A logo for college computing

Description automatically generated

**Crop Yield Prediction**

| *Swan Saung Lwin* |  |
| --- | --- |
| *2025282* |  |
| *Machine Learning* |  |
| *CA 1* |  |
| *Sunday, 18th April 2025 23:59* |  |
| *Sunday, 18th April 2025* |  |



Contents

[Introduction 1](#_heading=h.cersum6jhy0z)

[1. Business Understanding 1](#_heading=h.kupod9gvud48)

[1.1. Problem Statement 1](#_heading=h.psi2dckxclou)

[1.2. Objectives 1](#_heading=h.wldec8ibzjfd)

[2. Data Understanding 2](#_heading=h.bz4ldudkli0h)

[2.1. About Datasets 2](#_heading=h.k6nt37bo9asl)

[2.1.1. Yield 2](#_heading=h.1sekf3z1syrn)

[2.1.2. Pesticides 2](#_heading=h.vznny0hlwgk0)

[2.1.3. Precipitation 3](#_heading=h.ip7r4up74etc)

[2.1.4. Temperature 3](#_heading=h.rth9tshkl37s)

[2.2. Merging Datasets 3](#_heading=h.txjwi4idmog4)

[2.3. Exploratory Data Analysis 3](#_heading=h.g111wdhj4odi)

[2.3.1. Data Cleaning 3](#_heading=h.mybrfwrgklzd)

[2.3.2. Top 10 Producing Crops in ASEAN 4](#_heading=h.d4c30vsigby3)

[2.3.3 Average Yield of Rice by Country 5](#_heading=h.q3nraew7hnle)

[2.3.4. Correlation of Factors 5](#_heading=h.slna1gyzserr)

[2.3.5. Distribution of values 7](#_heading=h.n2s3fonb2b6m)

[3. Data Preparation 8](#_heading=h.cxop9cih540k)

[3.1. Category Encoding 8](#_heading=h.ud97es317tql)

[3.2. Feature Scaling 9](#_heading=h.atw90gh4mp8a)

[4. Modeling 9](#_heading=h.6609l845gxfr)

[4.1. Cross validation on 3 splits 10](#_heading=h.hsvfjag9y95j)

[4.2. Testing on 3 splits 10](#_heading=h.ku5quqluwjv7)

[4.3. Hyperparameter Tuning 11](#_heading=h.n3ohswny59ms)

[4.4. Training models 12](#_heading=h.t4ki28jfpko8)

[5. Evaluation 12](#_heading=h.xa3gw8clg8cy)

[Conclusion 14](#_heading=h.2rbhbwxj0vyu)

[References 15](#_heading=h.5akkuwhp012l)

# Introduction

Agriculture plays the most important role in the world's economy and world’s food supply. Almost 45% of the population in the world lives in households where agricultural activities represent the main occupation of the head and a large share of this agriculture. (Masters et al., 2013). Therefore crop yield prediction and agricultural planning becomes more important for economic sustainability of farmers, securing food supplies, and also GDP growth for agriculture based countries. Traditional forecasting methods often struggle to handle the complexity of environmental and agronomic factors. Therefore, this project proposes a supervised machine learning approach to predict crop yield based on weather conditions, usage of pesticides and historical yield data for ASEAN countries which are similar in geographical traits, and have a considerable GDP in agriculture.

# 1. Business Understanding

## 1.1. Problem Statement

Agricultural productivity is challenged by unpredictable factors such as climate change and pest infestations. Also the methods to deal with those situations vary in different areas of the world. As global demand for food rises, farmers and agricultural businesses must enhance productivity while managing environmental and economic risks.

The use of a wide range of chemicals to destroy pests and weeds is an important aspect of agricultural practice in both developed and developing countries. Undoubtedly, this has increased crop yield and reduced post-harvest losses. However, the expanded use of such pesticides expectedly results in residues in foods. (al-Saleh, 1994). At the same time, weather variabilities such as changes in rainfall, and temperature can significantly affect crop performance and make crop yield harder to predict.

These uncertainties hinder effective planning and risk management, limiting the stakeholders’ ability to allocate resources, anticipate market trends and predicting yield. With a data-driven approach to find patterns within those uncertain variables, the agricultural businesses could be growing at incredible pace, helping the lives of everyone involved in the supply chain.

## 1.2. Objectives

The key objectives of this project are defined as follow

* develop a predictive model using different machine learning techniques to estimate crop yield based on weather conditions, pesticide usage, and historical yield data.
* identify the most influential features affecting crop yield to support better decision making.
* compare multiple machine learning algorithms to determine the most accurate and robust model for yield prediction.

# 2. Data Understanding

## 2.1. About Datasets

To predict crop yield based on weather conditions, pesticides and history of crop yield, I've collected datasets containing all the necessary information. The datasets contain information about pesticide usage in countries by year, average precipitation, average temperature, and historical yield data and agricultural land mass. Pesticides, and Yield data are collected from Food and Agriculture Organization of the United Nations (2025), Precipitation data is collected from Our World in Data (2025), and Temperature data is collected from Palinatx (2024) at Kaggle. I've referenced the datasets from Patel R. (2022) at Kaggle and collected the datasets from more updated sources.

As mentioned above, there are 4 datasets used in this project for predicting crop yield. They are

* Yield
* Pesticides
* Precipitation
* Temperature

I have listed the description of each dataset from the source website as below.

### 2.1.1. Yield

This dataset contains the historical yield data from 2000 to 2022, grouped by area and the crops they grew. The columns included in this dataset described as below

* Domain Code (Encoded number for Domain)
* Domain (The domain for which the data represents)
* Area Code (Encoded number for Area)
* Area (Country name)
* Element Code (Encoded number for the metric)
* Element (The measurement metric of the data)
* Item Code (Encoded number for each crops)
* Item (Crop names)
* Year Code (The year the data is from)
* Year (The year the data is from)
* Unit (Measurement unit of yield)
* Value (Yield value)

### 2.1.2. Pesticides

This dataset describes the amount of pesticides each area used by year. The columns included in this dataset described as below

* Domain Code (Encoded number for Domain)
* Domain (The domain for which the data represents)
* Area Code (Encoded number for Area)
* Area (Country name)
* Element Code (Encoded number for the metric)
* Element (The measurement metric of the data)
* Item Code (Encoded number for item used)
* Item (Item used)
* Year Code (The year the data is from)
* Year (The year the data is from)
* Unit (Measurement unit of pesticides)
* Value (Pesticides value)

### 2.1.3. Precipitation

This dataset contains information about average precipitation in each country by year. The columns included in this dataset described as below

* Entity (Country name)
* Code (Country Code)
* Year (The year the data is from)
* Annual precipitation (Average precipitation per year)

### 2.1.4. Temperature

This dataset showcases average temperature in each country by year. The columns included in this dataset described as below

* Country (Country name)
* Year (The year the data is from)
* Annual Mean (Average temperature per year)
* 5-yr smooth (5 year smooth temperature)
* Code (Country Code)

## 2.2. Merging Datasets

I have merged 4 datasets on Country and Year. After merging, Year is leveraged to only include results from 2000 to 2022. Countries that are not present in any of the dataset are removed and ASEAN countries with different names are renamed based on official naming from the UN. After merging, the dataset is left with 137,674 rows and 9 columns.

## 2.3. Exploratory Data Analysis

### 2.3.1. Data Cleaning

Dataset is filtered to include only the ASEAN countries which are similar in geographical traits, weather and the crops they grow. At the current stage, the dataset looks like Figure 2.1.

