STEM Fellowship | 2022 REO Blueprint Competition

Samantha B. Chong[1] and Jason Thai[2]

[1]University of Waterloo
[2]University of Ontario Institute of Technology

August 13, 2022

Contents

1 Introduction 1

2 Background 1

3 Design Question 2

4 Design 2

5 Materials 6

6 Methodology 6

7 Discussion 7

8 Conclusion 7

Abstract

As the effects of climate change continue to evolve, their effect on our food systems grow. Statistical models may be a solution for this growing concern. Using statistical models created in Python using Pandas and Selenium, predictions can be about crop yield or the best crop to grow can be made. The specific variable's considered within this proposal are weather, area of the field, pesticide use and strain of crop. Using a proposed models an app or web-based program could be created for use around the world. As technology evolves, it must be utilized to optimize food production to feed our growing population.

1 Introduction

Everything humans do affects the delicate balance of the climate around us. In turn, small shifts in the environment have the potential to cause significant after-effects within society and the ecosystem. Although some effects of climate change appear inconsequential, they can have significant consequences. This proposal will discuss the effects of climate change on food security and propose developing a model to determine the optimal crops for farmers to grow based on various variables.

2 Background

At first glance, the connection between climate change and food insecurity seems relatively obscure; however, the closer you look, the clearer the image becomes. Some of the effects of climate change directly affect the crops. In contrast, others affect the food system.[1] Decreased rainfall impacts the plant's ability to grow. In contrast, flooding could affect the system's ability to deliver food before it spoils. In the short term, the effects are primarily seen in 3rd world countries where famine is already a prevalent issue. As a result, they are more susceptible to changes in the food

chain.[2] In countries such as Canada, the results of these issues manifest themselves as increased prices, once again most affecting the poorest of the population.

The main issue that climate change presents in terms of the agriculture industry is in terms of the variability it presents. This increased variability presents challenges as it may either shorten or lengthen growing seasons with little to no forewarning. In the future, it is thought that climate change may affect the ability of the land to grow crops to the efficiency it once did.[2]

For this proposal, the effects of pesticides must also be examined alongside weather conditions and historical yield. Data on weather conditions, pesticides and historical crop yield are important for making decisions related to agriculture.[3]

One of the important aspects of the model deals with data regarding the protection genetic modification and pesticides can provide. These are known as genetically modified traits. The International Service for the Acquisition of Agri-biotech Applications (ISAAA) provides a database of countries that have been approved based on their modifications. It features crops approved for commercialization, plant processing, and importation for consumption.[4]

3 Design Question

Can a statistical model use past information to suggest the version of a crop most optimal to be grown, based on varying factors such as temperature and precipitation?

4 Design

In order to determine the best crop for food yields, there are several attributing factors: weather, area of the field, pesticide use and specific crop and strain of crop growth. In addition, it is essential to consider the types of pesticides used and the type of modified genome each crop has. Considering these factors, the model can recommend the optimal genetic modifications and pesticides to use based on the weather input. This project uses the formal programming language Python due to its wide use and support in data analysis and machine learning models. Within Python, a data analysis tool "Pandas"[5] will be utilized.

First, extract the downloadable content from Kaggle and perform data cleansing shown in Figure 1. Plotting the data shows which crops are most affected based on these attributes shown in Figure 2. This graph will be useful for considering how the model will decide the optimal crop based on the input of average temperature and pesticides. Another plot in Figure 3 shows how an increase in demand and population affects these attributes over the years. This can help the model determine what to suggest based on the demand throughout the years.

Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
Albania	Maize	1990	36613	1485.0	121.0	16.37
Albania	Potatoes	1990	66667	1485.0	121.0	16.37
Albania	Rice, paddy	1990	23333	1485.0	121.0	16.37
Albania	Sorghum	1990	12500	1485.0	121.0	16.37
Albania	Soybeans	1990	7000	1485.0	121.0	16.37

Figure 1: Sample of the Data Retrieved from Kaggle

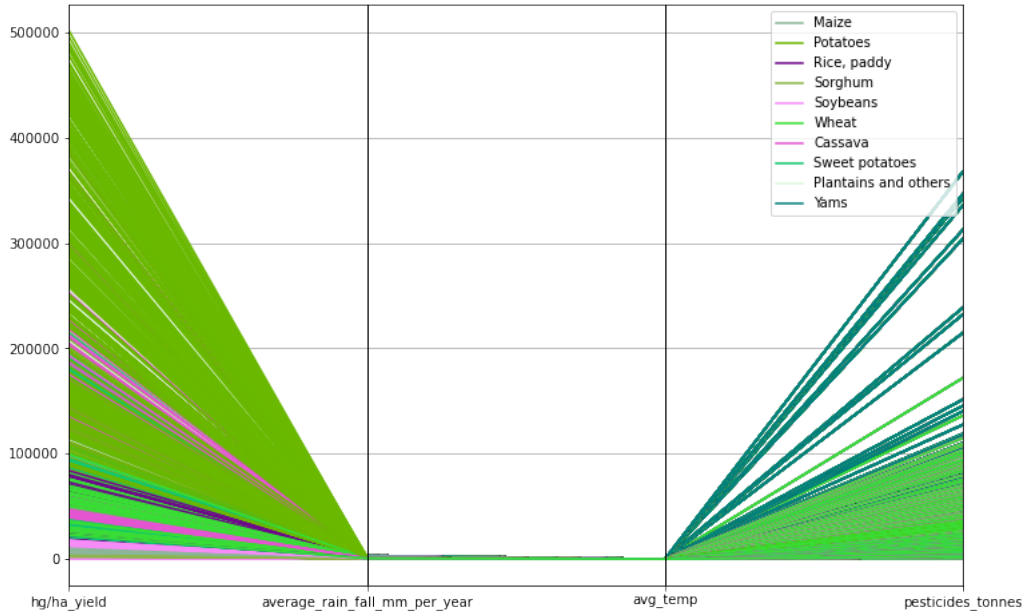

Figure 2: Parallel Coordinate Plot based on what Crop uses Based on Hectares, Rainfall in Millimeters, Average Temperature, and Pesticides in Tonnes

Afterwards, search for data relating to genetic modification and pesticides introduced to each crop based on the kinds of protection it offers, known as "Genetic Modified Traits". The "International Service for the Acquisition of Agri-biotech Applications" (ISAAA) provides a database (GM Approval Database) that lists countries that have been approved based on their modifications. Featuring crops approved for commercialization, plant processing, and importation for consumption and feed.[4] To utilize this information, lets introduce "Selenium"[6] which is a browser automation tool. Start by first retrieving the website link (URL) from ISAAA and begin the automation, scrape the URL and put it into a separate dataset shown in Figure 4.

The labels table of Figure 4 represents the plant name following the scientific name. Below are details relating to the plant: the genome name (Event Name) and a unique code identifier of the crop (Code). The next column tells the company that developed the genomes. Each event name in the database links to broader information about the specific traits and modifications of the crop. Therefore, we propose a proof of concept from then on.

In Figure 5. Following a basis for a flowchart diagram and a machine learning model, there is always a start and stop for any specified process in-between to show the system's main objective. From the previous design approach, steps a) and b) have been addressed, although a portion of b) has not. Detailing merging of the datasets is to demonstrate extracting common features of types of genetic modification and pesticides from ISAAA and "Crop Yield Dataset" and putting it into a single dataset; that way it is convenient to lookup information on. In addition to come up with the desired system that fits the proposed inputs and outputs will be a multi-classification model (categorization of more than two classes). Referring to Figure 4, the plant "Alfalfa - Medicago sativa" has several genetic modifications. Using this information allows the model to deduce which crop to recommend. Once the data has been assembled, the next steps c) and d) is to train and test the

