A black background with white text

Description automatically generated with low confidence

MSc Data Science Project

7PAM2002-0509-2023

Department of Physics, Astronomy and Mathematics

**Data Science FINAL PROJECT REPORT**

**Project Title:**

Crop yield prediction

**Student Name and SRN:**

Keerthana Valiyaparambu Dayanandan -21069118

Supervisor: Michael Kuhn

Date Submitted: 03/09/2024

Word Count: 7409

DECLARATION STATEMENT

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science at the University of Hertfordshire.

I have read the guidance to students on academic integrity, misconduct and plagiarism information at [Assessment Offences and Academic Misconduct](https://www.herts.ac.uk/__data/assets/pdf_file/0007/237625/AS14-Apx3-Academic-Misconduct-v17.0.pdf) and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project module or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6). I have not used chatGPT, or any other generative AI tool, to write the reportor code (other than where declared or referenced).

I did not use human participants or undertake a survey in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student Name printed: KEERTHANA VALIYAPARAMBU DAYANANDAN

Student Name signature: keerthana

Student SRN number: 21069118

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

Table of Contents

[Abstract 4](#_Toc176191110)

[Chapter 1: Introduction 5](#_Toc176191111)

[I. Overview 5](#_Toc176191112)

[II. Research Questions 5](#_Toc176191113)

[III. Objectives 5](#_Toc176191114)

[Chapter 2: Background 6](#_Toc176191115)

[I. Dataset 6](#_Toc176191116)

[II. Literature Review 6](#_Toc176191117)

[III. Algorithms 8](#_Toc176191118)

[a. Linear Regression 8](#_Toc176191119)

[b. Support Vector Machines 9](#_Toc176191120)

[c. Decision Trees 9](#_Toc176191121)

[d. XGBoost 9](#_Toc176191122)

[e. MSE/RMSE 10](#_Toc176191123)

[f. R-Squared 10](#_Toc176191124)

[g. ANOVA 10](#_Toc176191125)

[Chapter 3: Methodology 12](#_Toc176191126)

[I. Tools and Techniques 12](#_Toc176191127)

[II. EDA and Visualization 13](#_Toc176191128)

[III. Pre-Processing 17](#_Toc176191129)

[Chapter 4: Results and Conclusion 21](#_Toc176191130)

[I. Discussion 21](#_Toc176191131)

[II. Conclusion 25](#_Toc176191132)

[III. Future Work 26](#_Toc176191133)

[Chapter 5: Legal, Ethical and Professional Issues 26](#_Toc176191134)

[References 27](#_Toc176191135)

[Appendices 28](#_Toc176191136)

# Abstract

This paper examines the feasibility of utilizing machine learning models to enhance the utilization of pesticide and yield crop, using the theme of the increasing adoption of data analytics in the agricultural sector. Through the comparative analysis of such algorithms as Support Vector Machines, Linear Regression, Decision Trees, and XGBoost, the work defines the best models for using the historical crop yield data for the most optimum Pesticide levels. From the results it can be concluded that models Decision Trees and XGBoost yield a better performance and give high accuracy in predicting the yield. Also to illustrate the effects of geographical influences on crop production, the project employs the Analysis of Variance (ANOVA) where there is evidence of regional variation that affects agricultural productivity. Therefore, the study revived the need of incorporating geographical information in to models in order to improve on the quality and relevance of the models. The paper also deals with the legal and ethical issues in relation to utilization of agricultural data and the recommendations to the environment and farming revenues. The future work will then entail integrating data from the environment in real-time, and extending the approach to look into deep learning, as well as endeavor to see to it that the suggested models are both viable and fair for all farmers. This research is useful in narrowing down the generation and application of superior and more precise regional practices pertaining to cultivation that offer high yields with least detrimental effects on the environment.

# Chapter 1: Introduction

## Overview

Nowadays, the use of pesticides in agriculture is more common. The use of pesticides during farming and on plants still has multiple advantages and disadvantages. They represent a risk to biodiversity although they minimise the variety of pests and insects and therefore boost productivity. generally damaging the soil and water, as well as messing with various organisms that don't threaten agriculture. Using chemicals optimally is crucial in preventing such situations. This research can be very helpful to tackle that issue.

The intention of this research is to solve the prediction challenges of increasing crop yields through effective pesticide usage. Modern methods are outdated, ineffective in dealing with complex options, and fail to account for several factors. In data where numerous factors affect a single variable, machine learning can be applied to manage complicated trends (Machine learning: learn, develop, and evolve from data sets, 2021).

The project will examine several kinds of machine learning algorithms to try to find the most effective and profitable approach in finding the ideal manner to use pesticide. This study will additionally use a statistical approach known as ANOVA to analyse harvest variability across multiple variables (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)?

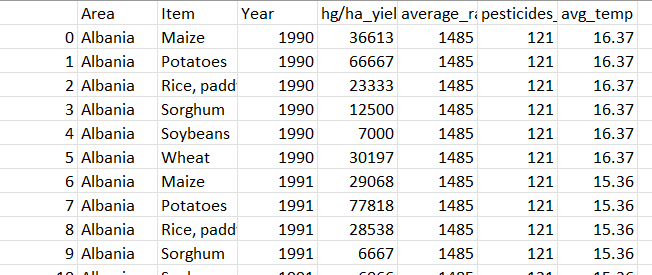
## Objectives

1. To identify how machine learning algorithms—such as Support Vector Machines, Linear Regression, XGBoost, and Decision Trees—perform better at predicting the ideal use of pesticide levels required to achieve the highest harvests.
2. Using multiple performance indicators, including R-Squared and RMSE (Root Mean Squared Error), in evaluating the techniques described above.
3. Implementing a pipeline for pre-processing data in Python to boost its quality and preparing it for predictive modelling..
4. Obtaining an unused dataset that duplicates real-world data, test the trained model.

# Chapter 2: Background

## Dataset

Furthermore to other indicators like pesticide use, rainfall, and temperature data through time for various countries, the dataset used in this research offers a picture of the agricultural yields. Rice, potatoes, wheat, soybeans, and additional crops belong in this dataset. With seven columns containing geography, rainfall, and other details, this dataset has about 29,000 records split over several years and countries. The graphic below provides a sample of the data's items.



This data came from the open-source dataset website Kaggle. Users are free to use any of the datasets on this website for personal purposes. So, it is not required to get ethical authorisation from any control.

Source: <https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset?select=yield_df.csv>

## Literature Review

The Indian agricultural sector, that offers income for two-thirds of the its population, has major crop yield challenges. Standard calculations of crop yields have produced uneven results because they have mostly depended on knowledge of certain farmers and certain crops. Several efforts have been put out to overcome these issues, focused on using data analysis and 20 techniques to better accurately predict yields in agriculture. The goal of Rajkumar and Mukunthan's (2023) suggested technique is to improve crop yield prediction through the application of supervised learning techniques and the four primary machine learning algorithms, Random Forest, SVM, Naive Bayes, and Decision Tree. The system analyses crucial variables like pH, humidity, rainfall, phosphorus, sulphur, and nitrogen to attempt to calculate optimal crop yields.

Important phases in this system include data collection and preparation, which include obtaining information from several sources and removing null and redundant values to produce an appropriate and clean dataset. After that, the pre-processed data is subjected to the four machine learning algorithms by the system. The Decision Tree conduct is suitable for problems with classification as well as regression because it represents decisions and their potential outcomes using a structure analogous to a tree. Based on the Bayes theorem, the Naive Bayes method effectively organises huge datasets with high density under an assumption of predictor independence. Support Vector Machine (SVM) is a powerful tool for identifying issues since it represents data points in an n-dimensional space and determines which hyperplane best divides various groups. An approach to team learning called Random Forest offers several decisionRandom 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.

Typically, an 80-20 split serves for dividing the dataset into sets for training and testing. Each algorithm is then fitted to the training data and its performance is assessed on the test data. The algorithm chosen for final implementation is those offering the highest accuracy. The best approach in this case is Random Forest, which is done using Python and 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, sulphur, 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.