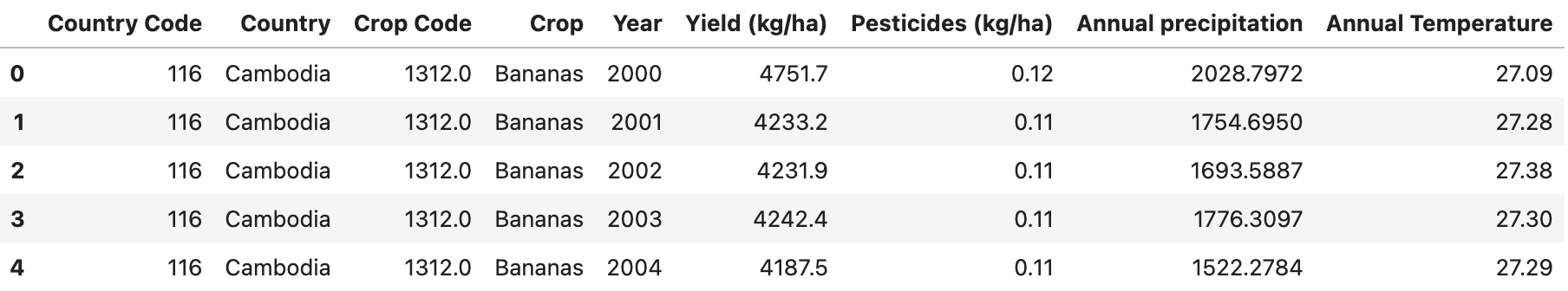


Figure 2.1. ASEAN yield dataset - First 5 rows

Country Code and Crop Code are dropped as they will not be usable during both analysis and training stages. Now there are 8,204 rows and 7 columns in the dataset. There are neither null values nor duplicates in the dataset as they are already cleaned during the merging process.

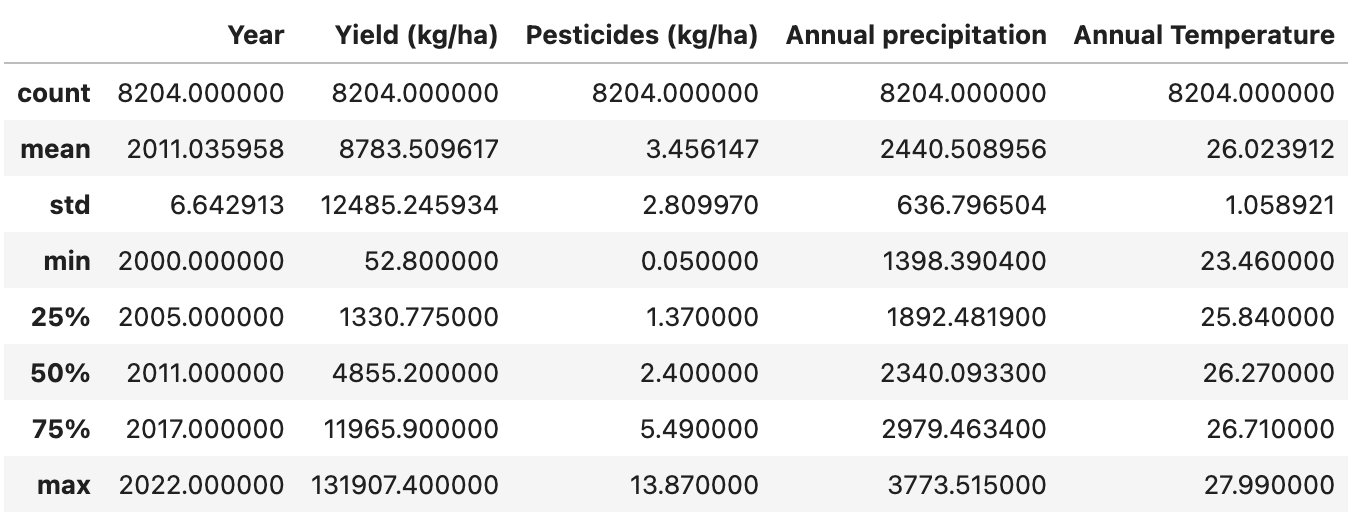
When I checked the data description shown in Figure 2.2, I noticed that standard deviation for Yield is high. It can be due to the fact that production numbers for each crop are different. 

Figure 2.2. ASEAN yield dataset - Data Description

### 2.3.2. Top 10 Producing Crops in ASEAN

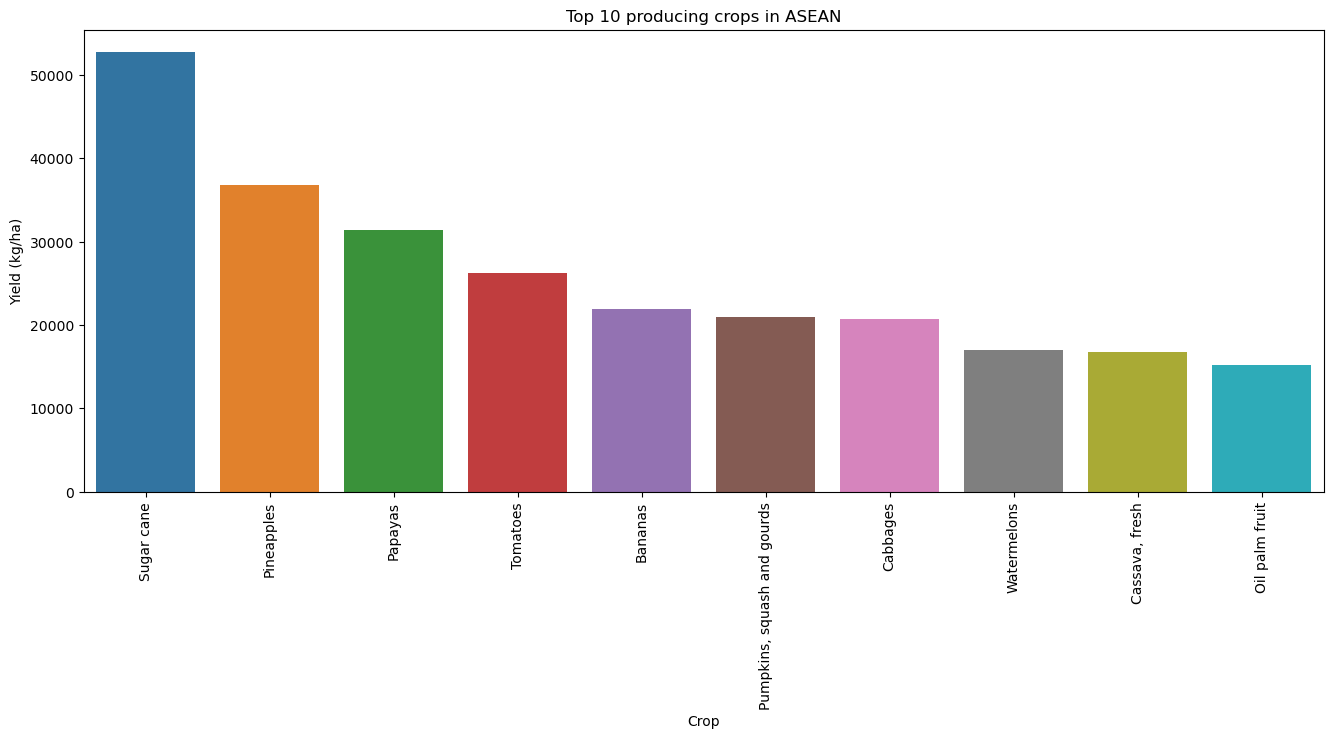
There are 99 unique crops produced in ASEAN so I try to see which are the top producing crops in ASEAN.

Figure 2.3. Top 10 producing crops in ASEAN

Sugar cane is significantly leading the yield in ASEAN, followed by Pineapples and Papayas. This also has to do with the measurement as the unit is on kg/ha which favors heavier crops to be on top.

