Data Mining for Business (BUDT758T)

Project Title: Smart Farming – Predicting Crop Yield to Enhance Agricultural Efficiency and Reducing Food Insecurity

Team Members: Chidubem Nwabunze

Madathil Geetanjali Menon

Jasmine Hill

Aishwarya Sadagopan

Sai Arjun Madikonda

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	Signature
Contact Author	Madathil Geetanjali Menon	Madathil Geentanjali
	Chidubem Nwabunze	Chidubem Nwabunze
	Jasmine Hill	Jasmine Hill
	Aishwarya Sadagopan	Aishwarya Sadagopan
	Sai Arjun Madikonda	Sai Arjun Madikonda

Executive Summary

Crop yield prediction is a crucial factor in ensuring the profitability of farmers. Predicting future crop yields can help farmers make informed decisions about planting and harvesting and optimize their crop yield. In this project, we used R to analyze historical data on crop yields and weather patterns to predict future yields. To best advise farmers, we focused on predicting the crop yields of the ten most consumed crops in the world. Doing so will make our predictions applicable to a larger population as these are produced in great quantities to fulfill worldwide demand.

Four models were used to make predictions and their RMSE values were calculated: linear regression (with an RMSE value of 46,238.99), regression tree (with an RMSE value of 84,181.4009), XGBM (with an RMSE of 47008.54), and GBM (with an RMSE of 27325.66). The results suggest that GBM had the lowest RMSE, followed by XGBM, linear regression, and regression tree, indicating that the gradient boosting method was the most accurate model for making predictions.

Data Description

Our dataset is sourced from Kaggle, which contains data from the FAO and the World Data Bank. We collected additional data from the FAO to add to the dataset. The columns included from the additional data added are potassium levels in soil (measured in tonnes), nitrogen levels in soil (measured in tonnes), phosphate levels in soil (measured in tonnes), and biomass burned in organic soils (measured in tonnes). All of these variables will be classified as numerical in our analysis. Our dependent variable is crop yield (measured in hg/ha) as our goal is to predict the crop yield based on agriculturally related variables. This variable is classified as numerical. The other independent variables in our dataset include the country the crop was grown in, the year the crop was grown, the type of crop grown, average rainfall (measured in millimeters per year), pesticide use (measured in tonnes), and average temperature.All of these variables will be treated as numerical except for country and crop, which are going to be classified as categorical/factor variables. Our dataset includes 26,092 data points (sample size *n*) and consists of 11 variables total, which has been previously mentioned.

Research Questions

As our goal is to predict crop yield, there are several questions that come to mind regarding this. We are primarily interested in knowing the factors that result in the highest crop yield. For example, which country (or set of countries) produces the crops with the highest yield? How many millimeters of rainfall contribute to a larger crop yield? How does pesticide use, nitrogen, phosphate and potassium levels affect crop yield? Does biomass burned on organic soils decrease crop yield? Does the biomass burned decrease crop yield alone or does it decrease it by also depleting the soil of necessary nutrients needed for crop growth (nitrogen, phosphate, potassium, etc.)? What combination of each of these variables contributes to high crop yield?

Methodology

Linear Regression

We decided to use linear regression as one of the methods because the algorithm is a straightforward and understandable one. It presumes that there is a linear relationship between crop production and the input factors (such as weather patterns and soil properties). It is simple, which makes it easy to comprehend and use.

The linearity of the connection between the input and output variables is taken for granted in linear regression. While the relationship may not always be approximately linear or capable of being translated linearly, this assumption can nonetheless offer some circumstances where reasonably good predictions can be made.

Clustering

Clustering techniques can be used to predict the yield of crops because they allow us to identify patterns and relationships among data points without necessarily knowing the outcome or target variable in advance. By clustering data on the basis of similarities or dissimilarities between data points, we can group crops with similar characteristics, such as similar weather conditions, soil types, fertilizer regimes, and crop varieties. These groupings can then be used to identify patterns and relationships that can be used to predict the yield of crops.

Regression Tree

We decided to use regression trees because our data contains many explanatory variables. Regression trees are easy to interpret and can handle both categorical and numerical variables, which is why we opted to use regression trees for further analysis. We also have a primary goal of predicting crop yield for each country and crop, so naturally, a regression tree would be useful as it will help determine yield quantity based on certain values of each variable included within the tree. Regression trees utilize subset selection by repeatedly partitioning the data in smaller subsets until a stopping criterion is reached based on the values of the predictor variables. At each of the tree's splits, we are given a variable and the value that has the most predictive power and continue down the tree until the subset is sufficiently homogeneous or the maximum tree depth is reached. At the end of the tree creation, we have a model that helps us predict the yield of a crop based on several predictor variables given.

