Global Analysis of Food Crops and Climate Change II

Andrea Bendayan
Jihoon Yun
Devyani Srivastava

***Abstract*— Previously, we concluded that the atmosphere temperature has been linearly increasing during the last twenty-three years, there is evidence that human activities like the burning of fossil fuels have caused soil degradation and decrease soil fertility. The increase in heat in the atmosphere has caused changes in regional temperatures and weather patterns. There is evidence that geographic ranges are shifting, and plants and trees are blooming in late winter. The increase in soil degradation, volatility in weather patterns and increase in temperature might pose a future challenge to the growing of nutritional food crops. The last project analyzed food crops which include potatoes, maize, soybeans, and wheat to provide adequate caloric intake. The second part of the analysis will have a greater focus on crops that can help fulfill micro and macro nutrients for various sample populations across the globe. The statistical analysis will determine the optimal place to grow each individual crop. In this sample we will be evaluating the yield of Cassava, Sweet Potatoes, Sorghum, Rice, and Plantains. The aim of this research is to reduce nutritional inequalities across the globe which can be attributed to climate change.**

Introduction

There is clear evidence that the temperature has been increasing steadily throughout the decades. The increase in temperature can be attributed to the increase in carbon emission causes the greenhouse gas effect which helps to trap heat in the atmosphere. The increase in heat in the atmosphere have caused changes in temperatures and weather patterns. The impacts of these changes are drastic: glaciers and ice sheets are shrinking, freshwater ice sheets in rivers and lakes are disappearing at an alarming rate, plant and animal species are being extinguished. The devastating environmental changes could potentially affect the food supply which can lead to inequality, poverty, and human suffering. We concluded previously that there are four main crops that can adapt to rising temperatures and provide adequate caloric needs. The crops are maize, potatoes, wheat, and soybeans. These crops have been heavily subsidized by governments across the world because they are heat resistant, high in calories and can be efficiently cultivated. However, those four main crops don't address the micro and macro nutritional deficiencies that are mostly shared by the economic disadvantaged inhabitants of developing nations. Data science and statistics can provide us with insight to mitigate the negative effects climate change. Statistics can help us predict climate changes over years, and its impact on farming and crop yields. This information can be used by scientists, statisticians, policy makers and global leaders to take vital measures to provide more effective instruments and policies to ensure that the global population has a wide variety of foods that can sustain human life. In context to the above motive, we decided to work on a sample with 23-years of study in temperatures and crop yield changes over various countries, which includes all the major continents.

.

DATA GATHERING AND DATA PREPARATION

Data Gathering

A. *The raw data has been collected from Kaggle. Further, the Kaggle data set is a dataset of Climate change with cropland in various countries, we decided to conduct analysis and predict the produced crop yield.*

B. *Dataset Details:*

C. *Dataset Years: provides the date of observation starting from 1990 to 2013.*

D. *AverageTemperature: details about local average temperature in Celsius.*

E. *Item: provides the specific crop that was grown*

F. *Area: details about country names.*

G. *Average_Rainfall: provides values on the quantity of collected rainfall per year in millimeters.*

H. *Pesticide_tonnes: Provides values on the quantity of pesticide tons per year.*

I. *Hg/Ha_yield: Yield output of the item or crop dependent on all the above factors.*

1. Data Cleaning and Data Preparation

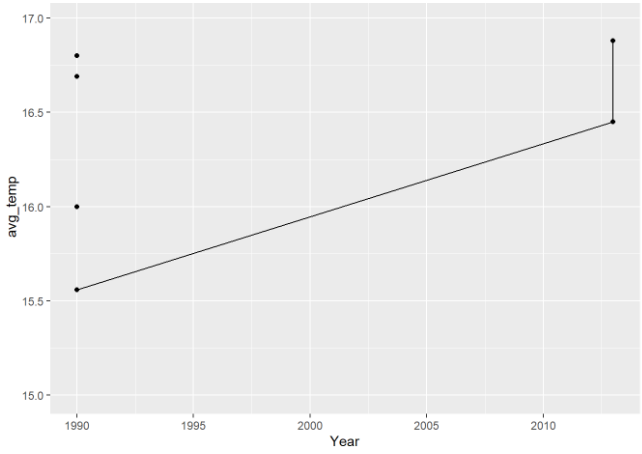
To clean and prepare the data, we performed the following steps:

- We are limiting our analysis to certain crops: cassava, sweet potatoes, sorghum, rice, and plantains and other bananas.
- We are curtailing to countries: Argentina, Mexico, Ukraine, Bangladesh, South Africa, and Australia
- Created data frames to perform comparative analysis of year 1990 and 2013, as well as analysis on a continuous basis throughout the years.

Data Preparation Steps:
Packages used in R

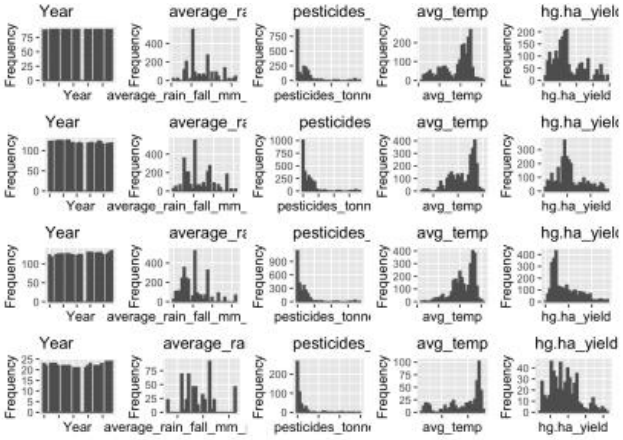
```r
```{r}
library(tidyverse)
library(dplyr)
library(readr)
library(ggplot2)
library(gridExtra)
library(caret)
```
```

Before conducting the exploratory data analysis on the desired crop yields, we were able to demonstrate that time in years and increase in average temperatures have a positive linear relationship. The data shows that from 1990-2013 there has been an average of an increase of 2 degrees in the 6 countries which are in different regions and continents. Therefore, we can infer from the results that there has been increase in global temperature in the last 23 years.



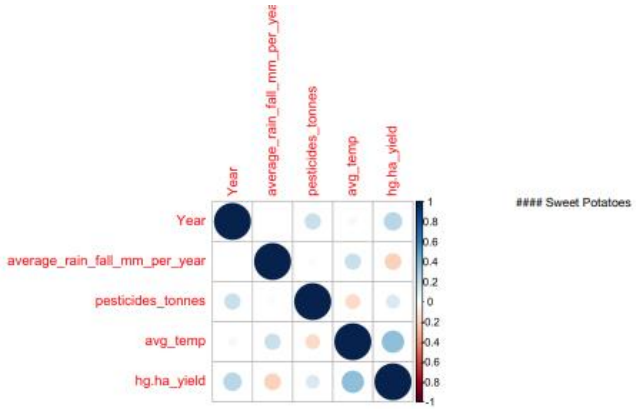
DATA PREPROCESSING

The data frame has been scanned for duplicate values. Duplicates and non-applicable values have been removed and excluded from the data frame. The summary function has been applied to show the statistics in the 4 desired crops. Categorical values of two of our input features, Item and Area will be converted into numerical values for easier processing. In preparation for data processing all the outliers will be identified and removed. Additionally, data will be processed by using one-hot encoding. One-hot encoding is a technique used to convert categorical variables into a format that can be used for regression analysis. The Area variable is a categorical variable that contains information about multiple countries. Using one-hot encoding, the function will convert this variable into a set of binary dummy variables, one for each level of the Area variable. The reason for performing one-hot encoding on a categorical variable is that regression analysis requires all predictor variables to be numeric. Since categorical variables are not numeric, they cannot be used directly in a regression model. One-hot encoding converts a categorical variable into a set of binary dummy variables that can be used as numeric predictor variables in a regression model.



