A PROJECT REPORT

on

"CROP YIELD PREDICTION"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN COMPUTER SCIENCE AND COMM. ENGINEERING

BY

PRATYUSH KUMAR PRASAD	2128035
KALYANBRATA GIRI	2128075
LAVANYA UPADHYAY	2128077
SUDESHNA RATH	2128101

UNDER THE GUIDANCE OF DR. AMIYA RANJAN PANDA

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESHWAR, ODISHA - 751024

March 2024

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

"CROP YIELD PREDICTION"

submitted by

Pratyush Kumar Prasad	2128035
Kalyanbrata Giri	2128075
Lavanya Upadhyay	2128077
Sudeshna Rath	2128101

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2023-2024, under our guidance.

Date: 10/4/2024

(Dr. Amiya Ranjan Panda)
Project Guide

Acknowledgment

We really appreciate DR. Amiya Ranjan Panda for his knowledgeable counsel and unceasing support in ensuring that this project meets its goals from the outset to the end.

Additionally, I would like to express my gratitude to each and every one of the study participants who helped us out with insightful and useful data. This initiative would not have been achieved without their willingness and participation.

In addition, I want to express my gratitude to my friends and coworkers for their encouragement and support during the project. Their kind criticism and recommendations have been very helpful in raising the caliber of this report.

Lastly but not the least, I want to sincerely thank my family.Their unwavering understanding, constant support, and unwavering encouragement were my pillars of strength.

PRATYUSH KUMAR PRASAD
KALYANBRATA GIRI
LAVANYA UPADHYAY
SUDESHNA RATH

ABSTRACT

The capacity to forecast crop yields with sufficient accuracy is the foundation of a successful agriculture industry. Despite their value, traditional methods frequently fall short when faced with the complex web of variables that affect crop development. This paper explores machine learning's emerging potential as a ground-breaking crop yield forecast tool. With the help of machine learning, one may analyze large datasets and uncover the intricate linkages that control crop health and productivity.

This revolutionary strategy is based on a plethora of machine learning algorithms, including several regression approaches such as support vector regression, lasso regression, ridge regression, decision tree regression, K-nearest neighbors regression, random forest regression, and linear regression. To improve forecast accuracy, these algorithms make use of the abundance of historical data on weather trends, soil properties, and yields. Through the process of choosing the best crops for a given growing environment and making prudent resource allocation decisions, machine learning models enable farmers to make better decisions at every stage of the agricultural cycle.

Furthermore, machine learning-driven forecasting ushers in a new era of unmatched accuracy in agricultural yield estimates, beyond the constraints of traditional methodologies. Moreover, ML-based forecasting enhances the precision of crop forecasts and allows farmers to assess the resilience and adaptability of different crops.

With this increased understanding, farmers are better equipped to evaluate the adaptation and resilience of different crops, allowing them to make decisions that support agricultural sustainability.

In summary, the use of ML to predict crop production promotes productivity, helps make well-informed decisions, and keeps the agriculture industry sustainable.

Contents

List of Figures

Chapter 1

Introduction

Agricultural productivity and food security are critical global concerns, particularly in the face of a growing population and changing environmental conditions. Predicting crop yield accurately is fundamental for effective agricultural planning, resource allocation, and risk management. In recent years, advancements in technology, data analytics, and interdisciplinary research have opened new avenues for enhancing crop yield prediction.

This research project aims to contribute to the ongoing efforts in improving crop yield prediction through an interdisciplinary approach that integrates agronomy, environmental science, data science, and machine learning techniques. By leveraging historical data, environmental factors, and cutting-edge predictive models, we seek to address key challenges in crop yield prediction and provide actionable insights for farmers, policymakers, and stakeholders in the agricultural sector.

Crop yield prediction plays a pivotal role in agricultural decision-making processes, including crop selection, planting strategies, resource allocation. Accurate predictions enable farmers to optimize their practices, mitigate risks associated with environmental factors, and ultimately enhance productivity and profitability. Furthermore, with the increasing pressure to meet global food demand sustainably, there is a growing need for innovative approaches that consider the complex interactions between agro ecological factors, climate variability, land management practices, and socio-economic dynamics.

This project seeks to address these challenges by harnessing the power of data-driven methodologies and interdisciplinary collaboration.It investigates the relationship between environmental factors (such as weather patterns) and crop yield variability.It develops and evaluates predictive models using machine learning algorithms to forecast crop yield with high accuracy and precision.This provides actionable insights and recommendations for farmers, policymakers, and stakeholders to optimize agricultural practices, enhance resilience, and promote sustainable food production systems.

The research will employ a combination of quantitative analysis and machine learning algorithms to analyze large-scale datasets encompassing historical crop yield data, environmental variables. Various predictive models, including regression analysis, random forests, and support vector regressions are trained and validated using evaluation metrics to assess their performance and generalization capabilities.

Chapter 2

Basic Concepts/ Literature Review

2.1 Crop Yield

Crop yield measured as hectogram per hectare (Hg/Ha), is the amount of crops harvested from a given area and relies on a complex interplay of various factors.Factors include weather conditions like temperature, and rainfall. Even agricultural practices like irrigation techniques and fertilizer application significantly influence crop yield. Optimizing these parameters through proper management practices is crucial for maximizing crop yields.

2.2 Linear Regression

Linear regression is a statistical tool used to understand how one factor (independent variable) influences another (dependent variable). In crop yield prediction, this translates to examining how factors like temperature, rainfall, and farming practices affect the final harvest. It assumes a straight-line relationship between these factors and yield. The model aims to find the best-fitting line that minimizes the difference between actual and predicted yields.

2.3 Ridge Regression

Regression models with multicollinearity can be managed using ridge regression, a regularization method. By effectively reducing the magnitude of the coefficients to zero, the penalty term stops overfitting. When predicting agricultural production based on tightly correlated predictor variables such as temperature and rainfall, ridge regression proves to be highly advantageous.

2.4 Lasso Regression

Similar to ridge regression, Lasso regression uses regularization to lessen overfitting and enhance the readability of regression models.In contrast to ridge regression, it uses the L1 norm for regularization and adds a penalty component to the OLS objective function.Some coefficients in the coefficient vector become completely zero as a result of Lasso regression's increase in sparsity.Because it can automatically identify and select the most relevant predictors for agricultural yield prediction, lasso regression is advantageous for feature selection.

2.5 Gradient Boosting Regression

Gradient Boosting is a machine learning technique that uses gradient descent optimization and decision trees to produce precise crop yield predictions. For regression tasks, it is an ensemble machine learning approach. Gradient descent is used to reduce the errors of the previous model, resulting in the successive construction of an ensemble of weak regression models, typically decision trees. It is often used to combine predictions from multiple models to boost accuracy in a range of industries due to its great predictive performance and longevity.

2.6 K Nearest Neighbours Regression (KNN)

K Nearest Regression is a proximity-based method that forecasts values or labels for incoming data points based on the proximity of prior data points. In regression, it predicts the average value among the K nearest neighbors; in classification, it assigns the majority class label. The K parameter controls the number of neighbors that are considered and influences the performance of KNN. While KNN is simple to use and intuitive, it can be computationally demanding, especially when working with large datasets, and it might not perform well with features that are irrelevant or noisy. Still, it's a decent beginning model, especially when applied to smaller datasets or in conjunction with ensemble approaches.

2.7 Decision Tree Regression

For regression tasks, decision tree regression is a non-parametric supervised learning method.The feature space is divided recursively into regions, with each division denoting a choice made in response to a function value.With each leaf node representing a training instance of an item, decision tree regression forecasts the average target value of the object.Decision trees work well for performance prediction problems when explainability is crucial since they are simple to understand and intuitive.

.

2.8 Random Forest Regression

Random Forest is an ensemble learning technique used for classification and regression issues. It builds multiple decision trees using bootstrapped samples of the data and features randomness. The final prediction in classification is determined by voting, while in regression, it is determined by averaging over all trees. Because Random Forest is robust and less likely to overfit, it performs well when applied to high-dimensional data.

2.9 Support Vector Regression

Support Vector Regression (SVR), a supervised machine learning technique, is applied to regression issues. It finds a hyperplane that maximizes the margin between data points within a specific margin (epsilon) in order to reduce error. SVR addresses non-linear connections by using kernel functions and is robust to outliers. Because of its memory efficiency, it is widely used in finance, economics, and engineering for applications like stock price prediction and function approximation.

Chapter 3

Problem Statement / Requirement Specifications

3.1 Project Planning

The successful execution of a machine learning project requires careful planning and systematic implementation. In the planning phase of our project, we outlined the following key steps to ensure the smooth progression of the project:

3.1.1. Define Objectives: The foremost step is to determine clear objectives and futuristic scopes of the study. Our objective was to develop a Crop Yield Prediction Model using Machine Learning by implementing the Crop Yield Prediction Dataset obtained from Kaggle which was originally taken from the World FAO. Our main goal was to evaluate various models used in predicting crop yield on the basis of various factors like Rainfall, Temperature and Pesticides used.

3.1.2. Data Acquisition: After defining our objectives, we go through the required dataset. For this problem statement, we obtained a dataset from Kaggle, which was originally taken from the FAO website, which analyzes the crop yields of ten "most consumed" crops all over the world. Confirming data quality and relevancy was crucial in selecting a suitable dataset for analysis.

3.1.3. Data Preprocessing: Once we have chosen the most appropriate dataset, we apply preprocessing techniques in order to clean, transform and modify the data, After choosing the most appropriate dataset, for a precise analysis. Preprocessing includes normalizing the data, deals with missing values or null values, detects outlier and manages missing values.

3.1.4. Feature Selection: By employing various techniques like Correlational Analysis, Statistical tests, we analyze the relation of all the features with the output. Hence it plays an important role in the analysis of the dataset. Feature selection plays a vital role in model evolution, as it helps in identifying the most instructive variables for forecasting the yield. We employed various feature selection techniques, including correlation analysis, and statistical tests, to recognize relevant features for model training.

3.1.5. Model Selection: The preprocessed data set along with the vital features, we select and implement various machine learning models for predicting the yield of crops. The models that we've chosen are- Linear Regression, Gradient Boosting Regression, Ridge Regression, Lasso Regression, k-Nearest Neighbor Regression, Decision Tree Regression, Random Forest Regression & Support Vector Regression.

3.1.6. Model Evaluation: Once we've trained the models, we evaluate their performances

using appropriate evaluation metrics. The error metrics that we have enforced are Pearson Correlation Coefficient(R), R Square Score, Root Mean Squared Error(RMSE), Mean Absolute Error(MAE) & Mean Absolute Percentage Error (MAPE).

3.1.7. Result Analysis: The conclusive step involved is analyzing the results obtained from model evaluation and concluding the effectiveness of different algorithms in predicting the yield of any crop. Understandings accumulated from the analysis were used to report future research directions and potential applications in agriculture.

Since the day of undertaking this project, commitment to a well-defined timeline, systematic progress monitoring, and executing improvement methods accordingly were essential for meeting project milestones and guaranteeing punctual fulfillment. Effective communication and collaboration among team members also promote knowledge-sharing and problem-solving, thereby contributing to the overall success of the project.

3.2 Project Analysis

During our project's analysis phase, we carefully examined and interpreted the outcomes of applying machine learning algorithms to provide forecasts. In order to support decision-making, this phase sought to gather information about how various algorithms performed, pinpoint important variables affecting forecast accuracy, and derive insightful conclusions.

3.2.1. Model Performance Comparison: We initiated by analogizing the performance of the eight machine learning algorithms employed in our study. Evaluation metrics including Pearson Correlation Coefficient (R), R square score, MAE, MAPE, and RMSE were computed for each model on both the training and testing datasets. The outcomes revealed deviations in predictive performance across different algorithms, with some models exhibiting exceptional performance compared to others.

3.2.2. Finding the Best-Performing Models: We identified the best-performing models based on their accuracy in yield prediction by consulting the assessment measures. Given its maximum accuracy and least error, the Random Forest Regression model is considered the best method.

3.2.3. Feature Importance Analysis: We carried out a feature importance analysis for the best-performing models in order to gain understanding of the factors influencing yield prediction. We were able to identify the most useful elements or related risk factors thanks to this investigation.

3.2.4. Limitations and Challenges: Although we obtained promising results, our analysis reveals constraints and difficulties associated with the prediction models. Among these is the

need for thorough feature refinement. Accurately interpreting the results and planning future research projects depend on an understanding of these limitations.

In conclusion, the project's study provided insightful information on the efficiency and applicability of machine learning algorithms for crop yield prediction.

3.3 System Design

3.3.1. Design Constraints

The following limitations will apply to the software system:

I. a. Budget: The project's budget has a set amount of money that must be adhered to.
II. b. Time: The software system must be designed and implemented within the allotted time limit because the project has a limited timetable.

3.3.2. System Architecture

The crop yield prediction model's system architecture consists of a number of parts and procedures intended to use machine learning models to accurately estimate yield status. There are several steps in the architecture, such as preparing the data, training the model, testing, and deployment. An overview of the system architecture can be found below:

- Data Collection and Preprocessing:

  Preprocessing techniques include addressing missing values, encoding categorical variables, and normalizing the data in order to clean it up and make it ready for analysis.

- Feature Selection and Engineering:
  I. To enhance the models' predicting ability, pertinent elements are selected or engineered from the dataset.
  II. One may use feature selection algorithms like principal component analysis, correlation analysis, or recursive feature elimination.
- Model Training:
  I. The preprocessed data is used to train multiple machine learning algorithms.
  II. Decision tree regression, logistic regression, k-nearest regression, random forest regression, support vector regression, and gradient boosting regression are among the models.
- Model Evaluation and Testing:
  I. Performance measures including the Pearson correlation coefficient (R), R Square Score, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) are used to assess the trained

models.

II. Techniques for cross-validation can be applied to guarantee the algorithms' resilience.
III. A different test dataset is used to evaluate the algorithms' capacity for generalization.

- Model Deployment:
  I. The top-performing model is then implemented for real-time prediction in a production setting.
  II. Developing standalone apps or integrating the model into already-existing software systems may be required for deployment.
  III. To track the performance of the deployed model and make appropriate changes or adjustments, monitoring mechanisms are implemented.

- Scalability and Performance Optimization:

I. The system architecture is crafted to be scalable, capable of handling extensive datasets and growing demands for prediction.
II. Methods for enhancing performance, such as utilizing parallel processing, distributed computing, or hardware acceleration, could be utilized to improve the efficiency of the system.

The yield prediction project's system architecture is generally designed to make effective use of machine learning models, guaranteeing scalability, dependability, and efficiency. Its objective is to provide accurate yield production projections by leveraging relevant attributes extracted from the data.

Chapter 4

Implementation

4.1 Methodology OR Proposal

4.1.1.Data Exploration and Preprocessing

The first phase of the project involves Data Exploration and applying preprocessing methods to analyze and revise the dataset. After reviewing the dataset's structure, it is observed that there are 28,242 rows and 7 columns. Each column represents a crop-related attribute. The final output or the target variable is the Yield of the crops (denoted by hg/ha_yield).

4.1.2.Extensive data analysis and feature selection

Following the initial data breakdown, we dive deeper into comprehending the interrelationships between attributes and the target variable. To obtain a comprehensive knowledge of the correlations within the dataset, we utilized a correlation heatmap. The heatmap showed several significant correlations among variables.

Later, we conducted various relational tests like plotting scatter plots, counterplots, and histograms to get an even clearer association among the attributes. After that, we preprocess the data and clean them by removing null values and modifying missing values.

Via these relentless feature selection techniques, we obtain the clean data that will ensure that our predictive models are built on a straightforward yet exhaustive set of input variables, thereby improving their interpretability and predictive accuracy.

4.1.3.Data balancing, splitting and scaling

We proceeded to split the dataset into feature variables and the target variable. Subsequently, we further divided the dataset into training and testing sets, allocating 80% of the data for training and dedicating the remaining 20% for testing purposes. This division assures that our models are trained on an adequately extensive amount of the data while maintaining a distinct set of observations for evaluating model performance and generality.

In rehearsal for modeling, we conducted data scaling to standardize the feature variables. This scaling process is crucial for enhancing the confluence speed of distinctive machine learning algorithms and enhancing the general stability and performance of our models. By homogenizing the feature variables, we facilitate a more sufficient comparison of their contributions to the predictive outcome, thereby

improving the robustness and interpretability of our models.

4.1.4.Data Modeling

For the predictive analysis, we employed a diverse ensemble of eight regression models to train and evaluate their performance on the training and testing sets.

These standards enclose a range of algorithmic techniques, individually presenting distinguishable advantages and knowledge in managing the elaborateness of the dataset and expecting the target variable, Yield.The ten classifier models utilized in our study are K-nearest Regression, logistic regression, Ridge regression, Lasso Regression, decision tree regression, random forest regression, support vector regression, and gradient boosting regression.

All the models were trained on the training set, which includes 80% of the dataset, and subsequently evaluated on the testing set, constituting the remaining 20% of the data. This approach enabled us to estimate the generalization performance of the techniques and determine their effectiveness in predicting the yield for any unseen data. We aimed to exhaustively estimate the performance of classifier models by employing a diverse set of them and identify the most effective model for predicting crop yield according to the dataset's features. This stringent evaluation procedure encourages informed decision-making concerning the preference of the most suitable model for real-world applications.

4.2 Testing OR Verification Plan

This testing strategy aims to verify that the crop yield prediction algorithm is operating accurately and yielding accurate results. Unit tests for the system's constituent parts and end-to-end tests for the system as a whole are included in the testing strategy.

✔ Unit Tests

1. Data Pre-processing:
   - Test case 1: Verify that there are no missing values following data processing.
   - Test case 2: Verify that categorical variables are encoded correctly.

2. Linear Regression:
   - Test case 1: Examine the linear regression model's accuracy using the training dataset.
   - Test case 2: Examine the linear regression model's accuracy using the testing dataset.

3. Ridge Regression
   - Test case 1 : Examine the ridge regression model’s accuracy using the training dataset.
   - Test case 2: Examine the accuracy of the ridge regression model with the testing dataset.

4.Lasso Regression
- Test case 1 : Examine the lasso regression model's accuracy using the training dataset.
- Test case 2: Examine the accuracy of the lasso regression model with the testing dataset.

5. Decision Tree Regression :
- Test case 1: On the training dataset, make sure the decision tree model is constructed correctly.
- Test case 2: Using the testing dataset, validate the decision tree model's ability to accurately predict crop yield.

6. K-Nearest Neighbors Regression (KNN):
- Test case 1: Verify the KNN regressor's accuracy using the training dataset.
- Test case 2: Verify the KNN regressor's accuracy on the testing dataset.

7. Random Forest Regression :
- Test case 1: On the training dataset, make sure the random forest model is constructed correctly.
- Test case 2: Using the testing dataset, validate the random forest model's ability to accurately predict crop yield.

8. Support Vector Regression (SVR):
- Test case 1: Verify the SVR regressor's accuracy using the training dataset.
- Test case 2: Verify the SVR regressor's accuracy using the testing dataset.

9. Gradient Boosting Regression :
- Test case 1: On the training dataset, make sure the gradient bossting model is constructed correctly.
- Test case 2: Using the testing dataset, validate the gradient boosting model's ability to accurately predict crop yield.

10. Performance Metrics Calculation:
- Test case 1: Verify accurate calculation of evaluation metrics like RMSE, R2 Score , MAE , MAPE, et cetera.
- Test case 2: Verify the evaluation metrics were calculated correctly.

✔ End-to-End Tests

1. Data ingestion:
- Test case 1: Verify error free loading of dataset into the system.

2. Data Pre-processing:

- Test case 1: Verify that there are no missing values following data processing.
- Test case 2: Verify that the categorical variables are properly encoded.

3. Model Prediction:
- Test case 1: Ensure accurate prediction of crop yield for the test dataset.
- Test case 2: Verify the model's generalization to new data.

4. Results Analysis:
- Test case 1: Verify the computation accuracy of the various evaluation metrics.
- Test case 2: Verify precision of evaluation metrics that were calculated

5. Performance Testing:
- Test case 1: Calculate the model training process's execution time.
- Test case 2: Analyze how much memory is used for model training and prediction.

4.3 Result Analysis

We evaluated the performance of eight different regression models on the training and testing datasets in our outcome analysis. The assessment covered a wide range of measures to give a thorough understanding of each technique's capacity for prediction. Here, we go into detail about the evaluation results and learnings from our research.

A. Evaluation Metrics:

1. **Pearson correlation coefficient(R)**: As the absolute value of the r score approaches 1 (either favorably or negatively), the strength of the linear relationship between the variables grows. The closer the r score is to 0, the weaker the linear relationship.

2. **R Square Score** : This metric indicates the extent to which the variance in the dependent variable (goal) can be explained by the variance in the independent variables (features). On a scale of 0 to 1, a greater number indicates a better match.

3. **Root Mean Squared Error (RMSE):** This is the square root of the MSE. Interpretation is made easier by using the same units as your target variable. A lower RMSE suggests a better match.

4. **Mean Absolute Error (MAE):** This statistic is used to calculate the average absolute difference between the expected and actual values. Compared to MSE, it is less prone to anomalies. A lower MAE indicates a better match.

5. **Mean Absolute Percentage Error (MAPE)**: Lower MAPE values are indicative of better model performance. A MAPE of 0% indicates perfect accuracy, i.e., the model's predicted values exactly match the true ones.A higher MAPE score corresponds to a higher prediction error.

B. Insights and Comparative Analysis:

The following table compares the evaluation metric values for the eight regression models that are implemented.

MODELS	R	R2 Score	RMSE	MAE	MAPE
Linear Regression	0.86451514	0.747312327961836	42630.324526694	29907.491754632363	0.911746079628310 1
Ridge Regression	0.86449494	0.747304233744846 2	42631.007299088 57	29864.887583074 08	0.907828721 8818285
Lasso Regression	0.86451789	0.747326175620723 5	42629.156408211	29893.997624505 49	0.909442181 3124439
Gradient Boosting Regression	0.97676114	0.953318181 0670842	18323.165169751 66	10837.581316159703	0.342079068 16599815
K -Nearest Neighbours Regression	0.99243046	0.984902176 2487353	10420.385530907171	4620.037285521 496	0.100574066 05524502
Decision Tree Regression	0.98976682	0.979558785 9983795	12124.949935282 29	3926.906497011 76	0.084134787 37 7784882
Random Forest Regression	0.99306713	0.986162070 9 2107	9976.131423430 012	3991.277385772 1227	0.094200838 0 2563672
Support Vector Regression	0.55635755	-0.205835479 3 077959	93125.933798496 13	57810.269009312 54	1.379110927 8 814748

After analyzing the performance indicators of all eight models, a number of conclusions were drawn. The Random Forest Regression performs best in an overall sense with the lowest RMSE value of 9976.131423430012 followed by K Nearest Regression with RMSE value of 10420.385530907171 and Decision Tree Regression with RMSE value of 12124.94993528229.

Conversely, models like Gradient Boosting Regression showed competitive performance in terms of R, R2 score, and RMSE, among other measures.

Even though some models performed exceptionally well in key metrics, their overall performance differed depending on the evaluation criteria. Support vector regression, for example, performed the poorest for the dataset due to its high RMSE error. Comparatively speaking, the performance of Lasso, Ridge, and Linear regressions was comparable. These subtleties highlight how crucial it is to take into account a variety of assessment criteria in order to thoroughly evaluate model performance and choose the best strategy for crop prediction.

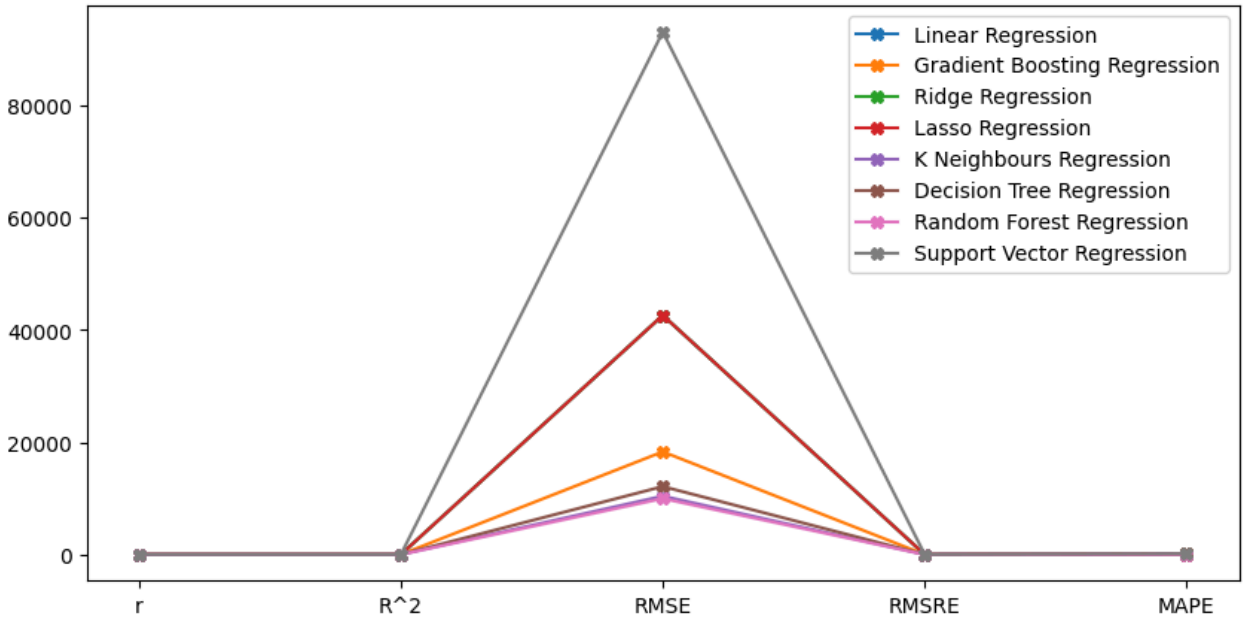

Fig 4.3.1: Comparison of evaluation metrics of all models

The above figure shows the comparison of different evaluation metrics where :

- r indicates the Pearson correlation coefficient between the predicted and actual yield value
- R^2 indicates the R score value of the model
- RMSE indicates the Root Mean Square Error between the actual and predicted yield
- RMSRE indicates the Root Mean Squared Relative Error which takes the error (difference between predicted and actual value) and divides it by the actual value
- MAPE indicates the Mean Absolute Percentage Error of the prediction.

