**Impact of Maize Harvest on**

**Undernourishment in Africa**

Assignment 6

Final Report

April 24, 2021

By: Lanya Xiang

Data Analytics

ITWS 6600

Professor Thilanka Munasinghe

**Abstract & Introduction**

Hunger is one of the most significant hindrances to poverty reduction and global development around the world. 1 in 9 people globally is currently undernourished. Of these 795 million, 98% live in developing countries (“Hunger in Developing Countries”). The UN declared Hunger to be #2 in its list of sustainable development goals.

The project focuses on linking food insecurity in Africa to crop yield trends for maize. I chose this topic because of my interest in the initiative to end world hunger. I focused my project on Africa since it is currently the poorest continent on Earth. For the purpose of the project, I chose to concentrate my analysis using one of Africa’s top produced crops— maize. This is a highly-popular crop that is often produced and consumed within the continent, making it a suitable choice for my project. The food security index I picked for the project is prevalence of undernourishment, which is given as a percentage for how likely that a randomly selected person from a country is malnourished. The goal of the project is to analyze this index against maize harvest trends to check if a relationship does actually exist. Therefore, the null hypothesis for the project is: there is no correlation between food insecurity and maize harvest yields in Africa.

**Data Selection Process**

When choosing the data for this project, I started off by browsing the UN database for relevant data on my topic. I found the FAOSTAT database, from which I sourced the [Maize](https://www.fao.org/faostat/en/#data/QCL) (Area Harvested) and [Undernourishment](https://www.fao.org/faostat/en/#data/FS) (3-Year Average) datasets for Africa. Both datasets had metrics that were constantly documented each year, which meant that the information would be constant and highly available so I would not have to worry about having issues with missing or inconsistent data. These datasets were also organized according to country code and year, which is excellent because I could easily merge the two sets together using these two common columns.

The Maize dataset is composed of 6 columns: the FAO area code, the area name, the Item (crop) code, the name of the crop, the year, and the area harvested in hectares (ha).

Table

Description automatically generated

Figure 1: Maize Dataset Snippet

The Undernourishment dataset is composed of 5 columns: the FAO area code, the area name, the year code, the year, and the value of the prevalence of undernourishment.

Table

Description automatically generated

Figure 2: Undernourishment Dataset Snippet

**Analysis**

**General observations:**

Overall, the quality of the original datasets was decent. I did not have to perform much cleaning other than using na.omit() and no real issues occurred during the importation of the data into RStudio.

However, some points of consideration for the project include sources of error, uncertainty, and bias for the datasets used in the project. Since this data comes from a single organization, it may have some issues in terms of the collection method and verification of the data. There could be missing information for certain countries which would impact the predictive ability of the models I create. For example, countries could be missing either maize harvest data or undernourishment data, or both, which would cause them to not be included in the model. There could also be differences in the interpretation of “Area Harvested”. For example, one country could interpret it as all harvest from the field, while another could interpret it as total usable harvest. Bias-wise, there could be more records of data for countries that are easier to gather data from. For example, rural countries like Niger and Rwanda could have less data compared to more technologically advanced countries like South Africa and Egypt. This would lead to an imbalance of records between the countries, which would also impact the models.

**Creating classification labels for the undernourishment dataset:**

For the purpose of the project, I needed to label the undernourishment values with classes.

To identify the best separation values, I first created a boxplot to look at the distribution of values.

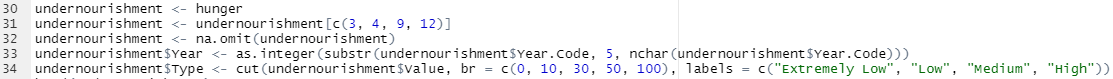
Chart, box and whisker chart

Description automatically generated

Figure 3: Boxplot for Undernourishment Prevalence

From the boxplot, I chose the 25% (10) and 75% (30) percentiles to be the main split. Since there were very little values after 60, I chose 50 as the value for the top class.

As a result, I created 4 classes (“Extremely Low”, “Low”, “Medium”, and “High”) to use in my classification models. Using cut(), I added a new column called “Type” to store the labels for the datapoints.

****

**Creating the joined dataset:**

After identification of the two datasets I would be using for the project, I aggregated them together using a left join on the Area and the Year.