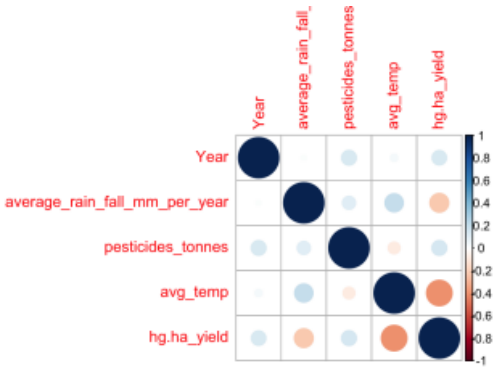
CORRELEATION ANALYSIS

To understand the correlation between the features we used correlation analysis. To execute this in R created a function that defined the correlation table and used the corrplot package. Below is the correlation tables for the 4 crops of interest.



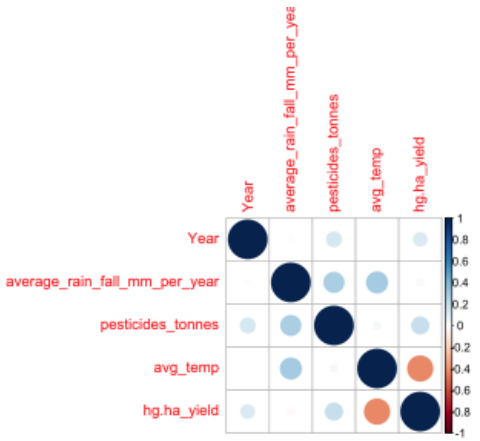
Cassava

There is a weak positive correlation between Year and yield (correlation coefficient = 0.238). This indicates that as the year increases, the yield of the crop tends to increase. There is a weak negative correlation between rainfall and yield (correlation coefficient = -0.193). This indicates that as the average rainfall increases, the yield of the crop tends to decrease. There is a very weak positive correlation between pesticide amount and yield (correlation coefficient = 0.127). This indicates that as pesticide usage increases, the yield of the crop tends to increase slightly. There is a weak positive correlation between temperature and yield (correlation coefficient = 0.354). This indicates that the yield of the crop tends to increase as the average temperature increases.



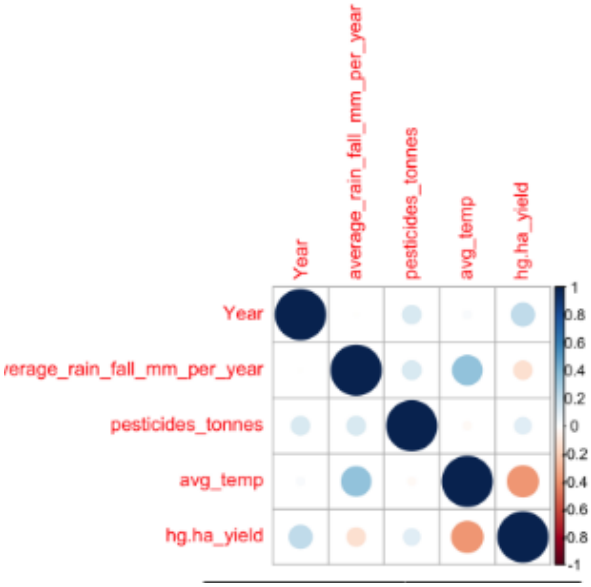
Sweet Potatoes

There is a weak positive correlation between Year and yield (correlation coefficient = 0.133). This indicates that the yield of the crop tends to increase slightly as the year increases. There is a weak negative correlation between rainfall and yield (correlation coefficient = -0.223). This indicates that as the average rainfall increases, the yield of the crop tends to decrease. There is a very weak positive correlation between pesticide amount and yield (correlation coefficient = 0.140). This indicates that as pesticide usage increases, the yield of the crop tends to increase slightly. There is a weak negative correlation between temperature and yield (correlation coefficient = -0.407). This indicates that as the average temperature increases, the yield of the crop tends to decrease. from the matrix we infer that there is no significant correlation between the features.



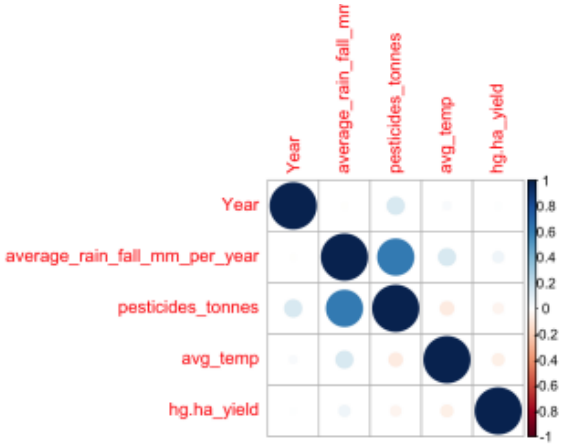
Sorghum

There is a very weak positive correlation between year and yield (correlation coefficient = 0.123). This indicates that the yield of the crop tends to increase slightly as the year increases. There is almost no correlation between rainfall and yield (correlation coefficient = -0.029). This indicates that the average rainfall has little effect on the yield of the crop. There is a weak positive correlation between pesticide amount and yield (correlation coefficient = 0.196). This indicates that as pesticide usage increases, crop yields tend to increase. There is a weak negative correlation between temperature increase and yield (correlation coefficient = -0.427). This indicates that as the average temperature increases, the yield of the crop tends to decrease



Rice

There is a weak positive correlation between Year and yield (correlation coefficient = 0.212). This indicates that the year has a slight effect on the yield of the crop. There is little correlation between rainfall and the yield (correlation coefficient = -0.133). This indicates that the average rainfall has little effect on the yield of the crop. There is little correlation between pesticide amount and yield (correlation coefficient = 0.108). This indicates that the amount of pesticides used has little effect on the yield of the crop. There is a weak negative correlation between temperature and yield (correlation coefficient = -0.389). This indicates that the average temperature has a slight negative effect on the yield of the crop.



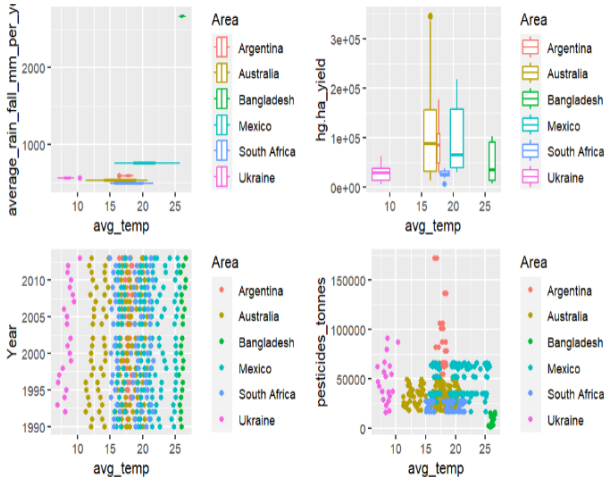
Plantains and Others

There is almost no correlation between Year and yield (correlation coefficient = 0.009). This indicates that the year has little effect on the yield of the crop. There is little correlation between rainfall and yield (correlation coefficient = 0.056). This indicates that the average rainfall has little effect on the yield of the crop. There is little correlation between pesticide amount and yield (correlation coefficient = -0.050). This indicates that the number of pesticides used has little effect on the yield of the crop. There is little correlation between temperature and yield (correlation coefficient = -0.066). This indicates that the average temperature has little effect on the yield of the crop.

II. Exploratory Data Analysis

We analyzed the trends of average temperature, pesticides, rainfall for the list of countries producing cassava, sweet potatoes, sorghum, rice and plantains in the dataset, resulting in the crop yields.

