Table of Contents

[Abstract 1](#_Toc133253237)

[Acknowledgements 1](#_Toc133253238)

[Chapter 1: Introduction 1](#_Toc133253239)

[I. Overview 1](#_Toc133253240)

[II. Problem Statement 1](#_Toc133253241)

[III. Research Questions 1](#_Toc133253242)

[Chapter 2: Background 1](#_Toc133253243)

[I. Dataset 1](#_Toc133253244)

[II. Literature Review 1](#_Toc133253245)

[III. Research Gap 1](#_Toc133253246)

[IV. Algorithms 1](#_Toc133253247)

[a. TFIDF 1](#_Toc133253248)

[b. Count Vectorizer 1](#_Toc133253249)

[c. Logistic Regression 1](#_Toc133253250)

[d. SVM 1](#_Toc133253251)

[e. XGBoost 1](#_Toc133253252)

[f. Decision Tree 1](#_Toc133253253)

[g. Artificial Neural Network 1](#_Toc133253254)

[Chapter 3: Methodology 1](#_Toc133253255)

[I. Tools and Techniques 1](#_Toc133253256)

[II. EDA and Visualization 1](#_Toc133253257)

[Chapter 4: Results and Conclusion 1](#_Toc133253258)

[I. Critical Discussion 1](#_Toc133253259)

[II. Conclusion 1](#_Toc133253260)

[III. Future Work 1](#_Toc133253261)

[Chapter 5: Legal, Ethical and Professional Issues 1](#_Toc133253262)

[References 1](#_Toc133253263)

[Appendices 1](#_Toc133253264)

# Abstract

# Acknowledgements

# Chapter 1: Introduction

## Overview

In the present age the usage of pesticide is very common in agriculture. But there are many pros and cons in using pesticides on plants and agriculture. While they control a lot of insects and pests thereby increasing productivity, they also pose a risk to environment. Mainly polluting the water and land, also meddling with other species that don’t cause harm to agriculture. To avoid such scenarios ideal usage of pesticide is a major requirement. To address this issue this research can be of great help.

This project aims to tackle the prediction challenge of optimizing pesticide usage such that crop yields are maximized. Present methods are old and don’t take into account multiple factors and can’t handle complex conditions. This is where Machine Learning can be used to use its potential of handling complex patterns (*Machine learning: learn, develop, and evolve from data sets*, 2021) in the data with multiple factors affecting a single variable.

This project will explore many ML algorithms to funnel the best performing and most effective algorithm for predicting the optimal usage of pesticide. This project will also delve into the statistical analysis to understand the variability in crop yields across various factors using a statistical method called ANOVA (Singh, 2018).

## Research Questions

1. What machine learning models most effectively predict optimal pesticide levels for maximizing crop yields based on historical data?
2. Are there significant differences in crop yields between countries using an Analysis of Variance (ANOVA)?

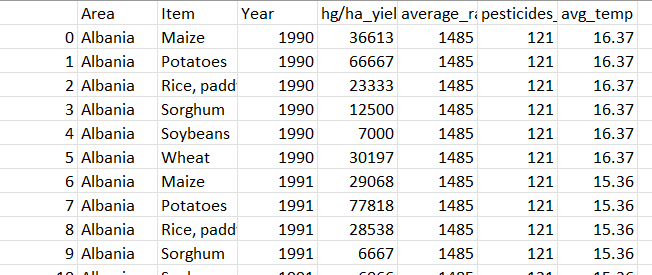
## Objectives

1. To identify and compare multiple Machine Learning algorithms like Support Vector Machines, Linear Regression, XGBoost, and Decision Trees to compare their performance in predicting optimal pesticide usage for maximizing crop yields.
2. Evaluating the above-mentioned algorithms using multiple performance metrics such as RMSE (Root Mean Squared Error) and R-Squared.
3. To develop a python pre-processing pipeline to clean and pre-process data to enrich the quality for predictive modelling.
4. To test the trained model on a holdout dataset that mimics the real-world data.

# Chapter 2: Background

## Dataset

The dataset used in this research provides an overview of agricultural yields along with other factors such as pesticide usage, rainfall, and temperature data over the years for multiple countries. The crops this dataset features are rice, potatoes, wheat, soybeans and more. This dataset contains almost 29k records spread across multiple years and countries along with 7 columns that contain geographical, rainfall and other features. Below is the image that shows the sample records in the data.



The data is taken from an open-source dataset website called Kaggle. All the datasets available in this website are free to use for non-commercial and purposes. Hence, ethical approval from any source is not needed.

## Literature Review

India's agriculture sector, crucial for the livelihood of two-thirds of its population, faces significant challenges with crop yield. Traditional crop production predictions have relied heavily on the expertise of individual farmers and specific crops, leading to inconsistent results. To address these challenges, various methodologies have been proposed, focusing on leveraging data analysis and machine learning techniques to predict crop yields more accurately. The proposed system aims to enhance crop yield prediction by employing supervised learning techniques through four main machine learning algorithms: Decision Tree, Naive Bayes, SVM, and Random Forest. The system focuses on identifying the best crop yields by analysing key factors such as nitrogen, phosphorus, sulphur, humidity, rainfall, and pH levels.

Data collection and preprocessing are crucial steps in this system, involving the gathering of data from various sources and the elimination of null and redundant values to ensure a clean and suitable dataset. The system then applies the four machine learning algorithms to the pre-processed data. The Decision Tree algorithm uses a tree-like structure to represent decisions and their possible consequences, making it suitable for both classification and regression problems. The Naive Bayes algorithm, based on Bayes theorem, assumes predictor independence and effectively handles large datasets with high dimensionality. Support Vector Machine (SVM) represents data points in an n-dimensional space and finds the hyperplane that best separates different classes, making it highly effective for classification tasks. Random Forest, an ensemble learning method, constructs multiple decision trees and merges them to improve accuracy and prevent overfitting, adeptly handling both categorical and continuous variables.

The dataset is divided into training and testing sets, typically using an 80-20 split, and each algorithm is applied to the training data, with performance evaluated on the test data. The algorithm that provides the highest accuracy is selected for final implementation. In this case, Random Forest is identified as the most accurate and is implemented using Python, integrated with a web interface developed using HTML, CSS, and Flask, allowing users to input data and receive crop yield predictions. The system incorporates additional internal attributes such as nitrogen, potassium, sulfur, humidity, and water level, critical for crop growth, achieving high prediction accuracy even with limited data. By using these factors and advanced machine learning algorithms, the system ensures that farmers can make informed decisions about crop selection, optimizing productivity and reducing the risk of crop failure. This machine learning-based system offers a reliable tool for farmers, demonstrating significant improvement over traditional methods and existing models, providing a more precise and efficient solution for crop yield prediction.

Agriculture is vital to the Indian economy, with over half the population dependent on it. Machine learning techniques can predict crop production based on parameters like rainfall and meteorological conditions. Random Forest, a powerful supervised machine learning algorithm, performs both classification and regression tasks, aiding crop selection to reduce yield losses. Environmental factors like weather and climate pose significant risks to agriculture's sustainability. Machine learning (ML) offers decision-support tools for Crop Yield Prediction (CYP), helping decisions about crop cultivation and management. The primary goal is to boost agricultural production using well-established models. Yield prediction is crucial for farmers to make informed decisions before planting crops, necessitating timely forecasts and analysis to maximize yield.

The proposed system uses machine learning to enhance crop yield predictions by analysing key factors such as rainfall, temperature, area (in hectares), season, and soil composition. Data is collected from various districts in India, with climate data sourced from government websites like data.gov.in and imd.gov.in. The system employs the Random Forest algorithm, a supervised learning technique effective for both classification and regression tasks.

The system begins with data collection and preprocessing, where data from multiple sources is cleaned to remove null and redundant values, ensuring a high-quality dataset for analysis. The Random Forest algorithm is then applied, generating decision trees from distinct data samples. Each tree provides predictions, and the final output is determined by majority voting, which enhances model accuracy. The pseudocode involves randomly selecting features, calculating nodes using the best split points, and creating multiple trees by repeating this process for each tree. For prediction, test features generate outputs from each tree, aggregate votes, and determine the most popular predicted outcome.

Model evaluation is conducted by training and testing the model using an 80-20 data split. The Random Forest algorithm's performance is evaluated against other algorithms to ensure the highest accuracy. The system effectively predicts crop yields by analysing multiple environmental factors, providing farmers with insights to optimize crop selection and maximize agricultural productivity. The user-friendly interface allows farmers to input climate data and receive crop yield predictions, aiding in informed decision-making. This approach leverages historical data to offer accurate, practical guidance for enhancing crop yield outcomes.

Accurate crop yield estimation is essential for strategic agricultural planning, including import-export policies and increasing farmers' incomes. Machine learning algorithms are crucial for crop yield prediction, addressing one of the agricultural sector's significant challenges. This article reviews the application of machine learning in predicting crop yields, particularly palm oil yield. It presents the current global status of palm oil yield, discusses widely used features and prediction algorithms, and critically evaluates state-of-the-art machine learning techniques in crop yield prediction. The review highlights the advantages and challenges of machine learning-based crop yield prediction, identifying current and future challenges in the agricultural industry and proposing potential solutions. The article emphasizes future perspectives on machine learning-based palm oil yield prediction, discussing remote sensing applications, plant growth and disease recognition, mapping, and tree counting. It proposes a prospective architecture for machine learning-based palm oil yield prediction, aiming to address new research challenges in crop yield prediction.

A lack of standardized feature sets is a significant challenge in crop yield prediction. Widely used features include climatic information, historical crop yield data, vegetation index, satellite data, soil properties, irrigation information, and crop management data. The optimal sub-features for specific crops under climatic information are not clearly identified, requiring further research. Satellite-based Solar-Induced Fluorescence (SIF) features have potential for improving yield prediction performance. Integrating multi-band satellite data with weather parameters can enhance crop yield forecasting at regional scales.

Various classification and regression algorithms, such as ANN, RF, LR, CNN, SVM, and LASSO, have been used for crop yield prediction. ANN and RF are the most utilized, but optimizing hyperparameters is crucial for improving performance. The review suggests that ensemble algorithms could increase prediction model robustness. In ecological studies, RF effectively handles variable collinearity, but may overfit predictions outside the training range.

Implementing new agricultural technologies is crucial for feeding a growing global population. Machine learning frameworks provide valuable insights by analyzing extensive data sets and predicting future outcomes. This review underscores the importance of selecting appropriate features and algorithms for crop yield prediction, particularly in palm oil yield. Further research with a larger number of features and diverse prediction algorithms is necessary to improve yield prediction accuracy and develop effective strategies for maximizing crop yields.