Gradient Boosting and Extreme Gradient Boosting

Gradient Boosting and Extreme Gradient Boosting (XGBoost) are powerful machine learning techniques that have been successfully used in a wide range of applications, including predicting crop yields. These techniques are particularly well-suited for predicting crop yields due to several factors. Firstly, crop yield prediction is a complex task that involves multiple variables such as weather, soil conditions, fertilizer usage, and more. They are well-suited to handle these types of complex, non-linear relationships between variables. Secondly, crop yield data is often noisy and may contain outliers or missing values. Gradient boosting and XGBoost are robust to these types of data issues and can still make accurate predictions even when some data is missing or erroneous. Thirdly, crop yield prediction involves both categorical variables (such as crop type) and numerical variables (such as temperature and rainfall). The boosting approaches work well with both category and numerical data.

Results and Findings

Linear Regression

Linear regression is commonly employed for predicting area yield for different crops due to its suitability for modeling the linear relationship between independent variables (e.g., fertilizer usage, weather conditions, rainfall) and the dependent variable (crop yield). The assumption of linearity, interpretability of coefficients, simplicity of implementation, and scalability make linear regression an appealing choice in agricultural research. Additionally, the model's ability to provide accurate predictions, aid in variable selection, and offer data-driven insights contributes to its widespread usage in crop yield prediction.

As a part of linear regression, data was partitioned before and after 2004 into training and testing data respectively. The model was built based on independent variables that remained after removing the variables due to multicollinearity, i.e. biomass burned and pesticides (in tonnes). The model yielded an RMSE value of 37935.87 for the training dataset and 46238.99 for the test set.

Regression Tree

A regression tree analysis was used to predict the yield of crops based on several explanatory variables. The variables used in the regression tree were potassium, average temperature, and average rainfall (mm per year). The tree had six terminal nodes. The training RMSE was calculated to be 69,937.74, indicating that the model had moderate accuracy in predicting crop yield. This is especially true when considering the range of yield within the data. The minimum yield in the data set was 50 hg/ha, and the maximum yield was 501,412 hg/ha. An RMSE of 69,937.74 is relatively low when looking at its placement within this range. The test RMSE of the unpruned tree was found to be 83,607.50, which is slightly higher than the training RMSE. This suggests that the model has more prediction error when dealing with new, unfamiliar data (the test set), which is normal. The best-pruned tree was created using cross-validation. The test RMSE of the pruned tree was found to be 83,607.50, the same as that of the unpruned tree. This indicates that the pruning did not improve the accuracy of the model. We can also see that pruning did not improve the model because the regression tree looks exactly the same after pruning.

We used the pruned tree to predict the crop yield of farmers in the test data set based on the rule that crops are only bought from individuals with predicted yield greater than 0. The total crop yield in hg/ha from all farmers with predicted yield greater than 0 was found to be 414,535,951. Overall, the analysis suggests that the model may be useful in predicting crop yield to some degree, but there is room for improvement given the RMSE values and how pruning did not increase accuracy.

XGBoost

An XGBoost model was run to predict crop yield. The variables used for this model are country, crop, year, average rainfall (in mm per year), pesticides (in tonnes), biomass burned, nitrogen, phosphate, potassium, and average temperature. The original data was split into training and testing sets based on the year. Unique countries in the testing data were identified, and those countries not present in the training data were removed. The XGBoost parameters, such as the objective, evaluation metric, maximum depth, and learning rate, were set.

The RMSE on the training data was determined to be 32832.28 while the RMSE on the test data was found to be 47008.54. Overall, the train and test data accuracies suggest that there is a moderate level of accuracy.

Gradient Boosting Model

A gradient boosting model was run to predict crop yield. The variables used for this model are Country, Crop, Year, average rainfall (in mm per year), pesticides (in tonnes) , biomass burned, nitrogen, phosphate, potassium, and average temperature. The original data was split into training and testing sets based on the year.

The GBM model was trained using the training data, with a specified distribution of "Gaussian" and a total of 5000 boosting trees, interaction depth of 4, and shrinkage of 0.1. The RMSE on the training data was determined to be 10759.89 while the RMSE on test data was found to be 27325.66.

The absence of strong regularization in gradient boosting allowed it to effectively capture the complex patterns and nuances within our dataset, leading to superior performance compared to the more regularized xgboosting method.

