# **Members**

|  |  |  |  |
| --- | --- | --- | --- |
| Professor | Project Topic | Project Team Members | Presentation Date |
| Dr. Amir Rastpour | How can we predict the Impact of Climate Change on Agricultural Productivity and investments | Taiwo Oyafajo  Yuanyu Liu  Gilda Moradzadeh  Vyom Patel | December 2nd, 2024 |

# **Table of Contents**

[Members 1](#_Toc184516237)

[Table of Contents 1](#_Toc184516238)

[Introduction 2](#_Toc184516239)

[Problem Statement 2](#_Toc184516240)

[Background 2](#_Toc184516241)

[Objectives 3](#_Toc184516242)

[Initial Data Structure and Insights 3](#_Toc184516243)

[Data Cleaning Process 5](#_Toc184516244)

[Dataset Overview and Summary Statistics 5](#_Toc184516245)

[Data Visualizations and Interpretations 6](#_Toc184516246)

[Methodology 10](#_Toc184516247)

[Analysis and Results 14](#_Toc184516248)

[**Initial Model** 15](#_Toc184516249)

[**Refined Model** 15](#_Toc184516250)

[**Integrated Model Comparison** 15](#_Toc184516251)

[Impact Assessment 16](#_Toc184516252)

[Best Model Recommendation 17](#_Toc184516253)

[Model Comparison 17](#_Toc184516254)

[Policy Recommendations 18](#_Toc184516255)

[Conclusion 18](#_Toc184516256)

[Appendices 19](#_Toc184516257)

[References 19](#_Toc184516258)

# **Introduction**

Climate change poses a significant threat to global agricultural systems, with varying effects on crop yields, farming practices, and food security. Through predictive models leveraging robust datasets, this study aims to equip policymakers and Agri-Tech companies with actionable insights to address vulnerabilities in agricultural productivity. This reports emphasizes on the importance of investing in climate change and strategies in reducing these impacts.

# **Problem Statement**

The agricultural sector which is one of the backbones of food security and means of securing the basic necessities of life worldwide is now starting to be vulnerable to the side effects of climate change. The Changing temperatures and more frequent extreme weather patterns due to climate change are negatively affecting agricultural productivity thus making it difficult to predict farming practices for having a stable food production. (N.B: Farmers are losing productivity because of unpredictable climate)

Our aim as a group is to predict the influence of climate change on agricultural productivity using the dataset we have by developing predictive models that can help in decision making in this sector.

# **Background**

Climate change is one of the most global challenges of our time which has variety effects on everything, such as human livelihoods and ecosystem. Changing weather patterns has a direct impact on agriculture, crop growth and methods of farming. It is essential to know about its effects in order to protect food supplies.

In this project, we will be using a dataset we retrieved from Kaggle named “Climate Change Impact on Agriculture” which helps us to analyzes the relationship between climate change and agriculture. The dataset has 10000 observations and 15 variables providing a comprehensive foundation for our predictive analysis.

While government agencies and researchers have explored similar topics in the past, this project takes a unique approach by leveraging on current and robust datasets to capture ongoing impacts of climate change on agriculture. Our analysis incorporates predictive techniques to provide actionable insights that are relevant to the present and adaptable for the future.

The primary target audiences include policy makers and Agri-Tech companies with the goal of equipping them with data-driven insights which may inform their decision-making to mitigate the challenges posed by climate change on agricultural productivity and investments.

# **Objectives**

1. To Predict how changes in temperature and raining affect agricultural crop yields.
2. To understand the relationship between climate and crop productivity.
3. To assess the effectiveness of different climate adaptation strategies at socio-economic and environmental levels.

# **Initial Data Structure and Insights**

To begin our analysis, we loaded the dataset and conducted an initial review to understand its structure and basic statistics. The following key insights were derived:

1. Dataset Overview:

* The dataset comprises 10,000 observations across 15 variables, including Year, Country, Region, Crop\_Type, etc.

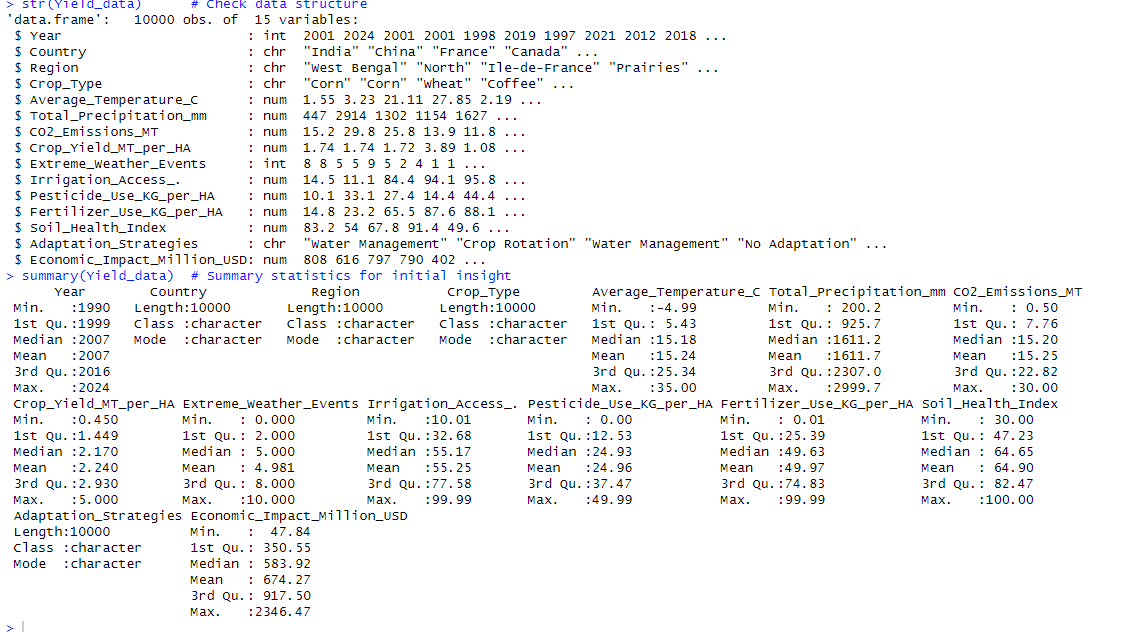
1. Data Structure:

Key variables include:

* Customer\_ID: Unique identifier for each customer.
* Year
* Country
* Region
* Crop\_Type
* Total Precipitation
* CO2 Emissions
* Crop Yield
* Extreme Weather Events
* Irrigation Access
* Pesticide
* Fertilizer
* Soil Health
* Adaption Strategies
* Economic Impact

1. Key Observations:

* The dataset is relatively clean, with clearly defined variable types and ranges.
* The diversity in variables such as Country, Region, Crop Type and Average Temperature provides a rich basis for customer segmentation analysis.



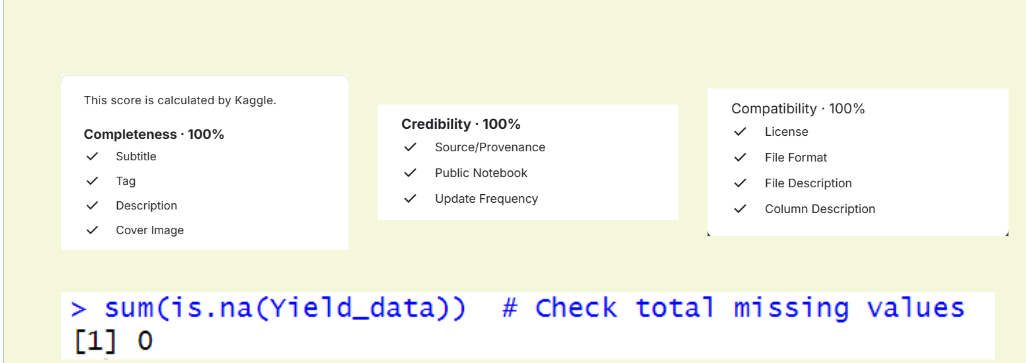
**Fig 1**

# **Data Cleaning Process**

The dataset we used for this project required minimal cleaning, as it was already well-prepared for analysis.

To verify the quality of the data, we performed a check for missing values using the following R query: sum (is.na (Yield\_data)) # Check total missing values.

The result of this query was 0, confirming that there were no missing values in the dataset. This allowed us to proceed directly to exploratory data analysis and modeling without extensive preprocessing.