C. Prediction and Residual Analysis

Error patterns and trends can be seen by plotting the residuals, or the difference between the values that were predicted and the actual values.

Residues from a successful model will be dispersed at random around zero (the red dashed line).

1. Linear Regression

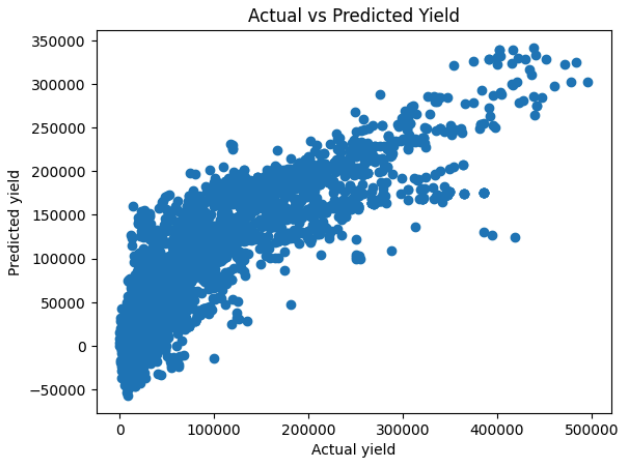

Fig 4.3.2: Scatter plot for actual vs predicted values for Linear Regression

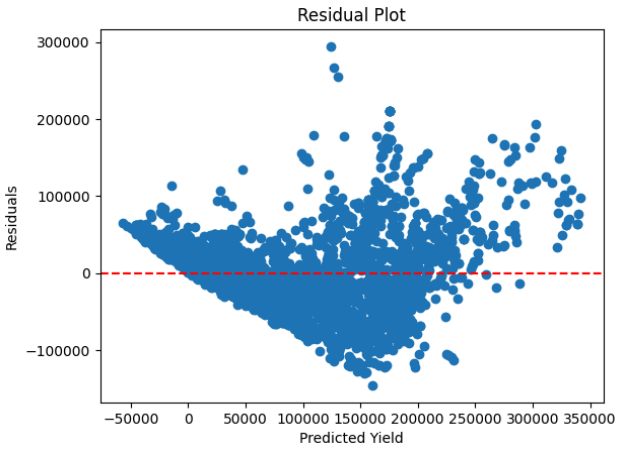

Fig 4.3.3: Scatter plot for residual values for Linear Regression

2. Ridge Regression

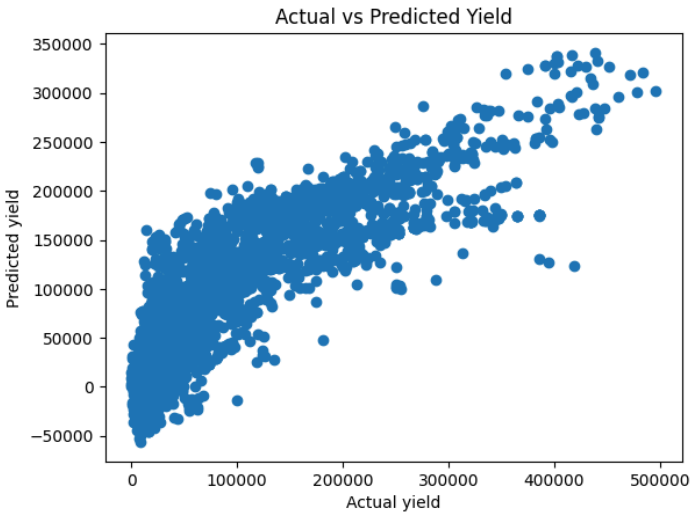

Fig 4.3.4: Scatter plot for actual vs predicted values for Ridge Regression

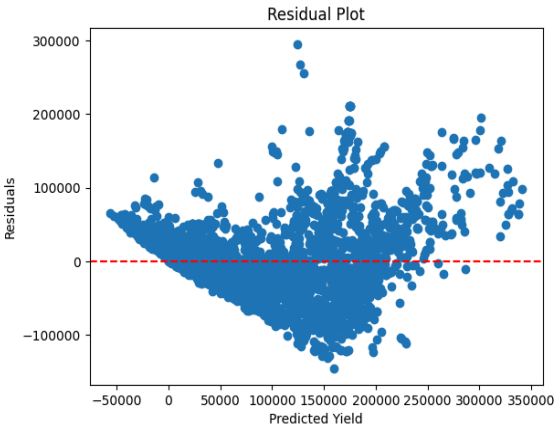

Fig 4.3.5: Scatter plot for residual values for Ridge Regression

3. Lasso Regression

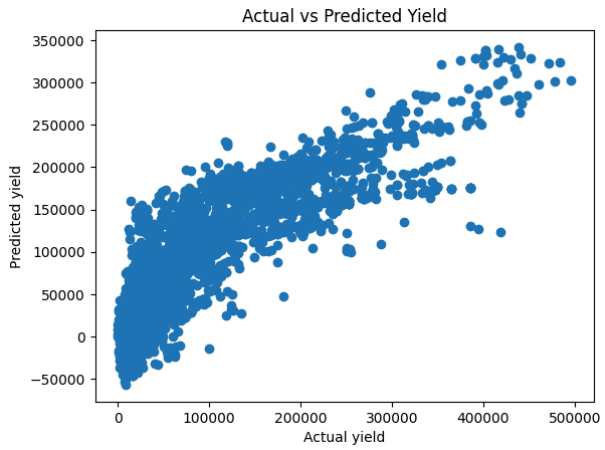

Fig4.3.6: Scatter plot for actual vs predicted values for Lasso Regression

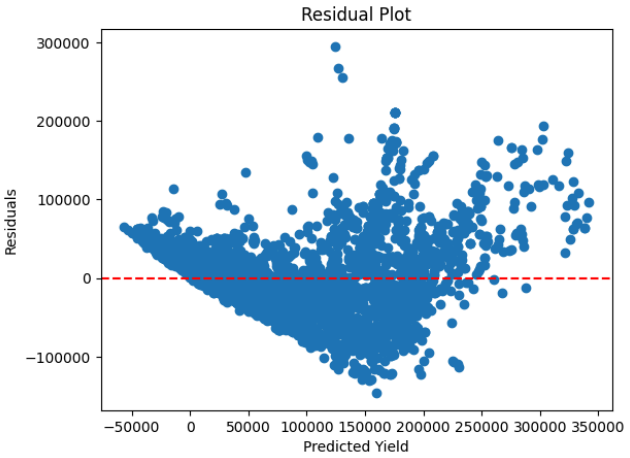

Fig 4.3.7: Scatter plot for residual values for Lasso Regression

4. Gradient Boosting Regression

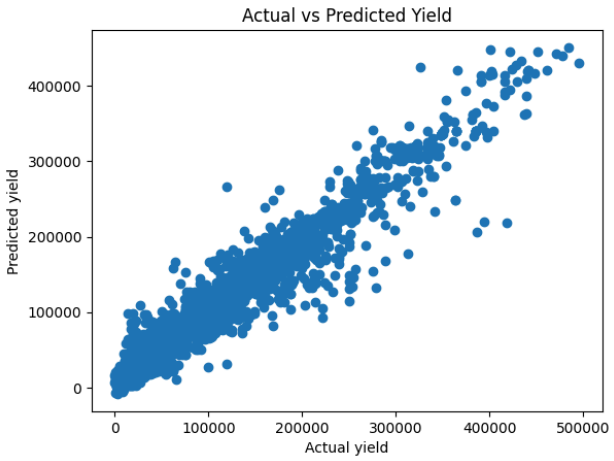

Fig 4.3.8: Scatter plot for actual vs predicted values for Gradient Boosting Regression

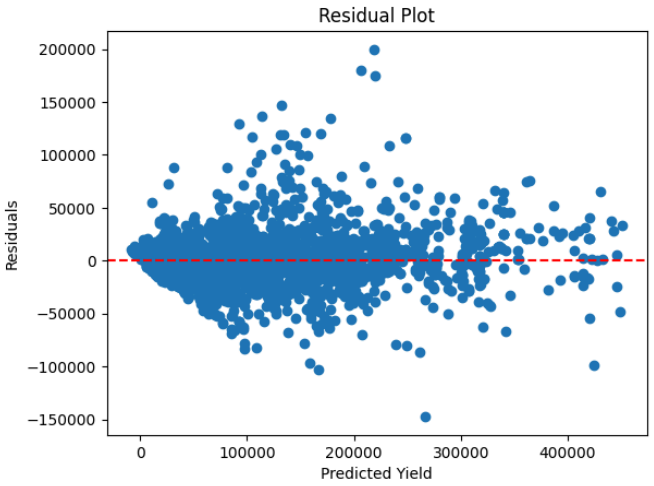

Fig 4.3.9: Scatter plot for residual values for Gradient Boosting Regression

5. K Nearest Neighbours Regression

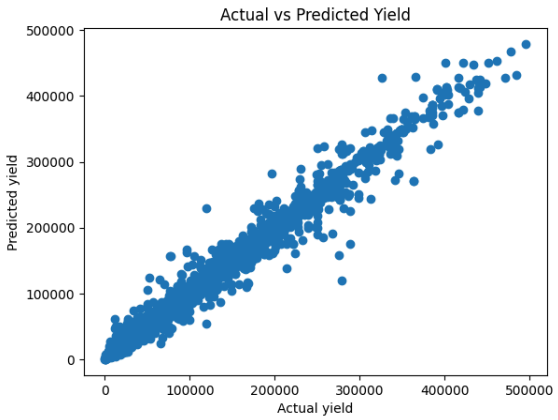

Fig 4.3.10: Scatter plot for actual vs predicted values for KNN Regression

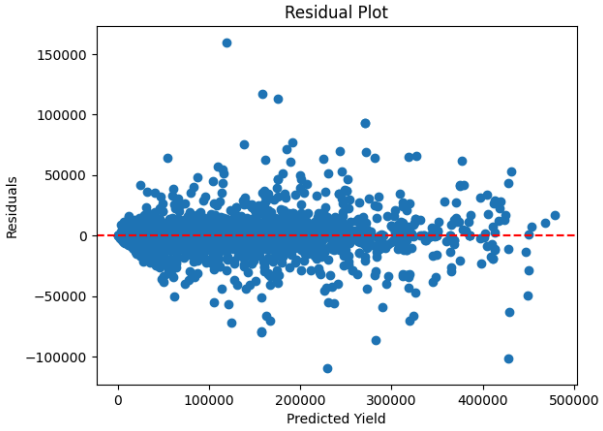

Fig 4.3.11: Scatter plot for residual values for KNN Regression

6. Decision Tree Regression

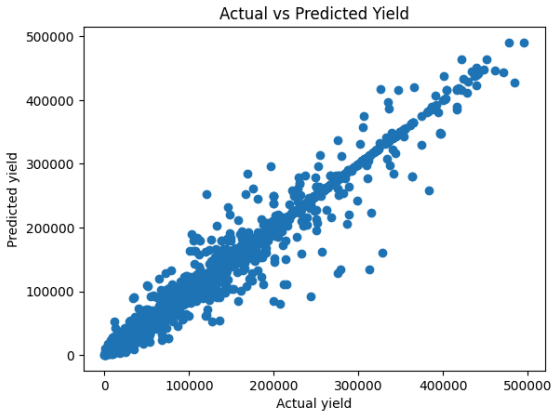

Fig 4.3.12: Scatter plot for actual vs predicted values for Decision Tree Regression