Clustering Analysis

Clustering techniques can be used to predict the yield of crops because they allow us to identify patterns and relationships among data points without necessarily knowing the outcome or target variable in advance. By clustering data on the basis of similarities or dissimilarities between data points, we can group crops with similar characteristics, such as similar weather conditions, soil types, fertilizer regimes, and crop varieties. These groupings can then be used to identify patterns and relationships that can be used to predict the yield of crops. We considered two clustering algorithms: K-means Clustering and Hierarchical Clustering.

K-Means Clustering

The goal of our analysis was using K-means Clustering Algorithm to predict average crop yield of different clusters and to identify the cluster with the highest average crop yield as well as the cluster members. The dataset we used contained crop yield from 1990 to 2013 for different crops and countries as well as other predictors. There were ten different crops in our dataset for the clustering analysis. We decided to analyze Maize, Potato, Wheat, and Cassava only.

For each of the crops, using the elbow method, K = 4 seemed to be the optimal number of clusters. We were able to use K-means Clustering Algorithm to predict average crop yield of different clusters and to identify the cluster with the highest average crop yield as well as the cluster members.

Hierarchical Clustering

We also looked at the data from a hierarchical perspective, to see how the different countries will be classified into different clusters using specified distance metric. For this report, we will only report on the complete-linkage or maximum distance between the different clusters, although our R-script contains other distance metrics. The complete linkage distance metric calculates the dissimilarity between clusters as the maximum distance between any two points in the clusters.

For the various crops considered we scaled the data except the countries and crops. We then assigned the country names to the row names so that it could appear as the labels in the dendrogram. We generated a plot of four clusters to go along what we had earlier in the K-means.

For the four dendrograms generated, we noticed that India and Brazil are in different clusters. Also from the plot we see that India and Brazil are at a greater distance when compared to the other two clusters. This is interesting since in using the K-means algorithm, India and Brazil were always in the same clusters. The reason for this difference might be because of the Algorithm differences between K-means and Hierarchical clustering. While K-means aims to minimize the sum of squared distances between data points and the cluster centroids, hierarchical builds a hierarchy of clusters by either merging or splitting based on a distance measure.

Also, the difference may be attributed to sensitivity to initial conditions. While K-means clustering is sensitive to initial conditions meaning that different starting positions for the centroids can result in different cluster assignments, Hierarchical clustering is not affected by this. We were able to use K-means Clustering Algorithm to predict average crop yield of different clusters and to identify the cluster with the highest average crop yield as well as the cluster members. We also implemented hierarchical clustering to view dendrograms and get a better picture. From the previously discussed results, we can clearly state that K-means clustering in R is a good decision for predicting crop yields. Making up such an analysis using previous data can be very useful especially for farmers that would like to implement precision in their crop yields' production.

This work thus helps to differentiate between countries and the crops that yield the most in the different countries. We also checked the internet to confirm that the clustering algorithm worked and can report that many of the countries in our preferred clusters belonged to the world top producing countries of the respective crops.

Conclusion

The goal of this project was to develop a predictive model that can help farmers make informed decisions about planting and harvesting by analyzing historical data on crop yields and weather patterns. To achieve this, four different models were used to make predictions, including linear regression, regression tree, XGBoost (XGBM), and Gradient Boosting Method (GBM). After building the models and calculating their Root Mean Squared Error (RMSE), it was found that GBM had the lowest RMSE of 27,325.66, followed by XGBM with an RMSE of 47,008.54. Linear regression had an RMSE of 46,238.99, and regression tree had the highest RMSE of 84,181.4009.

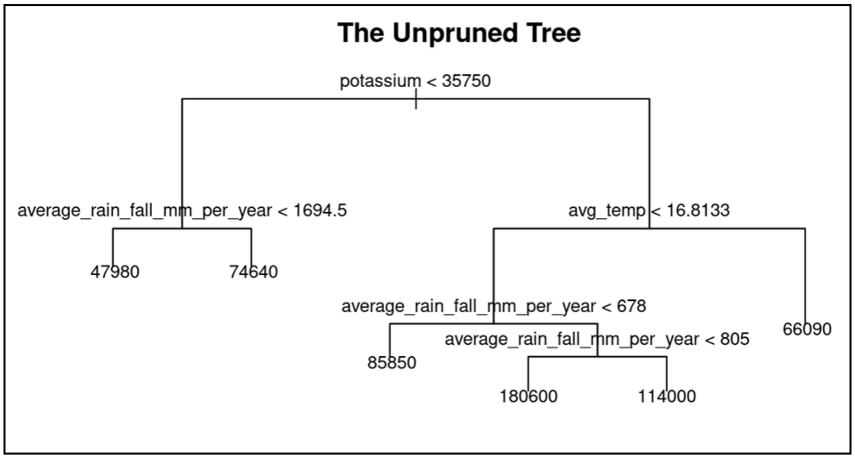
RMSE is a commonly used metric for evaluating the accuracy of regression models. The lower the RMSE value, the better the model's predictive performance. In this case, the results suggest that the GBM model performed the best and was the most accurate model for predicting crop yields based on the given data.