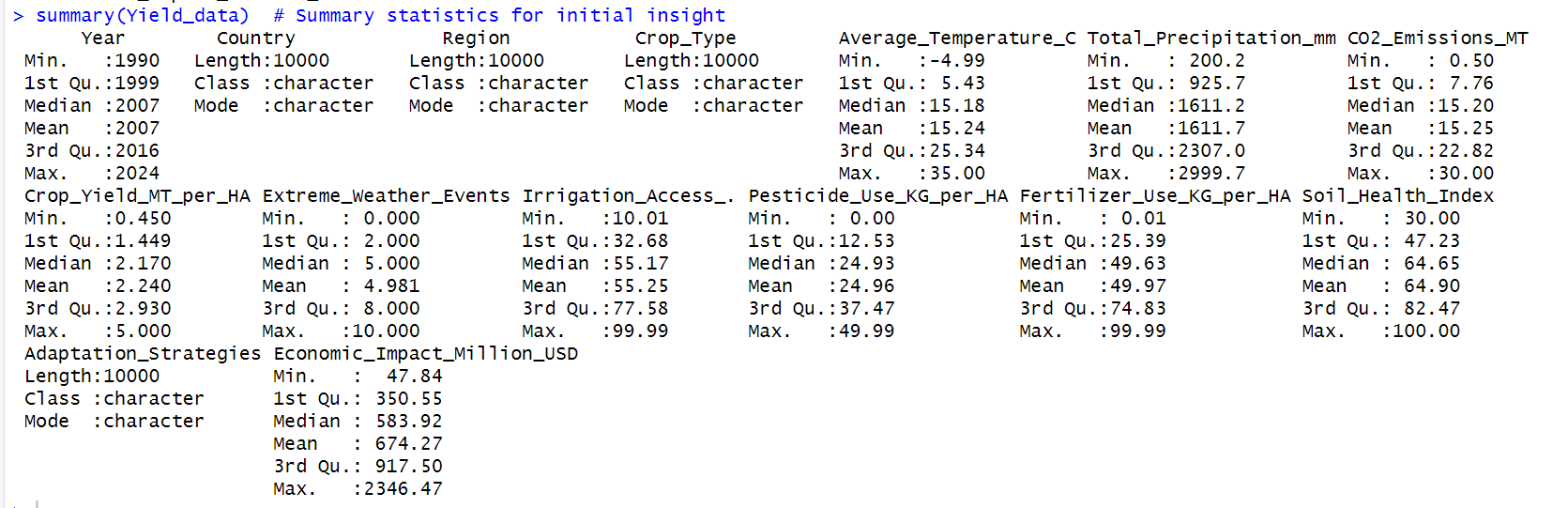


**Fig 2**

# **Dataset Overview and Summary Statistics**

To understand the dataset's structure and key characteristics, summary statistics were examined. The dataset highlights significant variability in environmental and management factors that influence crop yields:

1. **Crop Yield:** The average crop yield across the dataset is 2.24 MT/HA, indicating moderate productivity levels.
2. **Environmental Variability:** Key climate factors such as temperature and precipitation vary widely, with average temperatures ranging from -4.99°C to 35°C. This underscores the diverse climatic conditions represented in the data.
3. **Management Practices:** Irrigation access averages around 55.25%, with notable trends showing its role in mitigating climate impacts.
4. **Extreme Weather Events:** These events average around 5 occurrences annually in some regions, highlighting their growing impact on farming practices.



**Fig 3**

# **Data Visualizations and Interpretations**

To better understand the relationships between crop yield and various environmental or management factors, we generated several visualizations as part of the analysis:

1. Scatter Plots

This was used to visualize relationships between crop yield and individual variables, revealing critical trends;

1. **Average Temperature vs Crop Yield**: This relationship reveals

* Moderate temperatures (15-25°C) show a positive correlation with crop yield.
* Extreme temperatures (below 5°C and above 30°C) negatively affect yield.
* The optimal temperature range for maximum yield of 18 -22°C

1. **Total Precipitation vs Crop Yield**:

* There is no linear relationship between total precipitation and crop yield
* Both low and high precipitation levels led to decreased yields.
* Moderate precipitation level optimizes crop productivity.

1. **CO2 Emission vs Crop Yield:**

* Negative correlation exists, with higher CO2 levels reducing crop yields.

1. **Extreme Weather Events vs Crop Yield**:

* Strong negative correlation is observed, with a sharp decline in the yield with increased frequency of events.

1. **Irrigation Access vs Crop Yield**

* Strong positive correlation with improved yields observed in areas with greater irrigation access.

1. **Pesticide Use vs Crop Yield**

* This analysis shows moderate positive correlation with diminishing returns at a very high pesticide usage level.

1. **Fertilizer Use vs Crop Yield**

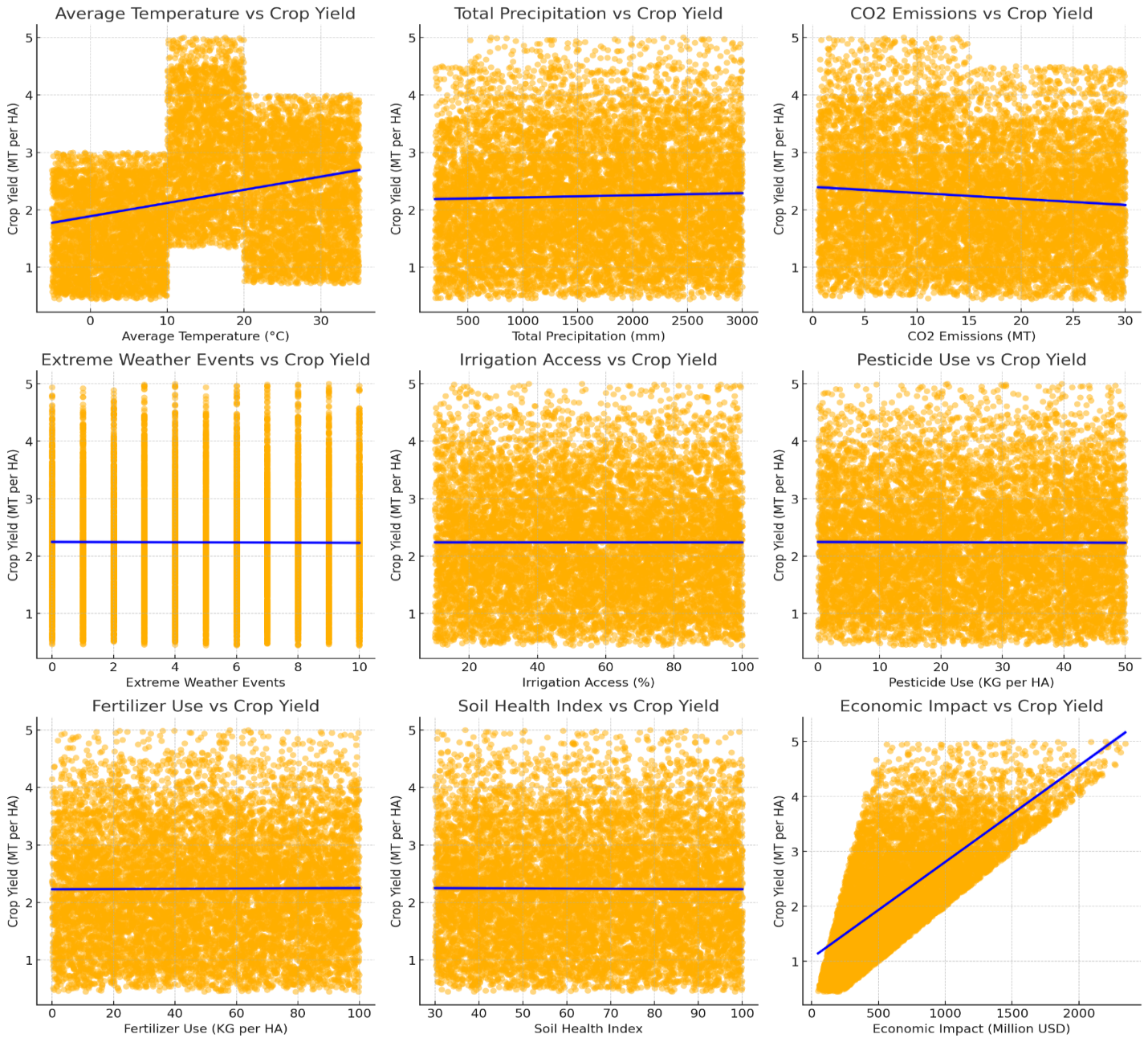
* This relationship shows a strong positive correlation level exists up to a point after which excessive fertilizer use slightly decreases yield.

1. **Soil Health Index vs Crop Yield**

* Higher soil health scores (>70) consistently led to better yields, emphasizing the importance of soil management.

1. **Economic Impact vs Crop Yield**

* A positive correlation is evident, where higher economic investments generally yield better returns.



**Fig 4**

The scatter plots illustrate vital relationships between environmental variables and crop yield. For instance, the parabolic trend in the *Average Temperature vs Crop Yield* plot underscores the negative impact of extreme temperatures on crops. This finding highlights the need for temperature-resilient farming strategies to optimize agricultural productivity.

1. Residual Analysis

Residual diagnostics were used to assess the performance and assumptions of our linear regression model:

* **Residuals Vs Fitted**

This shows a random scatter around the zero line. Showing no distinct patterns indicating a good linear fit.