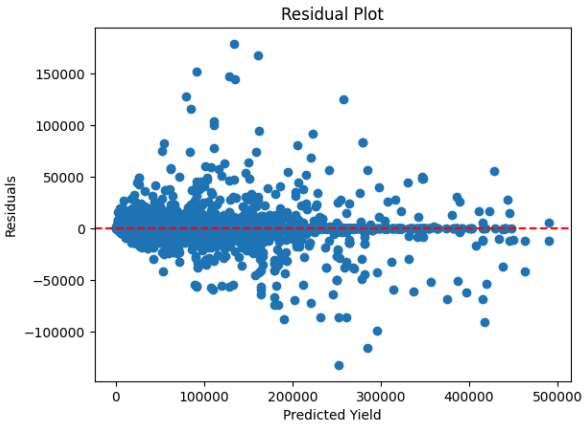

Fig 4.3.13: Scatter plot for residual values for Decision Tree Regression

7. Random Forest Regression

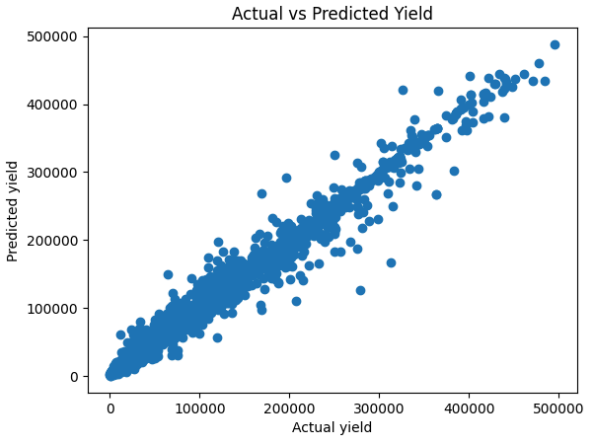

Fig 4.3.14: Scatter plot for actual vs predicted values for Random Forest Regression

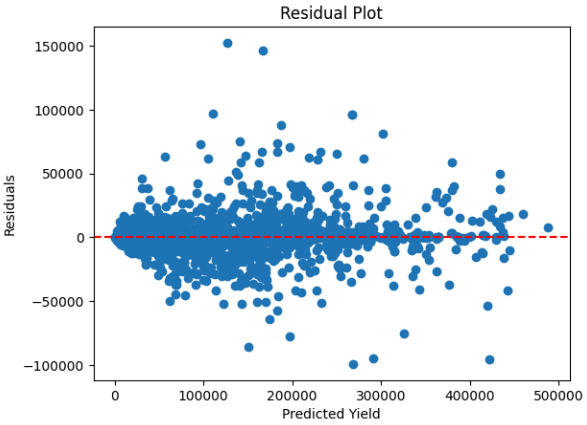

Fig 4.3.15: Scatter plot for residual values for Random Forest Regression

8. Support Vector Regression

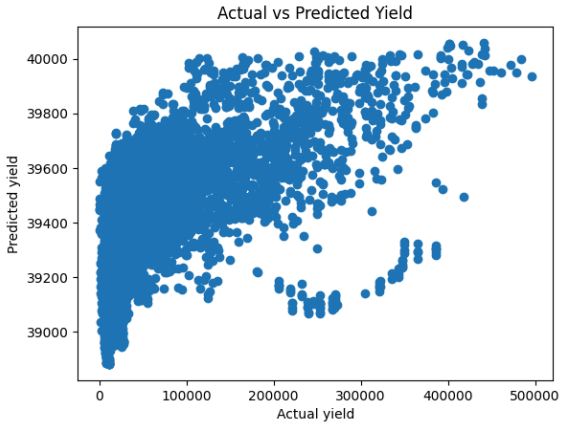

Fig 4.3.16: Scatter plot for actual vs predicted values for Support Vector Regression

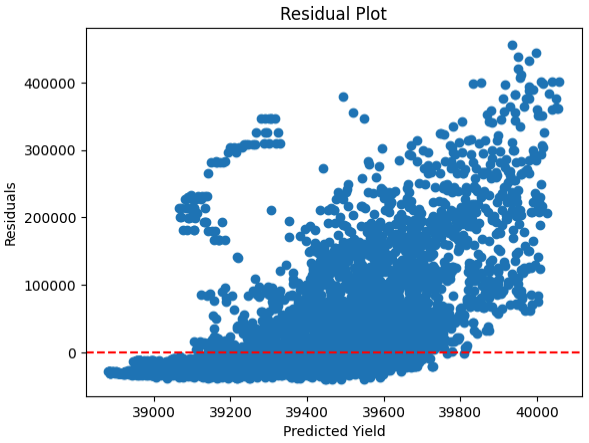

Fig 4.3.17: Scatter plot for residual values for Support Vector Regression

As shown in the figure 4.3.14 and 4.3.15, the scatter plots indicate a good distribution of residuals around the red line and linear relationship between actual versus predicted yield

values . This shows that Random Forest Regression is the best fit model amongst the models implemented in the study.

