Machine Learning Analysis of Food Scarcity and Surplus through Key Country Statistics

CPE/EE 695 Final Project Paper

Project Team 3: Christopher Morales, Michael Dasaro, Hamza Mustafa, Abdul Naeem Shaik

I. Introduction

Hunger is a serious problem worldwide. Technology has advanced incredibly far in recent years to the point where anyone with an internet connection has access to essentially all of human knowledge at their fingertips. With the advent of Uber Eats, Grubhub and the like, people can order food to be delivered straight to their door, even sometimes being delivered by a robot. This convenience in first world countries often leads to overordering and increased reliance on restaurants, with less home cooked meals. This in turn leads to more food waste as restaurants have to estimate order quantities and consumers discard leftovers, compared to home-cooked meals from groceries. With all that in mind, it is incredibly disconcerting to learn that roughly 11% of the world's population - that's over 830 million people - go to sleep hungry at night. Despite this, the issue is not necessarily one of scarcity but of distribution. Worldwide an estimated 17% of all food production is wasted, which adds up to an astronomical 1.3 billion tons of food wasted in a year. This is a great waste of resources, and we want to investigate this data and apply machine learning to gain some insight into this problem.

For this project, we are investigating food scarcity and food surplus to create a database that catalogs countries and compares data on scarcity and surplus. We will then collect data about these countries including their GDP per capita, infant mortality rate, and life expectancy. Lastly, we will use multiple machine learning methods to estimate the food situation of countries based on these three data sources to evaluate which is most directly related to the food scarcity problem.

II. Data

All of our work for this project was done in a GitHub repository, which can be found here. Work was done in Jupyter notebooks with comments allowing us to read over each other's work and stay organized. Raw data and processed data are stored as csv files and most of the processing is done in one file while the evaluation is done in another.

In order to approach this problem, the first thing that needs to be done is gathering all of the necessary data. First and foremost on the data front, we gathered food surplus and scarcity data by country. It goes without saying that if we intend to train a machine learning algorithm, we would need true values for training and testing. This preprocessing is done in a Jupyter

notebook, data_preprocessing.ipynb. We acquired food waste data from Kaggle by Joakim Arvidsson, updated in 2023. This database contains figures in tons/year of food wasted by households and retail stores. To start, we combine these figures and remove any unnecessary data, storing our total food waste in tons per year. Something of note here is that even relatively poor countries that we would expect to have food scarcity are estimated to have lots of food waste according to this database. Next, we load the food scarcity database from kaggle by Maryam Sikander to investigate further. This database tracks food scarcity over several years, but for a fair comparison we are using the most recent data (2020). This database also tracks food scarcity as a percentage of the population experiencing scarcity. In order to convert this into useful data, we load population statistics from this database and cross-reference country codes in order to estimate the number of people experiencing food scarcity in each country. Lastly we can estimate the amount of food needed by each of these people to estimate the number of tons of food scarcity that each country has. It is worth noting that lots of converting as estimating is occurring here, along with the fact that these databases are estimates themselves. For the purpose of this project, we are assuming that this data is valid, or at least a valid comparison between countries. This data can be found here.

With surplus and scarcity at hand, we had to next decide on what data can be best used to estimate food scarcity. Since data can vary greatly depending on the source, we decided to take all of the data from The World Bank. This could help to lessen some of the inherent error that inevitably comes from data acquisition. The first kind of data we decided on was GDP per capita. GDP per capita is a widely used measure of wealth, and it stands to reason that wealthier countries will tend to have more food while poorer countries will tend to have less. Next we decided on population size. Since the food scarcity and surplus data we gathered is simply in total tons of pounds for each country, we thought it would be a good idea to consider how many people this food would be for. Again, it stands to reason that generally, countries with more people will require more food, so this would be a good parameter to help the algorithm curb the effect of countries with massive population sizes. The penultimate set of data we decided on was infant mortality rate. Once again, infant mortality rate is a common metric for how ‘successful’ a country is. Countries with low infant mortality would likely correlate to having enough food or lesser food scarcity and vice versa. Finally, we decide on the average life expectancy of a country. Life expectancy is a good indicator of how long people live in a given country. It could

serve as a red herring, particularly for countries in conflict, as their life expectancy would be deflated by active conflict even if they did not have food shortages. Still, though, we believe this is another effective piece of information that can be used to estimate food shortage or surplus in a country.

III. Methodology

With all of the necessary data chosen, we then had to parse the data. This is an absolutely necessary step when using any kind of data. Real, raw data contains many holes and discrepancies and it is important to fix these in order to get reliable results from any machine learning algorithm you may decide to implement. First and foremost, we noticed that a lot of the data is out of order, and some of the data did not exist for specific countries. Additionally, some countries were counted twice in some datasets, namely countries that go by more than one name (e.g. United States or United States of America). The unfortunate fix for this was simply to parse the data manually. We first set the countries in alphabetical order. Finally, we checked each data set and removed any countries that did not appear in one or more of them.