## Research Gap

## Algorithms

### Linear Regression

### Support Vector Machines

### Decision Trees

### XGBoost

### MSE/RMSE

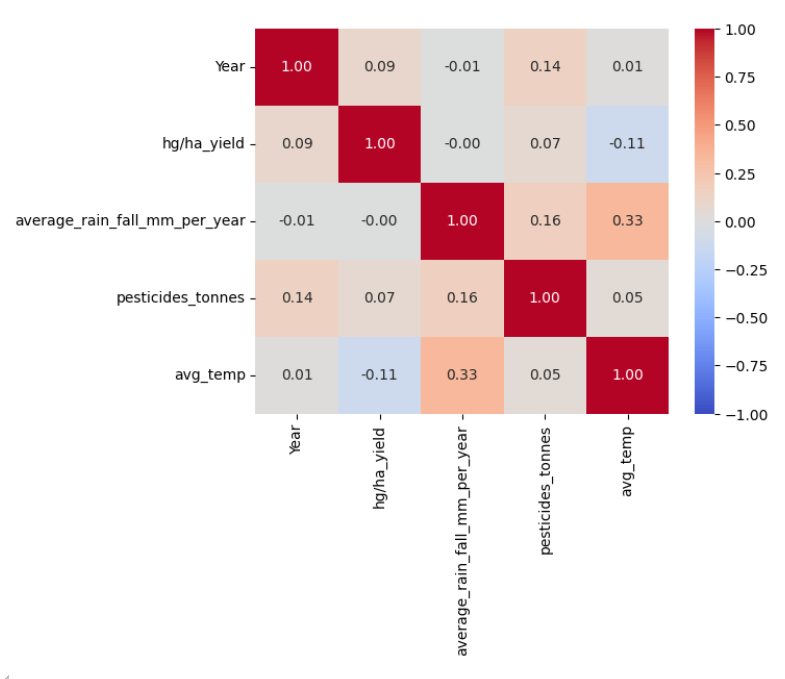
### R-Squared

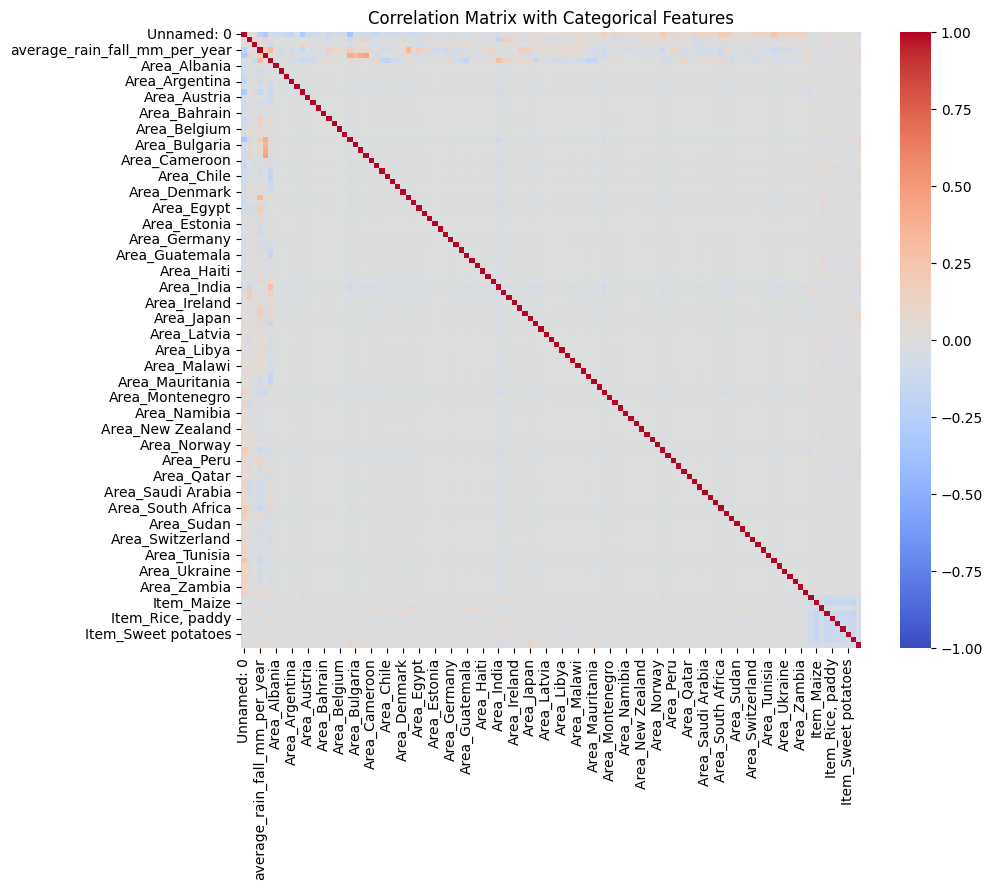
### ANOVA

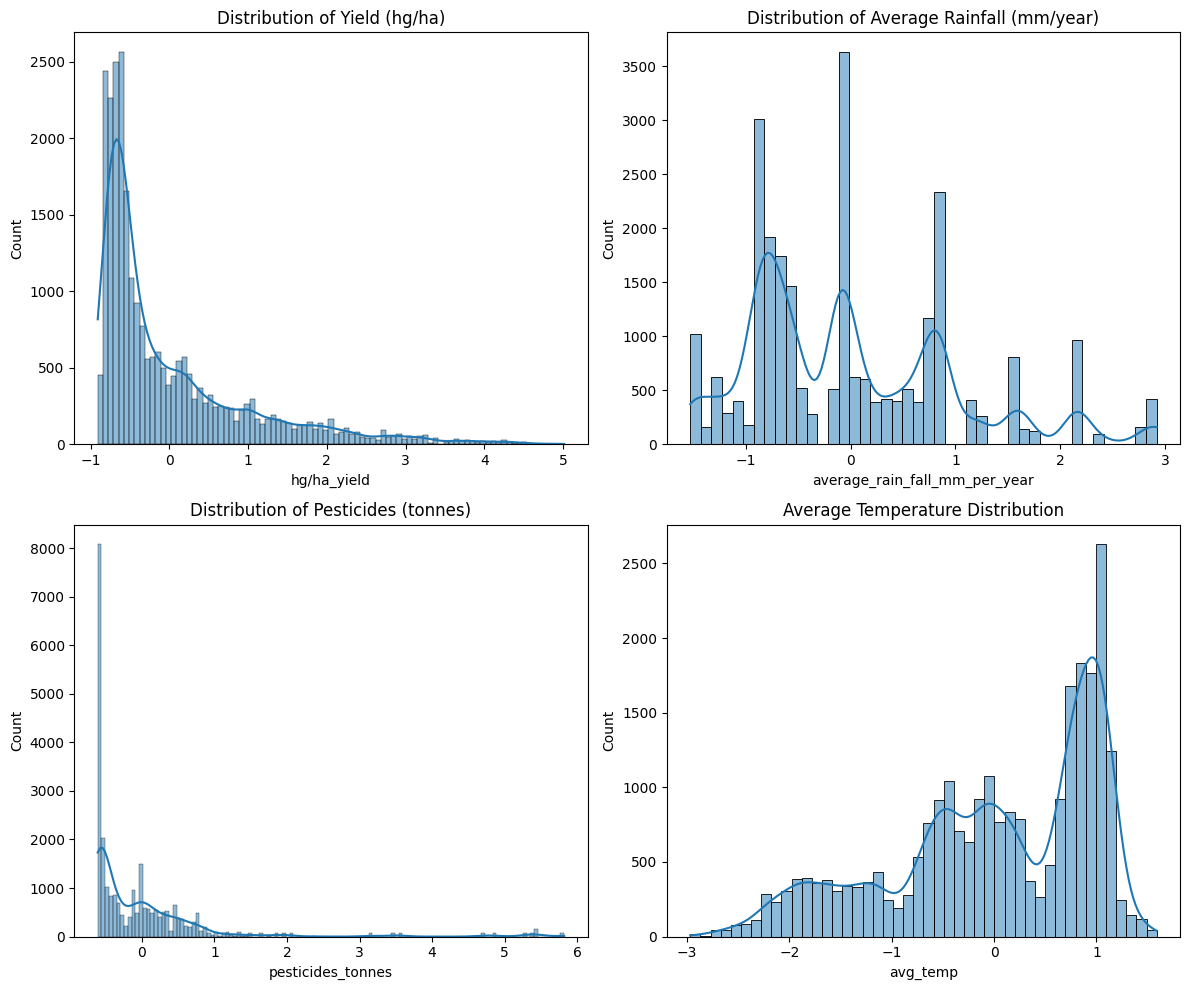