Exploratory Data Analysis Visualizations

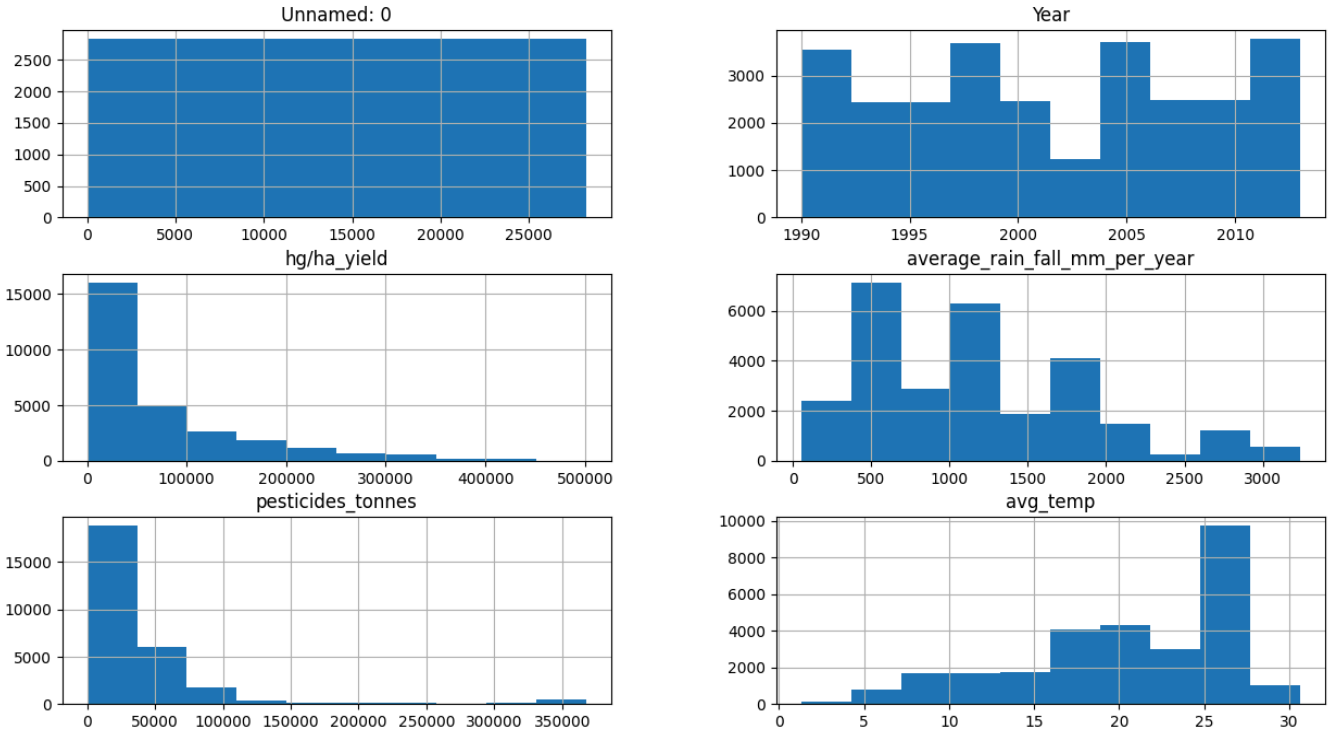

Fig 5.1 : Histograms to comprehend dataset

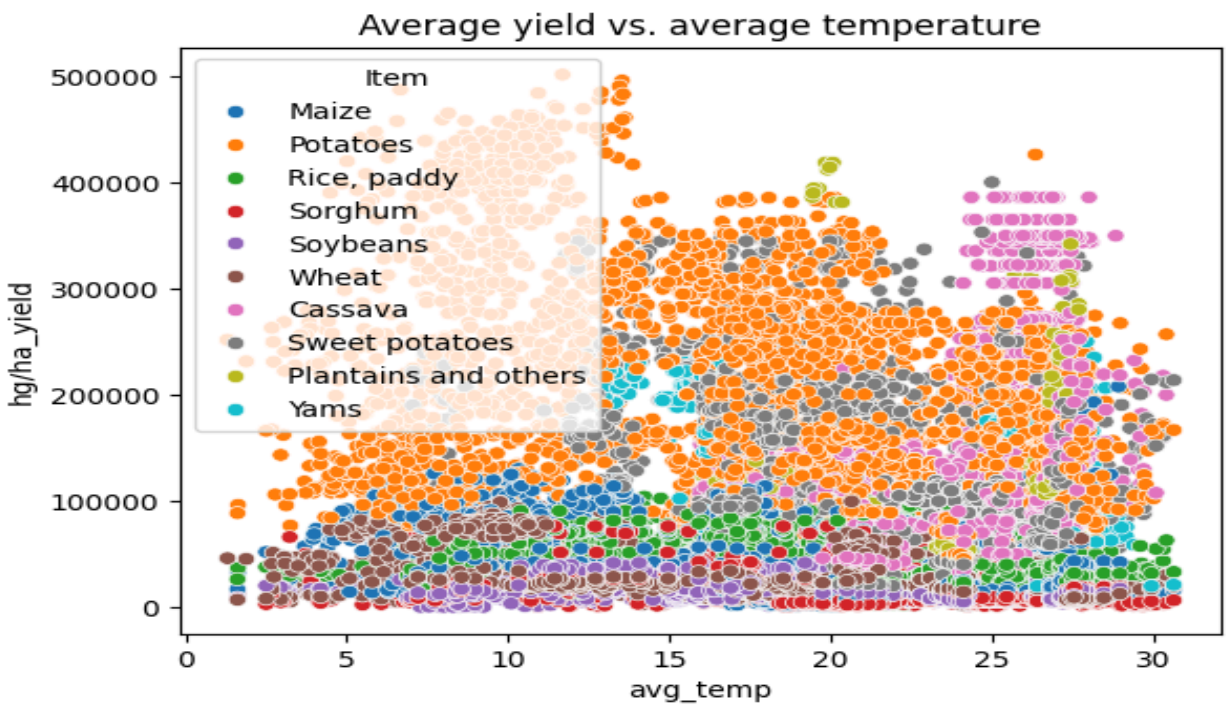

Fig. 5.2: Scatter Plot of Yield vs Avg Temperature by Crops

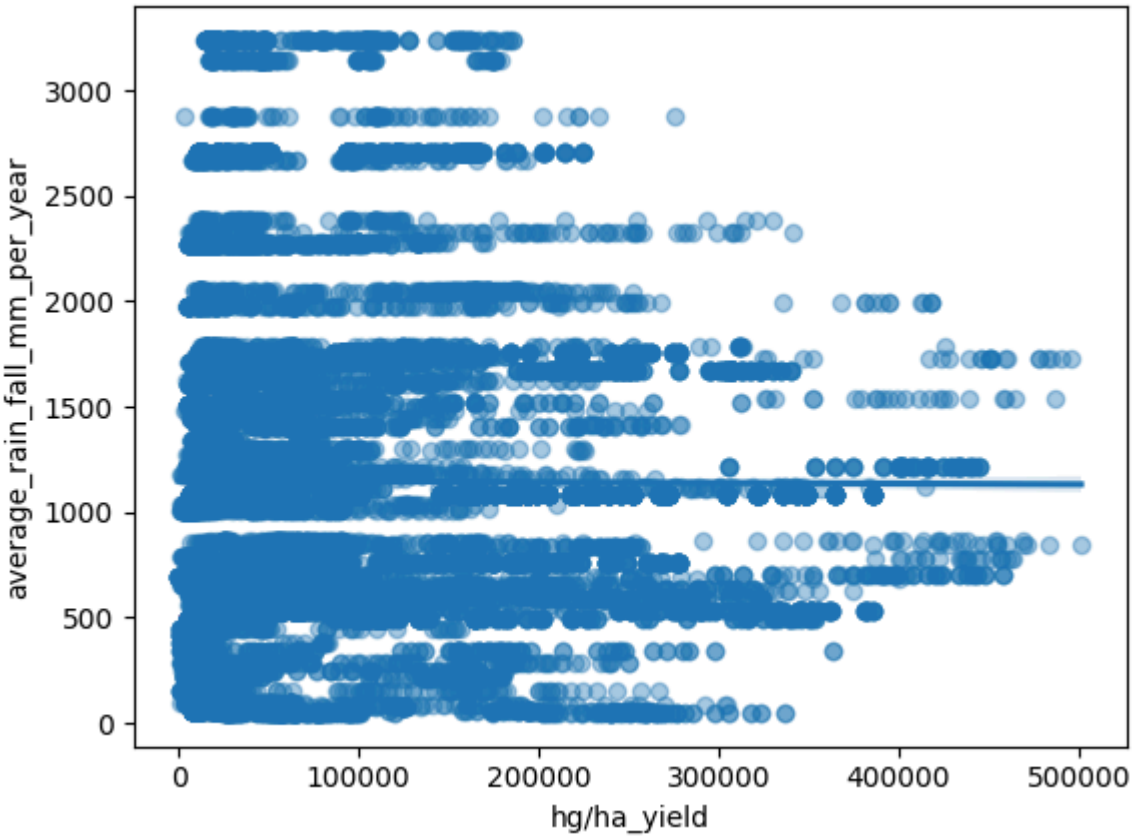

Fig. 5.3: Scatter Plot of Average Rainfall vs Yield

Fig. 5.4: Scatter Plot of Pesticides vs Yield

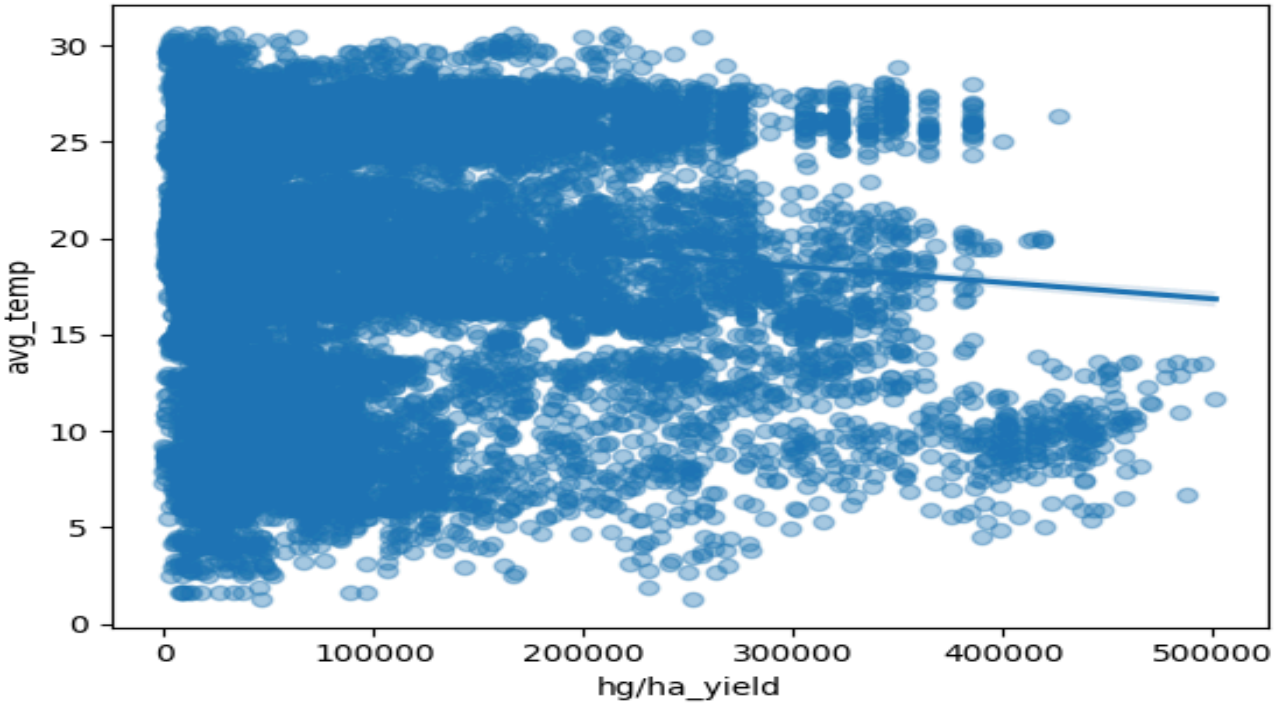

Fig. 5.5: Scatter Plot of Average Temperature vs Yield

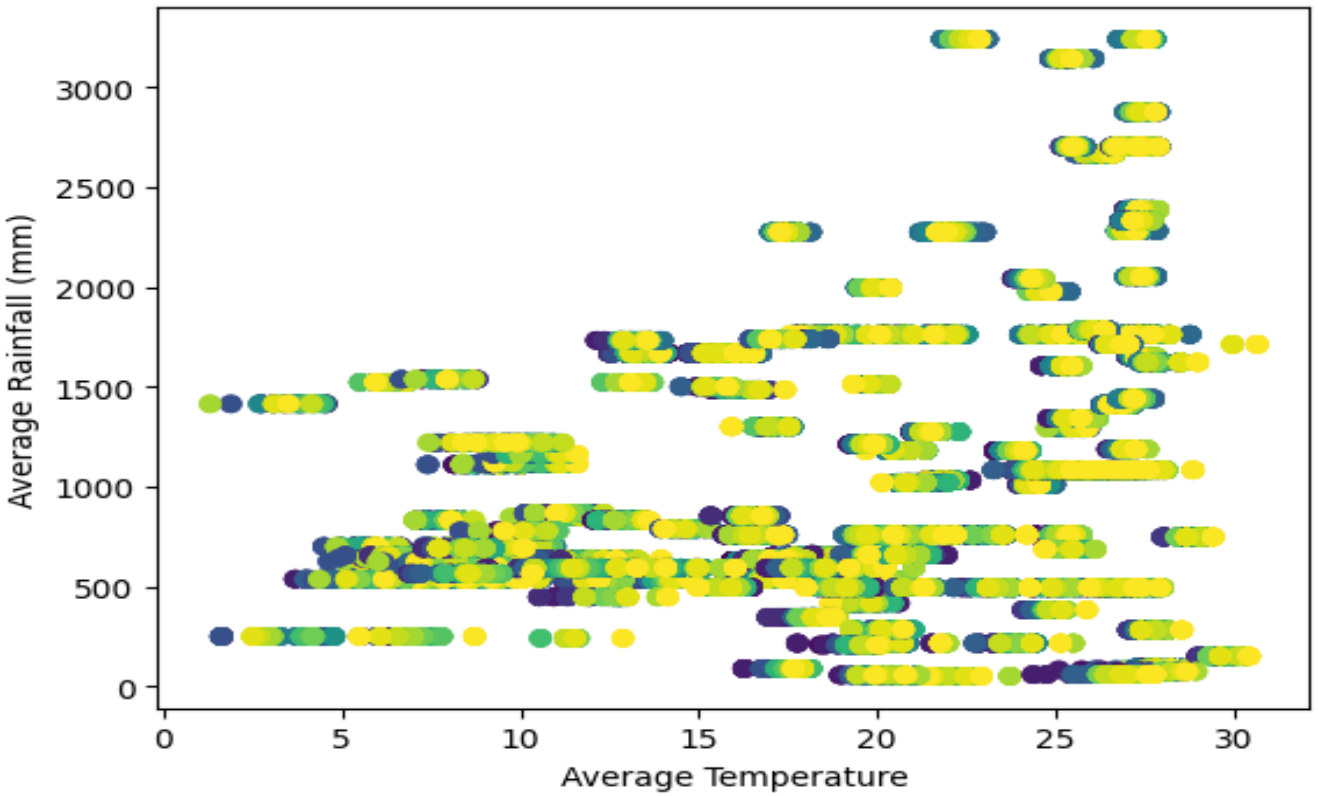

Fig. 5.6: Scatter Plot of Average Rainfall vs Average Temperature

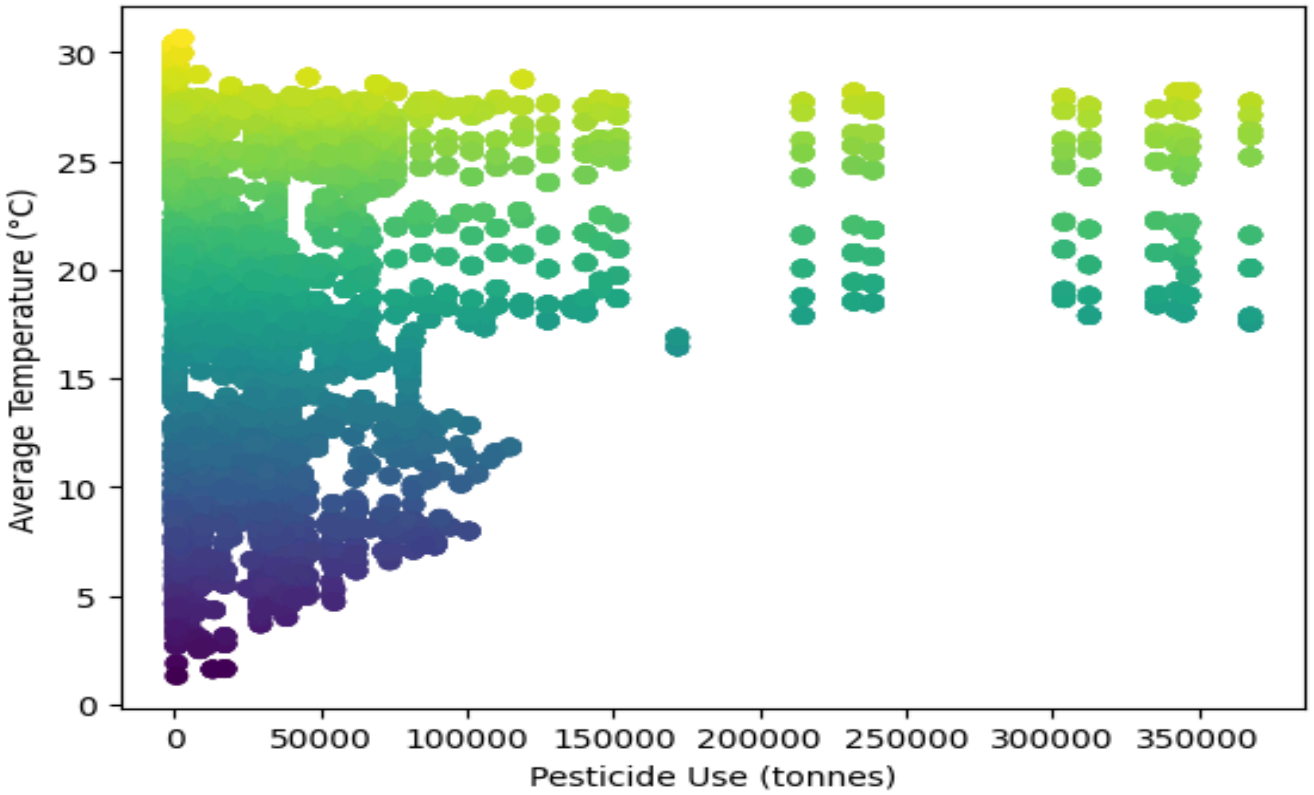