From there, being left with only countries that appear in all the datasets, we removed any double counted countries from the data and did a final check to be sure they were all in order with the associated countries. From here, it was simple to compile all of the data into one large .CSV file. Then all that was left was to fill any gaps in the data for the average value of that column. For example, if the country Afghanistan did not have a value for 'Life Expectancy,' the empty element would be replaced with the average life expectancy of all the countries in the dataset. This was done for all of the data. It is important to note that food surplus, food scarcity, and population size all did not have any holes in the data to fill. Our final dataset can be found here.

Once all of our data was loaded, organized, parsed, and verified, we were able to apply machine learning to it and gain some insights. This methodology can be followed along in the Machine_Learning_Analysis.ipynb. To simplify the food scarcity/surplus data, we first subtract them to find the net (positive or negative) food situation of each country. We then normalize the data to integers from -3 to 3. Finally we can begin our machine learning analysis by building a decision tree and random forest classifier for each of our three test variables. We begin by taking the GDP per Capita (along with population), get our training and testing data split, and fit a

decision tree. We then take the accuracy results and add it to the results table (seen and explained below). Following this we create a random forest classifier on the same data and add the accuracy to the table. We then repeat this process two more times, analyzing infant mortality rate and life expectancy in the same way, rebuilding and training the two models each time.

IV. Results

This table displays the percentage accuracy of each independent variable when predicting the food scarcity/surplus.

ML Model	GDP Per Capita	Infant Mortality Rate	Life Expectancy
Decision Tree	55.81%	58.14%	46.51%
Random Forest	69.77%	58.14%	53.49%

In this research we have used machine learning methods to study how GDP, per person infant mortality rate, life expectancy and net food availability in countries are connected. Our aim was to figure out which factors play a role in determining whether a country is likely to have surplus or shortage of food.

Model Performance Overview:

Here's a summary of the performance of the decision tree and random forest models;

Decision Tree Classifier:

GDP per Capita and Population; The model has achieved an accuracy of 55.81% indicating an association with food supply trends.

Infant Mortality Rate and Population; The model showed an accuracy rate of 58.14% suggesting an better correlation compared to the GDP

Life Expectancy and Population; The lowest accuracy rate of 41.86% was recorded here indicating that life expectancy may not be as predictive of food scarcity or surplus

Random Forest Classifier:

The Random forest classifier has showed the performance when analyzing ;

GDP per capita and Population: The model has showed the best accuracy with 69.77% indicating an strong predictive ability
Infant mortality rate and Population: When considering this , the model has decreased accuracy with 53.49% proving to be less effective compared to using GDP data
Life Expectancy and Population: Similar to the decision tree model the accuracy has dropped to 55.81% again indicating that life expectancy may not be as predictive of food scarcity or surplus

V. Conclusions

Our research demonstrates how utilizing machine learning can provide insights into the correlation between a nation's prosperity, wellbeing and its food situation. We discovered that a country's wealth as indicated by its GDP per capita plays a role in determining whether there is a food supply issue. Additionally insights from the analysis of infant mortality rates reveal a link between a nation's health status and its food availability. Performing this analysis manually would be tedious and unclear, and our use of Machine Learning here displays an interesting use case. Simply by gathering data on different variables and assessing the performance of Machine Learning models, we can quickly and effectively notice associations between them, which can lead to analysis of a cause and effect relationship. This method has broad applications, and we are satisfied that we were able to successfully use ML models on a global dataset and draw meaningful results.

These discoveries underscore the importance of implementing policies that prioritize growth and healthcare advancements to address hunger issues. The effectiveness of the Random Forest model indicates that machine learning tools can greatly benefit policymakers and

organizations striving to address challenges in food distribution. As efforts continue globally to tackle food insecurity our research contributes information that can support decision making. By leveraging data and machine learning methods we can explore avenues to enhance food security worldwide.

Bibliography/Works Cited

- [ ] GDP per capita (USD) by country: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
- [ ] Population by country: https://data.worldbank.org/indicator/SP.POP.TOTL
- [ ] Life expectancy (Years) by country: https://data.worldbank.org/indicator/SP.DYN.LE00.IN
- [ ] Infant Mortality Rate (per 1000 births) by Country: https://data.worldbank.org/indicator/SP.DYN.IMRT.IN
- [ ] Food Waste Database: https://www.kaggle.com/datasets/joebeachcapital/food-waste
- [ ] Food Scarcity Database: https://www.kaggle.com/datasets/maryamsikander/sdg-2-zero-hunger

  World Population Dataset: https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset