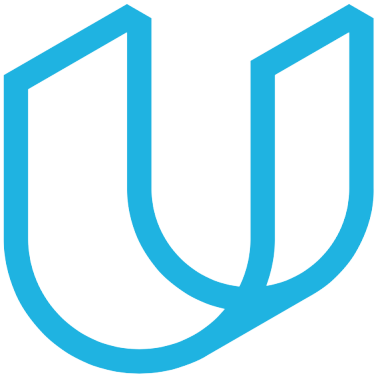
Capstone Project Proposal



Udacity Machine Learning Engineer Nanodegree

A capstone proposal for analyzing worldwide food production from the years 1961-2013.

Derek Helms

January 13^{th}, 2021

Table of Contents

I. Domain Background

The current world population is nearly 7.8 billion people, and this number is estimated to rise to around 9.7 billion in the year 2050. This means within the next 30 years, we will need to feed two billion more people without sacrificing the planet (1). This means we will need to double our crop production in order to feed that growing population. Agriculture is one of the greatest contributors of global warming, with farming consuming immense amounts of our water supplies and leaving major pollutants as its byproduct from fertilizer runoff (4).

This leads to the question, how do we supply the necessary amount of food for a growing world population without sacrificing the climate of our planet? There are many solutions to this question, but the focus of this capstone will be on analyzing the global production of consumables, as well as the ratio of food (human consumption) to feed (livestock consumption) produced by each country. Only 55% of the current world crop production is consumed by humans, with the remaining being fed to livestock. On top of this, nearly 25% of the world's food calories are wasted before they can be consumed (4).

The personal motivation for investigating this problem is my interest in applying machine learning to the field of agriculture. With an increasing need for sustainability, applying machine learning techniques to predict yield amounts or for detecting crop diseases has been a key factor in my motivation behind studying machine learning.

II. Problem Statement

As initially stated above, the problem we face is how do we produce the necessary amount of food for an exponentially increasing population without sacrificing our own planet.

One aspect to explore is which food items are the most produced and the use case for those crops, whether they are for human or animal consumption. This will allow us to understand where a majority of our food is being directed, and as suggested by National Geographic, could a shift in diet lead to more crops being used for human consumption rather than livestock feed.

Another aspect of the data to explore is separating the countries into clusters based on their yearly production. Analyzing which countries are responsible for feeding the majority of the population, and how to ensure they can continue to do so. From this, we can also see which countries are not highly producing, and how we can aid them in increasing production if possible.

III. Datasets and Inputs

Two different datasets will be used in the analysis, one coming from Kaggle and one coming from FAO (the Food and Agriculture Organization of the United Nations).

The dataset obtained through Kaggle is a Food Balance Sheet that originated from FAO's database, but has been formatted for easy of use, with a total of 63 features and 21477 observations. Food Balance Sheets represent the pattern of a country's food supply during a period of time, in this case it is yearly from 1961 to 2013, and measured in 1000 tonnes (2). Important features that this file contains are the country, item, element, and yearly production:

- Area: 174 unique country names from around the world.
- Item: 115 unique food items being globally produced.
- Element: food - human consumable food available at a given time.
  feed - livestock consumable food available at a given time.
- Y1961 - Y2013: item available at time of measurement per year.

The dataset obtained through FAO is a time series dataset that contains the estimated/projected population (for both sexes) in each country from 1961 to 2018. The estimates are based on data from the World Population Prospects and World Urbanization Prospects (3). An important note is that the population dataset contains countries and later years that are not included in the Food Balance Sheet dataset, but will be left in to account for global population and future estimates. Some important features to note within the dataset:

- Area: 245 unique country names (71 more than first dataset).
- Year: spans from 1961 to 2018 (5 years more than first dataset).
- Unit: population estimations are counted in units of 1000 persons.
- Value: estimate population count for each year.

The use cases for each dataset is as follows. The Kaggle dataset will be used in analyzing food production, seeing which countries produce the most and if it is for human or livestock consumption. This data will also be used to cluster countries together based on their production levels, ideally separating high and low producing countries. The FAO dataset will be used to compare the global population to the yearly total food production, as well as further analyzing the highest producing countries to see if they have the largest/fastest increasing populations.

IV. Solution Statement

V. Benchmark Model

In general, the majority of this capstone will be in exploring the data and finding insights about food production through graphing and statistical methods. Creating plots for top producing country's, top produced food items, and population changes will be the main focus of the work.

We will also create a KMeans model to cluster our data based on food production, partitioning the 174 original country's into k clusters. We will want to use the elbow method in selecting our number of clusters, with an initial guess of using 2 to 3 cluster to separate countries based on high and low production (and possible medium).

VI. Evaluation Metrics

With the data being unlabeled, it will be difficult to ensure that our clusters are properly partitioned. However, we can analyze the countries within the clusters and see where the algorithm decided to split based on production count.

One metric we can use to evaluate our model will be a silhouette score, which will measure on a scale from [-1,1] how close each point on one cluster is to points in the neighboring clusters (5).

VII. Project Design

VIII. Resources

[1] Oppenheim, Dor. "Who Eats the Food We Grow?" *Kaggle*, 30 Nov. 2017, www.kaggle.com/dorbicycle/world-foodfeed-production.

[2] "Food Balances (Old Methodology and Population)." *FAOSTAT*, Food and Agriculture Organization of the United Nations, 12 Dec. 2017, www.fao.org/faostat/en/#data/FBSH.

[3] "Annual Population." *FAOSTAT*, Food and Agriculture Organization of the United Nations, 16 Dec. 2019, www.fao.org/faostat/en/#data/OA.

[4] Foley, Jonathan. "A Five-Step Plan to Feed the World." *Feeding 9 Billion - National Geographic*, www.nationalgeographic.com/foodfeatures/feeding-9-billion/.

[5] "Selecting the Number of Clusters with Silhouette Analysis on KMeans Clustering." *Scikit*, scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html.