysis

Outliers can be detected by the Interquartile Range (IQR) method (**Q1−1.5IQR** or above **Q3+1.5IQR)** and boxplots (outliers, min, Q1, median, Q3, max, outliers) (Tukey, 1977). No outliers in base variables; derived metrics (profit, ROI, etc.) have outliers reflecting extreme cases.

```
=== Outlier Analysis ===
                      Count Percentage
Fertilizer                0       0.0%
Yield                     0       0.0%
Pest                      0       0.0%
Price                     0       0.0%
Cost                      0       0.0%
Satisfaction              0       0.0%
Yield_per_Fertilizer     48       5.4%
Cost_per_Ton             34       3.8%
Profit_per_ha             0       0.0%
Return_on_investment     28       3.1%
Yield After Pest          0       0.0%
Cost-to-Price Ratio      50       5.6%
```

Table 9: Outlier Percentage

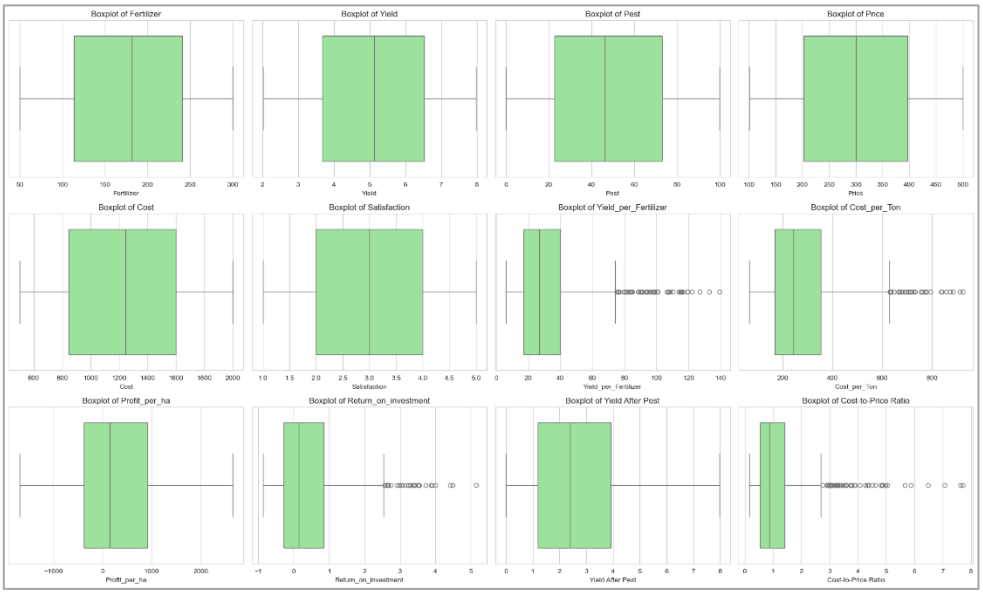

Figure 8: Boxplots of key variables

2.2.4 Confidence Interval

Compared North America and Australia on profit per hectare. They have substantial sample sizes in the data and represent distinct geographical/agricultural conditions.

2.2.4a Compute Sample Mean and Standard Error

We use the t-distribution because the population variance is unknown and sample sizes more than 30 and are moderate. The standard error formula for sample(ddof=n-1) is

$$SE = \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation (Moore, McCabe & Craig, 2017).

2.2.4b Compute 95% Confidence Interval

Interval calculation: The 95% CI for each mean is:

$$CI95\% = (mean\ A\ \pm t^* \times SE)$$

2.2.4c Compare Confidence Intervals

Comparing the two 95% CIs tells us about potential differences in population means. The 95% confidence intervals are:

North America: [218.50, 488.06]

Australia: [49.59, 367.87]

95% CIs overlap, so no statistically significant difference in mean profit per hectare between the two regions. To draw a stronger inference, a formal hypothesis test would be appropriate to determine.