With over fifty per cent of society dependent on agriculture, it is crucial to the Indian economy. Based on parameters like rainfall and weather, machine learning algorithms can predict crop yield. Robust supervised machine learning approach Random Forest helps in crop selection by supporting with classification and regression issues. There are major obstacles to the sustainability of agriculture due to environmental factors such as weather and climate. Crop Yield Prediction (CYP) decision-support tools are delivered by machine learning (ML), which aids in the cultivation and management of crops. Using reputable models, the main goal of (Ranjani et al., 2021) is to increase the efficiency of farming. In order to achieve maximum output, timely forecasts and analysis are necessary, and yield prediction is needed for farmers to make informed decisions before planting crops.

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.

Begin with gathering data and preprocessing, the system ensures a high-quality dataset for research by cleansing data from multiple places and removing null and redundant values. Decision trees are then created from various data samples using the Random Forest method. A simple majority defines the final output, which boosts the accuracy of the model, and each tree offers forecasts. Choosing features at random, calculating nodes using the optimal split points, and repeating this method for each tree are the stages performed in the pseudocode. The most widely-expected result is found by collecting votes, analysing the outputs created by each tree, and applying test attributes.

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.

For the reason of strategic agricultural planning, including import-export laws and boosting farmer incomes, accurate crop yield estimation is crucial. One of the primary problems affecting the farming industry is harvest prediction, which is where machine learning algorithms come into action. The use of machine learning to forecast crop yields—specifically, the production of palm oil—is reviewed in this article. It gives an overview of the current state of palm oil yield worldwide, talks about common features and predict algorithms, and offers an in-depth evaluation of cutting edge AI techniques to predicting crop yield. The review by Rashid et al. (2021) addresses present and coming challenges in agriculture and gives possible solutions. It also highlights advantages and obstacles of machine learning-based crop yield forecast. 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.

One of the main problems to crop yield prediction is the absence of regular feature sets. Climate, historical cultivation, vegetation index, satellite, soil parameters, irrigation, and agriculture information is a few widely use features. More research needed because it is unclear which sub-features are best for which crops given the climate data. The performance of yield prediction could be boosted by using satellite-based Solar-Induced Fluorescence (SIF) features. Regional harvest forecasts can be increased by merging weather parameters with multi-band satellite data.

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.

Placing fresh farming methods to use is vital for feeding the world's growing populace. Machine learning frameworks analyse large amounts of data and forecast future events to deliver helpful data. This review underlines the need it is to choose proper variables and algorithms to model crop output, particularly if it comes to palm oil yield. To increase yield forecast accuracy and develop viable strategies for optimising crop yields, more research utilising more kinds of features and algorithmic forecasting is needed.

## Algorithms

### Linear Regression

Linear regression is among the popular techniques of predictive modelling that identifies a linear relationship between an output variable and one or more input variables. Due to its operational amenability, it is widely used in numerous fields including economics, biology and engineering. These coefficients of the model assist also in defining the direction and intensity of such relations as a basis for a decision. Nevertheless, the linear regression assumes the linear connection only and depends on outliers as well as the variance of residuals. To apply the model, it is necessary for the assumptions such as linearity, non-correlation of errors, normal distribution of residuals and low collinearity to hold true. Failure to meet these assumptions results to bias, misleading results and wrong decisions. Therefore, it is critical to diagnose and validate the model in order to increase the likelihood of hypothesis testing as well as data analysis (Gawali, 2021).

### Support Vector Machines

Although, originally SVM is designed for classification problems, it can also be used for regression problems and we get Support Vector Regression (SVR). Concerning the model, SVR aims at identifying the level of association between the predictor and response variables while they allow variability in the prediction. Unlike the linear regression model that focuses on minimizing the sum of squared errors, SVR reduces the deviations outside an epsilon tube: a method used in avoiding over fitting. The selection of the forms of the kernel functions for the cases like linear, polynomial, RBF, and so on, is an important role in non-linear mapping. SVR is not very sensitive to outliers because there is the regularization parameter (C) and the epsilon margin. Nonetheless, it may be time and memory consuming when dealing with large sets and requester labelling accuracy also needs to be optimized. SVR finds its use in situations where variables in a system or relationships between the points present nonlinear models that cannot be modeled using a simple linear approach like the ones used in finance, bioinformatics and engineering (*Support vector machine (SVM)*, no date).

### Decision Trees

A decision tree is a widely used method for classification and regression, working by recursively splitting data based on features to create branches that represent decisions. Each node in the tree corresponds to a decision about a feature, and each branch represents the outcome, leading to a prediction or class label. The best attribute for splitting is chosen using measures like Gini impurity, information gain, or mean squared error. Decision trees are simple, interpretable, and can model non-linear relationships without assuming data distribution. However, they can easily overfit, especially with deeper trees, and are prone to instability, where slight changes in data can result in significantly different tree structures. Techniques like pruning and limiting tree depth are often employed to mitigate overfitting. Despite their limitations, decision trees are valuable in fields requiring model transparency, such as medicine and finance (Prasad, 2021).

### XGBoost

XGBoost also known as Extreme Gradient Boosting is considered sophisticated machine learning algorithm used in classification and regression problems. It improves the gradient boosting technique where various models are created through an iterative process to minimize errors, which the subsequent models correct for previous models’ faults. XGBoost it a highly efficient model with great parallel computation ability and performs very well with large datasets especially when it has missing values to handle. Vanilla XGBoost uses a smoother loss function than conventional L1 and L2 regularization to reduce overfitting for the best model performance. Likewise, XGBoost ensemble has other enhancements such as shrinkage (or learning rate) and second order optimization and the implementation of gradient and Hessian. However, it has its drawback such as hyperparameters which are so many that care has to be taken while setting them, and it performs poorly with unstructured data such as images and text as are handled with deep learning (Sonawane, 2023).

### MSE/RMSE

In machine learning, errors mean the deviations of the actual output from the predicted ones which are important for the assessment of the performance of the model. Mean squared error (MSE) is one of the most frequently used measures; it simply calculates the average of squares while it takes into consideration the outliers and it is useful when larger errors are needed to be emphasized more. Another measure derived from MSE is Root Mean Squared Error or RMSE since it’s easier to compare models and interpret since it keeps the units of the original data. MAE or the mean of absolute differences, because it averages out the simple difference while giving less consideration to the outliers and is usually recommended to be used when large errors are not very critical. MBD detects systematic bias, Log MSE appear to cater for large data ranger issues however they require log-normality. Depending on the specifics of the dataset and objectives of analysis, it is possible to select the proper metric; MSE or RMSE if one wants to penalize the errors with greater amounts, or if one has the aim to evaluate general errors without ignoring their relativity, the MAE could be used (Chugh, 2020).

### R-Squared

The coefficient of determination, or R-squared, measures how well a regression model explains the variance in the dependent variable. R-squared ranges from 0 to 1, with values closer to 1 indicating a good fit, meaning the model explains most of the variance. It is calculated by dividing the sum of squared residuals by the total sum of squares, reflecting the proportion of variance the model accounts for. However, a high R-squared doesn't guarantee model accuracy, as it doesn't address overfitting or validate assumptions. Additionally, R-squared is not suitable for comparing models with different dependent variables or datasets. Adjusted R-squared is a more refined metric that accounts for the number of predictors, offering a better comparison between models with varying numbers of variables. Despite its limitations, R-squared remains a key metric in regression analysis for assessing model performance (Moran, 2019).

### ANOVA

Key statistical procedures for comparing these groups of means include hypothesis testing and Analysis of Variance (ANOVA). Similar to t-test, ANOVA carries it a step further and enables the comparison across more than 2 means; is greatly used in experimental research to determine the impact of factors on an experimental variable. It partitions total variability into between group and within group variability. A large F ratio also show that there are differences between groups and thus impact of the independent variable.

Some of the varieties are One-Way ANOVA: Two-Way ANOVA, Repeated Measures, MANOVA every sort appropriate for the kind of the framework being used. Assumption used in ANOVA include normality, equal variance and independence of observation. Nevertheless, the results obtained are significantly different from each other, but it never indicates which groups are different from each other and which groups are not; for this, one needs to perform post hoc analyses like Tukey’s HSD. However, ANOVA is still significant across a range of scientific disciplines (*ANOVA (analysis of variance)*, 2010).