Fig. 5.7: Scatter Plot of Average Temperature vs Use of Pesticides

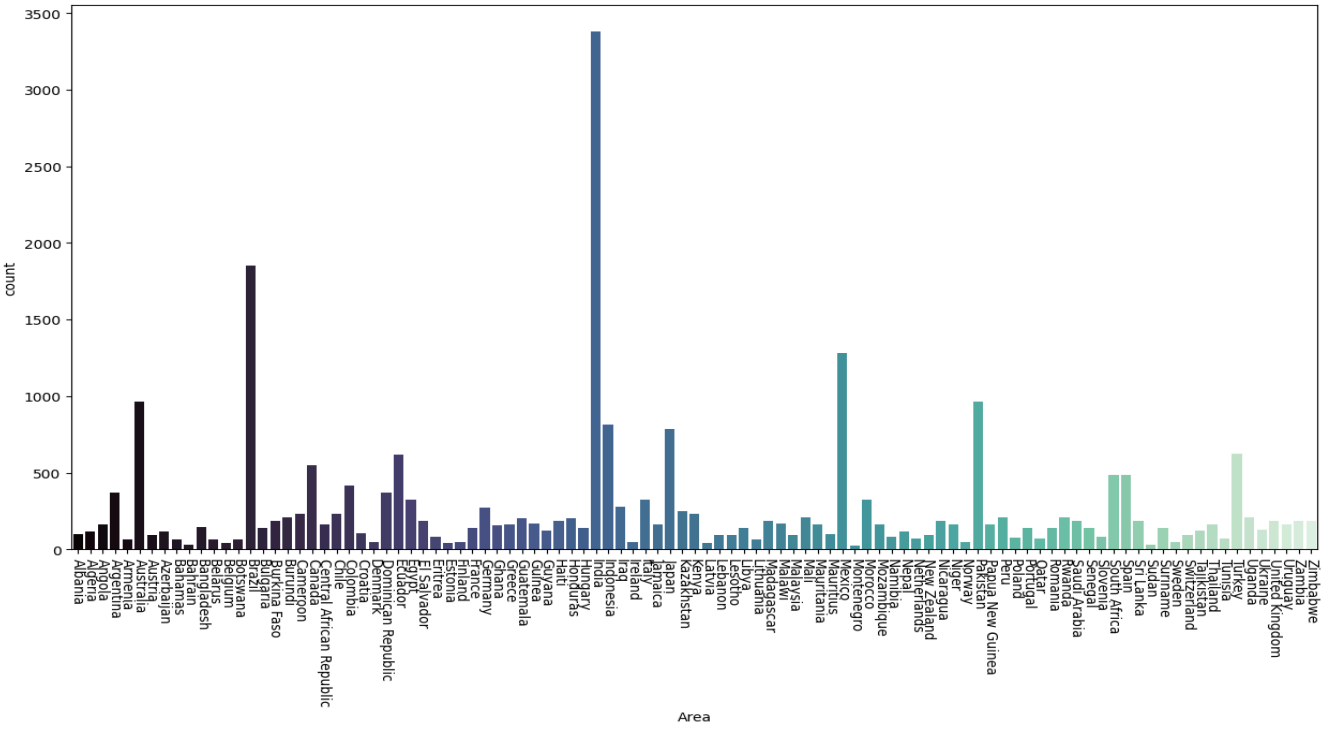

Fig. 1.8: Counter Plot of Average Temperature vs Use of Pesticides

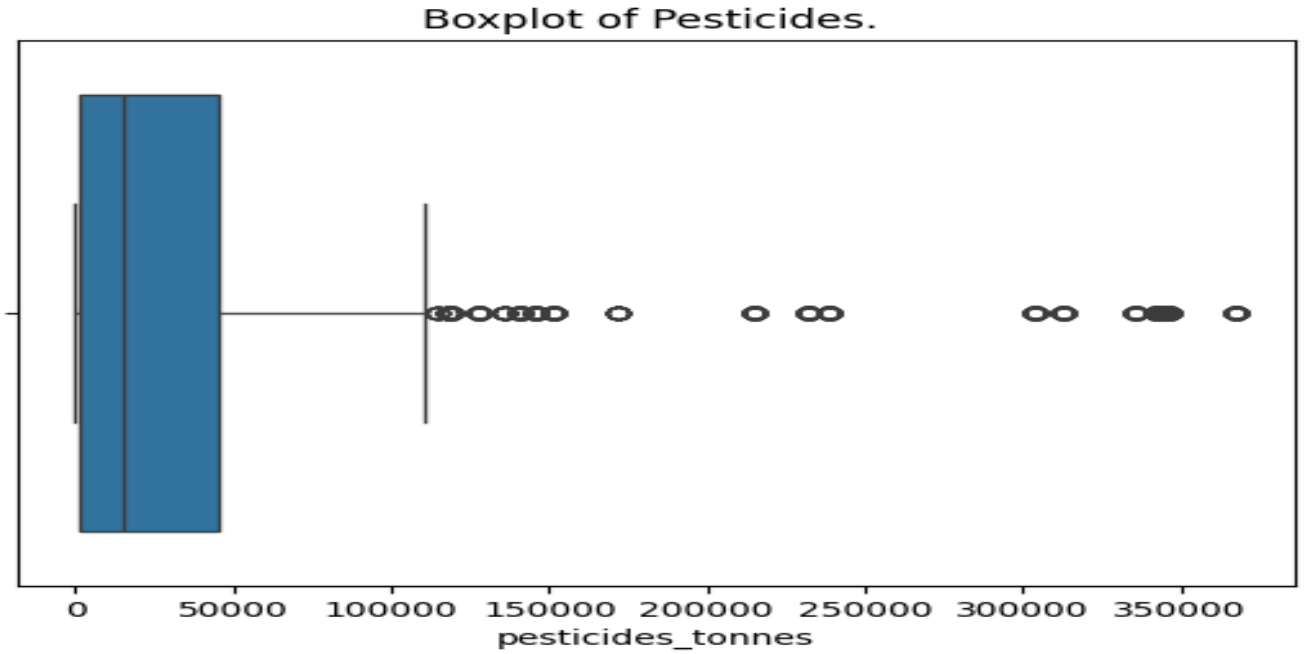

Fig. 5.9: Boxplot of Pesticides

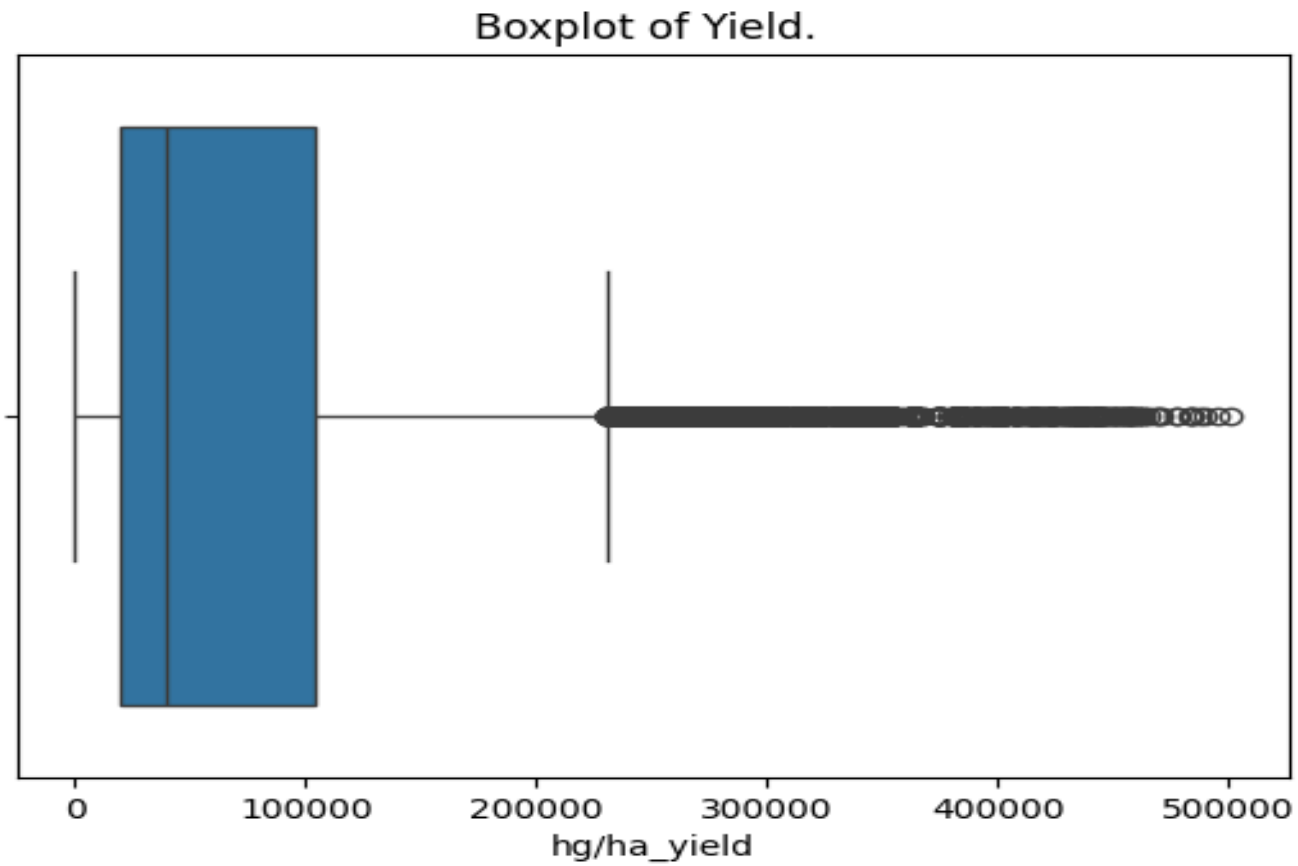

Fig. 5.10: Boxplot of Yields

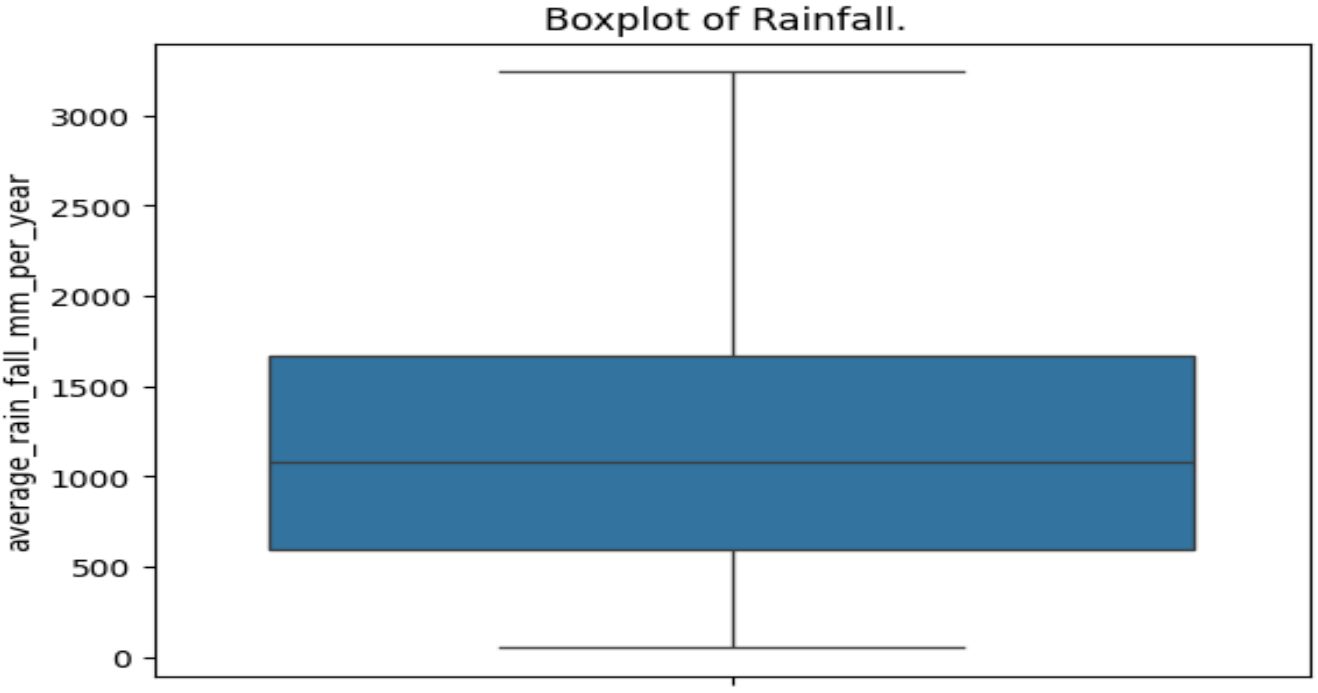

Fig. 5.11: Boxplot of Rainfalls

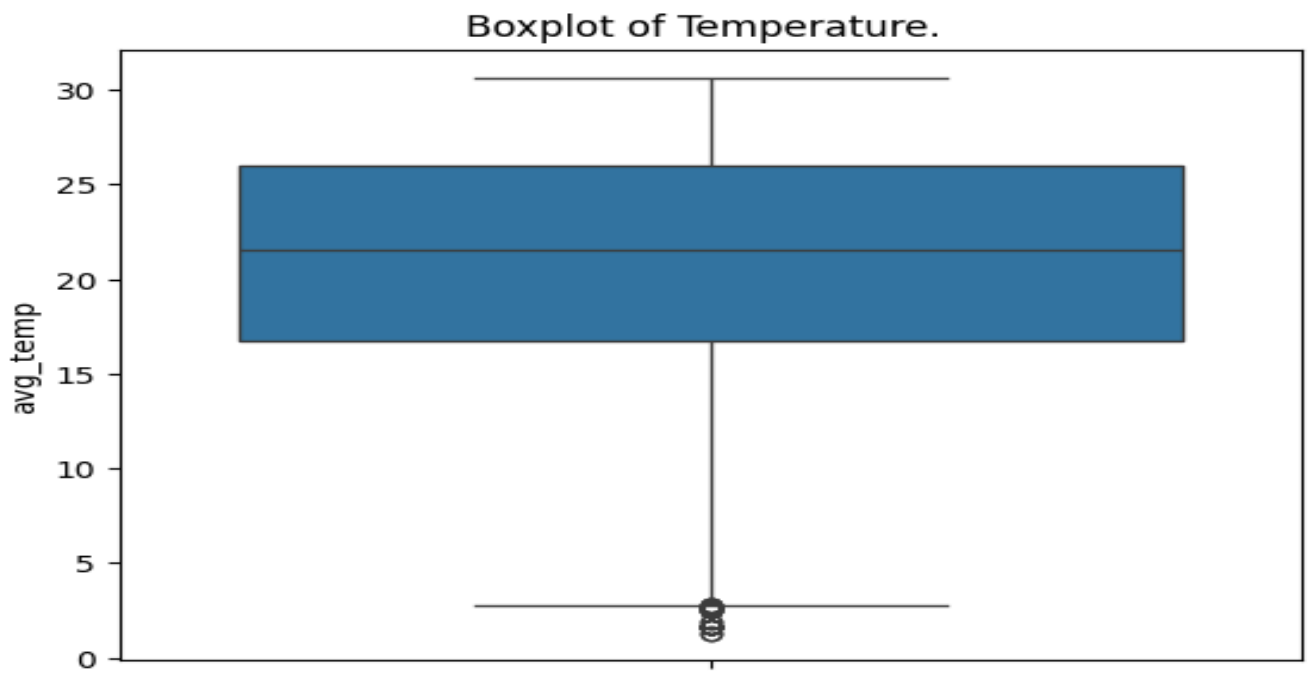

Fig. 5.12: Boxplot of Temperature

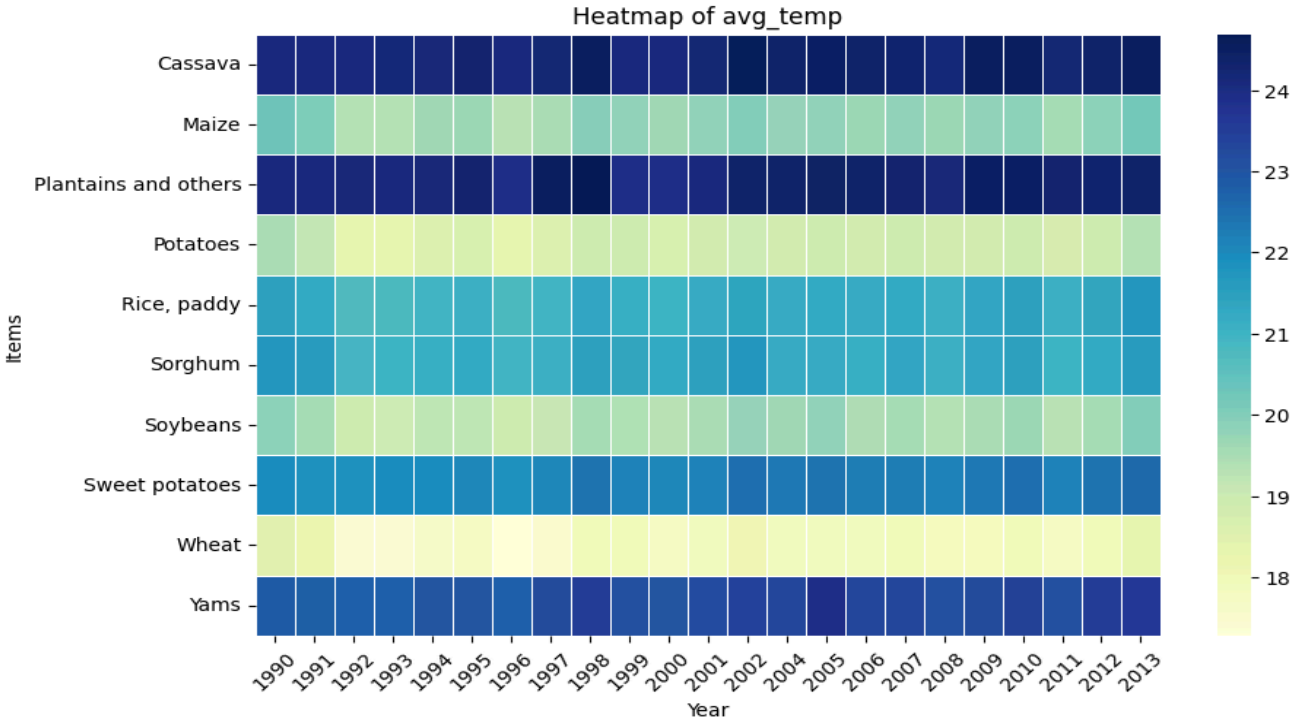

Fig. 5.13: Heatmap of Average Temperature

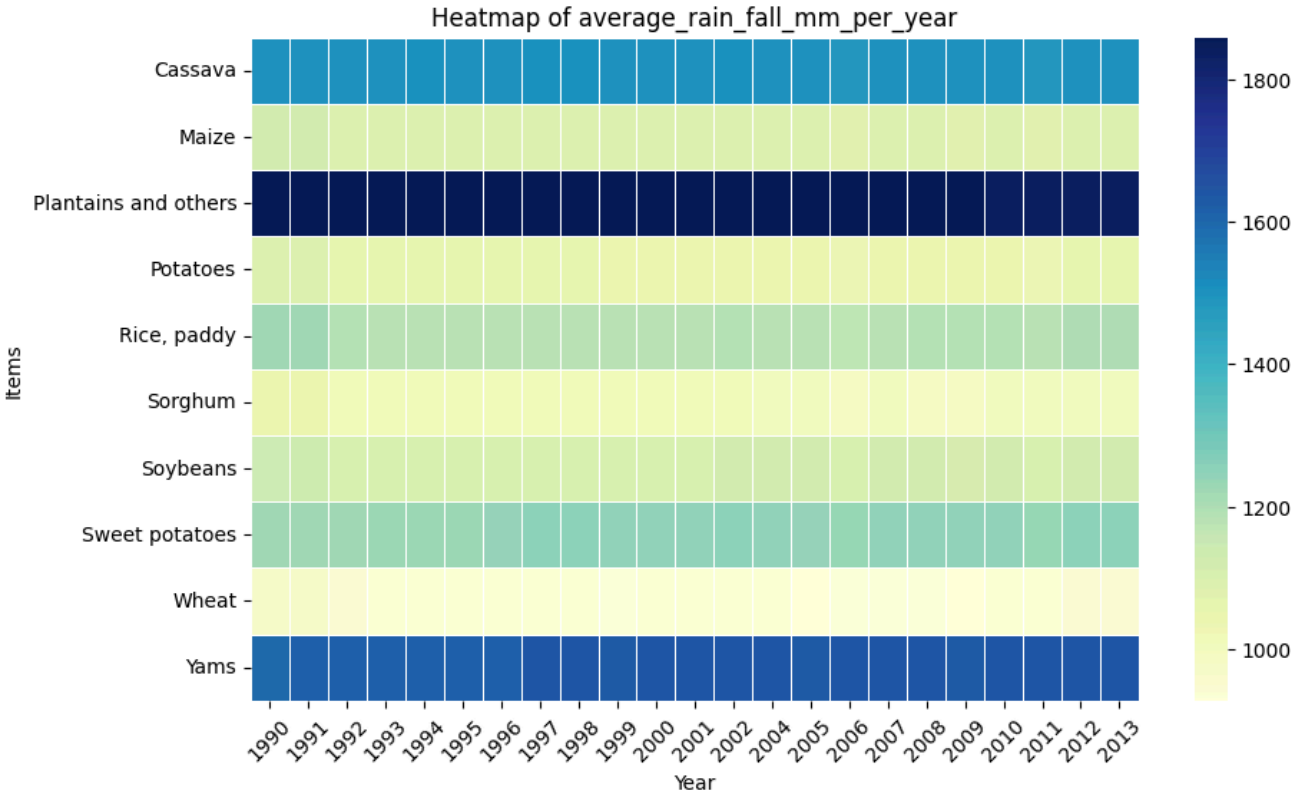

Fig. 5.14: Heatmap of Average Rainfall

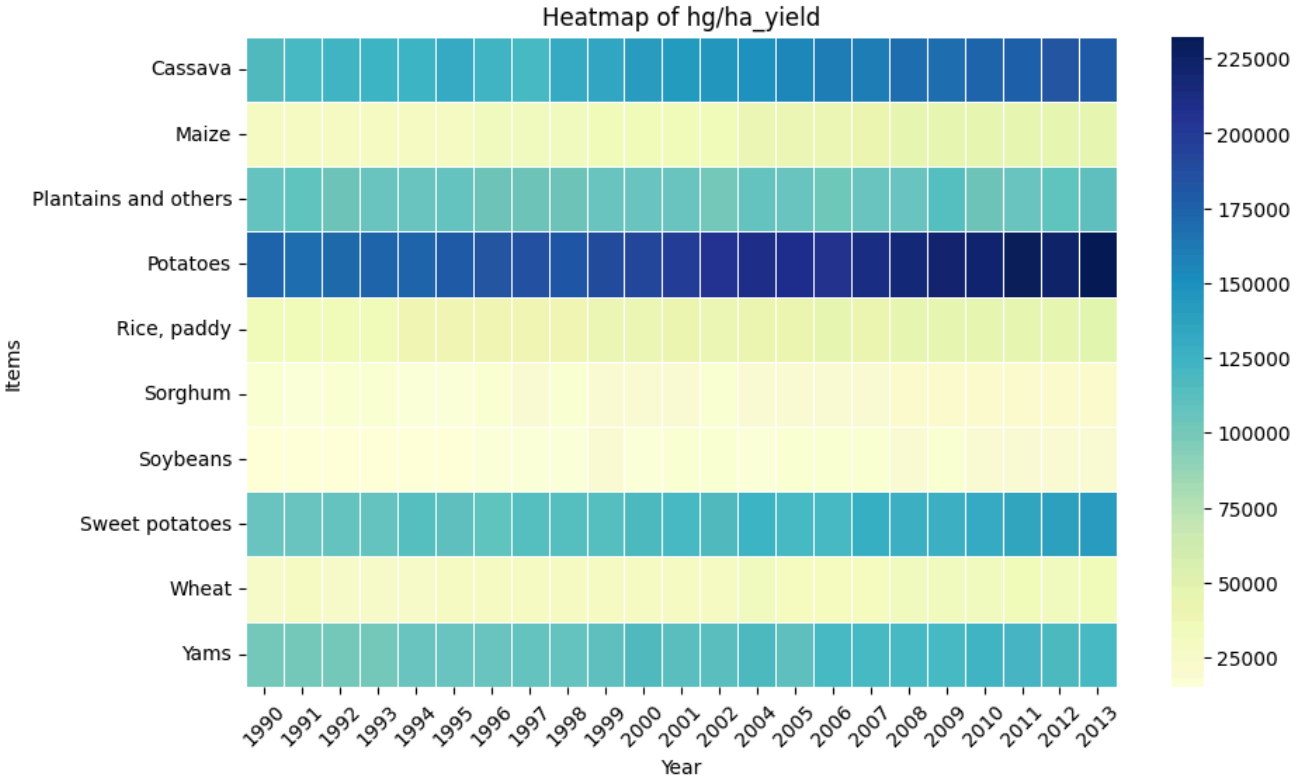