2.2.5 Inferential Statistics

Given none of the variables are normally distributed and assumption is violated for parametric tests (t-test, ANOVA), we conduct the following three non-parametric hypothesis tests. Significance level is set at 0.05 for null hypothesis (Gibbons & Chakraborti, 2011).

Yield vs Region

ANOVA non-parametric alternative Kruskal-Wallis test (p = 0.486) shows no significant difference in median yield across regions.

Satisfaction vs Yield

```
Level 1 vs 2: U = 19032.000, p = 0.822
Level 1 vs 3: U = 15875.000, p = 0.951
Level 1 vs 4: U = 18272.000, p = 0.437
Level 1 vs 5: U = 16501.000, p = 0.592
Level 2 vs 3: U = 15090.000, p = 0.851
Level 2 vs 4: U = 17344.000, p = 0.313
Level 2 vs 5: U = 15662.000, p = 0.761
Level 3 vs 4: U = 14086.000, p = 0.498
Level 3 vs 5: U = 12798.000, p = 0.645
Level 4 vs 5: U = 13382.000, p = 0.240
```

Table 10: Mann-Whitney U Test Results

Mann-Whitney U tests show fail to reject and no significant differences in yield across satisfaction levels.

Soil vs Yield

Kruskal-Wallis test (p = 0.529) shows no significant difference in median yield across soils.

2.2.6 Correlation Analysis

To identify potential predictors for our target variable, we start by calculating Pearson correlations between numerical features and yield (Benesty et al., 2009).