****

**Graphical user interface

Description automatically generated with medium confidence**

Figure 4: Combined Dataset Snippet

**Joined dataset distribution:**

I ran a summary on the newly combined dataset to take a look at the distribution of the variables I would be working with.

Text, letter

Description automatically generated

The data’s year ranges from 2002 to 2020. Area harvested varies heavily from 6 to 7822149 hectares. Undernourishment prevalence also has a large range from 2.5% to 81.7%. With the classification function that I ran, most of the datapoints are within the “Low” category, with a roughly even amount in “Extremely Low” and “Medium”, and very little in “High”. The points in “High” are likely to be from the same country.

**Visualization of the countries:**

To ensure that most of the countries had a balanced amount of data, I created a histogram to check the distribution. Fortunately, most of the countries except for one had equal amounts of data, so there is no issue with an imbalance in the data due to missing information from year to year.

****

Chart, bar chart

Description automatically generated

Figure 5: Histogram for Area Code

**Visualization of the maize dataset:**

I created a bar graph of the maize dataset to look at the distribution of the amount of maize that each country in the dataset harvested in total.

Text

Description automatically generated

Nigeria was by far the top producer, with the rest of the countries hovering around the middle.

Some countries such as Algeria, Comoros, and Mauritius did not have data for maize harvest, so they were excluded from the study.

A picture containing chart

Description automatically generated

Figure 6: Maize Production in Africa

**Model 1: Multivariate Regression**

**Model Rationale:**

I chose Multivariate Regression as my first model. This method uses regression to calculate the dependent value based on several independent variables. For this project, my multivariate regression model will aim to predict the undernourishment prevalence value based on country code, area harvested, and year.

**Creating the initial model:**

I used a standard multivariate regression function to create the model. I used the three independent variables with the goal to calculate the undernourishment value.



**Evaluating the initial model:**

The summary for the multivariate regression model shows that the p-value is 3.334e-06, which is highly significant. However, the multiple and adjusted R-squared values are still extremely low. This means that the model would only predict only approximately 3-4% of the data.

**Text

Description automatically generated**

As shown in the Added-Variable plot for the regression model below, there is a clear trend of significance for each of the 3 variables in relation to the independent variable.



Graphical user interface, application

Description automatically generated

Figure 7: Added Variable Plots for Multivariate Regression Model

**Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated**As shown by the correlation tests for the 3 independent variables, they all have some significant relation to the dependent variable. Therefore, I decided to keep all of them since there are no other better variables for the model.

**Performing Cook’s Distance:**

Since the model analysis showed that all 3 variables are significant to the independent variable, another method of improving the accuracy of the model would be to remove any outliers from the dataset.

I performed Cook’s Distance and identified a series of outliers in the data that should be removed. As shown in the figure below, the spikes represent the outliers according to Cook’s Distance.

A picture containing text

Description automatically generated

Chart, histogram

Description automatically generated

Figure 8: Cook's Distance for Multivariate Regression Model

Points were considered influential if they exceeded 3 times the mean of the Cook’s Distance model.

Graphical user interface, application

Description automatically generated

I went through the dataset and obtained the information for the outliers. The output is the row number for the point that is considered an outlier.

Table

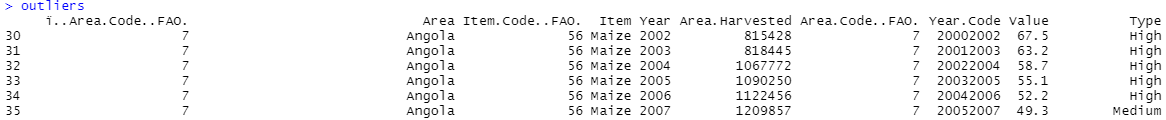
Description automatically generated

After identifying the outlier points, the next step is to store them in a dataframe, and then remove the elements in the dataframe from the project dataset using anti\_join().

Text

Description automatically generated

A total of 50 outliers were removed from the dataset as a result of Cook’s Distance. Below are some of the outliers that were removed.



**Recreating the model using clean data:**

I reran the same multivariate regression model once again but used the clean data this time. Compared to the old multiple and adjusted R-squared value, this run had a slightly better R-squared value. However, it was not a huge improvement.

Text

Description automatically generated

Looking at the Added-Variable plots, there is not much difference for the area code and area harvested variables. However, year has been slightly improved since its regression line shows a steeper slope.