Gradient Boosting is a machine learning algorithm that builds multiple decision trees iteratively, with each tree attempting to correct the errors of the previous tree. This method is effective in reducing bias and increasing the predictive power of the model. In comparison, XGBM is a variant of GBM that uses additional regularization and randomization techniques to further improve the model's accuracy.

In summary, the GBM model was found to be the most accurate for predicting crop yields based on the given historical data and weather patterns. Farmers can use this model to make informed decisions about planting and harvesting and increase their profitability.

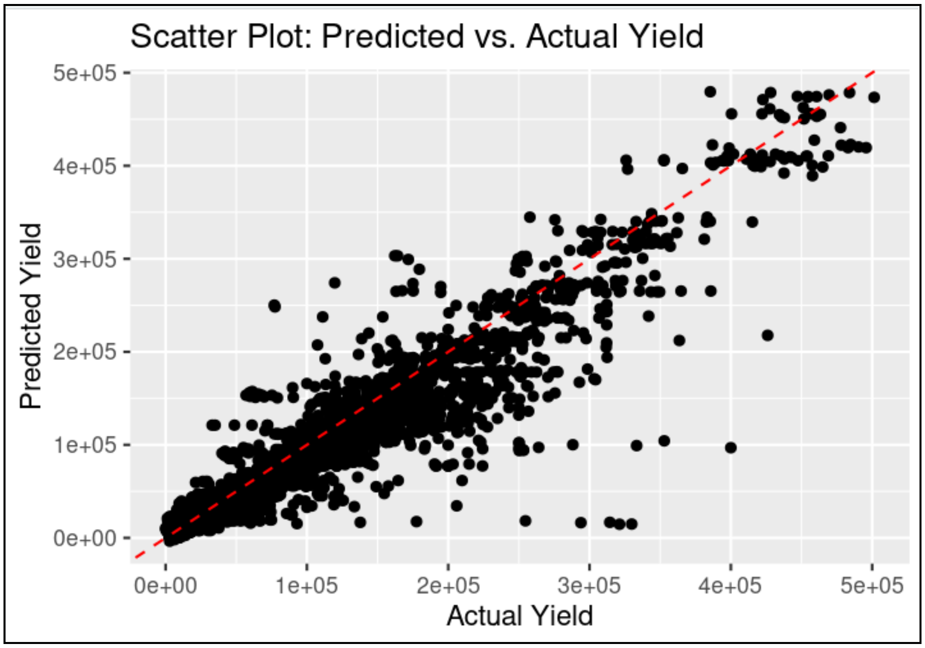
Appendix

Regression Tree: Pruned/Unpruned Tree

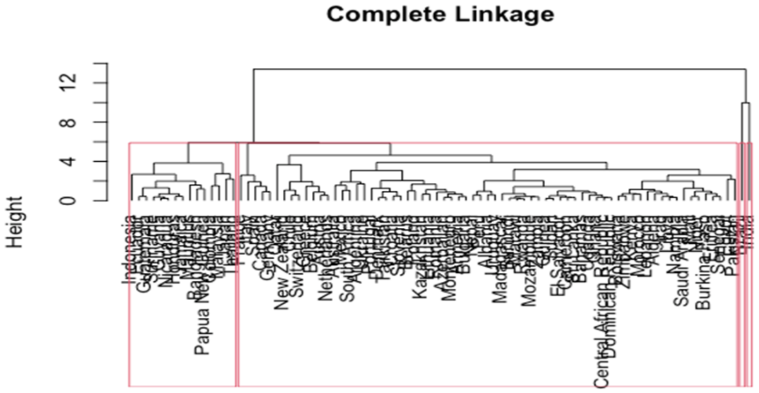

The Unpruned Tree
potassium < 35750
average_rain_fall_mm_per_year < 1694.5
avg_temp < 16.8133
47980
74640
average_rain_fall_mm_per_year < 678
average_rain_fall_mm_per_year < 805
66090
85850
180600
114000

The chart above shows the unpruned tree (which is the same as the best pruned tree). The variables used include potassium, average temperature, and average rainfall.