### 2.3.3 Average Yield of Rice by Country

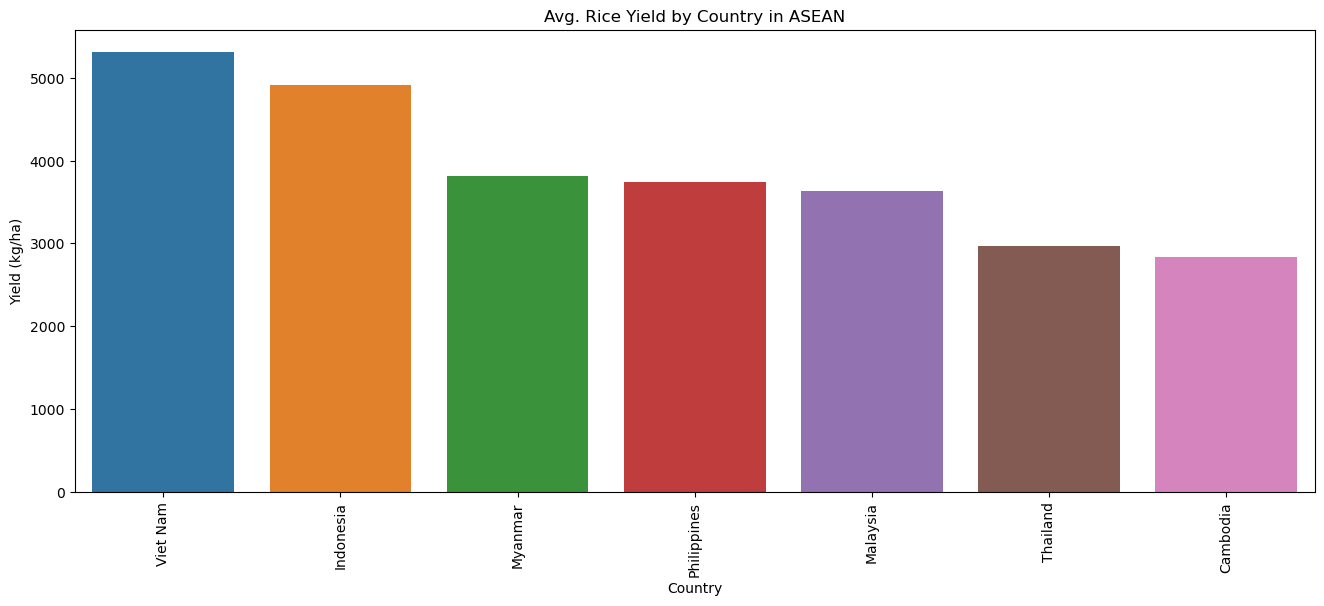
In Southeast Asia, rice is regarded as the main source of food. Therefore, I want to know which countries in ASEAN are leading in average yield for rice.

Figure 2.4. Average rice yield by country in ASEAN

Vietnam, and Indonesia have the best average yield in ASEAN. Myanmar, Philippines, and Malaysia have almost the same with Thailand and Cambodia having the least average yield per ha.

### 2.3.4. Correlation of Factors

I wanted to know how environmental factors and use of pesticides affect yield. Also the relation between the amount of pesticides used and the environmental factors.

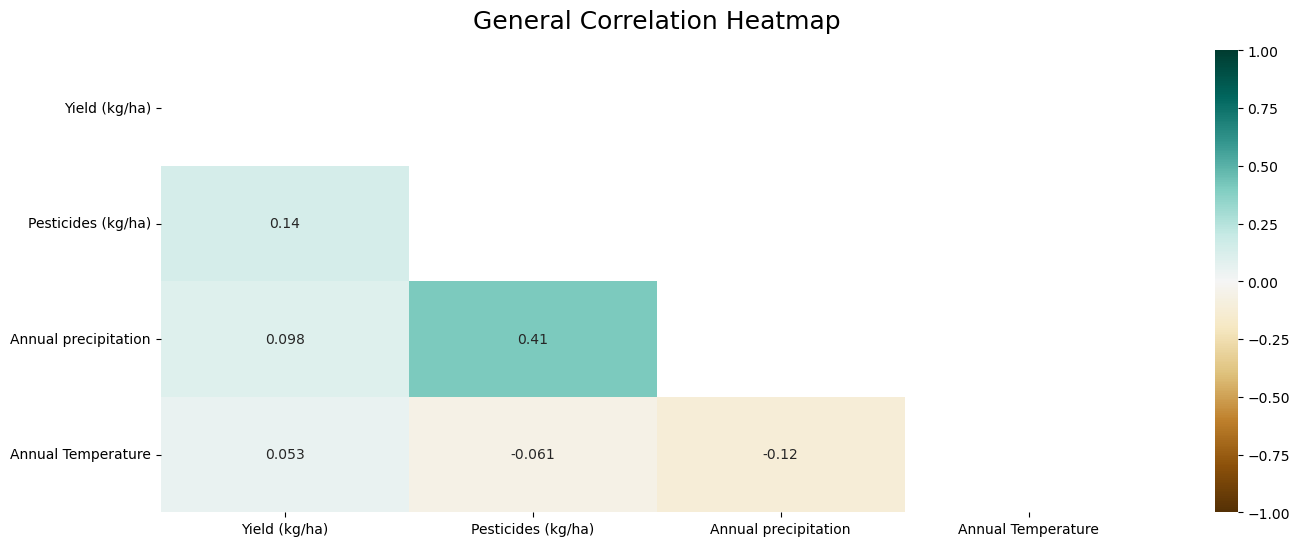


Figure 2.5. Crops general correlation heatmap

Figure 2.4. shows that the amount of pesticides used on crops depends moderately on the precipitation but only a negatively small amount on temperature.

These correlation of factors on yield could be different for different types of crops. According to an article by Ijaz S. (2023), sugarcane, maize, rice, and wheat are the most consumed crops in the world. Therefore, I want to see the correlation of factors for these crops. After plotting the correlations, varying influences of factors on yield can be seen as follows in Figure 2.5, 2.6, 2.7 and 2.8.