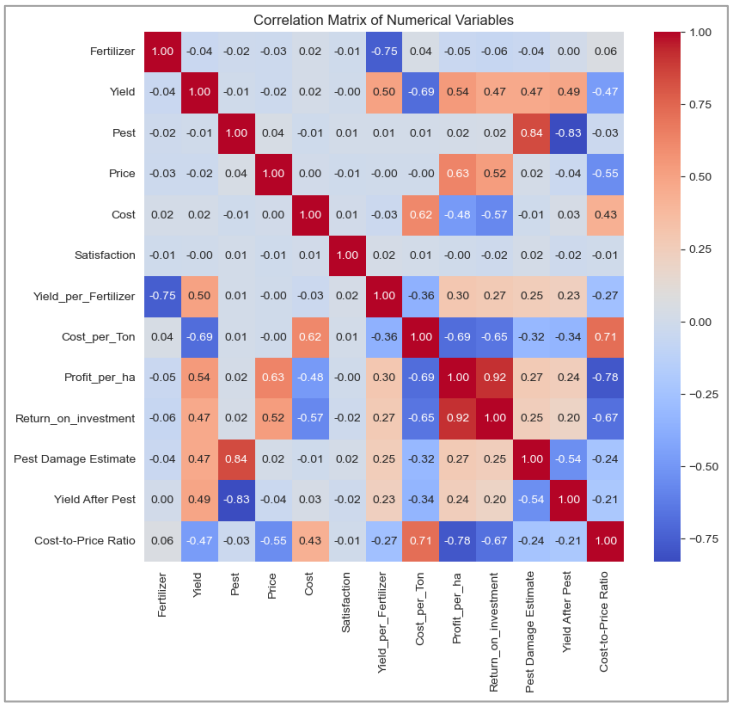

Figure 9:Correlation Headmap of numerical variables

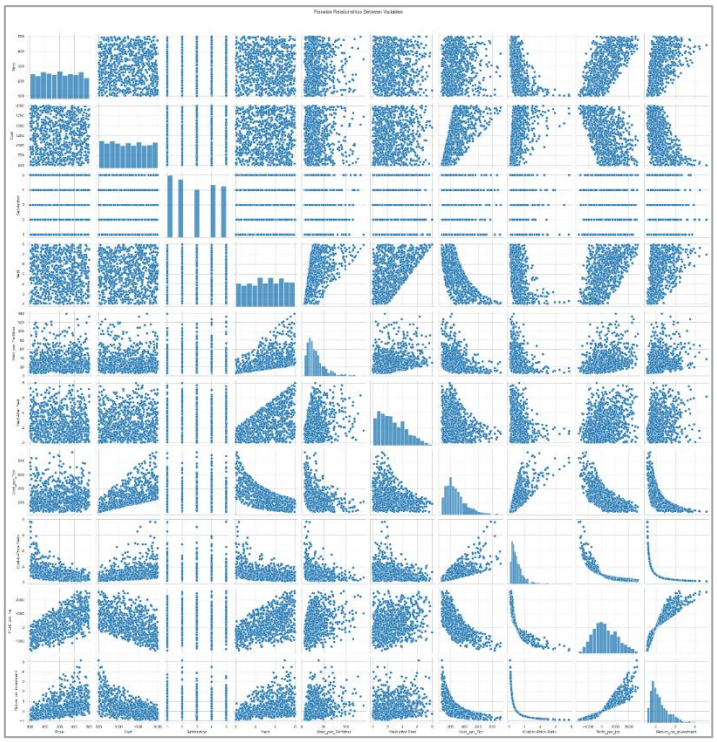

Figure 10: Pairplots of variables

- Fertilizer efficiency is inversely correlated with fertilizer use and moderately with yield.
- Cost per ton is negatively correlated with yield and positively with cost-to-price ratio.
- Profit per hectare and ROI are highly correlated.

2.2.7 Chi-square Test of Association for Categorical Features

```
Chi-Square Test Results:
Region: χ² = 7.80, p = 0.6484
Weather: χ² = 12.35, p = 0.1364
Soil: χ² = 7.78, p = 0.4550
```

Table 11:Chi-square Test Results

To examine whether categorical variables have statistically significant association with Yield Category, chi-square tests of independence were conducted (Agresti, 2013) and alpha value is set at 0.05.

No significant association between Region, Weather, or Soil and Yield Category (Low, Medium, High).

2.2.8 Baseline Linear Regression Model with original features

For initial linear regression model, we use the original features in the dataset first to explore the relationships and to prevent data leakage and overly optimistic performance.

OLS Regression Results

Dep. Variable:	Yield	R-squared:	0.019
Model:	OLS	Adj. R-squared:	-0.001
Method:	Least Squares	F-statistic:	0.9422
Date:	Thu, 15 May 2025	Prob (F-statistic):	0.527
Time:	12:11:37	Log-Likelihood:	-1723.5
No. Observations:	890	AIC:	3485.
Df Residuals:	871	BIC:	3576.
Df Model:	18		
Covariance Type:	nonrobust		

Table 12:OLS Regression Results

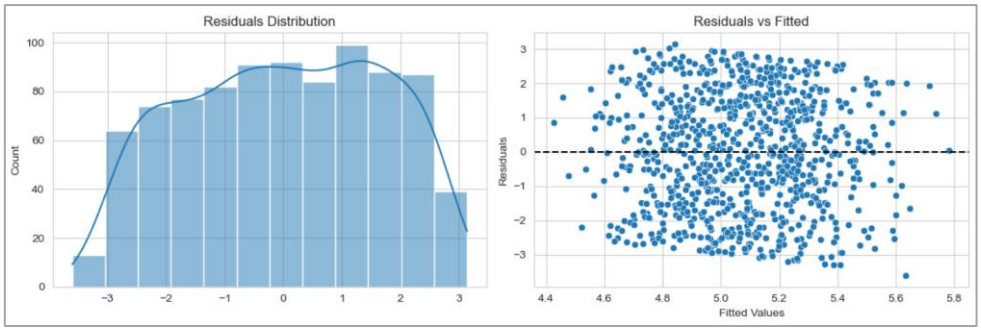

Figure 11:Residual Plots

Regression using original features yields low R^2 (0.019), indicating poor predictive power and the model explains only 2% of the total variance. Residuals are symmetrically distributed but scattered, showing the need for better features and non-linear models (Random Forest). Thus, we may need to use different approach (Classification) for predicting the level of Yield.