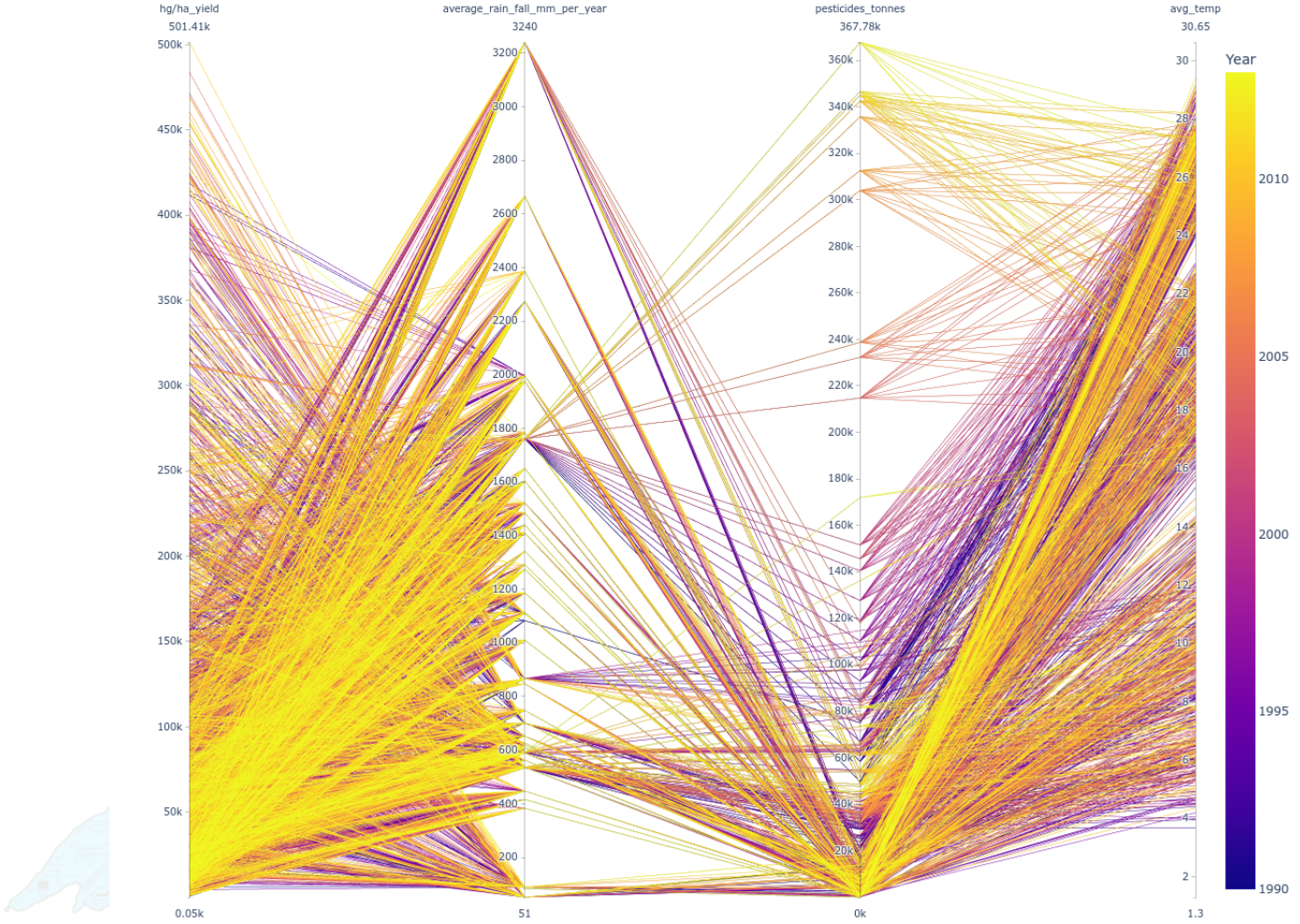

Figure 3: Crop Attributes Based on Year

Event Name and Code	Trade Name
Alfalfa - Medicago sativa :	Alfalfa - Medicago sativa :
Name: J101 Code: MON-ØØ1Ø1-8	Roundup Ready™ Alfalfa
Name: J101 x J163 Code: MON-ØØ1Ø1-8 x MON-ØØ163-7	Roundup Ready™ Alfalfa
Name: J163 Code: MON-ØØ163-7	Roundup Ready™ Alfalfa
Name: KK179 Code: MON-ØØ179-5	HarvXtra™