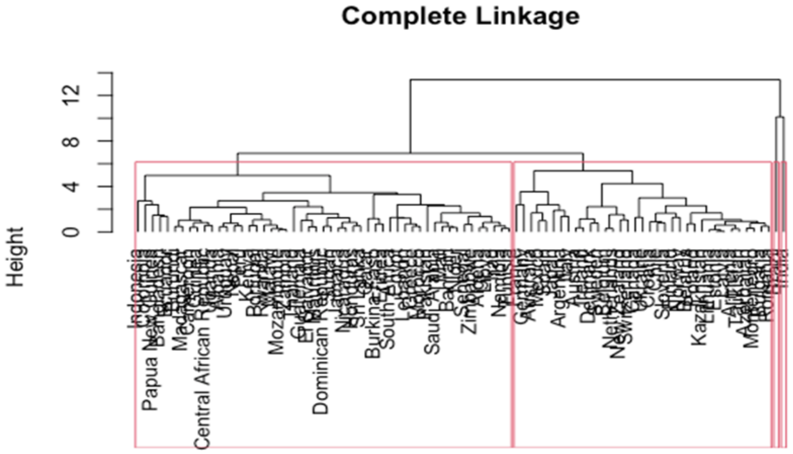
XGB Scatterplot



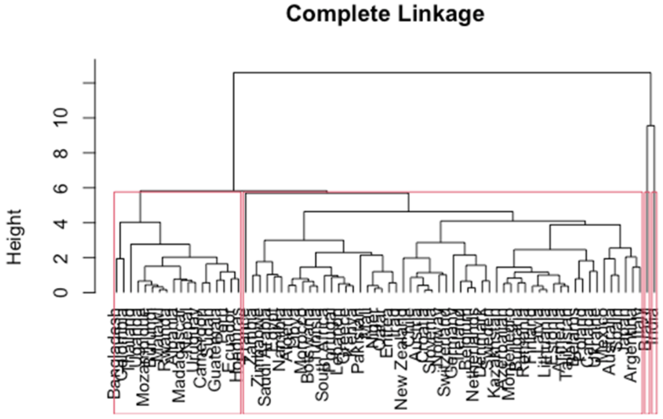
Complete Linkage Plot for Maize



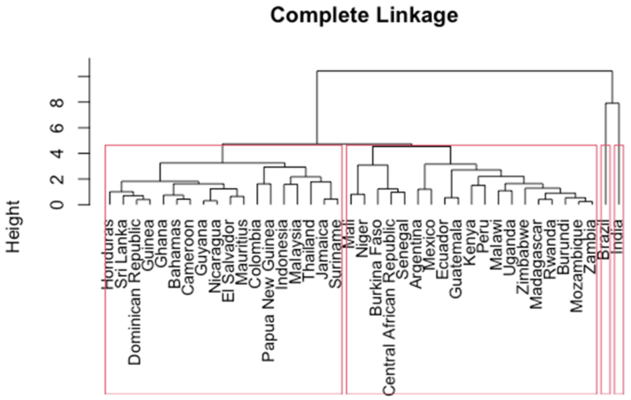
Complete Linkage Plot for Potato



Complete Linkage Plot for Wheat



Complete Linkage Plot for Cassava



K-Means Clustering Results of Cluster With Highest Average Yield

Crop	Cluster No.	Cluster Members	Average Yield
Maize	Cluster 3	Argentina, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bulgaria, Canada, Chile, Croatia, Denmark, France, Germany, Italy, Japan, Kazakhstan, Lithuania, Mexico, Montenegro, Netherlands, New Zealand, Poland, Romania, Slovenia, Spain, Switzerland, Tajikistan, Ukraine.	21308.21 hg/ha
Potato	Cluster 3	Argentina, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bulgaria, Canada, Chile, Croatia, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Japan, Kazakhstan, Latvia, Lithuania, Mexico, Montenegro, Netherlands, Norway, Poland, Portugal, Romania, Slovenia, Spain, Sweden, Switzerland, Tajikistan, Ukraine.	93151.62 hg/ha
Wheat	Cluster 3	Armenia, Austria, Azerbaijan, Belarus, Belgium, Bulgaria, Canada, Chile, Croatia, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Kazakhstan, Latvia, Lithuania, Montenegro, Netherlands, Norway, Poland, Romania, Slovenia, Spain, Sweden, Switzerland, Tajikistan, Ukraine.	16848.52 hg/ha
Cassava	Cluster 1	Bahamas, Burkina Faso, Cameroon, Central African Republic, Dominican Republic, El Salvador, Ghana, Guinea, Mali, Niger, Senegal, Sri Lanka, Thailand, Uganda.	35816.37 hg/ha