# Chapter 3: Methodology

## Tools and Techniques

1.Data Handling and Preparation

* Pandas (pd): A package employed for reading the CSV file and handling the data set by using the functions such as read. csv (), head (), str (), summary (), and duplicated (). This library is critical when performing data manipulation and analysis.
* Handling Missing Values: As for the missing values, the fillna(method=’ffill’) function was applied to the dataset to forward fill the missing values so as to avoid missing data in the dataset.
* Data Normalization/Standardization: The ‘StandardScaler’ class is obtained from the sklearn library. preprocessing was applied to remove multi-colinearity and the numerical factors were scaled, so that each feature has a zero mean, and a variance of one.
* One-Hot Encoding: Concerning the encoding of categorical feature, OneHotEncoder was applied to ‘Area’ and ‘Item’ features and provide binary column of it to feed into some machine learning algorithms.

2. Data Visualization

* Seaborn (sns): Used in generation of visual data such as the correlation heat maps, the distribution plots. This assists when trying to test the correlation between variables or when trying to analyze the distribution of data.
* Matplotlib (plt): Work in combination with Seaborn for the purpose of plotting for example modifying figure sizes whenever we want to display some plots. It is a fundamental library for building the static, interactive, and animated plots in Python.

3. Data Analysis

* Correlation Analysis: The corr() function was used to calculate the correlation matrix which depicts the numerical variables’ correlation using the Seaborn heatmap.
* Grouping and Aggregation: The groupby() function was applied to get mean yields concerning ‘Area’ and ‘Item’ to help determine the ‘Areas’ or the ‘Items ’ with the higher yields.

4 .Modeling

* Train-Test Split (train\_test\_split): Used to divide the dataset into the training and the testing set often important in assessing model performance since a portion of the dataset is not used in training the model.
* Linear Regression (LinearRegression): Used to forecast the target variable, which results to assessment measures such as Mean Squared Error (MSE) and Coefficient of determination (R-Squared).
* Support Vector Regression (SVR): Employed with the target variable target to be used in regression analysis. It successfully applies in situations where the relation between the variables is not a straight line.
* Decision Tree Regression (DecisionTreeRegressor): An algorithm which is not based on any assumptions about the shape of the unknown function, is included in the methods of ensemble, used for the regression problem.
* XGBoost (XGBRegressor): An advanced boosting algorithm used for regression this algorithm is efficient and effective in handling big data.

5. Model Evaluation

* Mean Squared Error (MSE): Utilized in order to calculate the mean of the square of the errors between the observed and predicted values of the dependent variable to ascertain the degree of goodness of the fit of the model.
* R-Squared: A calculated index for assessing the extent of the correspondence between the results of regression analysis and the actual data, calculated as the quotient of the variation between the sample mean and the calculated value and the variation of the sample from the average value.

6. Statistical Analysis

* ANOVA (Analysis of Variance): Carried out with the help of the stats model’s library with the OLS function to compare the variance of certain groups, namely ‘Area,’ to the yield of crops. This makes it possible to know whether there are reasonable differences in means in between the different groups or not.

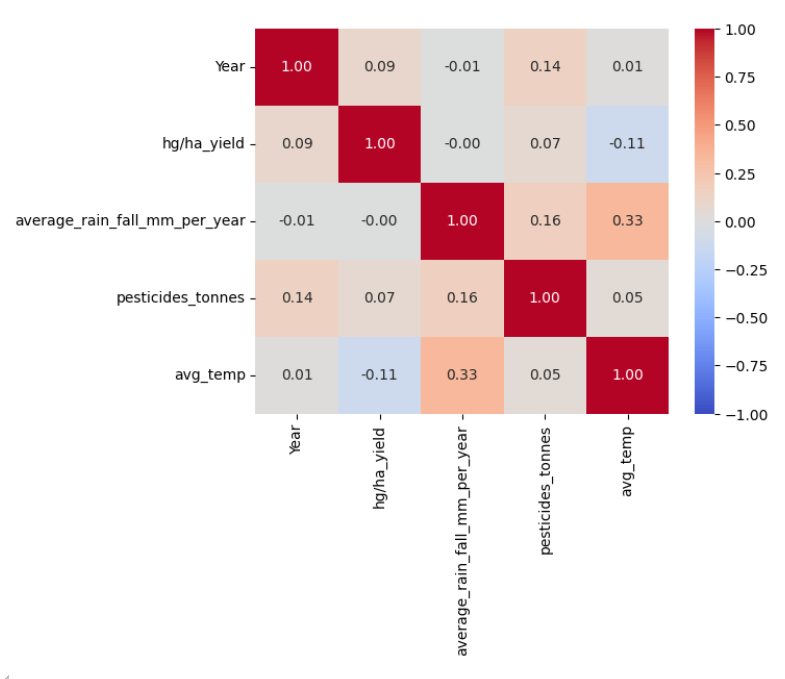
## EDA and Visualization

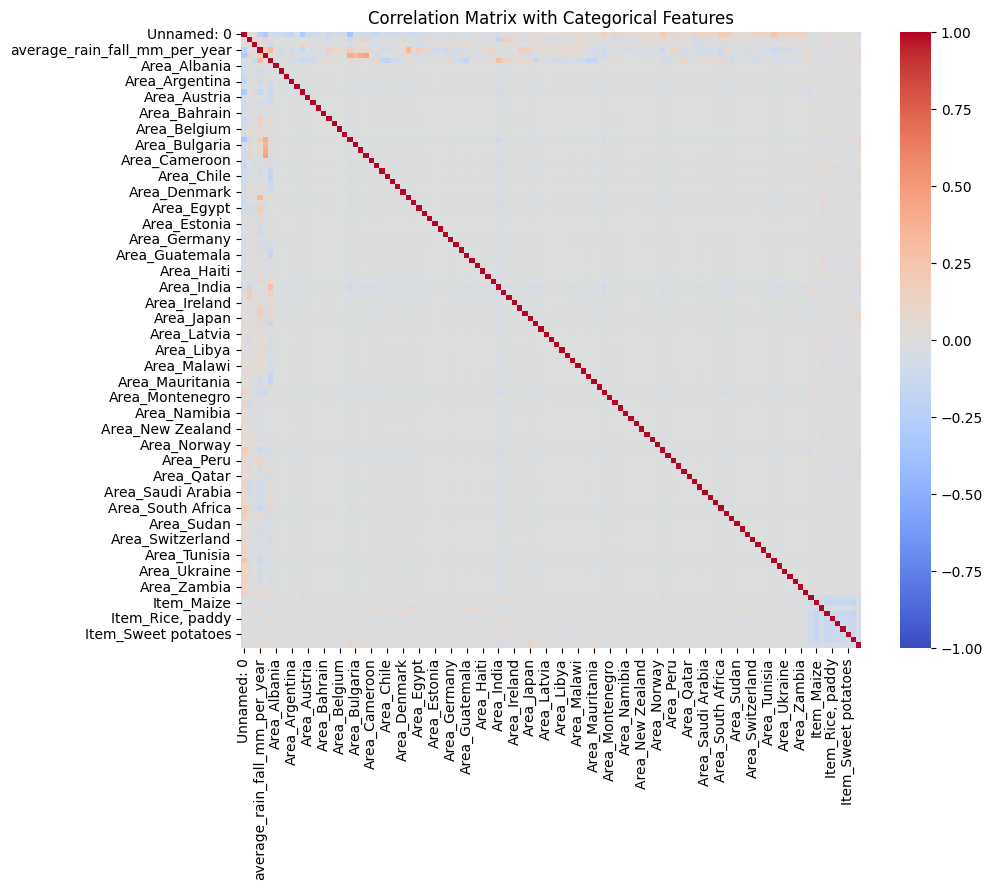
The EDA that was done on the crop yield prediction dataset incorporated distinct assessments in different visualizations to get insights of the general pattern, the distributions and dependencies of the variables. These analyses are paramount in finding out areas that need to be explored so that the modeling will benefit and also increase the extent of accuracy.

**Correlation Matrix Heatmaps**

The first figure shows two correlation heatmaps, which are indispensable to find interactions of numerical variables.