Figure 4: Sample of the Data Retrieved from ISAAA's GMO Approval Database

model. Simply validating the data for false positives and true negatives (over-fitting and under-fitting). The standard rule for dividing the data is 80% for training and 20% for testing. There are many algorithms to test and evaluate d). In this case, however, a random forest decision tree would be sufficient to determine the following attributes starting as a concept. Further explanation is shown in Figure 6. The last step f) saves the model for later use.

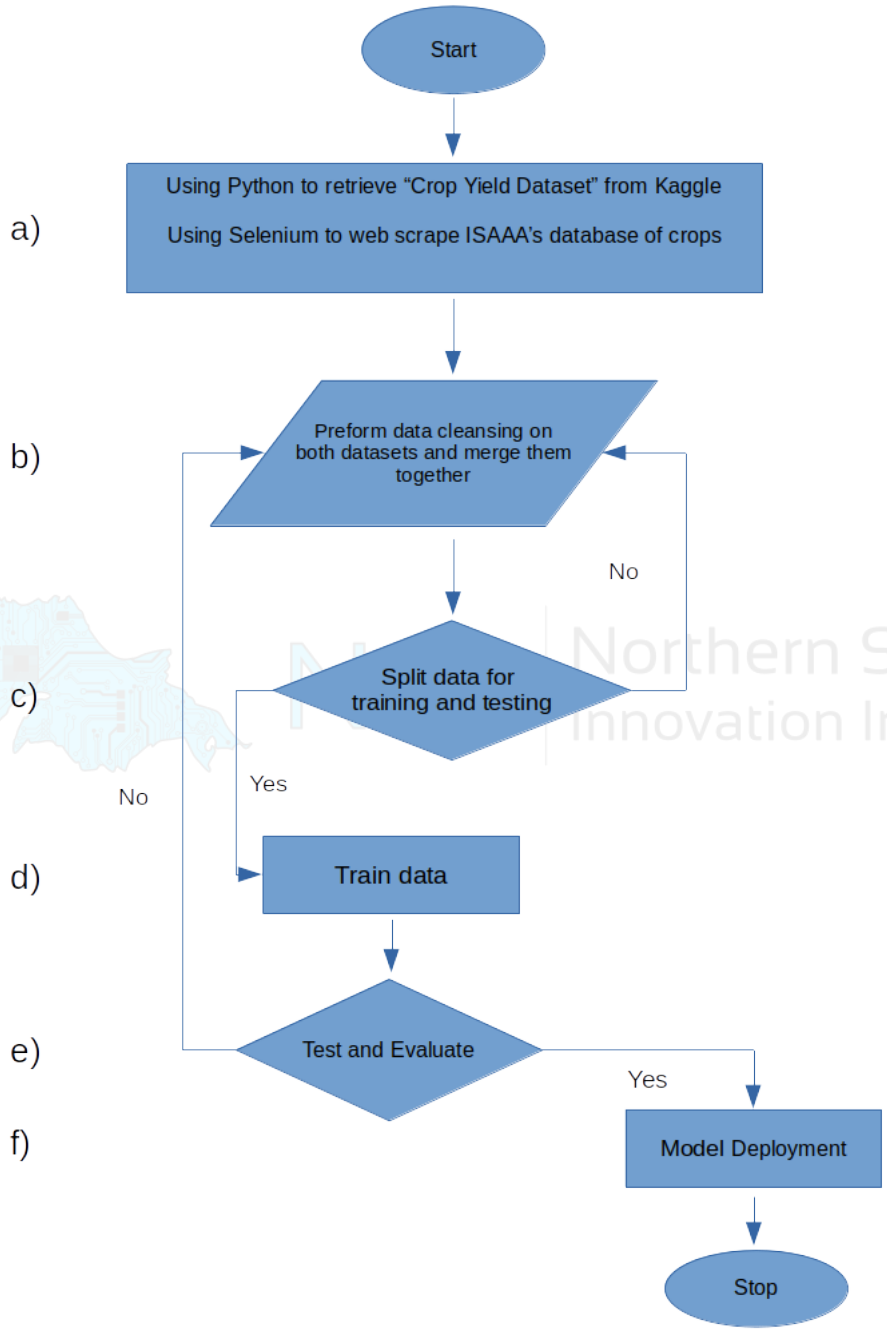

Figure 5: Flowchart of the Steps Labelled Including Additional Steps of Developing the Model

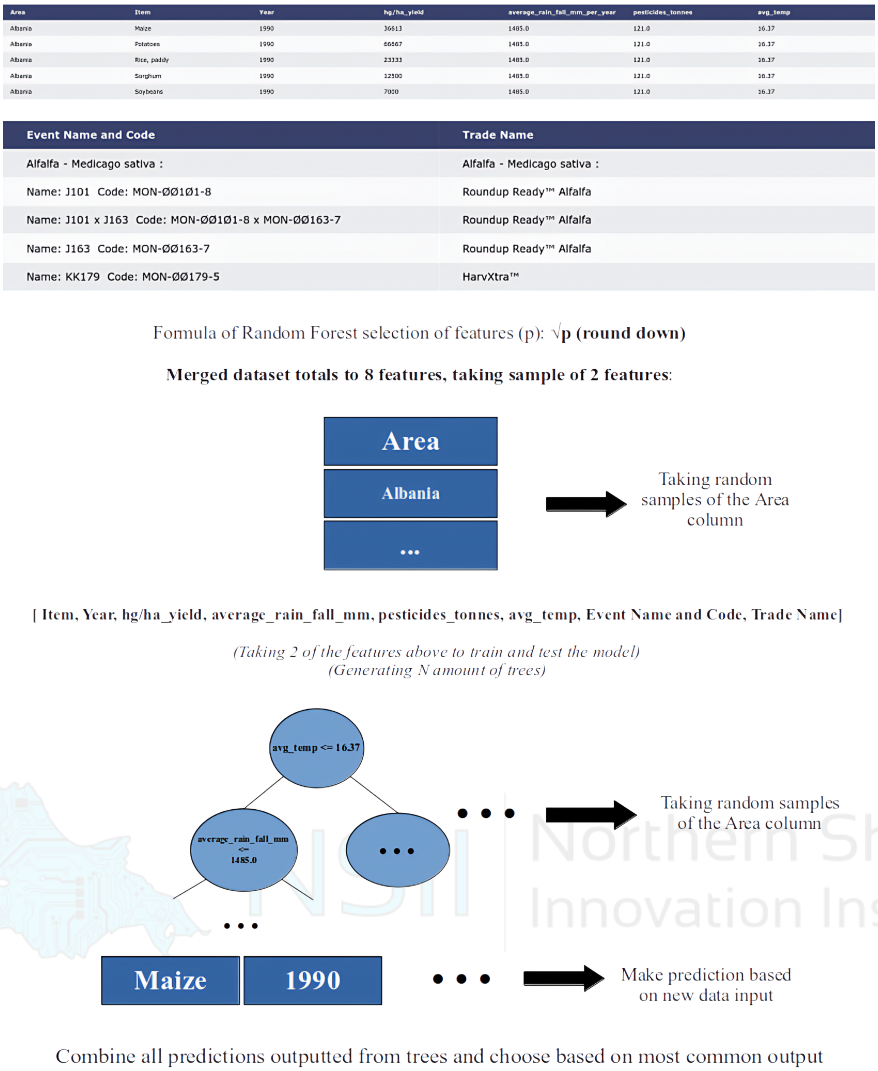

Figure 6: Example of a Decision Tree Based on the Data Provided

5 Materials

The main proposed source for this data, at least initially until other data sets can be compiled, is Kaggle.[7] More specifically, the Crop Yield Prediction Dataset this data set is reliable based on the specifications our statistical model needs.