# Chapter 3: Methodology

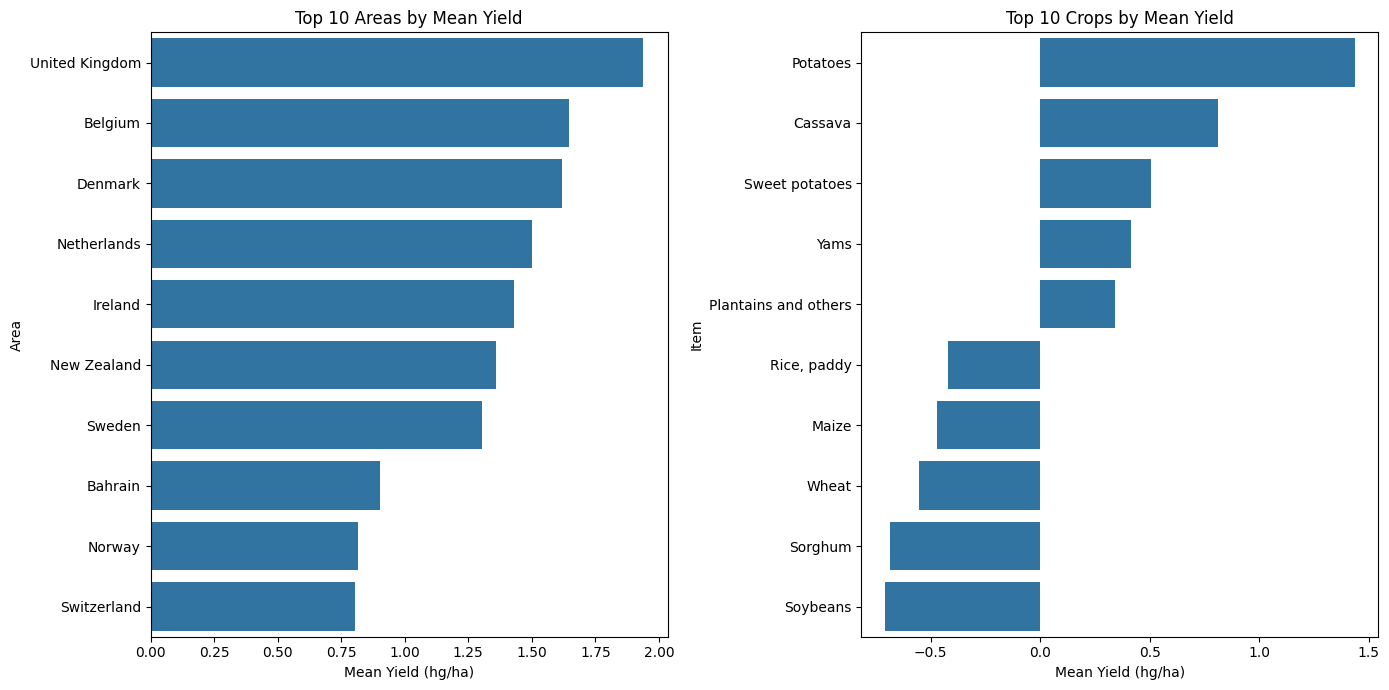
## Tools and Techniques

## EDA and Visualization

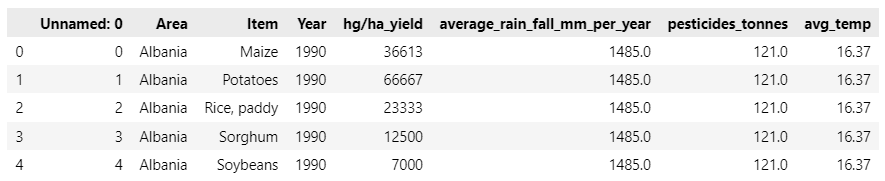


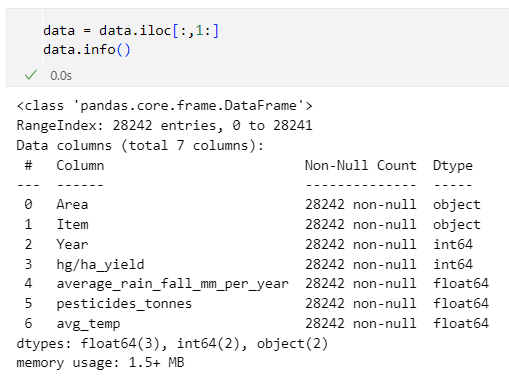




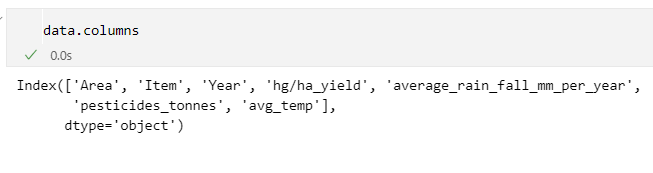