We can easily understand and compare the different temperature ranges, crop yields, pesticides tons used and average rainfall for the different countries using the below data visualizations.



1) Rainfall and average temperature relationship per area or country

The temperature varies from 10 to 25 degrees C or in Fahrenheit: 50 to 77 degrees, in the data set. The highest temperature is found in Bangladesh, while the lowest temperature is found in Ukraine. The graph shows a positive correlation between temperature and rainfall. The increase in temperature causes greater average rainfall, while lower temperatures decrease the amount of average rainfall. Bangladesh was removed because the amount of rainfall was an outstanding outlier in the dataset.

2) Relationship between temperature and average yield

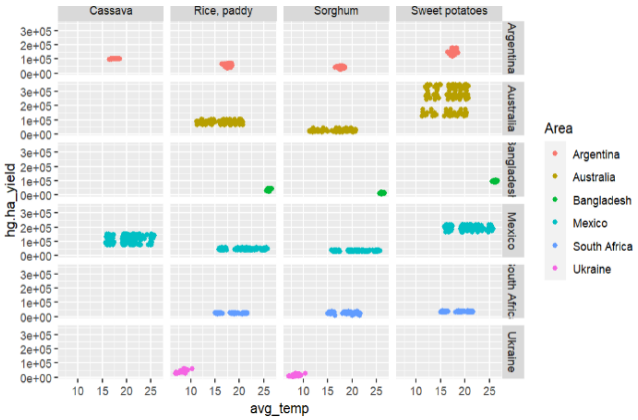
The boxplot shows the relationship between the average yield and temperature. The country with the highest temperature is Bangladesh, however the mean yield is as low as Ukraine. The countries with the highest mean yields of crops are Argentina and Australia, which have more of a temperate climate. The graph shows that an increase or decrease in temperature lowers the mean yield of crops.

3) Temperature variance from 1990 to 2013.

The graph shows the relationship of the variance between the countries in respect to increase in temperature throughout the last 23 years. Ukraine data point in 1990 and 1991 has been omitted in the dataset since it was part of the Soviet Union. The graph shows that warmer countries such as Bangladesh have had little to no change in respect to increase in temperature. However, in colder climates such as Ukraine it has been getting warmer throughout the decades, which can present an opportunity for cultivation in former colder countries.

4) Temperature and pesticide use

Pesticide use is clustered between the ideal temperatures of 15C and 23 C, which makes sense because within this temperature is where the most crop yields are. Bangladesh has the lowest pesticide use, and it is possibly due to being a developing nation, with little progress in farming industrialization. By contrast Argentina has the highest use of pesticides and it might be due to the decrease in government regulation to regulate the use of pesticides. Ukraine has also a high degree of pesticide use, however we can infer that pesticides are not the most important variable to achieve higher yields, since Australia which has a lower pesticide use has the highest yield of all the countries.



Yield per crop item and country compared to average temperature.

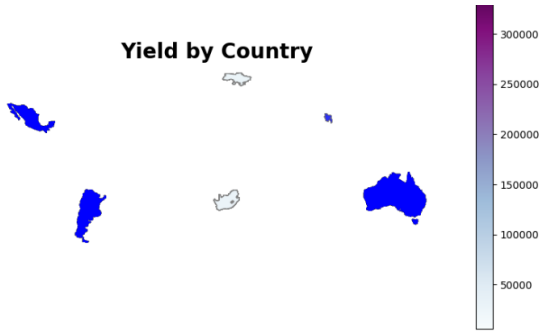
Cassava: Cassava is cultivated in Latin America, and it's a staple in many Latin American dishes. Cassava grows the best in temperatures that are over 15 degrees. Mexico currently has the highest yield since it is a warm tropical country in central America. Argentina has a moderate yield, which indicates that cassava grows better in warmer climates.

Rice: Rice is grown in all the sample population of countries. The yield is standard in every country, which infers that rice is very resilient and resistant crop regarding temperature changes.

Sorghum: Similarly, to rice, the yield is not affected by the temperature. There is a uniform and standard yield amount in all the samples, suggesting that sorghum is incredibly resistant to temperature volatility.

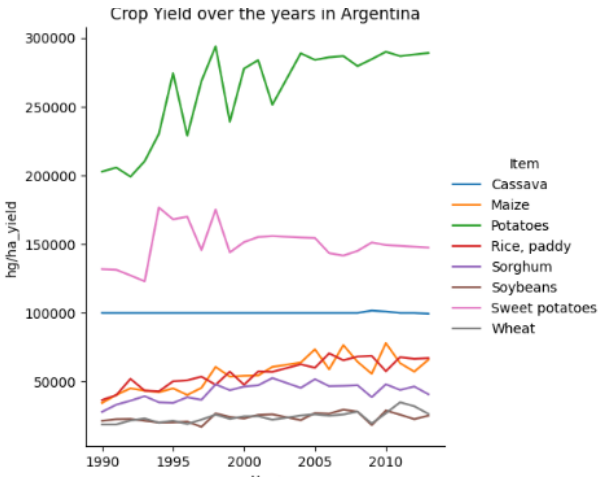
Sweet potatoes: Sweet Potatoes are not grown in Ukraine. The country with the lowest yield in sweet potatoes is South Africa, while the highest yield corresponds to Australia. Argentina and Mexico have a moderate yield, while Bangladesh has a smaller yield. In regard to temperature, the crop grows well between 15 and 25 degrees.

Plantains and others: There were no statistically significant yields in the sample population, and therefore it is omitted.

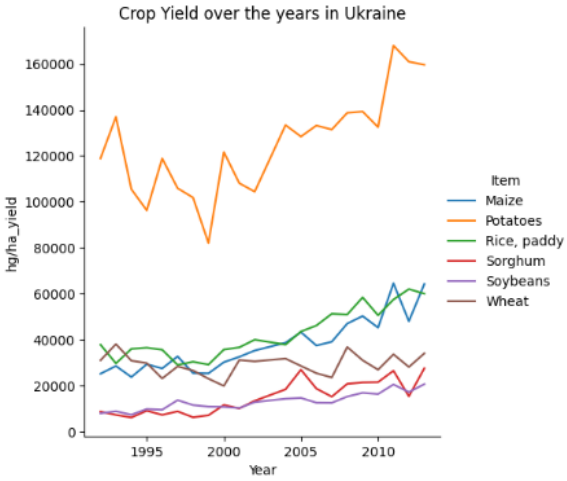


Countries with the highest yields, are large, industrialized nations such as Australia, Argentina and Mexico with temperatures that are mild.

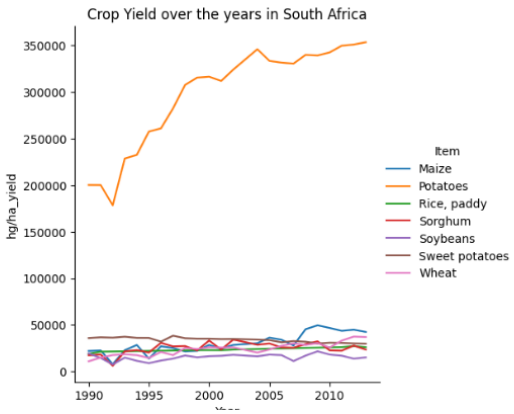
A closer look at Yields per Countries with all crops in the dataset



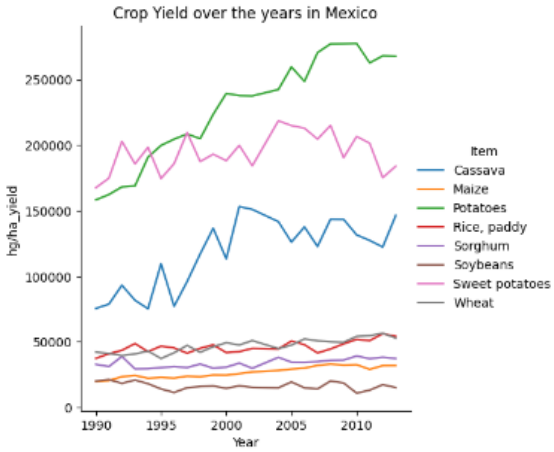
Sample population Argentina: Argentina seems to have the highest yield in potatoes, followed by sweet potatoes. Cassava, Maize, Rice, Soybeans and Sorghum have smaller yields compared to potatoes. Over the years there has been an increase in potato yields. There has been also a modest increase in the yields of rice and maize.