2.3 Data Encoding

Most machine learning models (especially linear regression, tree-based models and neural networks) **do not understand raw categorical data**. They require numerical inputs to perform mathematical operations. Encoding allows us to **transform categorical variables into numerical representations** (Zheng & Casari, 2018).

- Categorical variables were encoded for modeling.
- Weather and Soil were ordinally encoded based on agronomic relevance.
- Region was one-hot encoded.

2.4 Feature Scaling

StandardScaler was used to ensure all features are on comparable scales. Centering ensures that 0 represents the average level and unit variance makes each step in the ordinal scale equally weighted (James et al., 2013).

2.5 Separate Features and Target

Features (X) and target (Yield) were separated. Derived metrics based on Yield were excluded to prevent data leakage and wrong conclusion.

For PCA and LDA, we create target classes by also splitting into 3 splits by 33rd quantile and 66th quantile as Low, Medium and High.

2.6 Dimensionality Reduction Analysis: PCA vs LDA

The wheat production dataset along with new derived features contains numerous variables across multiple dimensions making visualization and analysis challenging. Thus in this report, we explore the possibility of dimensionality reduction techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) and we can identify underlying data patterns while preserving essential information (Jolliffe & Cadima, 2016).

2.6.1 Principal Component Analysis (PCA)

Unsupervised, transforms original features into new set of uncorrelated variables (principal components) and reduces dimensionality by maximizing variance. First two components explain 27.92% of variance only; 12 of 13 needed for 99.5% demonstrating complexity of underlying data structure.

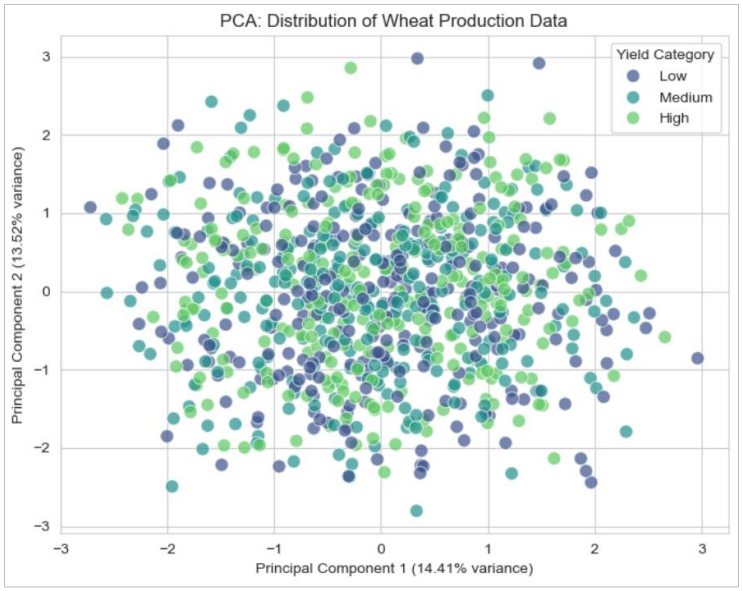

Figure 12: PCA 1 vs PCA 2

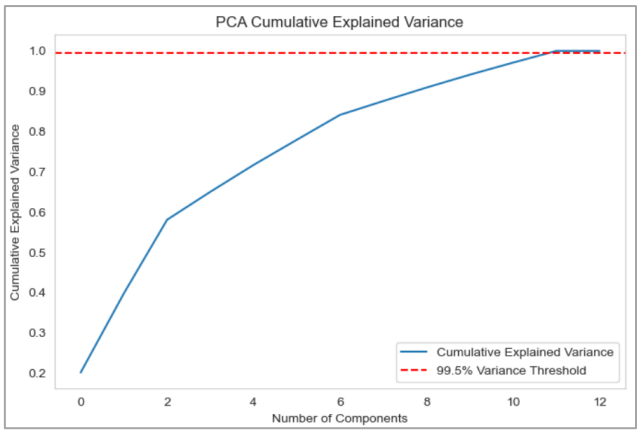

Figure 13: PCA Cumulative Explained Variance

2.6.2 Linear Discriminant Analysis (LDA)

Supervised, maximizes class separation while minimizing within-class variance. Unlike PCA, LDA explicitly uses class labels to find the optimal projection for classification. First discriminant explains 68.7% of variance and second 31.3%.

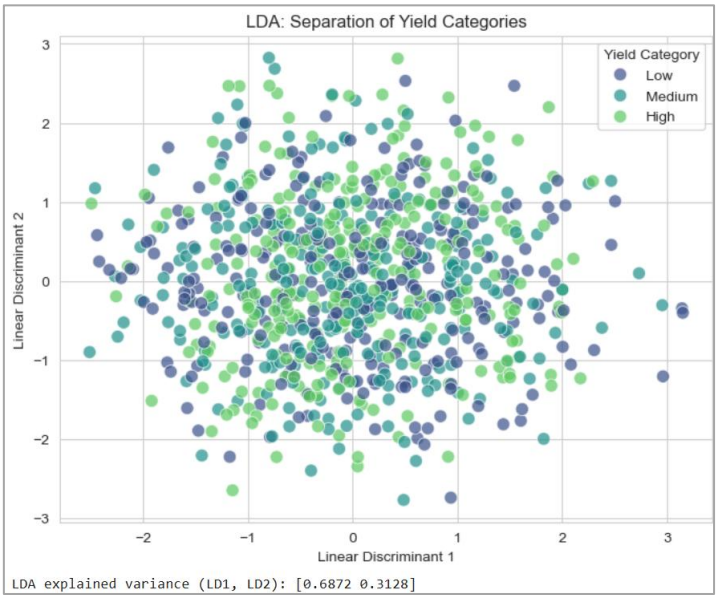

Figure 14: LDA 1 vs LDA 2

2.6.3 Visualization Comparison

Both PCA and LDA visualizations show substantial overlap with no clear boundary between yield categories, indicating weak class separation. Though LDA theoretically better suited for categorical distinction, the plot reveals minimal practical separation across all three yield classes.

2.6.4 Classification Performance

Classification accuracy is low: LDA-based Random Forest (35%), PCA-based (30%).

2.6.5 Key Differences Between PCA and LDA

PCA and LDA have several key differences and use cases as stated below.

Objective: PCA maximizes variance explanation whereas LDA maximizes class separation

Supervision: PCA works without labels (unsupervised) while LDA requires class labels (supervised)

Data transformation: PCA changes both the shape and location of the original data while LDA preserves location but enhances class boundaries

Application focus: PCA excels at general dimensionality reduction for visualization and noise reduction and LDA specializes in classification preparation

Interpretability: LDA components are explicitly oriented toward class labels (3 Yield bins), so their axes often have a clear meaning in terms of class separation. PCA's axes capture variance that might or might not align with the target hence interpreting them requires post-hoc analysis of loadings. If the classes are known and important to show, LDA can make them visually distinct. But if we want to explore the data structure without assuming classes, PCA provides an unbiased summary.

2.6.6 Implications for Analysis and Modeling

- Neither PCA nor LDA provided features with strong class-separating power for yield due to weak relationships between features.
- Regression is more appropriate for the continuous yield target.
- LDA is not suitable for regression, as it requires discrete labels and does not optimize for continuous outcomes.
- Underlying factors influencing wheat yield are complex, possibly non-linear, and distributed across many dimensions. Simple linear reductions may not capture these relationships effectively.
- Feature engineering and selection are critical for improving predictive performance.

3. PACE – CONSTRUCT STAGE

3.1 Machine Learning Approach

- Labeled data and a clear target variable (Yield) make supervised learning optimal.
- Predictive objective: Yield amount level prediction (classification – Low, Medium, High) derived from Yield by quantiles.
- Supervised models provide feature importance for actionable insights.
- Unsupervised learning is better for exploratory tasks like clustering, not direct yield prediction.
- No labels for validation in unsupervised learning (James et al., 2013).