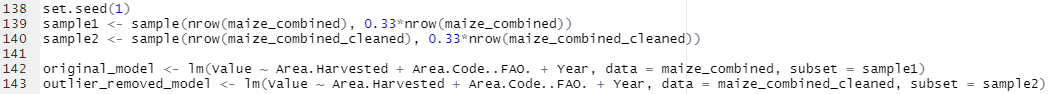
Graphical user interface

Description automatically generated

Figure 9: Added Variable Plots for Cleaned Multivariate Regression Model

**Performing mean-squared analysis:**

To evaluate the quality of the models, I chose to perform a mean-squared analysis. To set up, I first created a sample from the original and cleaned datasets. Then, I created the representative models using the samples as the subset.



Next, I created the mean square test for both of the models.



The results showed a mean squared value of 197.8 for the original model and a mean squared value of 112.0 for the cleaned model. Therefore, the removal of the 50 outliers allowed the MSE to drop greatly (by 85.8). This is excellent because the lower the MSE is, the more accurate the model is.

****

**Model 2: K-Means Clustering**

**Model Rationale:**

I chose K-Means Clustering (Classification) as my second model. This unsupervised method is an interactive algorithm that aims to sort data into k number of clusters by picking a centroid and then shifting until no change in the results in detected. The goal for k-means clustering is to create a model that is able to cluster the data into easily differentiable clusters (Dabbura). For this project, my k-means model will aim to cluster the data into clusters according to country, area harvested per year, and undernourishment.

**Preparing the model:**

To start, I first used the elbow method to calculate the ideal number of clusters for my model. I calculated the total within-cluster sum of squares using the 4 chosen independent variables (area code, year, area harvested, and value). Then, I plotted the results to create an elbow chart. Here, I identified the ideal number of clusters as 4 since the “bend” in the elbow appears to be on the 4th datapoint.



Chart, histogram

Description automatically generated

Figure 10: Elbow Method for K-Means Clustering

**Creating the model:**

After obtaining the ideal number of clusters, I created the k-means model.



The table below shows the positions of the datapoints within the 4 clusters as determined by the model.

Calendar

Description automatically generated

**Evaluating the model:**

There are 4 clear clusters on the model. Cluster 4 (purple) is the highest cluster with the most area harvested and a very low undernourishment value. Cluster 1 (red) is the higher area harvested with a low to medium undernourishment value. Cluster 3 (blue) has a medium harvest area with a low to medium undernourishment value. It slightly overlaps with cluster 1 towards the top. Cluster 2 (green) is the lowest and most densely populated cluster. This cluster consists of datapoints with very low area harvested and includes a range from low to high undernourishment values. Both clusters 2 and 3 have some outliers that are on the outside of the ellipses.



Chart, scatter chart

Description automatically generated

Figure 11: K-Means Clustering Model

When looking at the clusters plotted using area names and value, this model is able to assign all the datapoints for each country into at most two neighboring clusters, which means that it is relatively accurate at grouping each country into its respective areas.

****

A picture containing graphical user interface

Description automatically generated

Figure 12: Clustering Visualized by Area

To further analyze accuracy, I used a silhouette plot to validate the model. Silhouette analysis measures how well an observation is clustered as well as the average distance between clusters. In silhouette analysis, observations with a large silhouette Si (almost 1) are very well clustered. A small Si (around 0) means that the observation lies between two clusters, and observations with a negative Si are probably placed in the wrong cluster (“fviz\_silhouette”). For the 4 clusters in my model, they all have a high Si level (average 0.84), which means they are well clustered.

Text

Description automatically generated with medium confidence

**Chart

Description automatically generated**

Figure 13: Silhouette Analysis for K-Means Clustering Model

Overall, the K-Means Clustering model was highly accurate. It was able to sort the datapoints into 4 distinct clusters with unique characteristics. As shown in Figure 12, most of the well-off African countries such as Nigeria, South Africa, Kenya, and Ethiopia belonged in the same clusters while poorer and smaller countries such as Somalia and Sierra Leone were clustered together.

**Model 3: KNN Clustering**

**Model Rationale:**

I chose KNN Clustering (Classification) as my third model. This supervised method aims to sort data into pre-determined classes. The goal for knn clustering is to create a model that is able to correctly predict the target class of the datapoints given its features (the independent variables). For this project, my knn model will aim to predict the undernourishment class of each datapoint given the country, area harvested, and year.