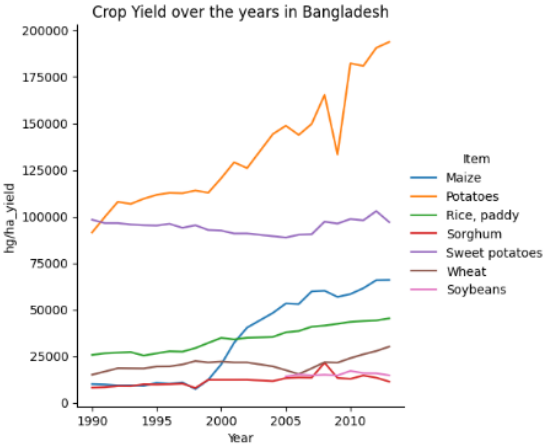
Sample Population Ukraine: Ukraine highest yield is represented in potatoes, and then followed by moderate yields in maize and rice. Lower yields correspond to the crops sorghum, soybeans, and wheat. Through the decades there has been an increase in the yields of potatoes, rice, maize, sorghum, and soybeans. Wheat has remained constant. The increase in yields for all crops could be attributed to the increase in temperature in Ukraine.



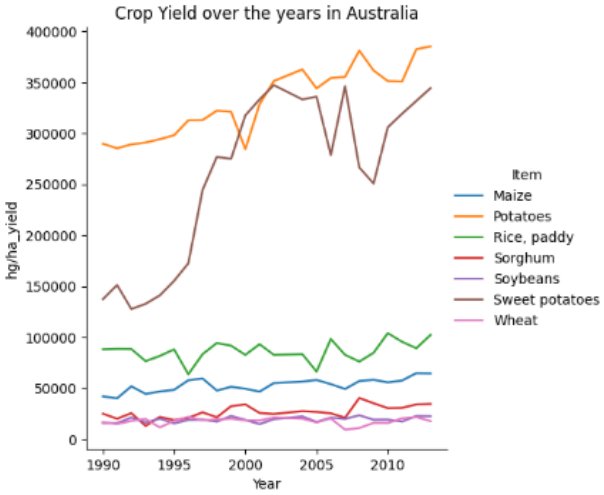
Sample Population South Africa: South Africa has the highest yield in potatoes, while having lower yields in all other countries. The crops that have increased their yield through the years include potatoes and maize. All other crops have remained constant.



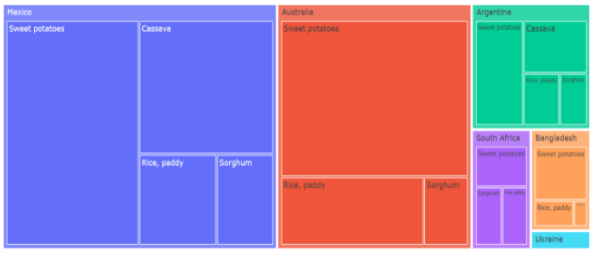
Sample Population Mexico: Mexico highest yield are made up by potatoes, sweet potatoes, and cassava. Mexico has lower yield for all the other crops which have remained constant over time. Throughout the years there is an increase in the yield of potatoes and cassava, while all the other have remained stable, with an exception in the decrease of sweet potatoes.



Sample Population Bangladesh: Bangladesh has the highest yields in potatoes and sweet potatoes. Bangladesh also has moderate yields for maize and rice. Throughout the years there has been a sharp increase in the yield of potatoes and maize, and a moderate increase in the yield of rice. All the other crops have remained constant.

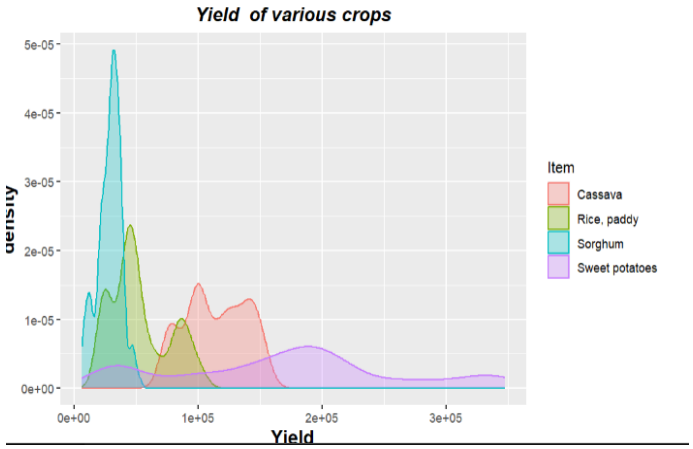


Sample Population Australia: Australia has equally high yields in potatoes and sweet potatoes. All the other crops have moderate yields and have been stable throughout the years. Yields have dramatically increased for both potatoes and sweet potatoes.



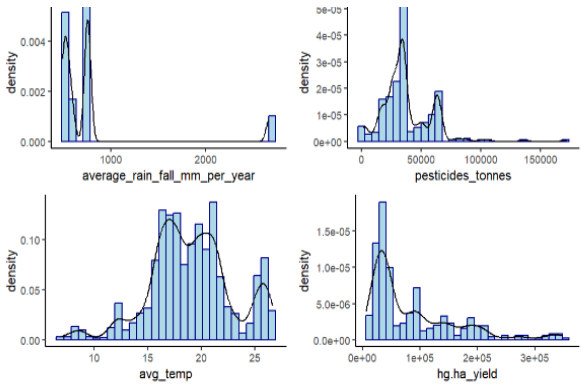
Sweet Potatoes seems to be the highest yielding crop for most countries.

Probability Density and Yield per crop



Sweet Potatoes have the larger spread while sorghum has the narrowest spread.

Probability density relative to variables



III. Multi Linear Regression Modeling

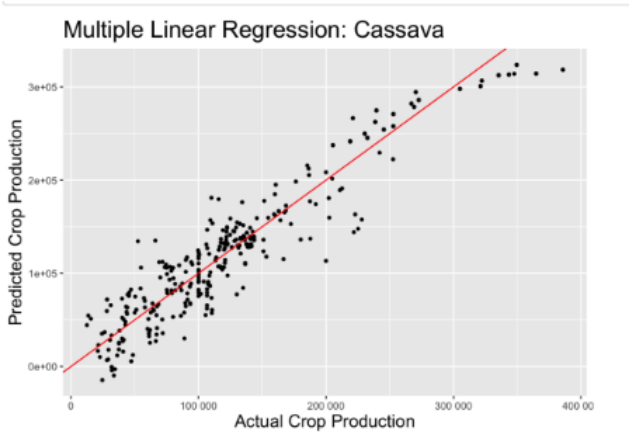
Multiple linear regression is used to assess the relationship between two variables while considering the effect of other variables.
Multiple linear regression models are defined by the equation

$$Y=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta pXp+\epsilon$$

The first model applied is multiple linear regression model since it is straightforward to interpret the coefficients of the model. This will help with the understanding of the relationship between the predictor and the dependent variable (yield).