3.1.1 Pros and Cons of Supervised and Unsupervised Machine Learning

Supervised Learning

Pros	Cons
High accuracy with labeled data	Requires extensive labeled data
Clear performance metrics (accuracy, F1-score)	Prone to overfitting with noisy data
Feature importance for actionable insights	Computationally intensive for large datasets

Table 13:Pros and Cons of Supervised Learning

Unsupervised Learning

Pros	Cons
Discovers hidden patterns (e.g., soil clusters)	No objective metrics for validation
Works with unlabeled data	Sensitive to noise/feature scaling
Reduces dimensionality (e.g., PCA)	Results are harder to interpret

Table 14:Pros and Cons of Unsupervised Learning

3.1.2 Features Selection

Independent variables: Fertilizer, Pest, Price, Cost, Region, Weather, Soil

Dependent variables: Yield Category (Low, Medium, High)

Derived metrics like Yield_per_Fertilizer or Cost_per_Ton are excluded to prevent data leakage.

3.1.3 Recommended Models

Since the features are weakly correlated and have no multicollinearity issue as we have seen in the Linear Regression model and statistical tests above, we use the following models (Chen & Guestrin, 2016).

Random Forest: Robust to outliers, handles numerical and categorical features well and non-linear relationships.

Gradient Boosting (XGBoost): Handles mixed types of features, captures complex non-linear interactions, high accuracy with feature interactions.

Logistic Regression: strong, interpretable baseline for multi-class classification. Useful for understanding the direction and strength of relationships between predictors and yield categories.

3.2 Modelling

3.2.1 Separate into 3 Training and Test sets of 10%, 20% and 30% Test

Split the data into 3 sets as follows and stratify the data so that class distributions is consistent across train/test sets. We use this method so that the model that performs well across different splits is likely to be more robust.

- 90% training and 10% test
- 80% training and 20% test
- 70% training and 30% test

3.2.2 Conduct Stratified 10-fold cross-validation of baseline model with 3 splits

We conduct **stratified 10-fold** and **cross-validation with 3 splits** using scikit learn's StratifiedKFold and cross_val_score. This ensures class distributions remain consistent across splits while balancing computational efficiency and robustness (Pedregosa et al., 2011).

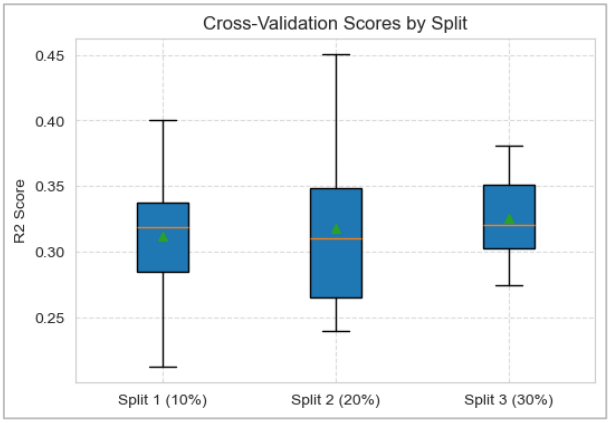

Figure 15: Cross-validation scores by split

Among all splits, Split 3 with a 30% test size achieved the highest mean cross-validation accuracy score of 0.326, indicating more stable and consistent model performance. This suggests that using a larger test set may provide a better estimate of generalization for this dataset.

3.2.3 Test the best split of 30% with holdout set to get baseline accuracy

Next, we test the best split with the actual test holdout set to compare the results. It looks like we have a poor baseline accuracy score of 35.2%. However, let's try hyperparameter tuning to see if we can improve the model performance more.