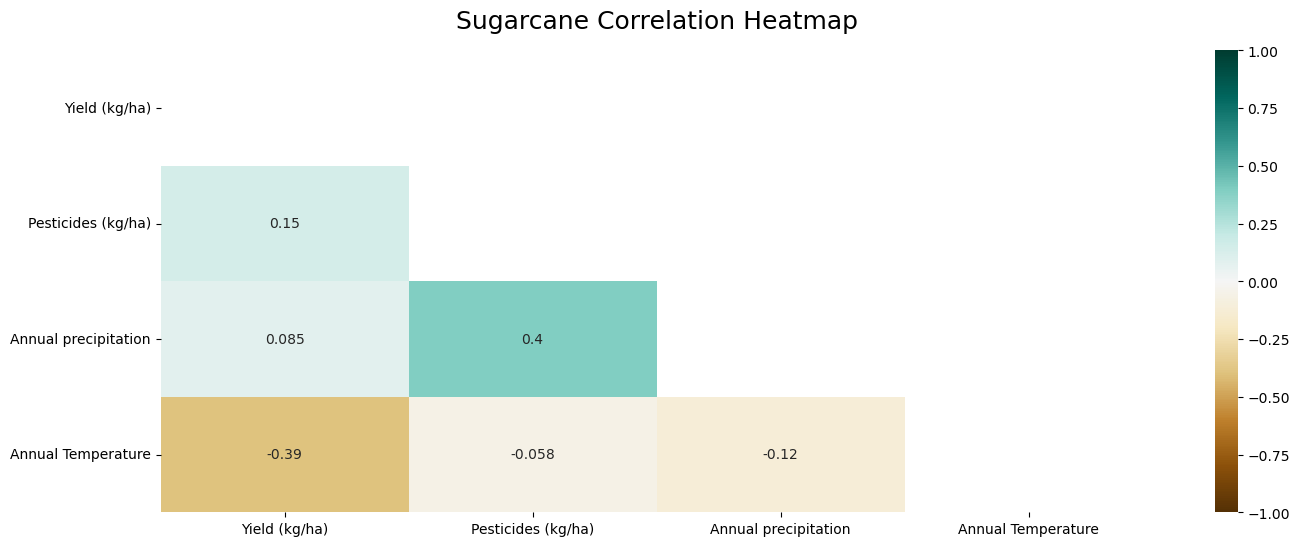


Figure 2.6. Sugarcane correlation heatmap

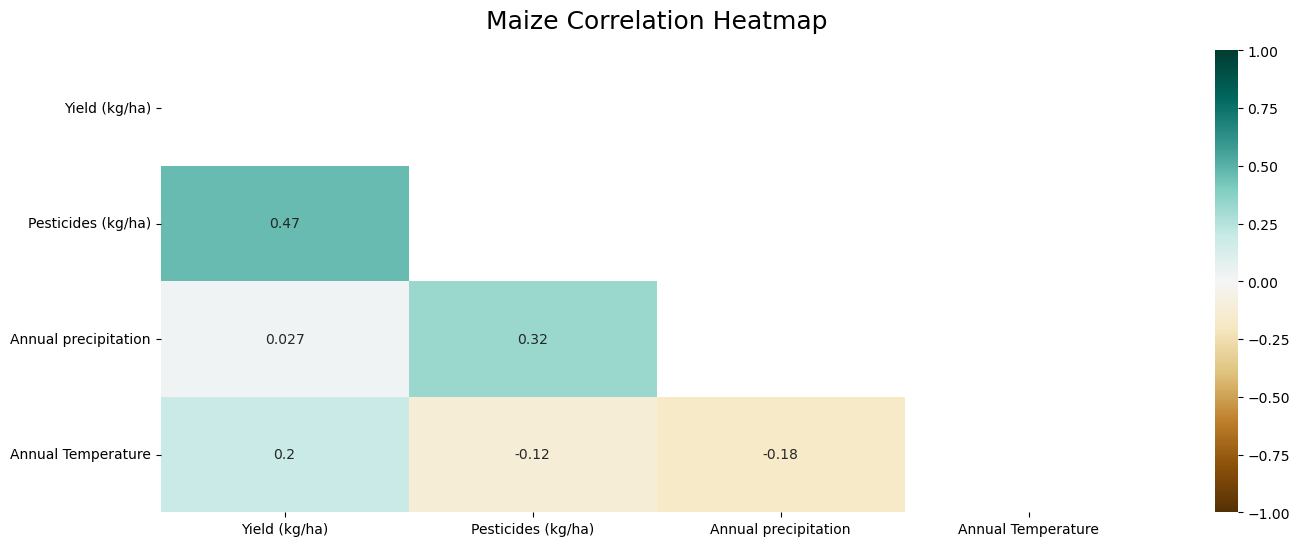


Figure 2.7. Maize correlation heatmap

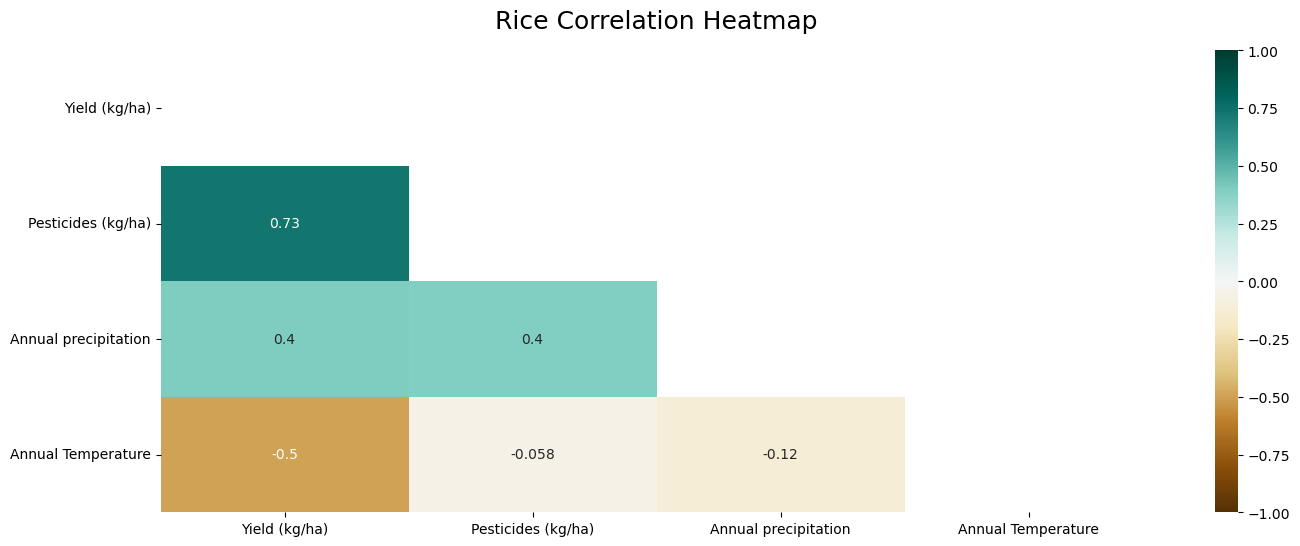


Figure 2.8. Rice correlation heatmap

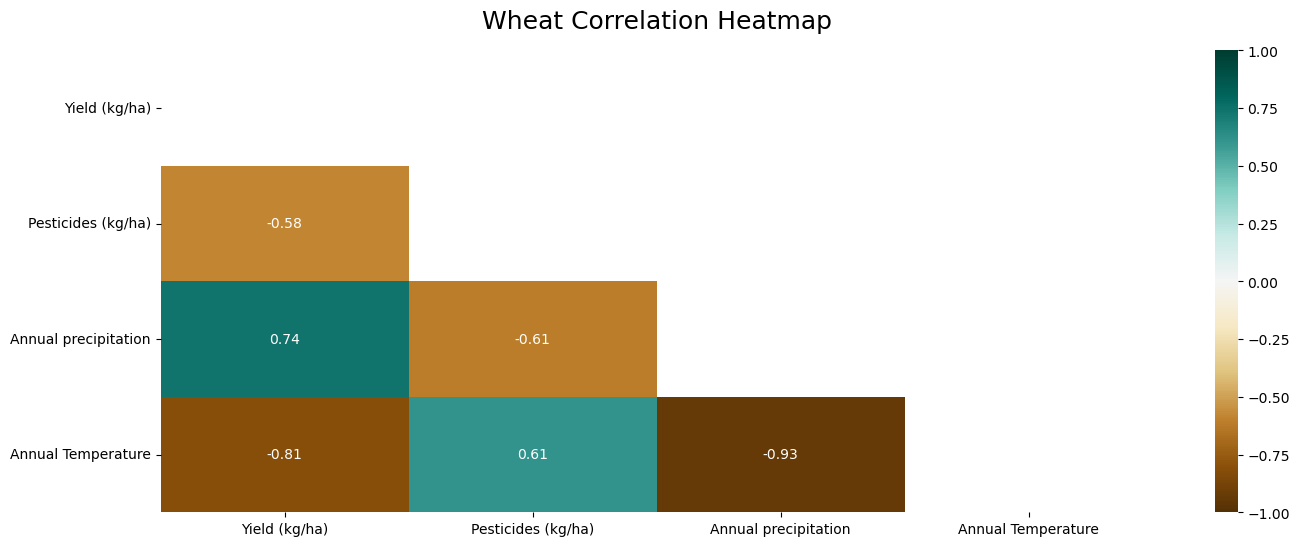


Figure 2.9. Wheat correlation heatmap

### 2.3.5. Distribution of values

After plotting the feature columns, I have noticed that Temperature is having a high number of outliers as shown in Figure 2.10. To solve the issue, I will be implementing feature scaling on the dataset.