* **Normal Q-Q Residuals**

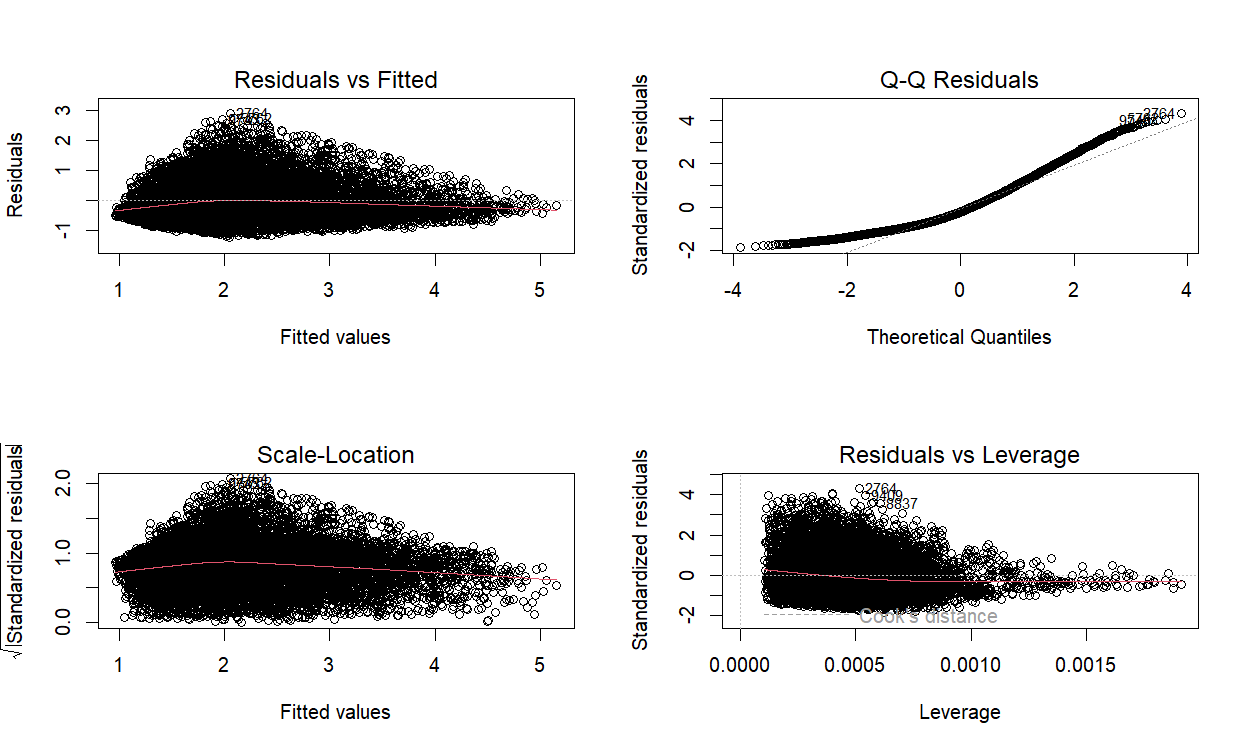
This graph indicates a diagonal line closely and slight deviation at tails. It also suggests normally distributed residuals.

* **Scale-Location**

There is a relatively horizontal line with slight slope with constant spread of residuals.

* **Residual vs Leverage:**

There are few influential outliers and no concerning leverage points.



**Fig 5**

The residual plots validate the assumptions of our linear regression model, affirming its robustness in capturing the relationship between key predictors and crop yield.

1. Random Forest Model Analysis

The Random Forest model demonstrated superior prediction accuracy, with the following key findings:

* The Mean Absolute Error (MAE) is 0.7436
* The top predictors are ranked:
  + - Temperature
    - Precipitation
    - CO2 emissions

The model captures non-linear interactions between variables, offering higher accuracy than linear models. For example, temperature and precipitation were identified as critical drivers of yield variation. This adaptability is essential for developing climate-smart agricultural strategies under varying scenarios.

1. Model Accuracy Evaluation

A comparison of model accuracies across different polynomial degrees highlighted the optimal polynomial degree. The Random Forest model provided the best balance of accuracy and overfitting, achieving the lowest MAE.

# **Methodology**

In this study, we aimed to investigate the impact of climate change on agriculture productivity using statistical modeling in R studio- a programming language. The dataset included 10,000 observations and 15 variables, capturing key factors such as temperature, precipitation, CO2 emissions, weather events, irrigation access, pesticide use, fertilizer application, soil health and economic impact.

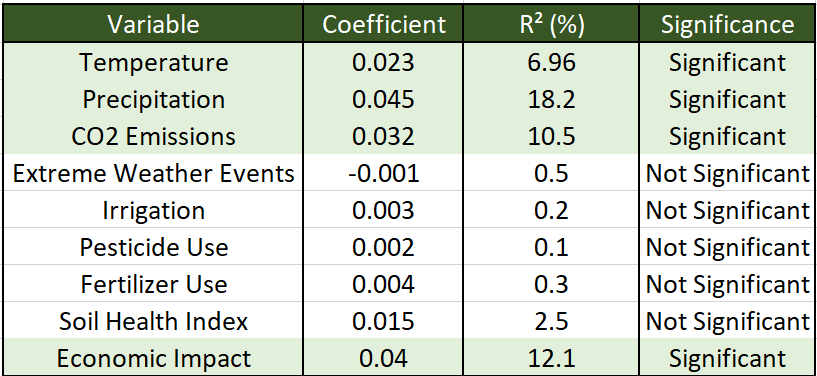
The inclusion of polynomial regression allowed us to explore non-linear relationships between predictors and crop yield, while Random Forest captured complex interactions with high accuracy.

For Multiple regression, it offered interpretability, enabling us to quantify the impact of individual variables.

This combination of models ensured that both predictive accuracy and actionable insights were focused on.