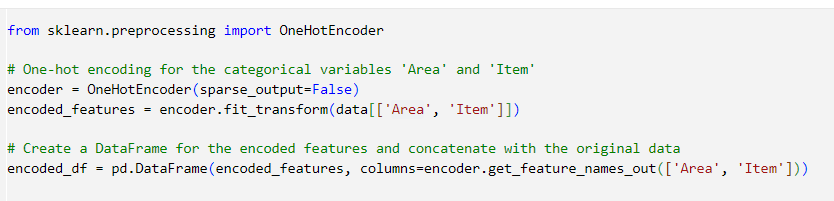
## Pre-Processing

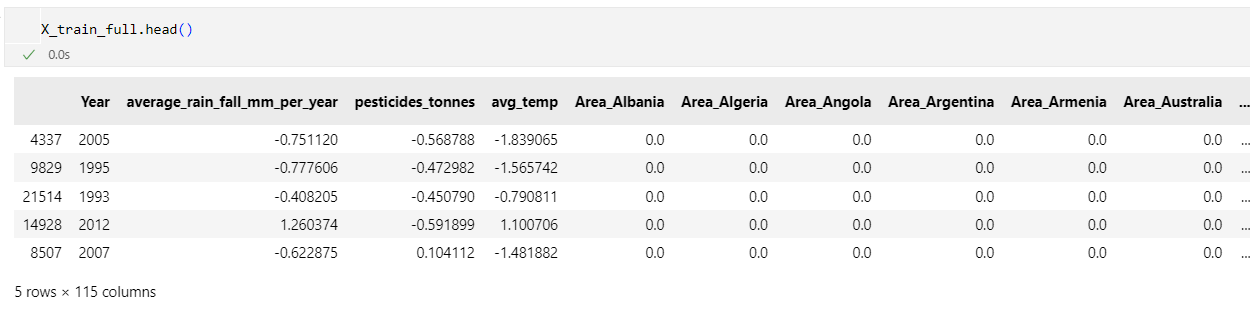


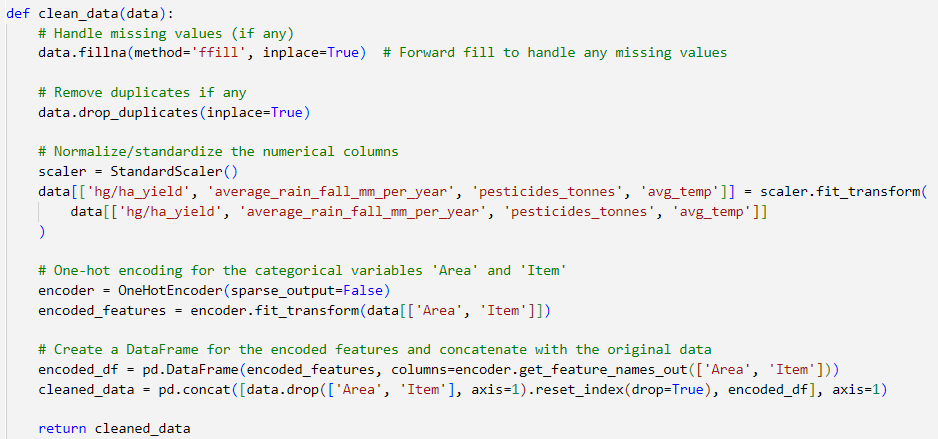






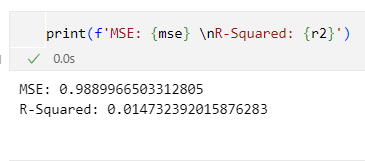


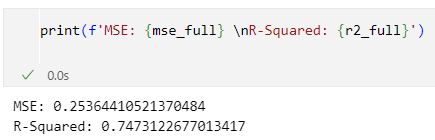


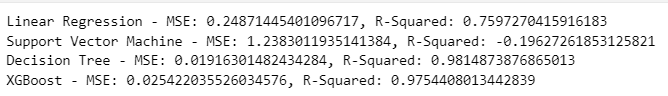


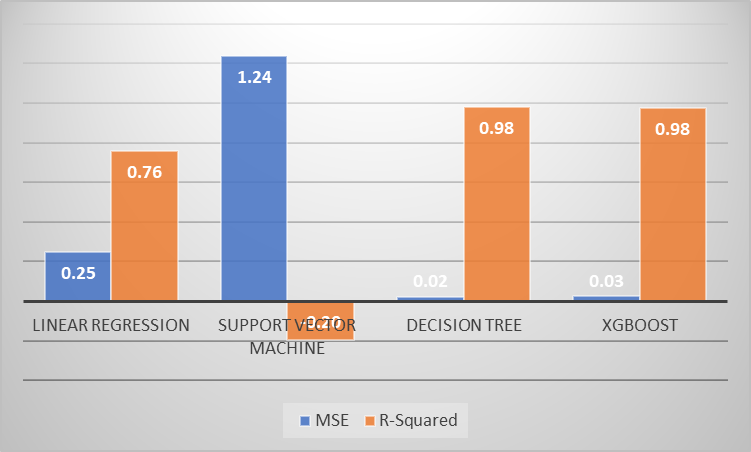
# Chapter 4: Results and Conclusion

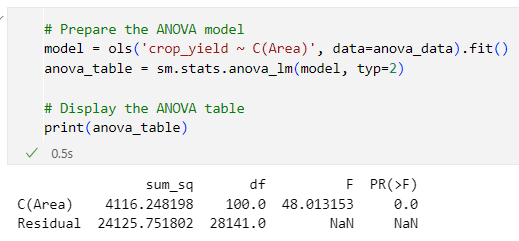
## Discussion







­­­­­­­­



## Conclusion

## Future Work

# Chapter 5: Legal, Ethical and Professional Issues

# References

# Appendices