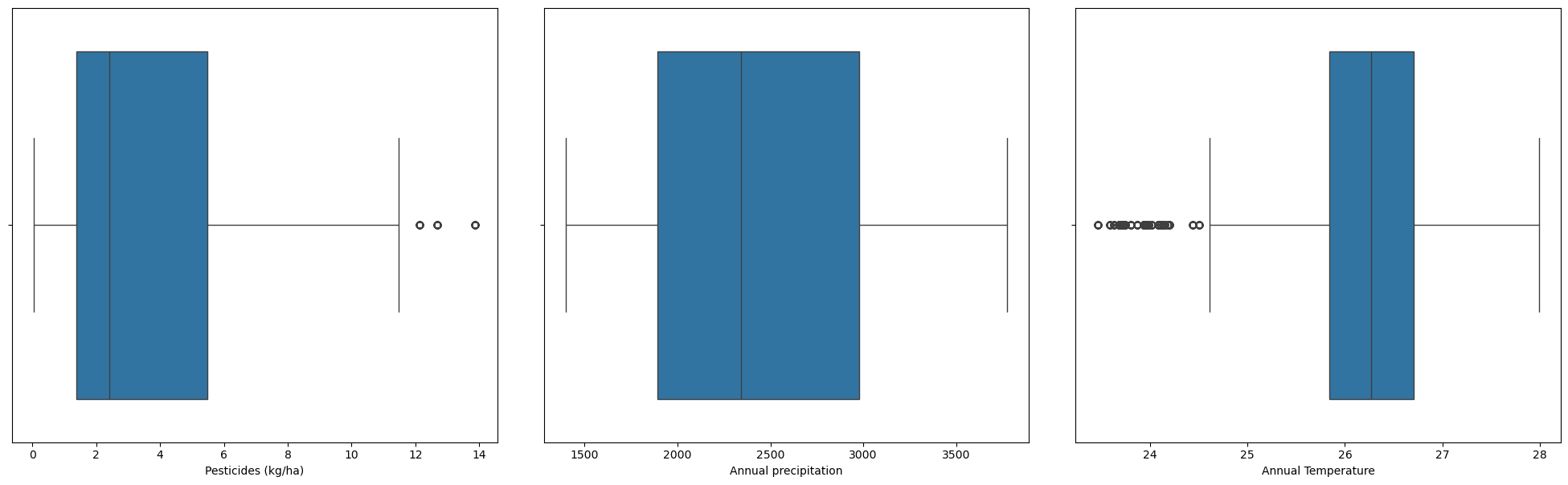


Figure 2.10. Boxplots showing distribution of feature columns

Both pesticides and precipitation have uniform distributions. Distribution for temperature kind of resembles normal distribution with some number of outliers on the left as shown in Figure 2.11.

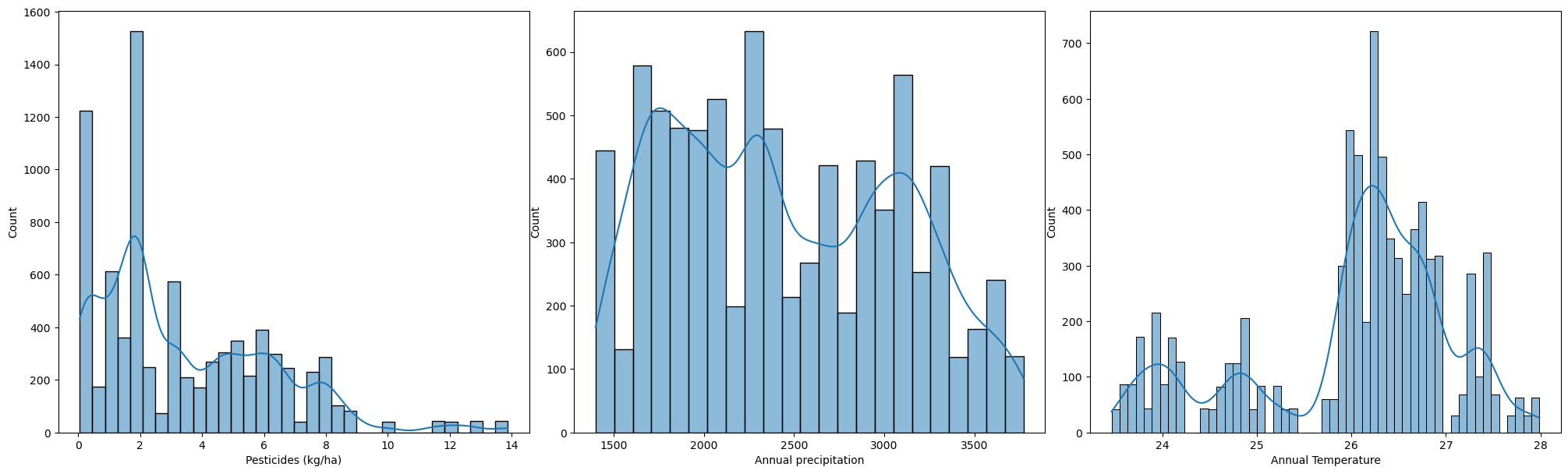


Figure 2.11. Histograms showing distribution of feature columns

# 3. Data Preparation

In this stage, preparation of data for modeling will be done. Year column is dropped first as it is irrelevant for training the model.

For training and testing the models, I will be initially testing with 3 splits (80-20, 90-10, 95-5). After comparing performances across 3 splits, I will choose one and proceed with it. However for preparing processing, I will be using 80-20 split and define functions to be flexible along the process with 3 splits.

## 3.1. Category Encoding

There are 2 categorical values in the dataset. To enable machine learning models to process the categorical data, the values have to be encoded. There are 7 unique values for the Country column and 99 unique values for the Crop column. I will avoid using label encoder as it will add ordinality to the variable and create bias upon training the models.

I will be applying one-hot encoding to Country values as it doesn't have many unique values. However, I cannot apply one-hot encoding to Crop as it will create 97 features and will face the curse of dimensionality because of this. Therefore, I will be applying target encoding to Crop. The main idea behind the target encoder is to encode the categories by replacing them for a measurement of the effect they might have on the target. The issue with target encoding is that information of the target variable will be feeded. To solve that, smoothing will be applied to the method.

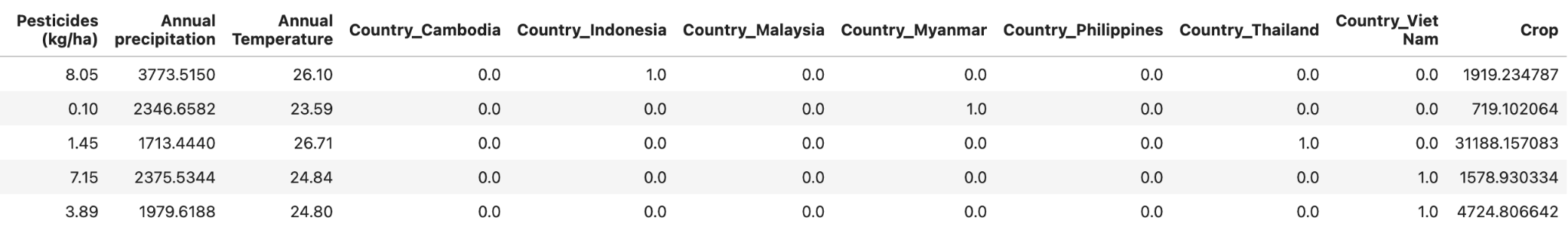


Figure 3.1. Dataset after category encoding

## 3.2. Feature Scaling

To avoid features with highly varying magnitude impacting the models' performance, feature scaling is applied using the min-max scaler. Min-max scaler transforms features into the range of 0 to 1.