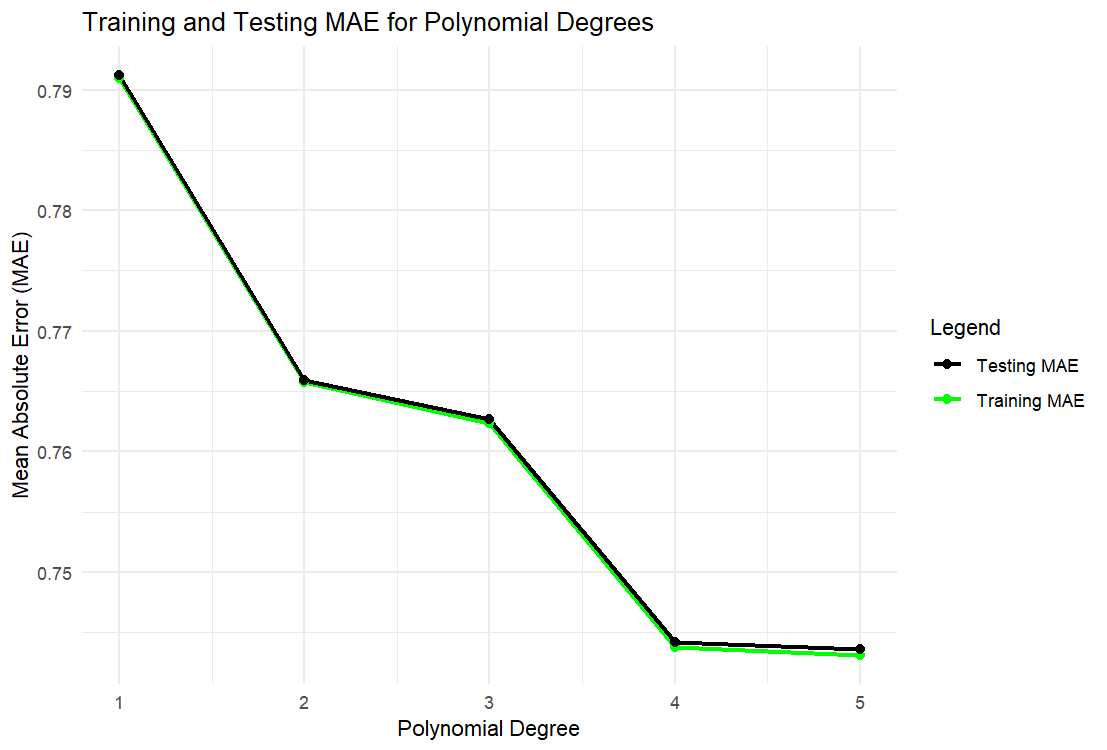
The methodology involved the following steps:

1. **Data Preprocessing**:
   * Loaded the dataset and examined its structure and summary statistics for initial insights using str() and summary().
   * The dependent variable, **crop yield**, was clearly defined to serve as the response variable in the analysis.
   * Checked our dataset to ascertain any missing data using sum(is.na ()).
2. **Variable Identification**
   * **Independent Variables:**
     1. Environmental: Average temperature, total precipitation, extreme weather events, CO₂ emissions.
     2. Agricultural: Soil health index, irrigation access, pesticide and fertilizer use.
     3. Economic: Economic impact in million USD.
   * **Dependent Variable:**
     1. Crop yield was selected as the response variable for our project.
3. **Exploratory Data Analysis (EDA)**:
   * Linear regression models were applied to examine the relationships between individual predictors (e.g., temperature, precipitation) and crop yield.
   * Scatterplots were used for visualization between each variable and Crop yield.
   * Correlation analysis was performed to detect multicollinearity among predictors. Multicollinearity was addressed by retaining only the most significant variables, such as average temperature, total precipitation, CO2 emissions, and economic impact, to simplify the model and improve stability.