3.3 Hyperparameter Tuning

Hyperparameter tuning optimizes model performance by finding ideal parameter combinations for accurately predicting crop yield categories. This systematic approach improves accuracy scores across diverse growing conditions, enhances model generalization to future seasons, and ensures reliable

identification of high-yield scenarios. Using **GridSearchCV** with 10-fold cross-validation delivers robust models ready for real-world agricultural planning and resource optimization. Parameters are carefully selected to balance underfitting and overfitting, model interpretability and complexity, performance and computational time (Sowmya & Prasad, 2024).

Model	Best Parameters	Best CV Score
Random Forest	max_depth: 10, max_features: 5, min_samples_leaf: 4, min_samples_split: 2, n_estimators: 100	0.385
XGBoost	gamma: 0, learning_rate: 0.3, max_depth: 3, min_child_weight: 3, n_estimators: 200	0.397

Table 15:Hyperparameter results

4. PACE: EXECUTE STAGE

4.1 Evaluation

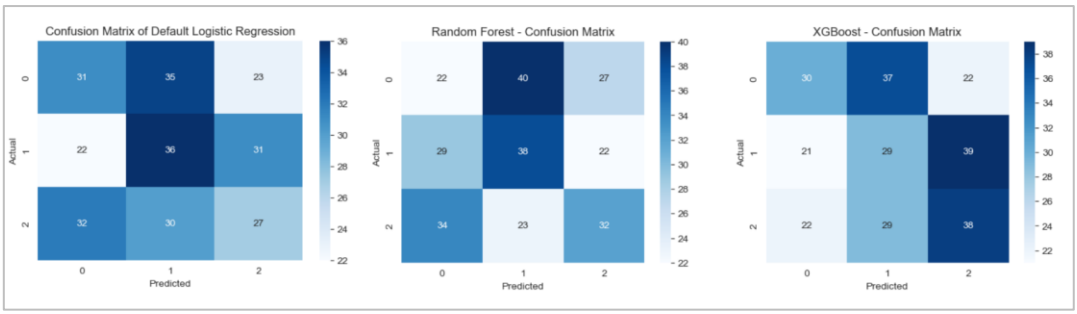

Figure 16: Confusion Matrices

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.352	0.351	0.352	0.351
Random Forest	0.345	0.343	0.345	0.343
XGBoost	0.363	0.367	0.363	0.363

Table 16:Results of the models

The confusion matrices show that all three models—Logistic Regression, Random Forest, and XGBoost struggle to accurately classify yield categories, with substantial misclassification across all classes. XGBoost achieves the highest accuracy (36.3%) and F1 score (0.36), slightly outperforming Random Forest (34.5% accuracy, 0.34 F1). Most predictions cluster around the central class, indicating difficulty distinguishing between categories. Overall, model performance is low, suggesting that the current features may lack strong predictive power for yield classification. Further feature engineering or alternative modeling approaches may be needed to improve results (Li et al., 2023).

4.2 Feature Importance

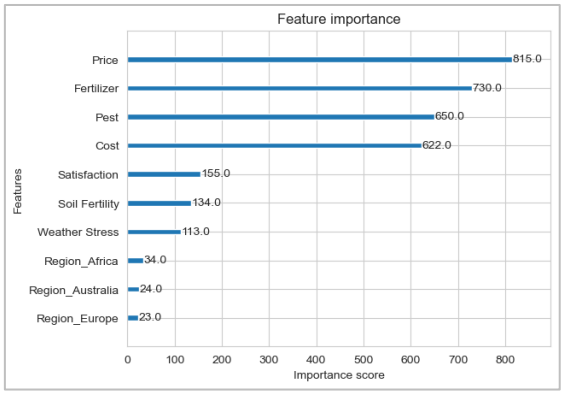

Figure 17: Feature Importance Graph

The feature importance chart shows that **Price**, **Fertilizer**, **Pest**, and **Cost** are the most influential variables in predicting the target outcome. Regional variables (Africa, Australia, Europe) contribute minimally to the model, suggesting that yield prediction is driven more by management and input variables than by geographic location in this dataset.