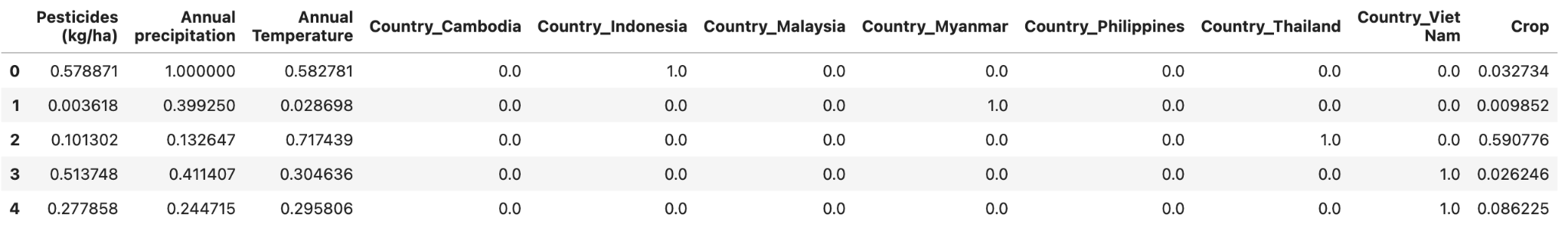


Figure 3.2. Dataset after feature scaling

# 4. Modeling

To predict crop yield, I have chosen a wide range of models such as tree-based algorithms, Random Forest, Gradient Boosting and XGBoost, Support Vector Machine, Ridge as a type of Linear Regression, and K-Nearest Neighbors. My strategy would be as follows

1. Cross-validate the above described models on 3 splits (80-20, 90-10, and 95-5).
2. Test the models on 3 splits and select the best performing models together with the best performing split.
3. Tune 3 best models using GridSearchCV on chosen split.
4. Train the models with the best parameters on the chosen split.

I will be evaluating the models based on 2 metrics, **Root Mean Square Error (RMSE)** and **R2 score**.

**Root Mean Square Error (RMSE)** shows the mean differences between actual and predicted values in the dataset in the unit of target variable.

**R-Squared Score (R2)** is a measure that provides information about the goodness fit of a model. It is the statistical measure of how well the regression line fits the model.

To evaluate the models, it is also important to understand the concepts of Generalization, Underfitting and Overfitting.

**Generalization** is the model's ability to understand and apply learned patterns to unseen data. Models with low variance also tend to underfit as they are too simple to capture complex patterns. However, low-bias models might overfit if they are too flexible. (IBM, 2024).

**Overfitting** means that there are less errors during training but when testing with unseen data, the amount of error is significantly higher. An overfit model can result in high model accuracy on training data but low accuracy on new data due to memorization instead of generalization

**Underfitting** describes the situation where performance of a model is bad and errors are significantly high across both training and testing datasets.

## 4.1. Cross validation on 3 splits

After testing the models across 3 splits, the training accuracy is collected as shown Figure 4.1.

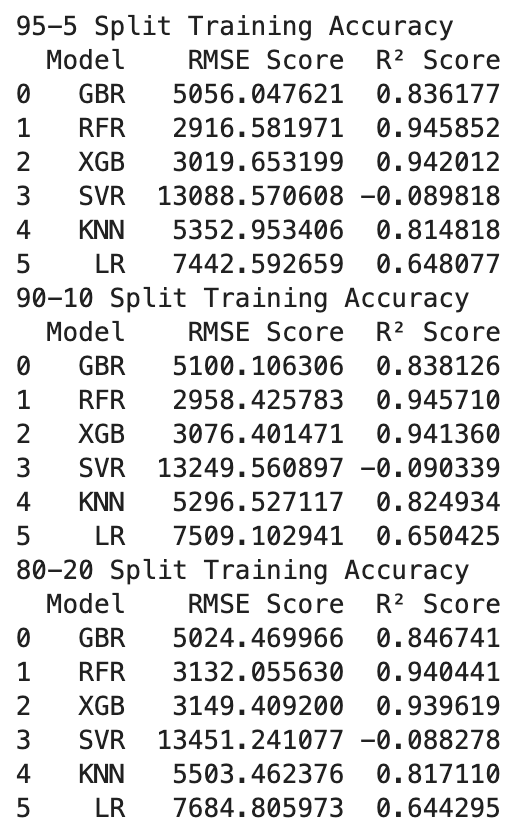


Figure 4.1. Training performance of models on 3 splits

Upon looking at the training accuracy, Random Forest and XGBoost are having the best r2 score of around 94% and GradientBoosting and KNN are having a decent accuracy of 83%. Ridge Regression doesn't perform very well with about 64% while Support Vector Machine is clearly underfitting.

All 3 splits have similar accuracies, so I will decide which split to choose on seeing the testing results.

## 4.2. Testing on 3 splits

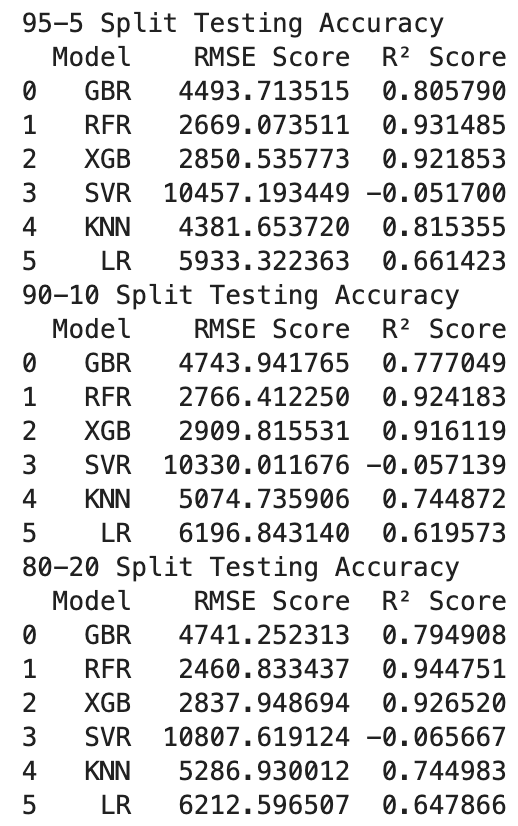


Figure 4.2. Testing performance of models on 3 splits

Random Forest and XGBoost have dropped off a little bit of r2 score but not enough to be called overfitting. However, GradientBoosting has about a 5% drop off in r2 score and can said to be a little bit overfitted.

Looking at the 3 splits, 90-10 seems to be the worst of all. To choose from 95-5 and 80-20, I will mainly look at the best 3 performing models. 95-5 has better R2 score for GradientBoosting of about 1.1% while 80-20 has better score for both Random Forest and XGBoost of around 1.3% and 0.5% edge.

Therefore, I will be choosing an 80-20 split for training and testing models. For the models, I will be choosing Gradient Boosting, Random Forest, XGBoost, and KNN as they are the best performing models.

## 4.3. Hyperparameter Tuning

When we're training machine learning models, each dataset and model needs a different set of hyperparameters, which are a kind of variable. The only way to determine these is through multiple experiments, where we pick a set of hyperparameters and run them through the model. This is called hyperparameter tuning. In essence, we're training our model sequentially with different sets of hyperparameters. Hyperparameters are external configuration variables that data scientists use to manage machine learning model training. (Amazon Web Services, 2024)

There are some popular hyperparameter tuning methods such as Grid Search, Random Search, and Bayesian Optimization. Out of all those methods, my choice is the Grid Search method.

**Grid search** is a sort of “brute force” hyperparameter tuning method. We create a grid of possible discrete hyperparameter values then fit the model with every possible combination. We record the model performance for each set then select the combination that has produced the best performance. (Navas, 2022)

I chose Grid Search for hyperparameter tuning because I want to define a grid of parameters that would make sense rather than randomly. The parameters I chose for the models are shown in Figure 4.3. and the best parameters after tuning are shown in Figure 4.4.