Fig. 5.16: Heatmap of Average Yield

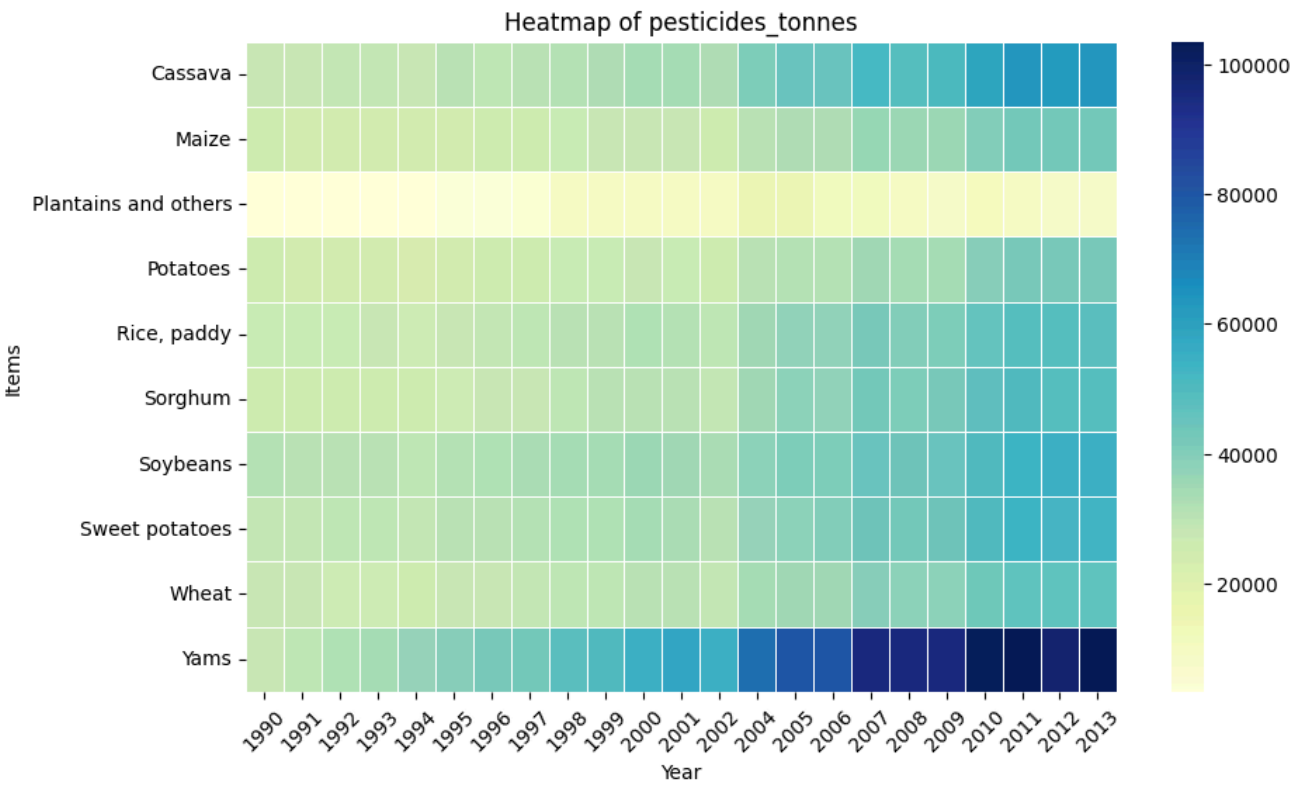

Fig. 5.17: Heatmap of Average Yield

Chapter 5

Standards Adopted

5.1 Design Standards

I. - To ensure data accuracy and quality, we adhered to the established guidelines for data pre-processing, which include handling missing data, identifying outliers, and normalizing data.
II. -For crop yield prediction, a new area of data science that is currently being applied in agriculture, we used Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, K-nearest neighbors (KNN) Regression, Random Forest Regression, Support Vector Regression, and Gradient Boosting Regression.