A. MULTIVARIATE LINEAR REGRESSION RESULTS

Cassava



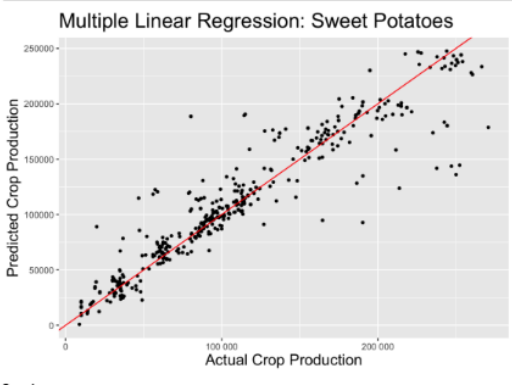

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91079 -16008  -3298  14940  90273
##
```

```
## Year                          3.467e+03  9.311e+01  37.236  < 2e-16 ***
## average_rain_fall_mm_per_year         NA         NA      NA       NA
## pesticides_tonnes            -2.071e-01  1.594e-02 -12.991  < 2e-16 ***
## avg_temp                      1.858e+02  3.654e+02   0.509  0.61116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25040 on 1593 degrees of freedom
## Multiple R-squared:  0.9252, Adjusted R-squared:  0.9232
## F-statistic: 458.1 on 43 and 1593 DF,  p-value: < 2.2e-16
```

Coefficients

The coefficient for the Year variable is 3.467e+03, which indicates that, on average, the response variable increases by 3.467e+03 units for each one-unit increase in the Year variable, holding all other predictor variables constant. The p-value for this variable is less than 2e-16, which indicates that this relationship is **statistically significant at the 0.001 level**. The coefficient for the pesticide variable is -2.071e-01, which indicates that, on average, the response variable decreases by 2.071e-01 units for each one-unit increase in the pesticides variable, holding all other predictor variables constant. The p-value for this variable is also less than 2e-16, which indicates that this **relationship is statistically significant at the 0.001 leve**l. The coefficient for the average temperature variable is 1.858e+02, which indicates that, on average, the response variable increases by 1.858e+02 units for each one-unit increase in the average temperature variable, holding all other predictor variables constant. However, the p-value for this variable is 0.61116, which **indicates that this relationship is not statistically significant** at conventional levels. The output also shows that there is no information available for the average rainfall per year variable. This could be due to missing data or collinearity with other predictor variables. The multiple R-squared value of 0.9252 indicates that approximately 92.52% of the variance in the response variable

can be explained by the predictor variables included in the model. **The F-statistic and its associated p-value indicate that the overall model is statistically significant at conventional levels**



Sweet Potato

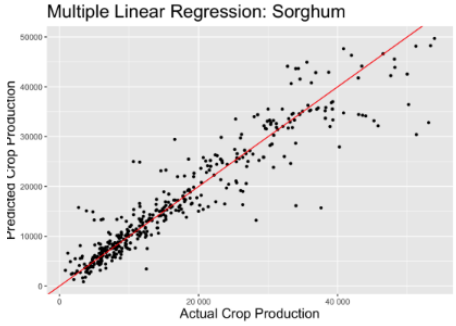
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118442   -5472    -100    5417  153008
##
```

```
## Year                          8.988e+02  6.196e+01  14.505  < 2e-16 ***
## average_rain_fall_mm_per_year        NA         NA      NA       NA
## pesticides_tonnes             1.902e-02  1.284e-02   1.481  0.13883
## avg_temp                      4.178e+01  2.349e+02   0.178  0.85888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19840 on 2172 degrees of freedom
## Multiple R-squared:  0.8896, Adjusted R-squared:  0.8868
## F-statistic: 318.1 on 55 and 2172 DF,  p-value: < 2.2e-16
```

Coefficients

The coefficient for the Year variable is 8.988e+02, which indicates that, on average, the response variable increases by 8.988e+02 units for each one-unit increase in the Year variable, holding all other predictor variables constant. The p-value for this variable is less than 2e-16, which indicates that this **relationship is statistically significant at the 0.001** level. The coefficient for the pesticides variable is 1.902e-02, which indicates that, on average, the response variable increases by 1.902e-02 units for each one-unit increase in the pesticides variable, holding all other predictor variables constant. However, the p-value for this variable is 0.13883, which indicates that this **relationship is not statistically significant at conventional levels**. The coefficient for the average temperature variable is 4.178e+01, which indicates that, on average, the response variable increases by 4.178e+01 units for each one-unit increase in the average temperature variable, holding all other predictor variables constant. However, the p-value for this variable is 0.85888, which indicates that this **relationship is not statistically significant at conventional levels**. The output also shows that there is no information available for the average rainfall variable. This could be due to missing data or collinearity with other predictor variables. The multiple R-squared value of 0.8896 indicates that approximately 88.96% of the variance in the response variable can be explained by the predictor variables included in the model. **The F-statistic and its associated p-value indicate that the overall model is statistically significant at conventional levels**.

Sorghum



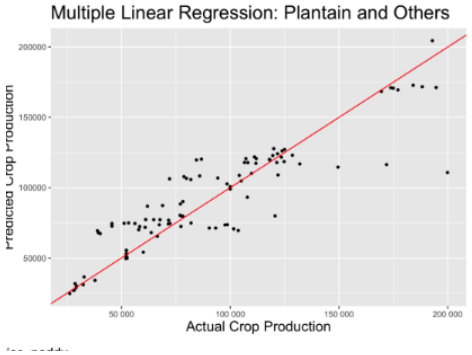
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18914  -1446    -132    1380   36352
##

## Year                           1.551e+02  1.251e+01  12.401  < 2e-16 ***
## average_rain_fall_mm_per_year         NA         NA      NA       NA
## pesticides_tonnes              2.470e-02  2.608e-03   9.471  < 2e-16 ***
## avg_temp                       4.152e+00  4.581e+01   0.091 0.927791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4006 on 2267 degrees of freedom
```

Coefficients

The coefficient for the Year variable is 1.551e+02, which indicates that, on average, the response variable increases by 1.551e+02 units for each one-unit increase in the Year variable, holding all other predictor variables constant. The p-value for this variable is less than 2e-16, which indicates that this **relationship is statistically significant at the 0.001 level**. The coefficient for the pesticides amount variable is 2.470e-02, which indicates that, on average, the response variable increases by 2.470e02 units for each one-unit increase in the pesticides variable, holding all other predictor variables constant. The p-value for this variable is also less than 2e-16, which indicates that **this relationship is statistically significant at the 0.001 level**. The coefficient for the average temperature variable is 4.152e+00, which indicates that, on average, the response variable increases by 4.152e+00 units for each one-unit increase in the average temperature variable, holding all other predictor variables constant. However, the p-value for this variable is 0.927791, which indicates that **this relationship is not statistically significant at conventional levels**. The output also shows that there is no information available for the average rainfall per year variable. This could be due to missing data or collinearity with other predictor variables. The multiple R-squared value of 0.8911 indicates that approximately 89.11% of the variance in the response variable can be explained by the predictor variables included in the model. **The F-statistic and its associated p-value indicate that the overall model is statistically significant at conventional levels.**

Plantains and Others




```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54596  -9308   -240    5559   89039
##