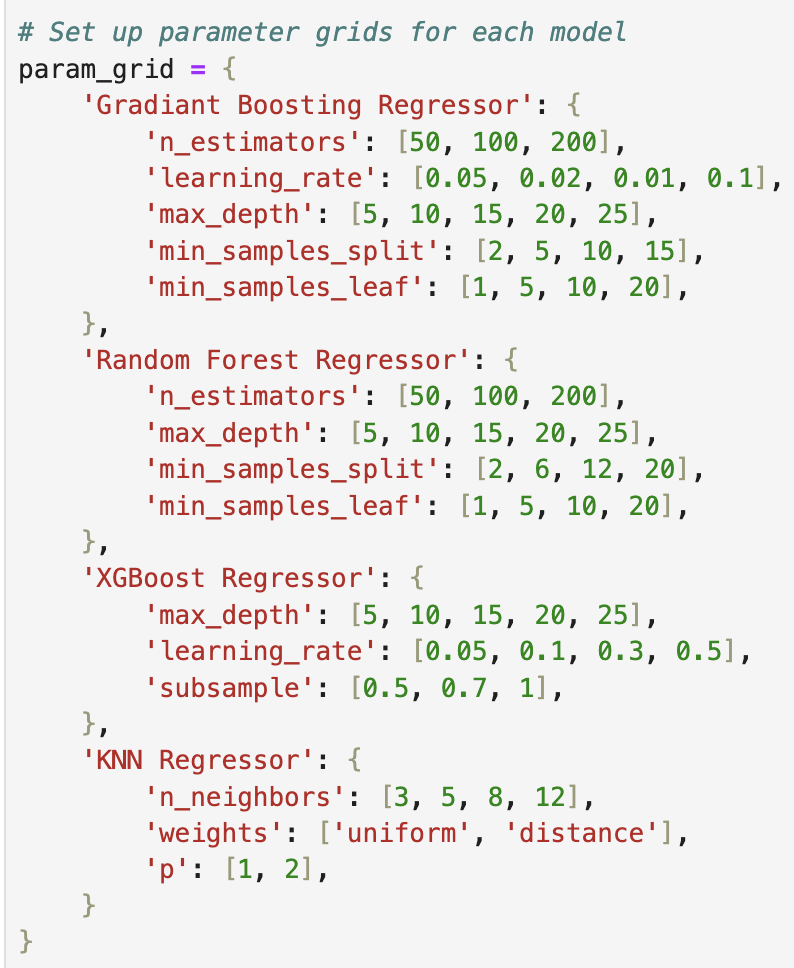


Figure 4.3. Hyperparameters for models

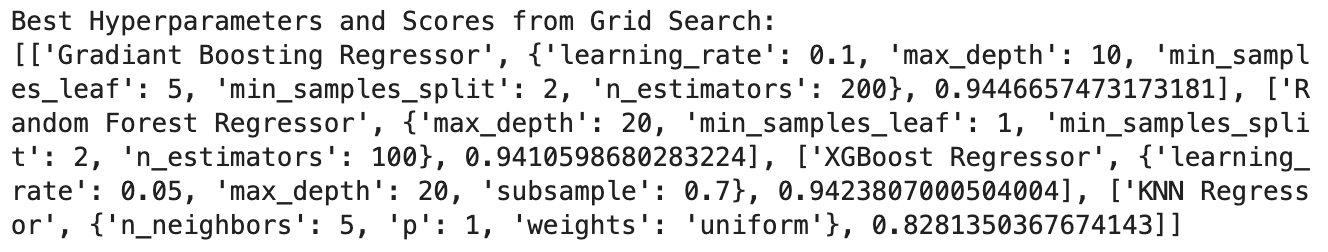


Figure 4.4. Best hyperparameters and scores from GridSearch

As the performance for KNN Regressor is significantly lower than others, it will not be included for the further steps.

## 4.4. Training models

The models are trained with the best parameters and all of them reached a 94% r2 score as shown in Figure 4.5.

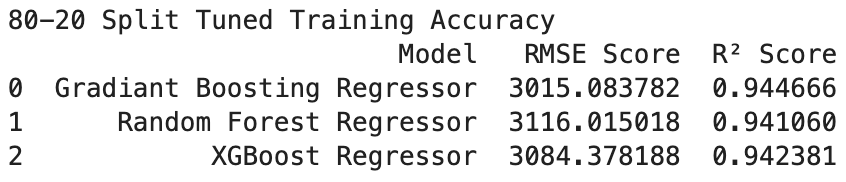


Figure 4.5. Tuned Training Performance

There's a significant increase in performance for GradientBoostingRegressor. R2 score for GradientBoostingRegressor improved by more than 9% and Root Mean Square Error decreased by about 2000. However, the performance doesn't increase much for both XGBoost and Random Forest. R2 score for both models only increased by about 0.2% and 0.06% respectively. The improvements for the models are shown in Figure 4.6.

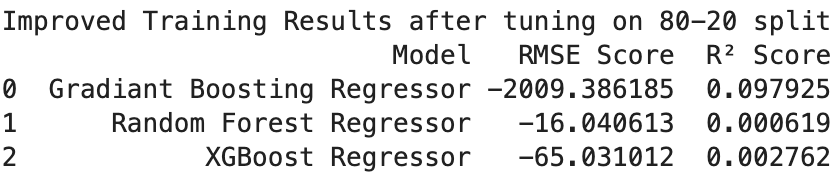


Figure 4.5. Improved Training Performance

# 5. Evaluation

All 3 models have around 94% r2 score and only varying about 0.5% as shown in Figure 5.1.

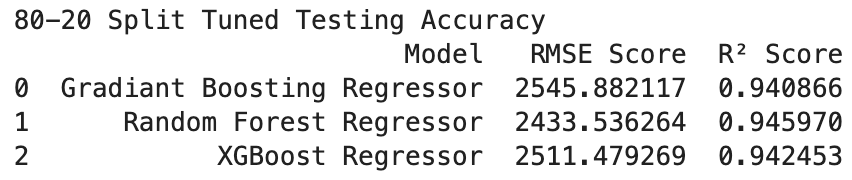


Figure 5.1. Tuned Testing Performance

R2 score of GradientBoostingRegressor increased by more than 14% and root mean squared error decreased by about 2,200. XGBoost Regressor also reduced its RMSE Score by about 320 and increased r2 score by 1.5%. However Random Forest Regressor has its r2 score decreased by about 0.2% due to the model becoming a bit less flexible. 0.002 r2 score and 50 RMSE is not a very high amount so I will recognize it as being still.

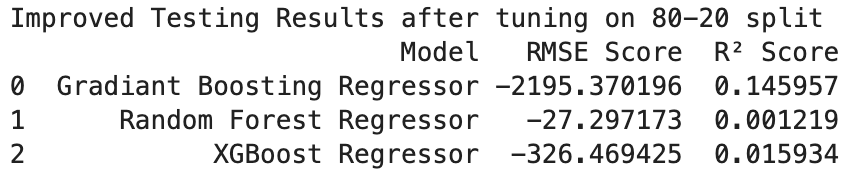


Figure 5.2. Improved Testing Performance

The generalization of the models are good with neither overfitting nor underfitting. Therefore, I will be deciding what will be the best model to use based on how the models determine the importance of each feature.