4.3 Recommendations to Improve Poor Model Performance

1. **Apply Advanced Feature Selection:**
   Use hybrid or embedded feature selection techniques (rank-based, or weighted rank-based algorithms) to identify and retain only the most relevant predictors, removing noisy or redundant features that can hinder model accuracy (Guha & Jabi, 2022).
2. **Enhance Data Quality and Diversity:**
   Increase sample size, incorporate data from multiple seasons and regions, and integrate additional variables (e.g., remote sensing, phenological, or environmental data) to improve generalizability and reduce overfitting (Thenkabail et al., 2021).
3. **Optimize Preprocessing Pipelines:**
   Implement data normalization, handle multicollinearity, and use transformation techniques (e.g., polynomial or power transforms) to better capture underlying patterns and manage outliers (Banachewicz & Massaron, 2022).
4. **Prevent Overfitting and Validate Robustly:**
   Employ cross-validation, data augmentation, and explainable AI techniques to ensure models generalize well to unseen data, and routinely analyze errors to guide further improvements (Arrieta et al., 2020; Pedregosa et al., 2011).

By focusing on these strategies, we can significantly enhance model accuracy, robustness, and reliability for agricultural yield prediction.

REFERENCES

Yaragal, S. (2023). PACE Framework for Effective GIS Data Analysis. LinkedIn.

McKinney, W., 2010. Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference, pp.56–61.

Ghosh, A., Nashaat, M., Miller, J., Quader, S. and Marston, C., 2018. A comprehensive review of tools for exploratory analysis of tabular industrial datasets. Vis. Inform., 2, pp.235–253. https://doi.org/10.1016/j.visinf.2018.12.004

Guyon, I. & Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research, 3, pp.1157–1182.

Kuhn, M. & Johnson, K., 2013. Applied Predictive Modeling. New York: Springer.

Field, A., 2013. Discovering Statistics Using IBM SPSS Statistics. 4th ed. London: Sage Publications.

Kim, H.Y., 2013. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restorative Dentistry & Endodontics, 38(1), pp.52–54.

Tukey, J.W., 1977. Exploratory Data Analysis. Reading, MA: Addison-Wesley.

Moore, D.S., McCabe, G.P. & Craig, B.A., 2017. Introduction to the Practice of Statistics. 9th ed. New York: W.H. Freeman and Company.

Gibbons, J.D. & Chakraborti, S., 2011. Nonparametric Statistical Inference. 5th ed. Boca Raton: CRC Press.

Benesty, J., Chen, J., Huang, Y. & Cohen, I., 2009. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Berlin: Springer, pp.1–4.

Agresti, A., 2013. *Categorical Data Analysis*. 3rd ed. Hoboken: Wiley.

Zheng, A. & Casari, A., 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. Sebastopol: O'Reilly Media.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R. New York: Springer.

Jolliffe, I.T. & Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202.

Chen, T. & Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785–794.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

Sowmya, P. & Prasad, A.V.K., 2024. Optimization of Crop Yield Prediction Models using Hyperparameter Tuning and Ensemble Learning. Nanotechnology Perceptions, 20(S7), pp.653–665.

Li, Y., Qian, J., Zhang, L. & Chen, Y., 2023. Challenges and opportunities in crop yield prediction: A machine learning perspective. *Computers and Electronics in Agriculture*, 214, p.108305. https://doi.org/10.1016/j.compag.2023.108305

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. & Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, pp.82–115.

Banachewicz, K. & Massaron, L., 2022. *The Kaggle Book: Data Analysis and Machine Learning for Competitive Data Science*. Birmingham: Packt Publishing.

Guha, R. & Jabi, M., 2022. A comparative study of filter, wrapper, and embedded feature selection techniques for classification problems. *Pattern Recognition Letters*, 159, pp.1–8.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

Thenkabail, P.S., Lyon, J.G. & Huete, A., 2021. *Remote Sensing of Water Resources, Disasters, and Urban Studies*. Boca Raton: CRC Press.