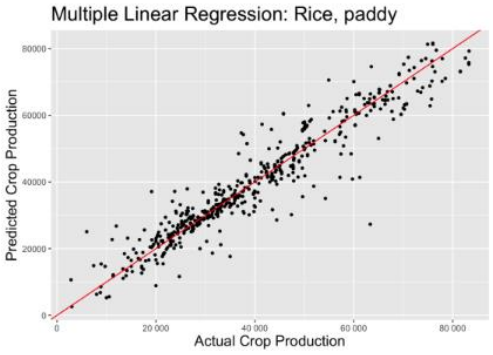
## Year                           3.028e+02  1.342e+02   2.257 0.024590 *
## average_rain_fall_mm_per_year         NA         NA      NA       NA
## pesticides_tonnes             -7.563e-02  9.254e-02  -0.817 0.414277
## avg_temp                       3.789e+02  8.169e+02   0.464 0.643077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18490 on 389 degrees of freedom
## Multiple R-squared:  0.7997, Adjusted R-squared:  0.7879
## F-statistic: 67.54 on 23 and 389 DF,  p-value: < 2.2e-16
```

Coefficients:

The coefficient for the Year variable is 3.028e+02, which indicates that, on average, the response variable increases by 3.028e+02 units for each one-unit increase in the Year variable, holding all other predictor variables constant. The p-value for this variable is 0.024590,

which indicates that this **relationship is statistically significant at the 0.05 level**. The coefficient for the pesticides variable is -7.563e-02, which indicates that, on average, the response variable decreases by7.563e-02 units for each one-unit increase in the pesticides variable, holding all other predictor variables constant.
The p-value for this variable is 0.414277, which indicates that this **relationship is not statistically significant at conventional levels.** The coefficient for the average temperature variable is 3.789e+02, which indicates that, on average, the response variable increases by 3.789e+02 units for each one-unit increase in the average temperature variable, holding all other predictor variables constant. However, the p-value for this variable is 0.643077, which indicates that this **relationship is not statistically significant at conventional levels.**
The output also shows that there is no information available for the average rainfall per year variable. This could be due to missing data or collinearity with other predictor variables.
The multiple R-squared value of 0.7997 indicates that approximately 79.97% of the variance in the response variable can be explained by the predictor variables included in the model. **The F-statistic and its associated p-value indicate that the overall model is statistically significant at conventional levels.**

Rice