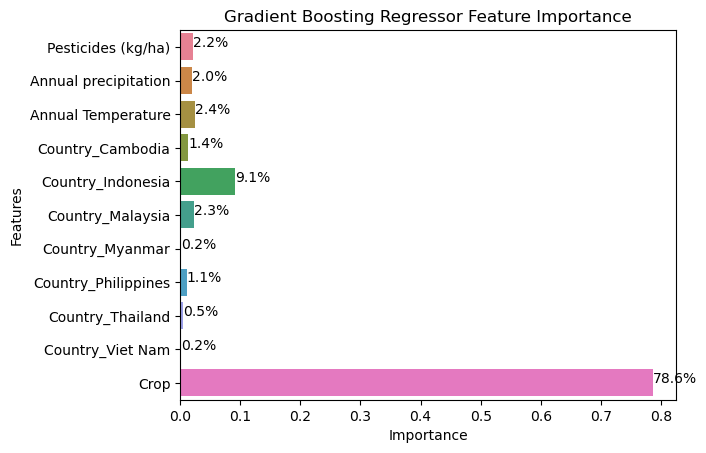


Figure 5.3. Gradient Boosting Regressor Feature Importance

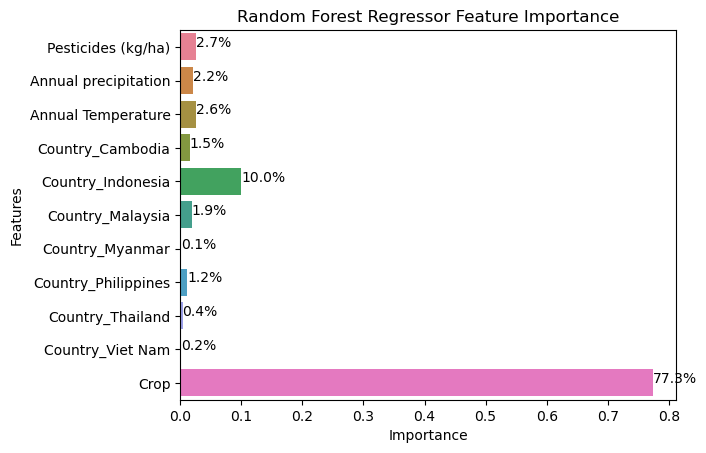


Figure 5.4. Random Forest Regressor Feature Importance

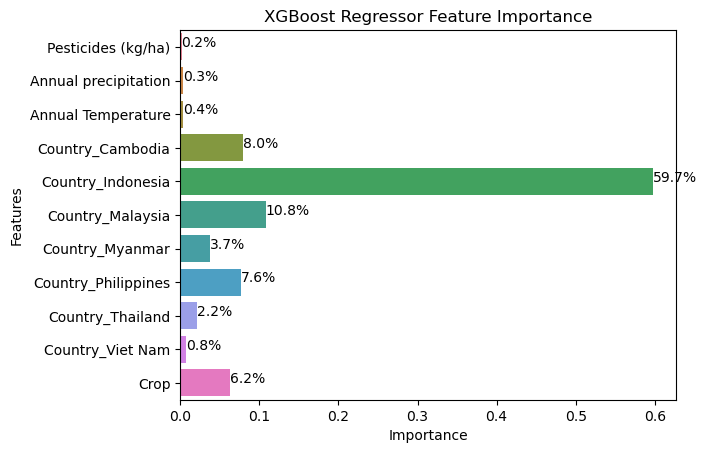


Figure 5.5. XGBoost Regressor Feature Importance

As the r2 score of the 3 models are not so different from each other, I have to decide based on how the model decides the importance of each feature. In the context of crop yield, the model should predict based on type of crop, the environmental and pesticide usage factors, and then the country producing it.

XGBoost does terribly in the context of data because it is focusing on countries rather than the other features. Gradient Boosting and Random Forest have a similar perspective on deciding feature importance. Therefore, I have to decide based on the distribution of small percentages. By comparing both models, Gradient Boosting has less emphasis on countries than Random Forest. Therefore I decided to choose Gradient Boosting Regressor as the best model to predict crop yield for ASEAN countries.

# Conclusion

In this project, I have gathered data from different sources and combined them to predict crop yield for ASEAN countries based on various factors. I have discovered insights on top-producing crops in ASEAN, and how the correlation between the yield, environmental and pesticide usage factors varies for each crop by analyzing 4 most consumed crops in the world.

Various preprocessing techniques are applied to the dataset to encode labels, and avoid bias towards high magnitude variables. I have also prepared 3 splits for the models to be trained and tested upon to find the best split where the models can generalize without overfitting nor underfitting.

Different types of regression models are cross-validated upon 3 splits and models such as Gradient Boosting Regressor, Random Forest Regressor, and XGBoost Regressor came out on top. I performed hyperparameter tuning for these models to find the best parameters for them. After training and testing the models with the best parameters, all the models perform almost the same with the r2 score of about 94%. Therefore, I decided to choose the best model based on how they predict yield by looking at feature importance. XGBoost Regressor is emphasizing a lot on countries rather than other features therefore, I decided not to choose it based on the context of the domain. Gradient Boosting and Random Forest have the similar feature importance but I decided to choose Gradient Boosting Regressor as the best model as it has less emphasis on countries than Random Forest Regressor.

# References

Masters, W.A., Djurfeldt, A.A., De Haan, C., Hazell, P., Jayne, T., Jirström, M. and Reardon, T. (2013). Urbanization and farm size in Asia and Africa: Implications for food security and agricultural research. Global Food Security, 2(3), pp.156–165. doi:https://doi.org/10.1016/j.gfs.2013.07.002.

al-Saleh, I.A. (1994). Pesticides: a review article. Journal of environmental pathology, toxicology and oncology, [online] 13(3), pp.151–161. Available at: https://europepmc.org/article/med/7722882.

Food and Agriculture Organization of the United Nations (2025). Pesticides Use. Available at: https://www.fao.org/faostat/en/#data/RP (Accessed 9th April 2025).

Food and Agriculture Organization of the United Nations (2025). Land Use. Available at: https://www.fao.org/faostat/en/#data/RL (Accessed 9th April 2025).

Food and Agriculture Organization of the United Nations (2025). Crops and livestock products. Available at: https://www.fao.org/faostat/en/#data/QCL (Accessed 9th April 2025).

Our World in Data (2025). Annual precipitation. Available at: https://ourworldindata.org/grapher/average-precipitation-per-year?time=latest#explore-the-data (Accessed 9th April 2025).

Palinatx (2024). Mean temperature for countries by year 1901-2022. Kaggle. Available at: https://www.kaggle.com/datasets/palinatx/mean-temperature-for-countries-by-year-2014-2022 (Accessed 9th April 2025).

Patel R. (2022). Crop Yield Prediction Dataset. Kaggle. Available at: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset (Accessed 9th April 2025).

Ijaz S. (2023). 15 Most Consumed Crops in the World. Yahoo Finance. Available at: https://finance.yahoo.com/news/15-most-consumed-crops-world-150822097.html?guccounter=1&guce\_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce\_referrer\_sig=AQAAAHxdfWdyulRxI52AX4MyLfc\_8WQf4HoZWRveEB3h59xQu78U4vWQHmSLz5uVV4ho2GmfpUTQieoy-6y8qIVygE-apowq3uhIBsLcQBccFbOhFj1OC-vAwGGp82d-2Pxa2WBYFYoOQRxKrjpcgq5A7MpMjmVtXj9JWv6meQrJcTZJ (Accessed 14th April 2025).

IBM (2024). Overfitting vs. Underfitting. Available at: https://www.ibm.com/think/topics/overfitting-vs-underfitting. (Accessed 18th April 2025).

Amazon Web Services (2024). What is hyperparameter tuning? - hyperparameter tuning methods explained - AWS. Amazon Web Services, Inc. Available at: https://aws.amazon.com/what-is/hyperparameter-tuning/. (Accessed 19th April 2025)

Navas, J. (2022). Anyscale - What is hyperparameter tuning? Anyscale. Available at: https://www.anyscale.com/blog/what-is-hyperparameter-tuning. (Accessed 19th April 2025)