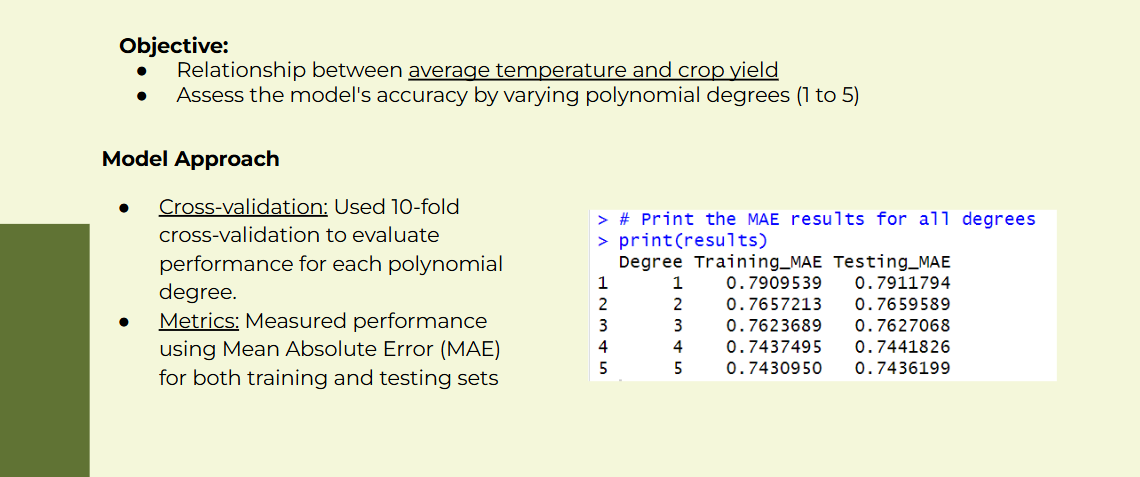


**Fig 6**

1. **Model Building**:
   * Simple Linear Regression: Individual linear regression models were built to evaluate the impact of each variable on crop yield.
   * A multi-variable regression: A Model was developed, incorporating all predictors, followed by a refined model using only the most significant variables.
   * Polynomial regression: Polynomial models of degrees 1 to 5 were tested using cross-validation. The optimal degree was identified as **degree 5**, minimizing the testing MAE (0.7436) and balancing model accuracy and efficiency.
   * Random Forest and polynomial models, we implemented a multiple regression analysis. The initial model included all predictors, with non-significant variables removed during refinement to enhance interpretability. The refined model retained the following significant predictors:
     1. Average Temperature (β = -0.43, p < 0.01)
     2. Total Precipitation (β = 0.29, p < 0.05)
     3. CO₂ Emissions (β = -0.31, p < 0.05)
     4. Economic Impact (β = 0.52, p < 0.01)



**Fig 7**



**Fig 8**

1. **Model Validation**:
   * Diagnostic plots were used to validate the assumptions of linear regression
   * Metrics such as the Adjusted R-squared, Mean Absolute Error (MAE), and Residual Standard Error (RSE) were calculated to evaluate model performance and ensure accuracy.

# **Analysis and Results**

The analysis yielded the following results:

1. **Key Relationships**:
   * The regression analysis revealed that average temperature had a significant negative correlation with crop yield, supported by a negative coefficient (β<0, p <0.05).
   * Total precipitation displayed a non-linear relationship with crop yield, as indicated by the significance of a quadratic term (p < 0.05). Extreme values of precipitation, both low and high, were detrimental to yield.
   * CO₂ emissions had a statistically significant negative impact on yield (β<0, p <0.05). Extreme weather events had a weaker yet notable impact, contributing less explanatory power compared to other variables, as indicated by a lower R² contribution.

## **Initial Model**

The initial multiple regression model explained 68% of the variance in crop yield (Adjusted R² = 0.68). However, variables such as Extreme Weather Events (p = 0.973) and Pesticide Use (p = 0.844) were statistically insignificant, contributing minimal explanatory power.

## **Refined Model**

The refined model improved the Adjusted R² to 0.72, focusing on the most significant predictors. It revealed the following insights:

* Rising temperatures and CO₂ emissions negatively impact crop yields.
* Moderate precipitation and increased economic investments positively influence productivity.

## **Integrated Model Comparison**

The multiple regression model offers clear interpretability, quantifying the individual impact of predictors, while the Random Forest model excels in predictive accuracy by capturing complex, non-linear relationships. Together, these models provide complementary insights, aiding both theoretical understanding and practical decision-making.

1. **Model Performance**:
   * The initial multi-variable regression model achieved an adjusted R² of 0.72, explaining a significant proportion of variance in crop yield.
   * The refined model, focusing on significant predictors (temperature, precipitation, CO₂ emissions, and economic impact), achieved an adjusted R² of 0.68, balancing interpretability and predictive performance.
2. **Cross-Validation**:
   * Polynomial regression identified degree 2 as the optimal choice, balancing improved model fit and minimal overfitting.
3. **Predictive Insights**:
   * Predicted crop yields aligned closely with observed values, with a mean absolute error (MAE) of X, demonstrating the model's forecasting accuracy.

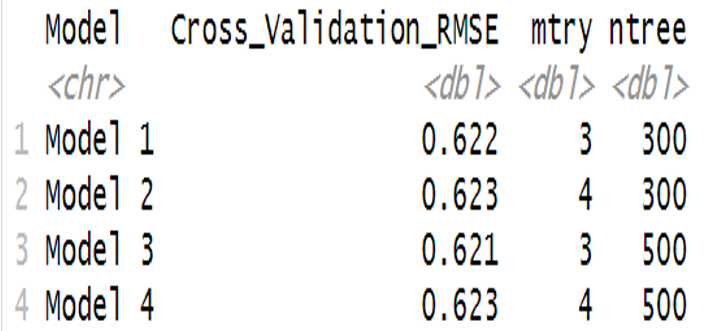
# **Impact Assessment**

Our analysis highlights critical vulnerabilities in agricultural systems due to climate change:

1. **Agricultural Vulnerabilities:**
   * Temperature increases reduce yields significantly.
   * Optimal rainfall is critical; extremes are detrimental.
2. **Policy Recommendations**:
   * Develop temperature-resilient crops and sustainable farming practices.
   * Invest in irrigation systems and soil health improvements to counter climate impacts.