**Preparing the model:**

I created the dataframe for the model to include the 4 variables for my model. Then, I ran an na.omit() function to make sure that there were no NAs in my data.



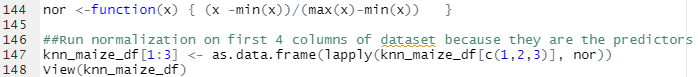
Table

Description automatically generated

Figure 14: KNN Dataframe

**Normalizing the dataframe:**

Since the data in the dataframe have relatively large ranges, I had to normalize the data in order to ensure that the KNN function would be able to run correctly.



Table

Description automatically generated

Figure 15: KNN Dataframe After Normalization

**Creating testing and training sets:**

Following the normalization, the next step was to create testing and training sets for the model. I used the general process of creating a sample and then separating them into training and testing sets with a 70/30 ratio.



**Creating the model:**

Before creating the model, I first had to calculate the k value for the knn function. This was done by finding the total number of rows and then square rooting it. Since there was a total of 511 rows in the dataframe, the k value was 23. I plugged this number into the function and created the KNN model.

A picture containing text

Description automatically generated

**Evaluating the model:**

To evaluate the model, I created a confusion matrix to compare the results of the prediction.

A picture containing text

Description automatically generated

Table

Description automatically generated

As seen in the matrix, the model was able to correctly predict a large majority of the datapoints correctly. For the “Extremely Low” class, it correctly predicted 45/46 of the results (97.8% accuracy). For the “Low” class, it correctly predicted 45/46 of the results (97.8% accuracy). For the “Medium” class, it correctly predicted 44/52 of the results (84.6% accuracy). For the “High” class, it correctly predicted 1/1 of the results (100% accuracy). With a combined accuracy rate of 85.5%, the knn clustering model was highly accurate at predicting the level of undernourishment prevalence that the particular country had.

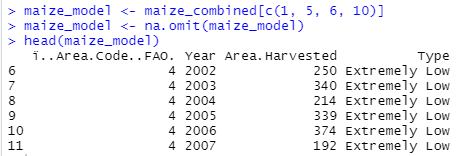
**Model 4: Random Forest**

**Model Rationale:**

I chose Random Forest (Classification) as my fourth and final model. Like knn clustering, this supervised method aims to sort data into pre-determined classes based on a labeled sample set. The goal for random forest is to create a model that is able to correctly predict the target class of the datapoints given its features (the independent variables) through averaging the results of many decision trees. For this project, my random forest model will aim to predict the undernourishment class of each datapoint given the country, area harvested, and year.

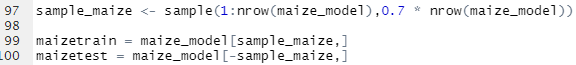
**Creating the dataframe:**

I created the model data frame and included the variables for the area code, year, area harvested, and type. I used the na.omit() function to clear any NAs out of the dataframe.

****

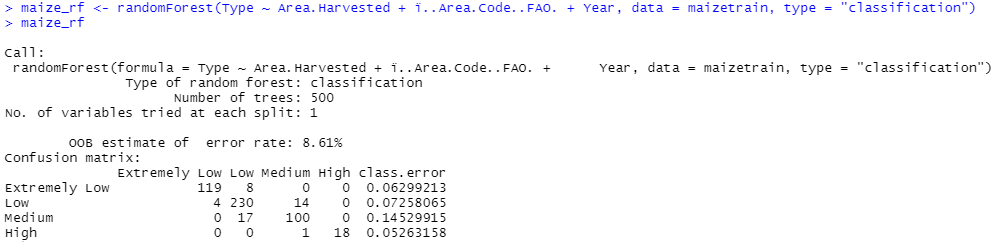
**Setting up the testing and training sets:**

Next, I created the training and testing sets using a 70/30 ratio.