* **First Heatmap (Small Matrix):** In this heatmap, the many-to-many relations describe the correlation of several numerical characteristics, which include Year, hg/ha\_yield (yield, hectograms per hectare), average\_rain\_fall\_mm\_per\_year, pesticides\_tonnes, and avg\_temp. Therefore, the diagonal indicates the perfect correlation of 1. 00, as expected. Likewise, rest of the variables seem to have moderate levels of inter-correlation with no visible linearity. For instance, average\_rain\_fall\_mm\_per\_year is barely related to hg/ha\_yield, so the increased amount of rainfall does not mean that yield would increase at a consistent rate in this data. Pesticides\_tonnes, and avg\_temp like fertilisers\_tonnes also depict weak negative relations with yield. This implies that perhaps these variables will come into play in predicting yield in more sophisticated ways and or even manners that are not linear.
* **Second Heatmap (Large Matrix with Categorical Features):** The following image is the heatmap based on all of the features with both encoded categorical features such as Area and Item included together with the numerical features. The diagonal one more time presents immediate correlations in the different categories. Regardless, given the number of variables, the structure of the matrix is not easy to analyze directly, although it shows how the results might be connected between regions (Area) and crop types (Item) to the target variable (hg/ha\_yield). The interdependencies of the analytically observed correlations guarantee that the connection between these peculiarities and yield may require interactions of several factors.

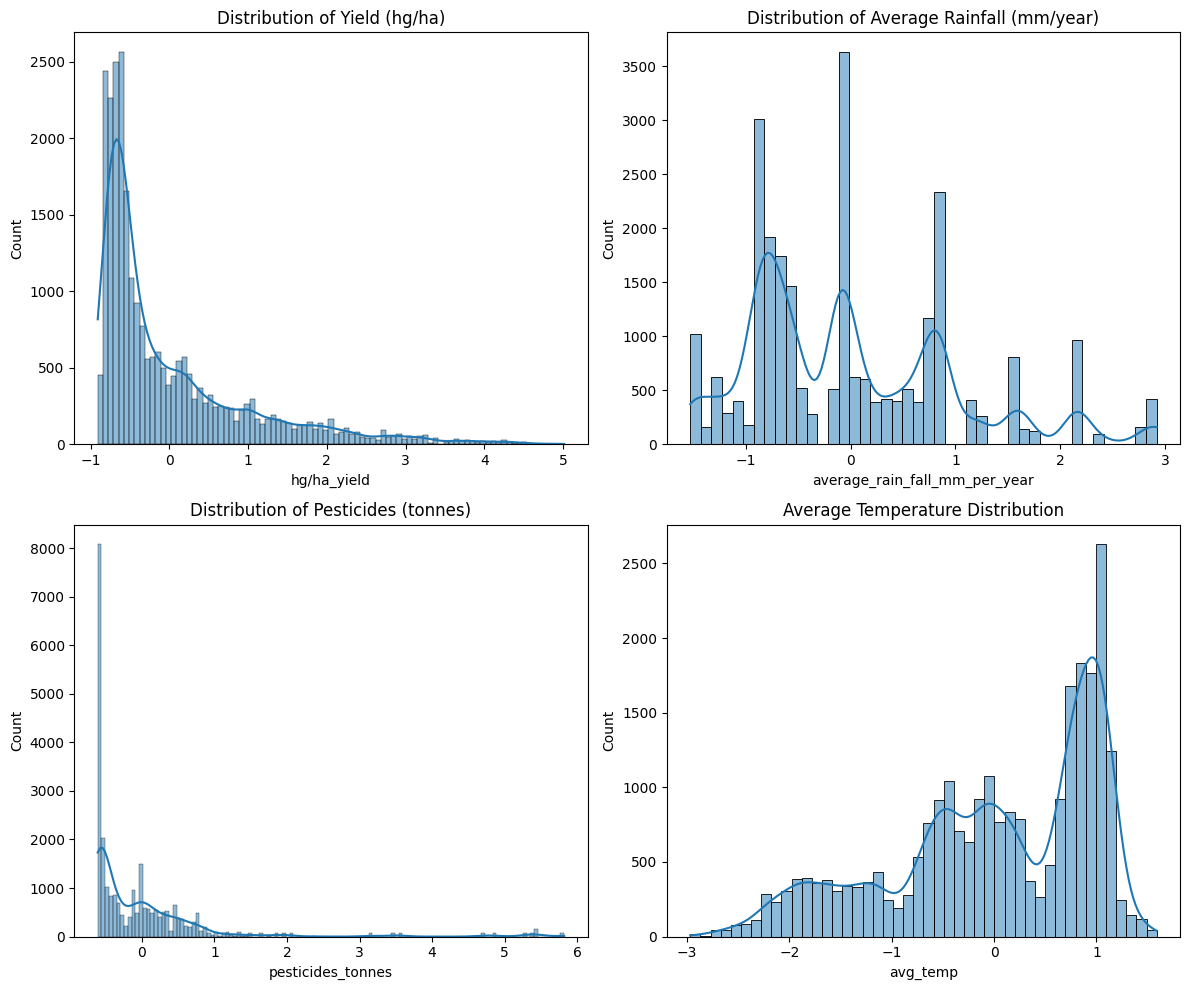




**Distribution Plots**

The second set of visualizations consists of distribution plots for key numerical features: And the variables are hg/ha\_yield, average\_rain\_fall\_mm\_per\_year, pesticides\_tonnes, and avg\_temp.

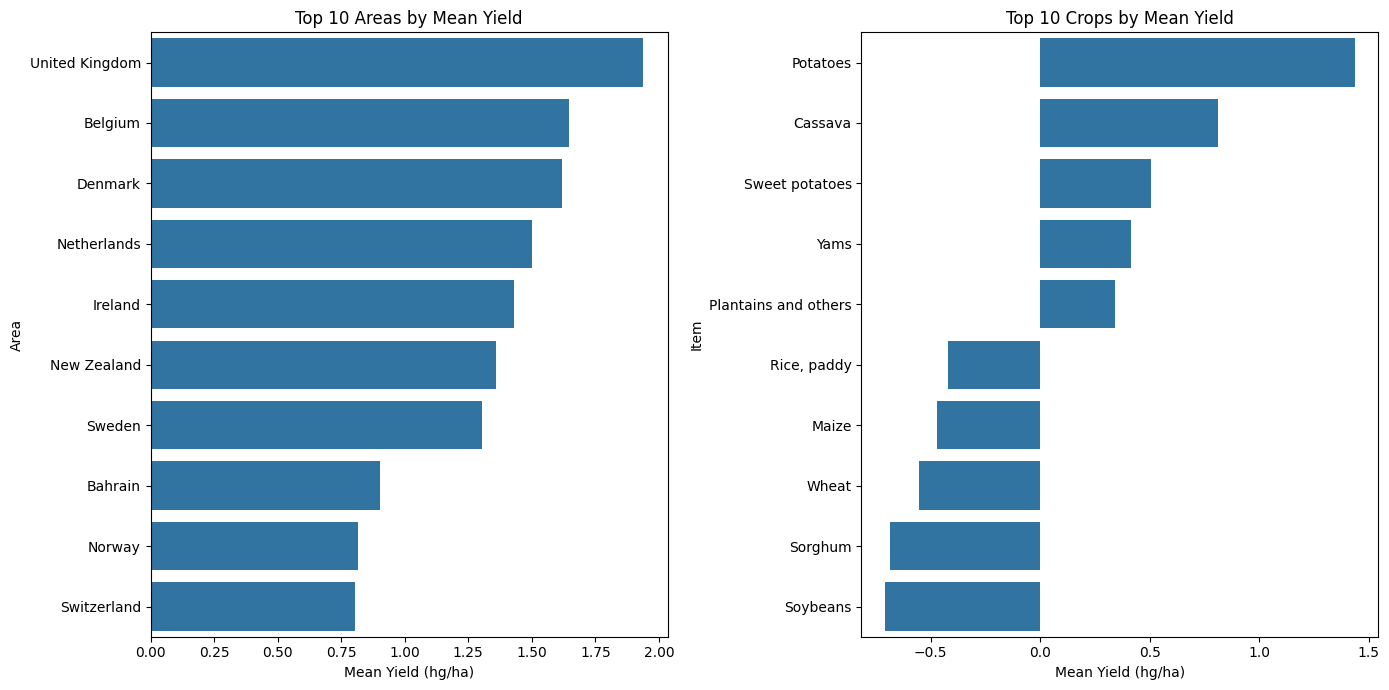
* **Yield Distribution (hg/ha\_yield):** It is also can be noted that the shape of the yield distribution to the right, having a long tail. Nearly all the yield values are on the lower side of the Yield Histogram, which means that higher yields are the exceptions. This skewness indicates that specific rates may be coming from certain conditions or regions which may include better farming practices and or good climatic conditions.
* **Average Rainfall Distribution (average\_rain\_fall\_mm\_per\_year):** This can be seen with the amount of rainfall distribution where one finds multiple humps indicating a possible existence of different clusters or areas that receive different amounts of rainfall. This multimodal distribution could be due to the many climatic regions from which the data was gathered and each region brought in his own yields.
* **Pesticides Usage Distribution (pesticides\_tonnes):** Another form of skewed distribution is observed in the case of pesticide usage; this is because the mean is influenced by very small amount of pesticide users who apply a large quantity of pesticides and the majority of the data is more or less clustered around the lower end of pesticide usage. The long tail suggests that most of the regions or crops using pesticides might be doing it in a limited manner hence might relate to more intensive farming.
* **Average Temperature Distribution (avg\_temp):** The distribution of the temperatures seems to be approximately bi-modal, probably due to collection of data from places with differing mean temperature possibly different season or place. These as maybe referring to areas such as the temperate and tropical regions characterized by temperature which influences growth of crops.



Before proceeding onto the next chapter, this chapter presents the bar plots of the top 10 areas and crops by mean yield.

The last visualization set belongs to the bar plots to show the first ten areas and the crops by the mean yield.

* Top 10 Areas by Mean Yield: The main idea unveils those countries like the United Kingdom, Belgium, Denmark have the highest average yield. These regions may be involving higher levels of technologies in agriculture, better climate for the production of crops among others leading to higher productivity. The fluctuation of the mean yields of these regions helps support the geographic and climate influences on agriculture.
* Top 10 Crops by Mean Yield: According to the plot for crops, Potatoes, Cassava, and Sweet Potatoes seem to rank high as far as yield is concerned. Such crops can be regarded as highly productive ones, which yields may be grown in optimal climatic conditions. This coaxes the notion that some crops are naturally more productive than others because of their growth gene or the technique used in farming.



The EDA is useful in showing the pattern of associations and distributions of the selected variables in the dataset. In the correlation analysis, it is revealed primary features may not have high correlation coefficient with yield or have low R-squared values although there may be interaction effects. The distribution plots focus on the variations in the environmental conditions and in the farm, management practices on a geographical basis. Lastly, bar plots of top areas and top crops by yield put emphasis on the fact that there is a strong relationship between the location and type of crop in farming. These insights then lead to a more in-depth analysis where the discovered patterns and relations are farther assessed and utilized in bettering the predictions of crop yields.

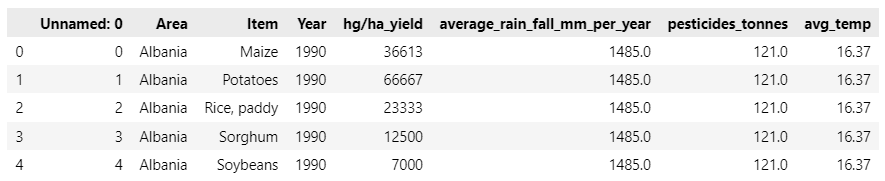
## Pre-Processing

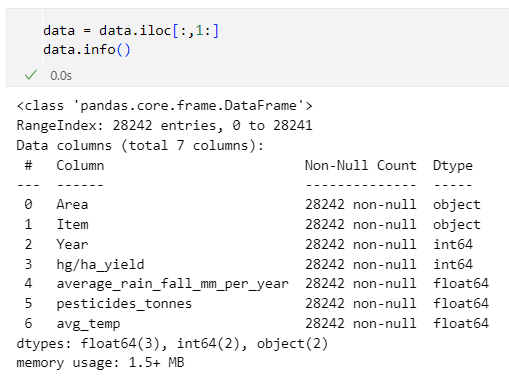
The given data description and information in the images are quite useful in the context of data preprocessing in the crop yield prediction.

Data Overview: It has 7 columns and 28242 records including Area, Item, Year, hg/ha\_yield – yield in hectogram per hectare, average\_rain\_fall\_mm\_per\_year, pesticides\_tonnes and avg\_temp. All of them are involved in the process of analysis and three of them are numerical which are Year, yield in hg/ha, average rain fall mm per year, pesticides tonnes and average temperature while two of them are categorical which are Area and Item.

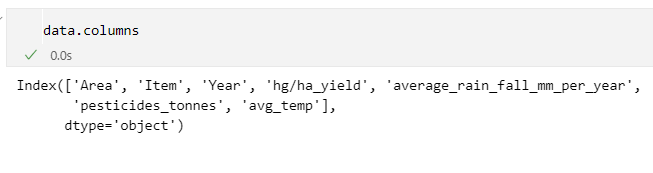
Data Types and Completeness: there are no null values in any of the columns and hence all the columns are non-null. This reduces the requirement of applying imputation approaches during the preprocessing stage. The data types of the variables are correct; numerical are of int64 or float64 type, and the categorical data type is object.

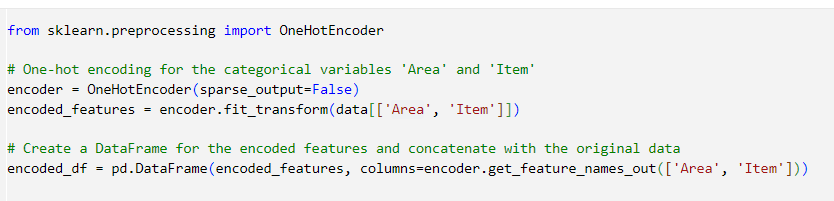
Statistical Summary: Essentially, the describe () function in statistics generates the quantitative summary of the data by indicating the mean, standard deviation, minimum, and maximum values. There is a considerable variability of the hg/ha\_yield from 50 to 501,412 and high stdev indicating substantial variations in yields starting from regional and temporal differences. The above distributions of rainfall, pesticide usage, and temperature also support the explanation of the dissimilar distribution of the environment and farming practices.











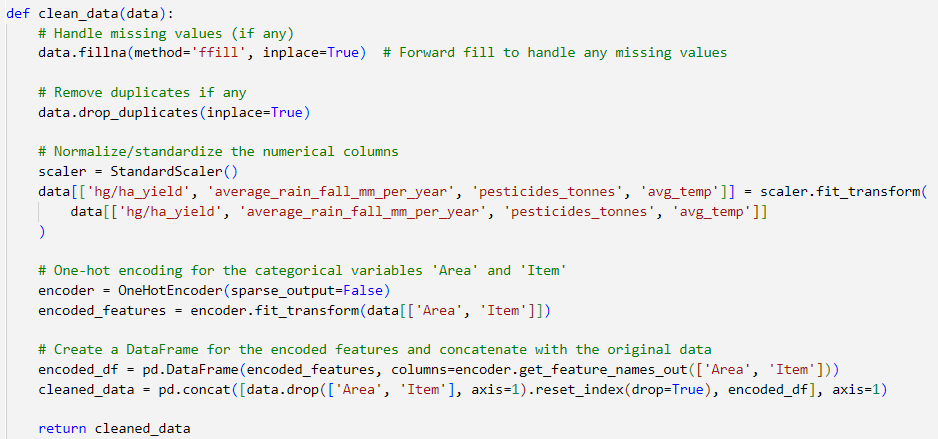
These are operations on the given data set aimed at having clean data, free from inconsistencies that can negatively impact the modeling process.

To address the missing values, forward fill (fillna(method='ffill')) is used which replaces the missing values with the last available value. This method is also valuable when working with time-series or sequential data because changes are made gradually, thereby preserving data coherence. After this, the same values are eliminated with the help of the drop\_duplicates () function so that the program does not tilt in favor of a particular portion of the data.

Subsequently, the independent variables including hg/ha\_yield, average\_rain\_fall\_mm\_per\_year, pesticides\_tonnes, and avg\_temp are normalized using StandardScaler from sci-kit-learn. This standardization process scales the features to have a mean equal to 0 and a standard deviation equal to 1 important that most algorithms that receive the input data are sensitive to the scale of the data they are receiving.

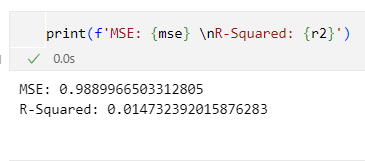
As for the transformations of categorical variables (Area and Item), OneHotEncoder is used. This transforms the categorical data into a numerical form whereby every category is a separate column and will therefore be good for feeding into to the machine learning models that use numerical data.