```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25559  -1704     147    1857   33370
##
## Year                          5.044e+02  1.463e+01  34.483  < 2e-16 ***
## average_rain_fall_mm_per_year        NA         NA      NA       NA
## pesticides_tonnes             5.227e-02  3.139e-03  16.654  < 2e-16 ***
## avg_temp                      6.843e+01  5.385e+01   1.271 0.203905
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5017 on 2543 degrees of freedom
## Multiple R-squared:  0.9115, Adjusted R-squared:  0.9091
## F-statistic: 374.2 on 70 and 2543 DF,  p-value: < 2.2e-16
```

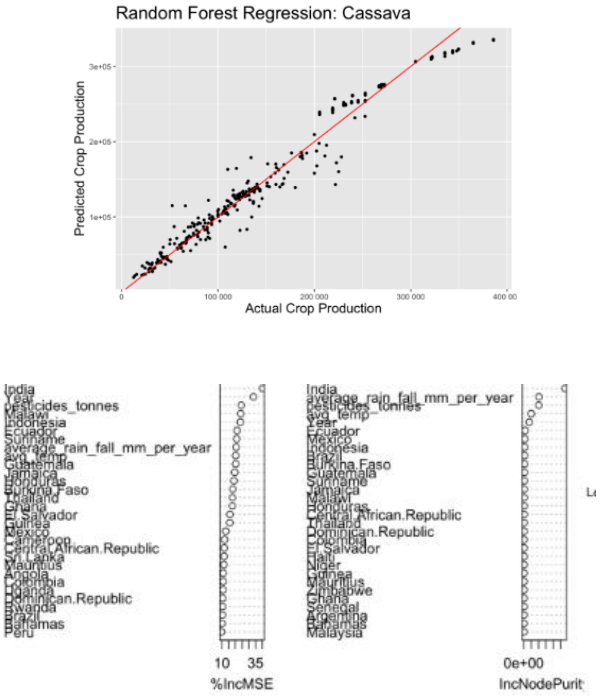
Coefficients
The coefficient for the Year variable is 5.044e+02, which means that for every unit increase in Year, the predicted yield increases by 504.4 hg/ha. This coefficient is also statistically significant with a p-value less than 0.05, which indicates that the Year variable is a good predictor of the yield. The coefficient for average rainfall per year is not defined because of singularities. This could be because this variable is perfectly correlated with another variable(s) in the model. The coefficient for pesticides is 5.227e-02, which means that for every unit increase in pesticides, the predicted yield increases by 0.05227 hg/ha. This coefficient is also statistically significant with a p-value less than 0.05, which indicates that the pesticides variable is a good predictor of the yield. The coefficient for average temperature is 6.843e+01, which means that for every unit increase in average temperature , the predicted yield increases by68.43 hg/ha. However, this coefficient is not statistically significant with a p-value greater than 0.05, which indicates that the average temperature variable **relationship is not statistically significant**. The multiple R-squared value of 0.9115 indicates that the model explains 91.15% of the variation in the yield. The adjusted R-squared
value of 0.9091 takes into account the number of predictor variables in the model and penalizes the model for including irrelevant variables.

The F-statistic of 374.2 with a p-value less than 0.05 indicates that the model is statistically significant and at least one of the predictor variables is a good predictor of the yield.

IV. RANDOM FOREST REGRESSION

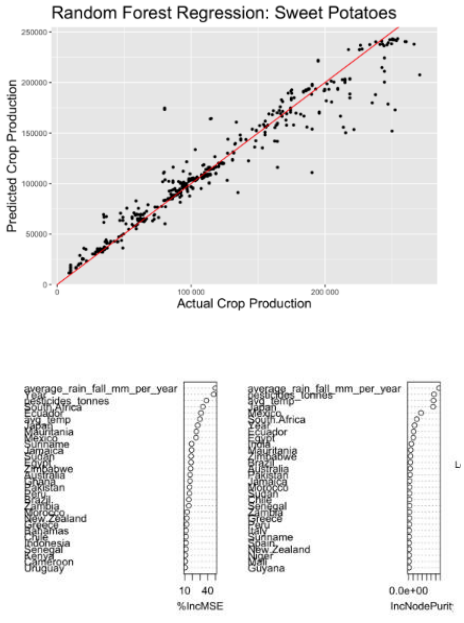
The Random Forest Regression model, does not directly calculate p-values. However, the model does provide variable significance. Variable significance or importance is a measure of how much each variable affects the model's predictive performance, determines if a variable is statistically significant. The Random Forest Regression model can also estimate the variable importance by measuring how much the model's predictive performance decreases when each variable is randomly transposed to residual or fitted.

The first metric is *%IncMSE* (Increased MSE), which indicates how much the MSE of the model increases when each variable is not selected. Therefore, a higher *%IncMSE* value means that the variable is important to the performance of the model. The second metric is *IncNodePurity* (increase in node purity), which indicates how much the purity of the nodes (percentage of samples belonging to one class) increases when the nodes are divided using that variable. Therefore, a higher value of *IncNodePurity* means that the variable is important for dividing the nodes.



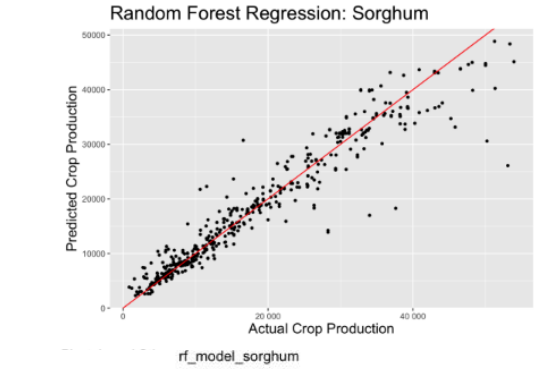
Cassava

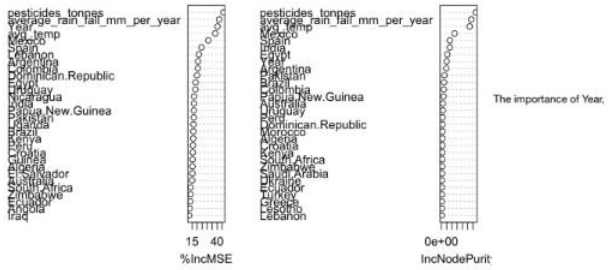
the variable rainfall per year has the highest importance, followed by Pesticides, average temperature, and Year.



Sweet Potato

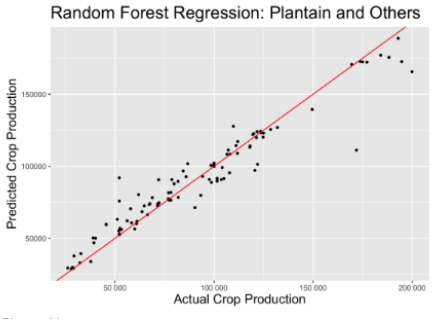
In this model among variables: year, pesticides, rainfall per year, and average temperature, the feature with the highest *%IncMSE* value is Year. This shows that this feature reduces the MSE of the model the most. Also, the feature that reduces node impurity the most is average rainfall per year. Therefore, Year is the most important feature and average rainfall per year also plays an important role. The other features, pesticides, and average temperature, also play an important role in the model.

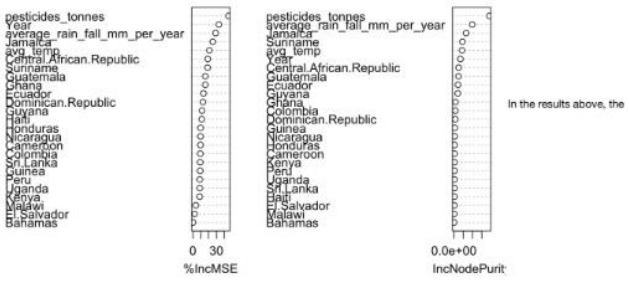




Sorghum

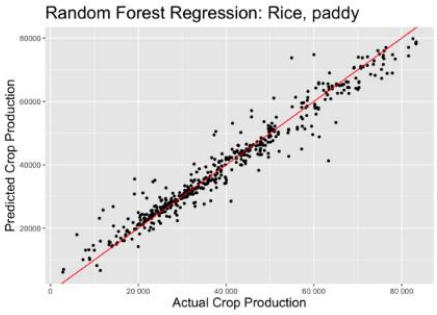
The following variables, with their respective results: pesticides, average rainfall per year, and average temperature are 41.98%, 47.14%, 44.53%, and 37.42%, respectively. Among them, pesticides has the highest importance, which could mean that the amount of pesticides and their impact on food production in the target region is large. In addition, year, rainfall per year and average temperature all have high importance, which indicates that climate and year have a great impact on food production in the target area.

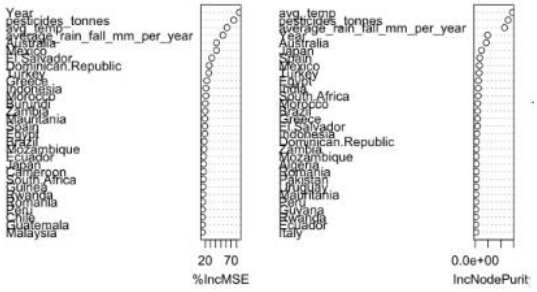




Plantains and Others

The variables pesticides, average rainfall per year, average temperature, and Year have the highest importance. The pesticides variable was found to have the largest impact on classification accuracy, which suggests that the variable is related to crop production and health issues. The average rainfall per year variable was found to be an important variable because it is directly related to crop production. The average temperature variable was found to be an important variable because it is one of many factors related to crop production and health issues. The Year variable was found to be an important variable because the agricultural environment changes significantly from year to year.





Rice

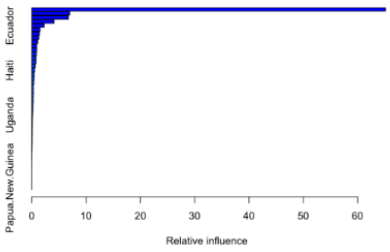
Year has the highest significance of 85.7%, which can be attributed to the fact that rice production varies greatly from year to year. The pesticides variable has a high significance of 70.0%. This is one of the important variables affecting rice production, and it is important to ensure that the right number of pesticides are used to increase rice production. The average temperature variable has a moderate
significance of 63.3% and is one of the factors affecting production. Higher or lower temperatures relative to a constant temperature can have a negative impact on rice production. The average rainfall per year variable has an importance of 55.8%, which is moderately influential. This shows that sufficient rainfall plays an important role in rice production.

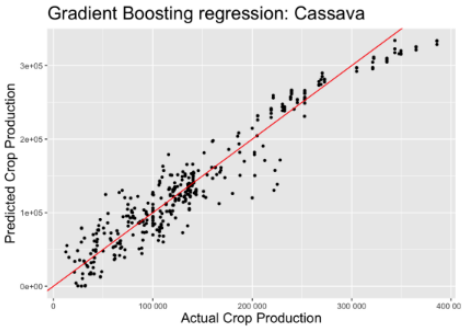
Findings

The random forest machine learning model has contributed to enhancing the classification accuracy for prediction using the variables. Therefore, these variables play an important role in the data and are the most useful variables for the model to predict.

V. Gradient Boosting Regression

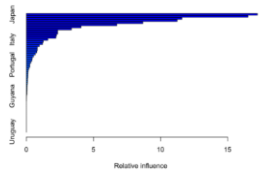