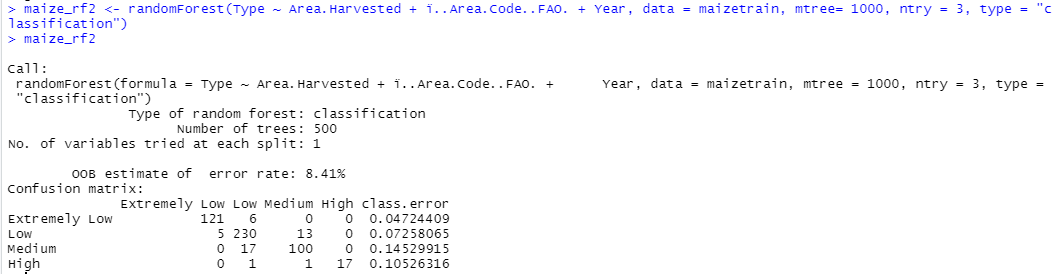


**Creating the random forest model:**

I created the random forest model using the 3 variables. I used the “classification” technique, which uses the variables to predict what the range of the value of undernourishment would be.

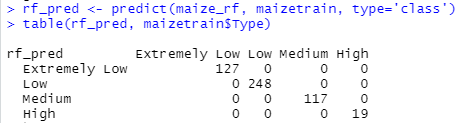


I created a second tree with different ntree and ntry values. When creating more trees and using a ntry of 3, I was able to minimally decrease the error rate from 8.61% to 8.41%.

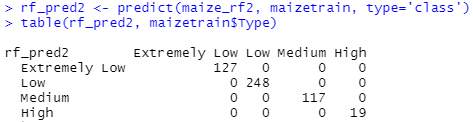


**Analysis of the random forest model:**

When testing the predictive ability of the first random forest model, the model was extremely accurate. The confusion matrix showed no error at all. The output of the random forest model showed a OOB estimate of 8.61%, which is excellent. All 4 of the categories have an extremely low class.error rate, which means that the model is equally good at predicting each of the different categories.



For the second model, it was also able to correctly predict the type of undernourishment for all of the datapoints. The error rate for this model was slightly lower than that of the first random forest, but the difference is extremely miniscule so both models can be considered to have performed excellently.



**Conclusion**

Data analytics is a powerful tool for solving global issues such as hunger. With the ability to interpret and process large amounts of data, we can tackle large scale issues using informed methods. By creating models and algorithms to interpret the data, we can quickly gain insight on the most important problem areas while planning ahead.

With this project, I was able to overturn my null hypothesis by proving that maize yields do in fact have an impact on the undernourishment prevalence levels in African countries. To visualize this relationship, I created a total of 4 models: multivariate regression, k-means clustering, knn clustering, and random forest. I chose a wide spread of models in order to find the one that best suited the project. Both the clustering models and the random forest proved to be more relevant and accurate compared to the multivariate regression. At their current state, the clustering models are able to correctly predict the undernourishment levels with reasonable accuracy. The random forest model is the best of the 4 models, with a roughly 92% accuracy rate.

When performing the research for the project, I initially wanted to use more than one crop in order to get a better idea of how crop harvest in general would affect undernourishment prevalence. I had to change my approach since many of the other crops I chose were lacking data. Since maize had the most thorough distribution of data, I ended up choosing it over all the other crops. Additionally, I had also wanted to include drought data as another point of reference for the crop yield. Since droughts would likely impact crop harvested, I wanted to see if it would also affect undernourishment prevalence. However, given the time constraint, I was unable to integrate this aspect of the situation into the project. Given more time, I would definitely also include this dataset into the models.

As for next steps, I would start by integrating the datasets for droughts and precipitation in order to better understand the relationship between crop yield and undernourishment. I would also look into lasso regression to see if it is better than the standard multivariate regression I performed. I would also do more analysis with the clustering such as analyzing using rand() and comparing it to another clustering technique such as meanshift.

Overall, I successfully proved the relationship between maize yield and undernourishment prevalence. Throughout the process, I gained a lot of basic R skills and familiarity with several different packages.

**Citations**

Dabbura, I. (2020, August 10). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Medium. <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>.

*fviz\_silhouette: Visualize Silhouette Information from Clustering*. (2020, April 2). rdrr.io. <https://rdrr.io/cran/factoextra/man/fviz_silhouette.html>.

Harrison, O. (2019, July 14). *Machine Learning Basics with the K-nearest Neighbors Algorithm*. Medium. Retrieved November 11, 2021, from <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>.

*Hunger in Developing Countries: Five Facts You Need to Know*. (2016, August 31). Borgenproject. <https://borgenproject.org/hunger-in-developing-countries-five-facts/>.

Yiu, T. (2021, September 29). *Understanding Random Forest*. Medium. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.