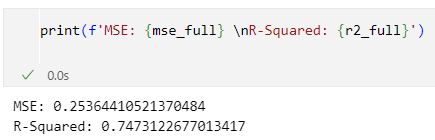
Finally, we stack them back together, thus, have a DataFrame with original features except the categorical ones and the new one-hot encoded features. This fully preprocessed dataset is now ready for further analysis and modeling; the data are preprocessed to specification to come up with reliable predictions.

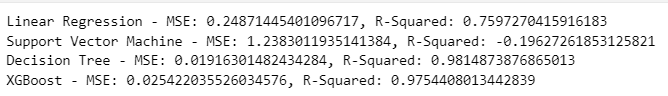


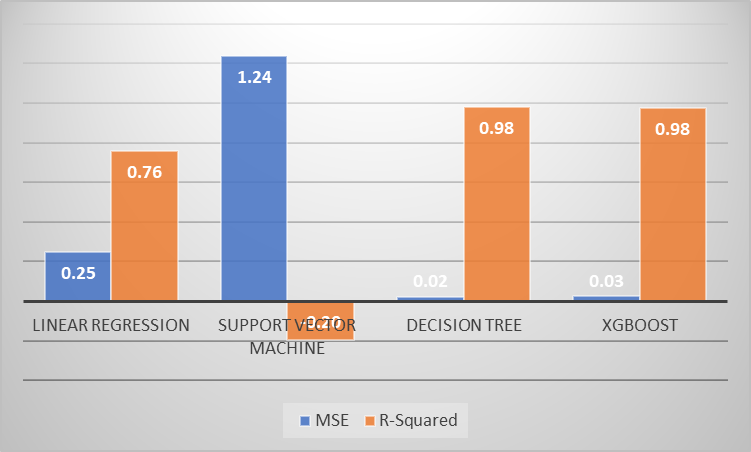
# Chapter 4: Results and Conclusion

## Discussion







­­­­­­­­

The performance of various machine learning models and their ability to predict crop yields are compared in the images shown above. The performance assessment indicators employed by this study are Mean Squared Error (MSE) and R-Squared (R²) since these are pertinent in the evaluation of Complex Graphics models.

In the first picture, the results of the initial model are presented with the value of MSE about 0. 989 and R-squared value of 0. 015. The high value of MSE together with a very low value of the R-Squared means that the model has a low accuracy of predictions which are far from the real values. From the above equations, the value of R-Squared is very close to zero and measures to 1. 5 %; this implies that the crop yield model about varies the data point; elaborately, it implies that the model is not an effective tool for prediction in this respect.

Instead, the second image shows how two different models perform, this second model gives a significantly smaller MSE of 0. 254 and a much higher R-Squared value of 0. 747. This is further improvement from earlier indication suggesting that the model is now in a position to explain over seventy percent of variation in the data. The decrease in MSE also point to the fact that the predictions are closer to the real results and therefore this model fit is more appropriate in practical applications of predicting yields of crops.

The third image provides a comparative analysis of several models: The algorithms used in development were Linear Regression, Support Vector Machine (SVM), Decision Tree, and XGBoost. The metrics for each model are as follows:

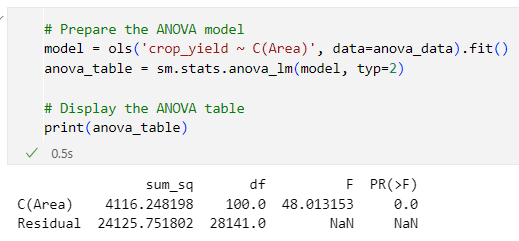
* Linear Regression: MSE of 0. 249 and values of R-Squared of 0. 760.
* Support Vector Machine (SVM): MSE of 1. 238 and R-Squared of negative 0. 196.
* Decision Tree: The Spectator What is the matter with stock markets? No, it was not the decline in MSE of 0. 019 and R-Squared of 0.981.
* XGBoost: MSE of 0. 025 and coefficient of determination of 0. 975.

Of all these models, Decision Tree has been considered most effective with the least MSE of 0. 019 and the highest coefficient of determination ‘R-Squared’ that equals to 1. 981. There is evidence that the Decision Tree model yields the highest accuracy level, with predictability levels standing at 98%. Explaining 1% of the variance in the crop yield data we obtain the following economic model. Similarly, we have achieved a very good result for the XGBoost as well, the values of MSE and R-Squared slightly lower and slightly higher, respectively than the Decision Tree making it another good contender for crop yield prediction.

On the other hand, the model which is the least accurate is the SVM, whose MSE stands at 3,579 and an R-Squared value of -0. 073. A negative R-Squared directs that the model is worse than a straight line that would sit parallel to the X-axis –that is it just predicts the average value of the target variable. This is further underscoring the inappropriateness of SVM for this particular prediction task, could be the nature of data or indeed how the model has been done.

For the Linear Regression model, therefore, it is significantly better than the SVM but just still below the performance score for Decision Tree and XGBoost. Thus, in income constant price expenditures Malaysia’s MSE of 0. 249 and the adjusted R-Squared of the model was 0. 760 Consequently, an understanding of this model which otherwise has been quite helpful may not capture all the details as elaborated by the more enhanced models.

The fourth figure is a bar chart where MSE and R-Squared have been used to represent the performance of the different models. This we can deduct from the visual comparison where the Decision Tree and XGBoost models stand out with lower MSE and higher R-Squared as compared to the other models. The chart also demonstrates why the SVM model is not suitable for the task; it also shows other models’ performance compared to the best models in the chart.



The results presented in ANOVA tables give a statistical possibility of comparing the areas and productions of different crops. The following indicators are obtained from the ANOVA table: Sum of Squares, Degrees of Freedom (df ), F-statistic, and p-value/alpha-level. All these parameters are instrumental in the context of analysing the impact that geographical variances impose on yield and testing the significance of these variances.

**Sum of Squares**

They include: The Sum of Squares (SS) that is the total amount of variation of the data that can be attributed to different sources. In this context, the Sum of Squares is divided into two components: the, between-group sum of squares which relates to variance between two or more groups – in this case, the different geographical areas and the within-group sum of squares that relate to residual error or variation within any one geographical area.

* C(Area) Sum of Squares: 2712. 248 shows how much of crop yield volatility can be attributed to the disparity in the geographical regions.
* Residual Sum of Squares: 24125. 752 amounts to the abstraction of variability in crop yield that is not accounted for by the geographical area factor, but is present within each area.

The greater the between-group sum of squares in proportions to the within-group sum of squares, the greater the chances of rejecting the null hypothesis with respect to departure in means of different groups.

**Degrees of Freedom**

Degrees of Freedom (df) as used in statistics and especially in analysis of variance pertains to the number of ‘‘free’’ values that can be picked to enter into a statistical distribution. In ANOVA, degrees of freedom are divided into two types:

* Between-group df (C(Area)): This value, which is 100 in the table presented above, refers to the number of areas minus one. It is defined by the number of pairwise comparisons that can be made between group means.
* Residual df: The value 28141 is the residual error regardless of degrees of freedom for it is has direct relation to the number of observations in a given data set minus the number of groups.

These degrees of freedom play important roles in calculation of the F-statistic for the reason that they affect the variance estimates that are used out of computation.

**F-Statistic**

The F-statistic turns between the between-group variance and within-group variance. It remains an important federal variable in the analysis of variance and signifies whether observed differences in group means are actual or random.

F-Statistic: 48. F(2, 013) = 78. 98, which is substantial, suggests that the standard deviation of the geographical areas’ means is considerably larger than the standard deviation of the groups’ means. A large F-value often indicates the possibility that the group means are not equal; or at least one group mean is different.

**P-Value**

P-value is a valuable measure in hypothesis testing as well as in ANOVA cases. It stands for the chances of getting such data (or something more outrageous) if the hypothesis that has been assumed to be true is in fact true in the population. In this regard, the null hypothesis is assumption that three factors do not have significant impacts on the crop yields of various geographical regions.