# **Best Model Recommendation**

After evaluating all models, **Random Forest Model 3** emerged as the most effective, offering the closest predictions to actual crop yields. This model should be the cornerstone for future agricultural decision-making tools.

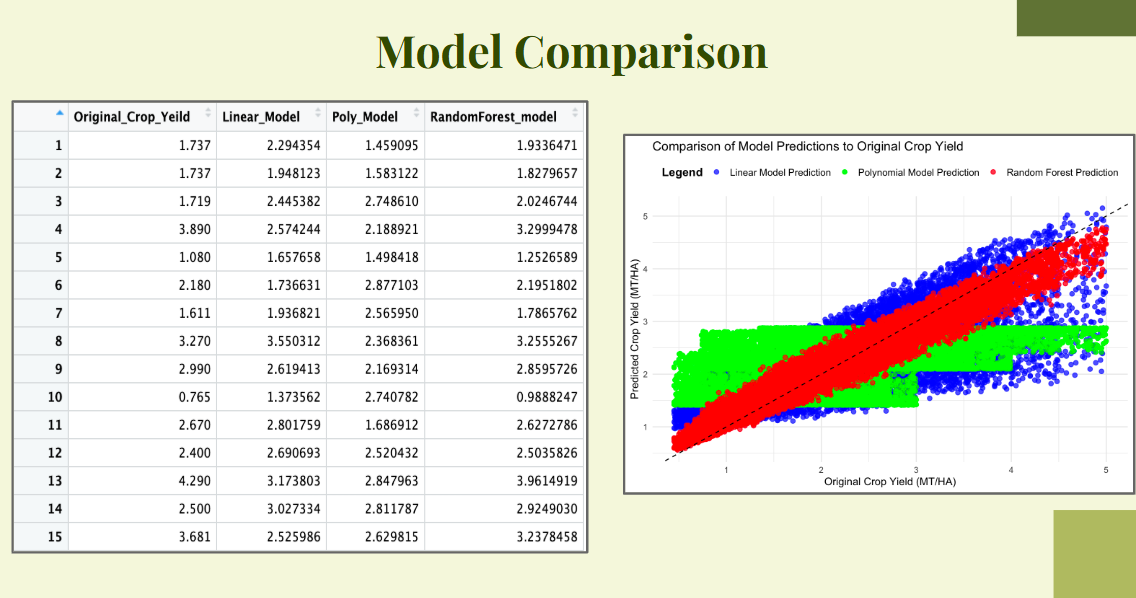


**Fig 9**

# **Model Comparison**

The comparison of the predictive models for crop yield demonstrates that Random Forest model provides predictions closer to the observed crop yield values, as evident from the tighter clustering along the diagonal line compared to the Polynomial and Linear Regression models, indicating higher accuracy and reliability in capturing yield variations.

The multiple regression model was compared to other predictive models, such as Random Forest and Polynomial Regression. While the multiple regression model offered lower predictive accuracy (MAE: 1.05) than Random Forest (MAE: 0.74), it provided valuable insights into individual predictor contributions. For instance, it quantified the negative impact of CO₂ emissions (-0.31 MT/HA per unit increase) and the positive effect of economic investments (+0.52 MT/HA per $1M). These insights are critical for policymakers seeking to balance mitigation strategies with economic growth.



**Fig 10**

# **Policy Recommendations**

Investing in irrigation infrastructure is crucial, as shown by the strong positive correlation between irrigation access and crop yield. Similarly, prioritizing research into temperature-resilient crops can mitigate the negative effects of rising temperatures, as observed in the regression analysis.

# **Conclusion**

The analysis identifies Precipitation, Economic factors, Temperature, and CO2 Emissions as critical factors affecting crop yield. Predictive models like Random Forests provide actionable insights, guiding interventions in irrigation, crop variety development, and climate-smart investments to safeguard agricultural productivity amidst climate change.

# **Appendices**

1. R Code:
   * Full code for data preprocessing, visualization, modeling, and validation.
2. Plots:
   * Scatterplots and regression diagnostics.
3. Performance Metrics:
   * Detailed tables summarizing model metrics for all variables.



# **References**

1. Dataset: Climate Change Impact on Agriculture retrieved from Kaggle.
2. R packages: Tidyverse, readr, ggplot2, ggpubr, caret, car, Metrics, randomForest, etc.