The gradient boosting regression model helps to obtain the importance of each feature, which can indicate how much influence a specific feature has on the future predictions. The gradient boosting model uses a procedure to obtain the calculations of the feature's importance. This approach is commonly used in tree-based algorithms. The procedure involves counting the number of times each feature is used and how much it contributes to the improvement in predictive power in the trees.

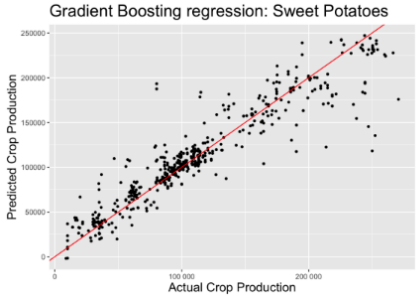




Cassava

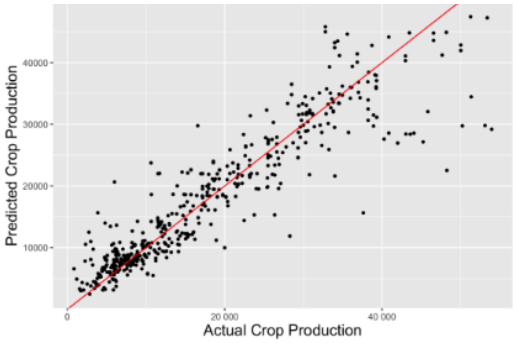
In this case pesticides have the highest importance, followed by average temperature and average rainfall per year. The Year variable also has a high importance, but not by much compared to the other variables. These results suggest that in this dataset, pesticide usage and climate conditions have a strong influence on the model predictions.

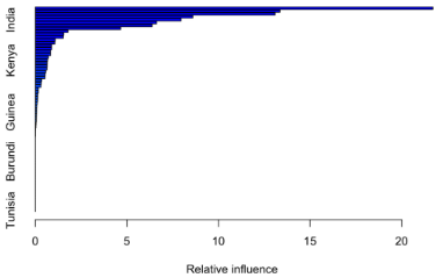




Sweet Potatoes

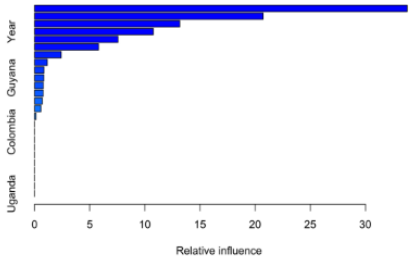
From the results above, average rainfall per year appears to be the most important feature in the model, followed by average temp and pesticides. These results suggest that climate and pesticide usage have a strong influence on the model's predictions. However, Year appears to be less important compared to the rest of the features impact in this model.

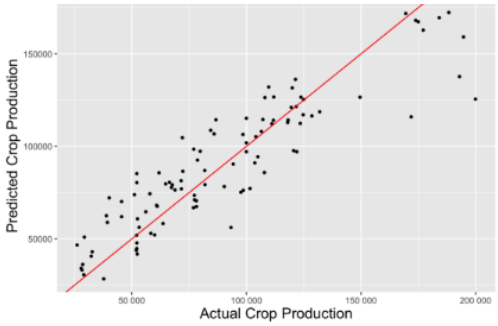




Sorghum

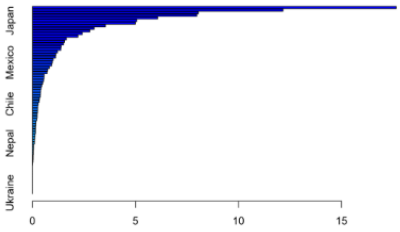
Average Temp' This is an important feature because the higher the average annual temperature in a region, the more likely it is that insects, diseases, etc. will occur and reduce yields. Next, average rainfall per year, is an important feature because the more annual rainfall an area receives, the better the crops will grow. Pesticides This is an important feature because the higher the amount of pesticides used in an area, the more productive the crops will be, but at the same time, the more likely it is that side effects will occur, such as pesticide residues affecting food safety. Year: The dataset contains data from 1961 to 2016, which is important feature because different years can have different effects on crop productivity or disasters. clustering which is due to different crop types we took. The combination of the exploratory data analysis and the linear regression, shows that the increase in global temperatures of 2 degrees in twenty three years, is not significantly correlated to a decrease or increase in yield output. However, rainfall is signficantly correlated to yield crop output, suggesting that the increase in temperature might be indirectly related to output yields.

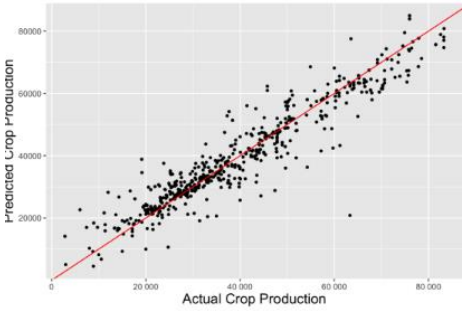




Plantains and Others

Pesticides has the highest importance at 33.7856, followed by average rainfall per year at 20.7252. Average temperature has a relatively low importance of 13.1604, and "Year" has an even lower importance of 10.7580. This importance value indicates the influence of the feature on the predicted outcome, with higher values having a greater impact on the prediction. Therefore, pesticides and average rainfall per year have the greatest impact on the model's prediction, while average temperature and "Year" have an impact on the prediction but are less influential.





Rice

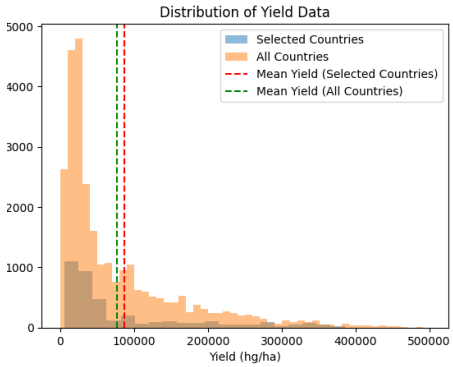
The importance indicates how important the variable is in predicting the dependent variable (rice production). For example, the variable avg temp has the highest importance of 16.89, indicating that it has the strongest relationship with the dependent variable pesticides and average rainfall per year also appear as significant variables, with significance values of 12.76 and 7.94, respectively. On the other hand, the variable "Year" appears to have a relatively low importance (6.04).

Findings and Results

Regression model has the best outcomes.

V. STATISTICAL TESTING

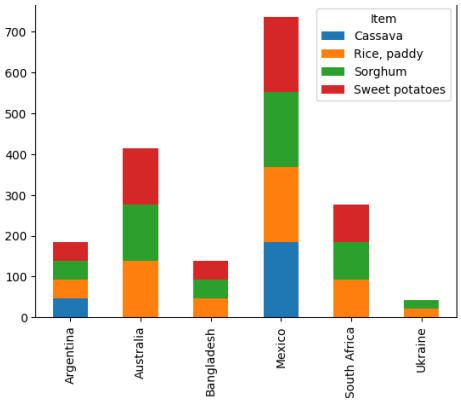
1) T Test:



The t-value indicates how many standard errors the sample mean is away from the overall mean, and the p-value indicates the probability of obtaining a t-value as extreme as the observed t-value, assuming the null hypothesis is true. If the p-value is less than a chosen significance level (e.g., 0.05), we reject the null hypothesis and conclude that the mean yield for the selected countries is significantly different from the overall mean yield. The t-test on the given data, assumes that the null hypothesis is that the mean yield for the selected countries (Mexico, Australia, South Africa, Argentina, Bangladesh, Ukraine) is equal to the overall mean yield for all countries, and the alternative hypothesis is that the mean yield for the selected countries is different from the overall mean yield.

p=1 h1 accepted

2) Chi squared Test:



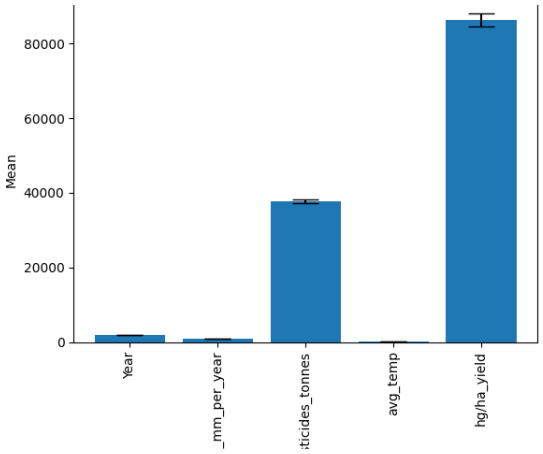
The data for the selected countries, calculates the overall mean and standard deviation of the yield, performs a one-sample t-test with the selected data and the overall mean, and prints the t-value and p-value. The t-value indicates how many standard errors the sample mean is away from the overall mean, and the p-value indicates the probability of obtaining a t-value as extreme as the observed t-value, assuming the null hypothesis is true. If the p-value is less than a chosen significance level (e.g., 0.05), we reject the null hypothesis and conclude that the mean yield for the selected countries is significantly different from the overall mean yield.

h0: No relationship between Countries and crop

h1: relationship exist.

null hypothesis h0 accepted.

3) One Way Anova:



The graph is showing the comparison of means, we can use a bar plot with error bars to represent the standard error of the mean. In the above graph each bar represents the mean of a variable and the error bars represent the standard error of the mean.

h0: no statistical difference between rainfall,year,temperature,pesticides,yield

h1: there is statistical difference between rainfall,year,temperature,pesticides,yield

p=0

h1 is accepted.

VI. Conclusion

The combination of the three models, statistical testing and exploratory data analysis shows that the increase in global temperature of 2 degrees in 23 years, is not significantly correlated to a decrease in output of the selected crop yield. In some cases, the increase in temperature benefits the yields of specific crops. However, there is statistical significance differences in the output of yield when related to average rainfall per year, pesticide use and yearly productivity. The increase in temperature relating to global warming can provide an opportunity for countries with colder climate conditions to increase their yield of the selected crops. However, crops might be affected by the amount of rainfall that is also caused by climate change. In conclusion, it is safe to suggest that governments could create policies to grow the above crops in various climates to ensure that the world population meets their nutritional needs.