P-Value: 0., and a value closer to 0 (rounded to this value) implies that the likelihood of witnessing the above-mentioned F-statistic in the context of the null hypothesis is practically nil. If the p-value is less than some conventional cutoff, which is often 0. 05, thus, it will be appropriate to reject the null hypothesis. However, it is important not to lose sight of the other aspects of the case: the p-value of 0 indicates that. It can therefore be concluded that deterministic variable of geographical area does have a statistically significant impact on crop yields given that the coefficient estimate obtained, 0 strongly suggest this.

All the ANOVA results taken collectively show that the geographical area has an effect whereby it influences the yields. The p-value being 0. 0 implies refusal to retain the null hypothesis hence signifying that the variations in the crop yields across the various areas are not attributed to chance. The high F-statistic also supports this conclusion, that greatly exceeding the F-tab, which indicates that the variation between different areas is much higher than the variation within them.

Such conclusions are invaluable for other agriculture decisions because they establish locale determination as the key hypothesis in crop yields forecasting. It also helps identify region-specific solutions, in areas like agriculture, budgeting, or policies that will conform more to the specifics of giving population’s needs. Furthermore, these results imply that providing models as a basis for agricultural yields the precision of geographical location must be taken into consideration and calculated based on the specific condition of the area in question.

## Conclusion

The goals of this research were to fill the following gaps in the use of shrubs/agricultural crops and pesticides: The research questions centered on the best suited machine learning models for estimating the right quantities of pesticides to use, and whether there are any systematic variations in crop productivity across different regions using Analysis of Variance (ANOVA).

Performance of several numerical machine learning models such as Support Vector Machines (SVM), Linear Regression, Decision Trees, & XGBoost was assessed using Mean squared error (MSE) and R². The results suggested a number of considerations: First, not all models are equally suitable for the work. In particular, Decision Trees and XGBoost classifiers turned out to be the most accurate when addressing the problem of crop yields prediction with the least MSE and the highest R-Squared rates. This proves that they are better suited to unravel the dynamics and the non-linear trends that are normally embedded in the data.

SVM turned out to be very unsuitable for this type of regression task since it had a high MSE and negative the R-Squared value; this could have been due to the data or perhaps the nature of the assumptions made by the model. Nevertheless, when it comes to Linear Regression, while outperforming SVM, it failed to capture the same amount of accuracy as of complicated models as the Decision Tree and XGBoost. This result suggests that for proper selection of machine learning models it is necessary to take into account the peculiarities of data and solved problems.

Further, the analysis of variance offered and additional perspective by considering the role that geographical characteristics play in crop productivity. The results showed also a statistical significance p = 0, 0, which confirms that geographical area plays a significant role in deciding the number of yields of crops. This finding addresses the second research question to the effect that crop yield variability across the different locations is not as a result of random variation but is as a result of some specific geography-related factors. The high F-statistic also substantiates this conclusion which indicates that the amount of variation between different areas is in fact much larger than the amount of variation within the same area.

When answering the research questions, the research was able to discover that Decision Trees and XGBoost are the best algorithms in machine learning to anticipate the right proportion of pesticide use required to get the best yields from the crops. These models’ high accuracy indicates that they are able to learn the structure of the data and can identify and work with all of the interaction between data points, thus being useful tools for precision agriculture.

ANOVA, indicated that geographical factors affected crop yields. For this reason, this insight is very important in the identification of appropriate approaches to region-specific agricultural practices and policies to improve productivity and sustainability.

Altogether, the results presented in this work can be considered as a valuable contribution to the field of agricultural analytics as it demonstrates the efficiency of state-of-the-art machine learning techniques for optimizing the pesticide consumption and predicting crop yields. It also draws attention to the issue of geographical location in the prediction of crop yields hence defining future agricultural interventions.

## Future Work

Possible recommendations for further research and development of crop yield predictions are listed below. Including additional types of data, including geospatial or temporal, for example, sensors recording conditions in real time from satellite imagery, levels of soil moisture, and weather predictions, may improve the model. Such dynamic data could help build models that would need update less frequently or use real time information to provide more accurate recommendations with regards to pesticide usage and crop status. Further, using and training of sophisticated models for deep learning such as CNNs and RNNs may reveal further layers in the data we are analysing. Perhaps, integrating some machine learning approaches with some of the existing agronomic knowledge could help improve the models even more. It would also be possible for authors of future studies to consider the projection on the expected production of crops and the use of pesticides and in turn give the farmer both the agricultural and the economic potential consequences. Finally, a replication of the study with various crops, and in other climate zones would extend the scope of the models and enhance transferability.

# Chapter 5: Legal, Ethical and Professional Issues

Data Privacy and Consent: One should always ensure data privacy when integrating data from agriculture, and especially when the data pinpoint a certain farm or region. All data must be open-source, or permission must be sought from the owner/owners in a written manner, and data cannot be used unfair treatment, it must be anonymized.

Environmental Impact: Some ethical issues may come up where the predictive models may suggest the use of pesticides and this may lead to their overapplication thus adding to environmental pollution. Recommendations should not exploit methods that are destructive to the soil, water, and other non-target organisms.

Impact on Farmer Livelihoods: The ethical considerations must address the following necessities of the small-scale farmers. There should be suggestions made by the models that can be financially implemented by all the associated parties without creating inequality and leave technology benefits to only the intended receivers.

Legal Compliance: Recommendations for pesticide use should adhere to provisions of the local and international laws in agricultural application in that it should not exceed certain limits, and also should not hamper the exportation of produce.

# References

*ANOVA (analysis of variance)* (2010) *Statistics Solutions*. Available at: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/anova/ (Accessed: August 24, 2024).

Chugh, A., 2022. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared—Which Metric Is Better? 2020.

Gawali, S. (2021) *Linear regression in machine learning*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/ (Accessed: August 24, 2024).

*Machine learning: learn, develop, and evolve from data sets* (2021) *Brunel*. Available at: https://www.brunel.net/en/management-guide/machine-learning (Accessed: June 12, 2024).

Moran, M. (2019) *R-Squared: Telling us what we know and what we do not know*, *Statistics Solutions*. Available at: https://www.statisticssolutions.com/r-squared-telling-us-what-we-know-and-what-we-do-not-know/ (Accessed: August 24, 2024).

Prasad, A. (2021) *Regression trees*, *Analytics Vidhya*. Available at: https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047 (Accessed: August 24, 2024).

Rajkumar, N. and Mukunthan, M.A., 2023, December. Efficient Crop yield Analysis Prediction in Modern Agriculture System using Machine Learning Algorithm. In *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)* (pp. 1-4). IEEE.

Ranjani *et al.* (2021) “Crop yield prediction using machine learning algorithm,” in *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*. IEEE, pp. 611–616.

Rashid, M., Bari, B.S., Yusup, Y., Kamaruddin, M.A. and Khan, N., 2021. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE access*, *9*, pp.63406-63439.

Singh, G. (2018) *What is Analysis of Variance (ANOVA)?*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/ (Accessed: June 12, 2024).

Sonawane, P. (2023) *XGBoost — How does this work - Prathamesh Sonawane*, *Medium*. Available at: https://medium.com/@prathameshsonawane/xgboost-how-does-this-work-e1cae7c5b6cb (Accessed: August 24, 2024).

*Support vector machine (SVM)* (no date) *Mathworks.com*. Available at: https://in.mathworks.com/discovery/support-vector-machine.html (Accessed: August 24, 2024).