5.2 Coding Standards

I. To make our code easier to comprehend and maintain, we added comments and documentation
II. To guarantee readability and uniformity of the code, we adhered to the PEP 8 style guide for Python code.
III. To increase maintainability and reusability, we divided the code into many functions and classes using a modular code design.

5.3 Testing Standards

We followed standard protocol and separated our data into training and testing sets to estimate our model's performance.

The main objective of our project was to maintain stringent standards in order to create a robust, stable, and scalable crop production forecasting system.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

Machine learning algorithms for crop prediction provide a comforting tool for farmers seeking to increase agricultural production. By examining elements like historical data and weather patterns, these models may estimate acceptable crops for a certain geographical area and predict crop yields.This information enables farmers to make more educated decisions regarding planting and resource allocation, eventually enhancing efficiency and profitability. Weather sensors can help improve machine learning models for crops.

6.2 Future Scope

While this study provides valuable perspicuity into crop yield prediction using machine learning models, there are numerous avenues for future research and improvement. Firstly, investigating more developed feature engineering methods and integrating domain-specific knowledge could further promote algorithm implementation. Moreover, uncovering ensemble methods and hybrid models may lead to better predictive precision and robustness. This not only improves agricultural output projections, but it also reforms farming by placing data in farmers' hands. Prediction models based on machine learning have enormous potential to improve agriculture by encouraging data-driven decision-making, ultimately helping to ensure food supply in the future.

References

[1] Van Klompenburg, T., Kassahun, A., & Çatal, Ç. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709.

[2] Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021, September 7). *An interaction regression model for crop yield prediction*. Scientific Reports.

[3] Anderson, W. A., Dippel, A. B., Maiden, M. M., Waters, C. M., & Hammond, M. C. (2020). Chemiluminescent sensors for quantitation of the bacterial second messenger cyclic di-GMP. In *Methods in enzymology on CD-ROM/Methods in enzymology* (pp. 83–104).

[4] Otchere, D. A., Ganat, T., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science & Engineering*, *208*, 109244.

[5] Ridge Regression as a Technique for Analyzing Models with Multicollinearity on JSTOR. (n.d.). *www.jstor.org*.

[6] Regression shrinkage and selection via the lasso on JSTOR. (n.d.). *www.jstor.org*.

[7] Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbor (KNN) algorithm and its different variants for disease prediction. *Scientific Reports (Nature Publishing Group)*, *12*(1).

[8] Anwla, P. K. (2021, December 1). *Decision Tree Algorithm overview explained*. TowardsMachineLearning.

[9] Nailman, A. (2023, February 12). *Exploring Machine Learning Models: Classification for data analysis*. Machine Learning Models.

[10] Evgeniou, T., & Pontil, M. (2001). Support Vector Machines: Theory and applications. In *Lecture notes in computer science* (pp. 249–257).

[11]Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768.

[12]Filho, D. B. F., Júnior, J. a. S., & Da Rocha, E. C. (2011). What is R2 all about? *Leviathan (São Paulo. Online)*, *3*, 60.

[13]Hodson, T., Over, T. M., & Foks, S. S. (2021). Mean squared error, deconstructed. *Journal of Advances in Modeling Earth Systems*, *13*(12).

[14]Hodson, T. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487.

[15]Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679.

[16] Juwono, F. H., Wong, W. K., Verma, S., Shekhawat, N., Lease, B. A., & Apriono, C. (2023, December 1). Machine learning for weed–plant discrimination in agriculture 5.0: An in-depth review. Artificial Intelligence in Agriculture.

[17] Crop Yield Prediction using Machine Learning Algorithm. (2021, May 6). IEEE Conference Publication | IEEE Xplore.

[18] Iniyan, S., Varma, V. A., & Naidu, C. T. (2023, January 1). Crop yield prediction using machine learning techniques. Advances in Engineering Software (1992).

Crop Yield Prediction

SUDESHNA RATH
2128101

Abstract:

To boost agricultural productivity and sustainability, crop output prediction accuracy must be improved. Often, the complexity of factors affecting crop growth is too great for traditional methods to fully comprehend. Machine learning (ML) algorithms employ historical data on weather, and past yields to increase prediction accuracy.

Individual contribution and findings:

She was in charge of implementing data preprocessing and conducting exploratory analysis.
To learn more about the dataset, she began with exploratory data analysis. She used correlational analysis (heatmap) and relational techniques including scatter plot, counter plot, and histogram to achieve it.
Throughout the entire process, she worked closely with the other team members to ensure that the regression results fit the overall project goals and objectives. She provided her opinions and recommendations based on data preprocessing and investigation methodologies.

Individual contribution to project report preparation:

She contributed to the following portions :
1. Introduction
2. Problem Statement / Requirement Specification
3. Implementation- Methodology
4. Exploratory Data Analysis Visualisation

Individual contribution to project presentation and demonstration:

She prepared the presentation slides that included the exploratory data analysis that included scatter plots, counter plots which helped us to establish the relation between the item count and the area they are produced. She also helped design the slide that involves the heat map through which we establish the correlation between the items and other attributes, per year.

Full Signature of Supervisor: Full signature of the student:

……. ……..

CROP YIELD PREDICTION

LAVANYA UPADHYAY
2128077

Abstract:

Crop yield prediction accuracy needs to be increased in order to increase productivity and sustainability in the agricultural sector. Often, conventional methods are unable to adequately capture the complexity of factors affecting crop development. To improve prediction accuracy, machine learning (ML) algorithms use historical data on prior yields, soil conditions, and weather. This helps utilize new age technology to overcome the challenges involved in the agricultural sector specially for agrarian economies like India.

Individual contribution and findings:

She was responsible for implementing the different regression models for predicting the crop yield based on several features. She calculated the different evaluation metrics used for assessing the performance of the different models of regression. After calculation of evaluation metrics , a comparison was made between the evaluation metric values of different models .
Scatter plot was used to determine the actual vs predicted yield values and residual plot was studied to identify the best fit model for the dataset .
After analyzing , the best fit model with least error and high correlation between actual and predicted yield value was determined, that is , Random Forest Regressor Model which was used to calculate the prediction in the prediction system.

Individual contribution to project report preparation:

She contributed for the following portions :

1. Basic Concepts / Literature Review
2. Testing/Verification Plan
3. Result Analysis

Individual contribution for project presentation and demonstration:

She prepared the “Evaluation Metrics ” and “Result Analysis” part of the presentation. This involves implementing the different regression models and analyzing their evaluation metrics . After comparing evaluation (performance) metrics of the models the best fit model is selected and used for prediction of crop yield for sample input.
Overall, she played an important role in determining the best fit model for crop yield prediction while closely working with her team members to obtain the desired results. Her input enabled the group to come up with significant outcomes.

Full Signature of Supervisor: Full signature of the student:

……. ……..

CROP YIELD PREDICTION

KALYANBRATA GIRI
2128075

Abstract:

This study aimed to evaluate the performance of diverse machine learning models using the FAO Kaggle-based crop yield dataset, which encompasses yields from the world's top ten consumed crops. Various regression models, including Linear, Gradient Boost, Ridge, Lasso, Decision Tree, Random Forest, SVC K Neighbours Regression, were employed for yield prediction. Rigorous testing and training were conducted to ensure model robustness, with evaluation based on metrics such as RMSE, MAE, R2 Score, among others. Interestingly, the Random Forest regressor emerged as the least effective in predicting crop yield within this dataset. SVC exhibited the highest RMSE (9976.13) with an error of 93125.

Individual contribution and findings:

He was responsible for researching and providing the necessary information regarding the regression models. He also provided relevant data sets that helped to implement the various regression models.

He started by going through the different research papers provided and found all the required information that is required for the project. He also helped in cleaning the data by removing any unnecessary or missing values and implemented techniques to detect outliers.

After the data was processed, he helped in carrying out exploratory data analysis and helped determine which characteristics were most important to obtain the correct prediction.

Individual contribution to project report preparation:

He contributed for the following portions :

1. Introduction
2. Problem Statement/ Research Specification.
3. Standards Adopted
4. Methodology- to clean, transform and modify data sets
5. Conclusion

Individual contribution for project presentation and demonstration:

He was responsible for crafting the "Basic Concepts" segment of the presentation. This involved defining each type of regression model to be employed in the project and employing a suitable regression algorithm on the prepared data for regression analysis. The process included selecting the optimal regression model and evaluating its efficacy.

Full Signature of Supervisor: Full signature of the student:

……. ……..

CROP YIELD PREDICTION

PRATYUSH KUMAR PRASAD
2128035

Abstract:

This study aimed to evaluate the performance of diverse machine learning models using the FAO Kaggle-based crop yield dataset, which encompasses yields from the world's top ten consumed crops. Various regression models, including Linear, Gradient Boost, Ridge, Lasso, Decision Tree, Random Forest, SVC K Neighbours Regression, were employed for yield prediction. Rigorous testing and training were conducted to ensure model robustness, with evaluation based on metrics such as RMSE, MAE, R2 Score, among others. Interestingly, the Random Forest regressor emerged as the least effective in predicting crop yield within this dataset. SVC exhibited the highest RMSE (9976.13) with an error of 93125.

Individual contribution and findings:

He researched potential topics, explored regression models (linear regression, random forests, etc.). He began by reading through the several study papers offered and finding all of the essential material for the report and compiled relevant resources with proper citations. He documented these models and assisted with data cleaning to ensure high-quality data for analysis.
He collaborated carefully with his peers to ensure that everything functioned perfectly.

He contributed for the following portions :
Implementation :

1. Introduction
2. Problem Statement/ Research Specification.
3. Standards Adopted
4. Methodology- clean
5. Conclusion
6. References

Individual contribution for project presentation and demonstration:

He prepared the "Regression models" section of the presentation. This involves performing a regression analysis on the produced data using an appropriate regression algorithm. This includes determining the optimal regression algorithm and evaluating the performance of the algorithm.Overall, he played an important role in ensuring that the data was ready for regression analysis and provided the team with high quality and accurate data for analysis. His contribution helped the team to find significant results..

Full Signature of Supervisor: Full signature of the student:

……. ……..

PLAGIARISM REPORT

CROP YIELD PREDICTION

ORIGINALITY REPORT

14%	11%	5%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	3%
2	fastercapital.com Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	"Sustainable Development through Machine Learning, AI and IoT", Springer Science and Business Media LLC, 2023 Publication	1%
5	Submitted to Banaras Hindu University Student Paper	1%
6	Submitted to Liverpool John Moores University Student Paper	<1%
7	www.iieta.org Internet Source	<1%
8	www.journaltocs.ac.uk Internet Source	<1%

9	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
10	Submitted to Liverpool Hope Student Paper	<1 %
11	ebin.pub Internet Source	<1 %
12	Timothy Pede, Giorgos Mountrakis, Stephen B. Shaw. "Improving corn yield prediction across the US Corn Belt by replacing air temperature with daily MODIS land surface temperature", Agricultural and Forest Meteorology, 2019 Publication	<1 %
13	Submitted to University of Exeter Student Paper	<1 %
14	todayfounder.com Internet Source	<1 %
15	Submitted to Army Institute of Technology Student Paper	<1 %
16	research.library.mun.ca Internet Source	<1 %
17	eurchembull.com Internet Source	<1 %
18	repositori.uji.es Internet Source	

		<1 %
19	Submitted to Manchester Metropolitan University Student Paper	<1 %
20	www.db-thueringen.de Internet Source	<1 %
21	www.faq.de Internet Source	<1 %
22	"Industrial Engineering and Industrial Management", Springer Science and Business Media LLC, 2024 Publication	<1 %
23	Submitted to University of Northampton Student Paper	<1 %
24	Submitted to Staffordshire University Student Paper	<1 %
25	Submitted to University of Greenwich Student Paper	<1 %
26	Submitted to Wageningen University Student Paper	<1 %
27	jneuroengrehab.biomedcentral.com Internet Source	<1 %
28	www.neuroquantology.com Internet Source	<1 %

29	www.reddit.com Internet Source	<1 %
30	ijsrcseit.com Internet Source	<1 %
31	jbc.bj.uj.edu.pl Internet Source	<1 %
32	pdfs.semanticscholar.org Internet Source	<1 %
33	www.ir.juit.ac.in:8080 Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		