The proposed programming language for this project is Python and within Python, the tool Pandas will be utilized. Selenium will be used to optimize data collection processes.

6 Methodology

1. Extract the downloadable content from Kaggle and perform the data cleansing shown in Figure 1.

2. Further analysis of the data shows which crops are most affected based on the chosen attributes in Figure 2. Using the data from this graph, the model can

determine the optimal crop based on the average temperature and pesticides.

3. Using will demonstrate how increasing demand and population affect these attributes, increasing the accuracy of the model.

4. Using Selenium, input the URL of the ISAAA website and begin the automation. This will be done by scraping the URL and placing it into a separate data set, as shown in Figure 4.

7 Discussion

This proposal focuses on a proof of concept using data obtained from Kaggle, an online data bank. Provided that the proof of concept is a success, we propose collecting and compiling a data set with more in-depth data that covers a broader breadth of crops, crop variations, pesticides and other variables that may be useful in the future. Several sources would add to this data set, but most importantly, it would be added to by the farmers who use the model.

Through the expansion of this idea individual farmers and larger corporations would be better able to use their land and effectively grow crops for our food system. Future development of this idea could result in creating an app or website in which farmers or other individuals could enter their specific details and receive an output showing a breakdown of the most effective crops for them to grow. Farming can be more effectively utilized by using a program like this, increasing production yields.

The possible extensions of this research are vast as similar technology could be used in other areas related to food production, such as raising livestock.

8 Conclusion

In recent years as technology has become increasingly integrated into our day-to-day lives, much of agricultural technology has stayed relatively constant. In order to keep up with our rapidly evolving climate crisis, the systems we use must evolve to allow them to be the most effective tool we can create. From year to year, the climate around us changes, with events that were once rarities becoming an everyday affair. For these reasons, it is of the utmost importance that we evolve with it. Statistical models must be created to allow us to utilize our diminishing farmland with the utmost efficiency.

Acknowledgements

We would like to thank the Northern Shores Innovation institute for supporting our entry. Additionally, the resources and mentorship provided by STEM Fellowship and the REO Blueprint team were excellent and appreciated in creating this submission.

References

[1] Gregory P J, Ingram J S I and Brklacich M. Climate change and food security;. [cited 2022 July 28]. [Internet]. Available from: http://doi.org/10.1098/rstb.2005.1745.

[2] Food and Agriculture Organization of the United Nations Rome. Climate Change and Food Security: a Framework Document;. [cited 2022 July 28]. [Internet]. Available from: https://www.fao.org/publications/card/en/c/f42f5d4d-df3b-504a-945c-fa2ad3afc658/.

[3] Patel R. Crop yield prediction dataset;. [cited 2022 July 12]. [Internet]. Available from: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset?select=yield_df.csv.

[4] ISAAA Inc . GM Approval Database;. [cited 2022 July 12]. [Internet]. Available from: https://www.isaaa.org/gmapprovaldatabase/default.asp.

[5] Pandas. Python Data Analysis Library;. [cited 2022 July 12]. [Internet]. Available from: https://pandas.pydata.org/.

[6] Selenium. Selenium automates browsers;. [cited 2022 July 20]. [Internet]. Available from: https://www.selenium.dev/.

[7] Kaggle. Your Machine Learning and Data Science Community;. [cited 2022 July 12]. [Internet]. Available from: https://www.kaggle.com/.