# Appendices

# Importing libraries for data manipulation, analysis, and visualization

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Importing libraries for data manipulation, analysis, and visualization

from google.colab import drive

drive.mount('/content/drive')

data = pd.read\_csv(r'/content/yield\_df.csv.zip')

#printing the first 5 rows

data.head()

# Removing the first column from the DataFrame

# Displaying a summary of the DataFrame, including the index, column names, non-null counts, and data types

data = data.iloc[:,1:]

data.info()

# Producing detailed statistics for the DataFrame's numerical columns

data.describe()

# Displaying the list of column names in the DataFrame

data.columns

# Extracting the target variable 'hg/ha\_yield' from the DataFrame

target = data['hg/ha\_yield']

# Handle missing values (if any)

data.fillna(method='ffill', inplace=True) # Forward fill to handle any missing values

# Convert data types if necessary (all seem appropriate here)

# data['Year'] = pd.to\_datetime(data['Year'], format='%Y')

# Remove duplicates if any

data.drop\_duplicates(inplace=True)

# Normalize/standardize the numerical columns if necessary

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

data[['hg/ha\_yield', 'average\_rain\_fall\_mm\_per\_year', 'pesticides\_tonnes', 'avg\_temp']] = scaler.fit\_transform(

data[['hg/ha\_yield', 'average\_rain\_fall\_mm\_per\_year', 'pesticides\_tonnes', 'avg\_temp']]

)

# Calculating the correlation matrix for the DataFrame (excluding the first two columns)

# Visualizing the correlation matrix using a heatmap with annotations and a color map

corr = data.iloc[:,2:].corr( )

sns.heatmap(corr,annot=True, cmap='coolwarm', fmt=".2f",vmin=-1)

# Distribution plots for numerical features

fig, axes = plt.subplots(2, 2, figsize=(12, 10))

sns.histplot(data['hg/ha\_yield'], kde=True, ax=axes[0, 0])

axes[0, 0].set\_title('Distribution of Yield (hg/ha)')

sns.histplot(data['average\_rain\_fall\_mm\_per\_year'], kde=True, ax=axes[0, 1])

axes[0, 1].set\_title('Distribution of Average Rainfall (mm/year)')

sns.histplot(data['pesticides\_tonnes'], kde=True, ax=axes[1, 0])

axes[1, 0].set\_title('Distribution of Pesticides (tonnes)')

sns.histplot(data['avg\_temp'], kde=True, ax=axes[1, 1])

axes[1, 1].set\_title('Average Temperature Distribution')

# Adjusting the layout of the plot

# Displaying the plot

plt.tight\_layout()

plt.show()

# Visualizing yield by Area and Item

# Top 10 areas and crops by mean yield

top\_areas = data.groupby('Area')['hg/ha\_yield'].mean().sort\_values(ascending=False).head(10)

top\_items = data.groupby('Item')['hg/ha\_yield'].mean().sort\_values(ascending=False).head(10)

plt.figure(figsize=(14, 7))

plt.subplot(1, 2, 1)

sns.barplot(x=top\_areas.values, y=top\_areas.index)

plt.title('Top 10 Areas by Mean Yield')

plt.xlabel('Mean Yield (hg/ha)')

plt.subplot(1, 2, 2)

sns.barplot(x=top\_items.values, y=top\_items.index)

plt.title('Top 10 Crops by Mean Yield')

plt.xlabel('Mean Yield (hg/ha)')

# Adjusting the layout of the plot

# Displaying the plot

plt.tight\_layout()

plt.show()

# Importing functions and classes for machine learning and model evaluation

# For splitting data into training and testing sets

# For creating and using linear regression models

# For evaluating model performance

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error, r2\_score

# Prepare the dataset for modeling

features = data[['average\_rain\_fall\_mm\_per\_year', 'pesticides\_tonnes', 'avg\_temp']]

target = data['hg/ha\_yield']

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(features, target, test\_size=0.2, random\_state=0)

# Initialize and train the linear regression model

model = LinearRegression()

model.fit(X\_train, y\_train)

# Predictions and model evaluation

y\_pred = model.predict(X\_test)

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

X\_train.head()

# Printing the Mean Squared Error (MSE) and R-Squared (R²) values to evaluate the model's

print(f'MSE: {mse} \nR-Squared: {r2}')

from sklearn.preprocessing import OneHotEncoder

# One-hot encoding for the categorical variables 'Area' and 'Item'

encoder = OneHotEncoder(sparse\_output=False)

encoded\_features = encoder.fit\_transform(data[['Area', 'Item']])

# Create a DataFrame for the encoded features and concatenate with the original data

encoded\_df = pd.DataFrame(encoded\_features, columns=encoder.get\_feature\_names\_out(['Area', 'Item']))

model\_data = pd.concat([data.drop(['Area', 'Item'], axis=1).reset\_index(drop=True), encoded\_df], axis=1)

# Correlation matrix for the new dataset with encoded categorical features

corr\_matrix\_full = model\_data.corr()

plt.figure(figsize=(10, 8))

sns.heatmap(corr\_matrix\_full, cmap='coolwarm',vmin=-1)

plt.title("Correlation Matrix with Categorical Features")

plt.show()

# Prepare data for modeling with categorical features

features\_full = model\_data.drop('hg/ha\_yield', axis=1)

target\_full = model\_data['hg/ha\_yield']

# Split the data into training and testing sets

X\_train\_full, X\_test\_full, y\_train\_full, y\_test\_full = train\_test\_split(features\_full, target\_full, test\_size=0.2, random\_state=0)

X\_train\_full.head()

# Fit the model with the new dataset

model\_full = LinearRegression()

model\_full.fit(X\_train\_full, y\_train\_full)

# Predictions and model evaluation with the new dataset

y\_pred\_full = model\_full.predict(X\_test\_full)

mse\_full = mean\_squared\_error(y\_test\_full, y\_pred\_full)

r2\_full = r2\_score(y\_test\_full, y\_pred\_full)

#printing the values for evaluate the model's

print(f'MSE: {mse\_full} \nR-Squared: {r2\_full}')

from sklearn.preprocessing import StandardScaler, OneHotEncoder

def clean\_data(data):

# Handle missing values (if any)

data.fillna(method='ffill', inplace=True) # Forward fill to handle any missing values

# Remove duplicates if any

data.drop\_duplicates(inplace=True)

# Normalize/standardize the numerical columns

scaler = StandardScaler()

data[['hg/ha\_yield', 'average\_rain\_fall\_mm\_per\_year', 'pesticides\_tonnes', 'avg\_temp']] = scaler.fit\_transform(

data[['hg/ha\_yield', 'average\_rain\_fall\_mm\_per\_year', 'pesticides\_tonnes', 'avg\_temp']]

)

# One-hot encoding for the categorical variables 'Area' and 'Item'

encoder = OneHotEncoder(sparse\_output=False)

encoded\_features = encoder.fit\_transform(data[['Area', 'Item']])

# Create a DataFrame for the encoded features and concatenate with the original data

encoded\_df = pd.DataFrame(encoded\_features, columns=encoder.get\_feature\_names\_out(['Area', 'Item']))

cleaned\_data = pd.concat([data.drop(['Area', 'Item'], axis=1).reset\_index(drop=True), encoded\_df], axis=1)

return cleaned\_data

data = pd.read\_csv(r'yield\_df.csv/yield\_df.csv')

cleaned\_data = clean\_data(data)

cleaned\_data = cleaned\_data.drop(columns=['Unnamed: 0'])

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_squared\_error, r2\_score

from sklearn.svm import SVR

from sklearn.tree import DecisionTreeRegressor

from xgboost import XGBRegressor

from sklearn.linear\_model import LinearRegression

def train\_and\_evaluate\_model(model, features, target):

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(features, target, test\_size=0.2, random\_state=0)

# Train the model

model.fit(X\_train, y\_train)

# Predictions and model evaluation

y\_pred = model.predict(X\_test)

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

return mse, r2

def evaluate\_models(data):

features = data.drop('hg/ha\_yield', axis=1)

target = data['hg/ha\_yield']

models = {

'Linear Regression': LinearRegression(),

'Support Vector Machine': SVR(),

'Decision Tree': DecisionTreeRegressor(),

'XGBoost': XGBRegressor()

}

results = {}

for model\_name, model in models.items():

mse, r2 = train\_and\_evaluate\_model(model, features, target)

results[model\_name] = {'MSE': mse, 'R-Squared': r2}

return results

model\_results = evaluate\_models(cleaned\_data)

for model\_name, metrics in model\_results.items():

print(f'{model\_name} - MSE: {metrics["MSE"]}, R-Squared: {metrics["R-Squared"]}')

# Renaming the column 'hg/ha\_yield' to 'yield' in the cleaned\_data DataFrame for clarity

cleaned\_data.rename(columns={'hg/ha\_yield': 'yield'}, inplace=True)

import statsmodels.api as sm

from statsmodels.formula.api import ols

data.rename(columns={'hg/ha\_yield': 'yield'}, inplace=True)

# Filter the relevant columns

anova\_data = data[['Area', 'yield']]

anova\_data.rename(columns={'yield': 'crop\_yield'}, inplace=True)

# Prepare the ANOVA model

model = ols('crop\_yield ~ C(Area)', data=anova\_data).fit()

anova\_table = sm.stats.anova\_lm(model, typ=2)

# Display the ANOVA table

print(anova\_table)