Relationship between Food Supply and Undernourishment/Obesity Prevalence

Samantha Jarrett - GTID: 903743596

ISYE6740 Final Project

06 December 2023

Table of Contents

ABSTRACT

The purpose of this study is to investigate the relationship between a country's food supply and its undernourishment/obesity rates to gain insight on the dietary composition of the population to be used to provide potential recommendations. A KNN model was implemented, resulting in 86% accuracy in prevalence of undernourishment classification and a Random Forest model was implemented, resulting in 75% accuracy in prevalence of obesity classification. These models indicate that food supply is indeed related to the prevalence of undernourishment and obesity within a country's population. This information allows for the focus of providing countries with high prevalence of undernourishment with more nutritionally dense foods such as fats, meat, and dairy. Additionally, it allows for focus of educating countries with high prevalence of obesity to produce and consume less fats and sweeteners and more healthy options such as vegetables, grains, and beans.

INTRODUCTION

Undernourishment, also known as underfed, is prevalent when an individual consumes an insufficient number of calories to cover their energy requirements for a healthy lifestyle. If an individual does not intake a minimum amount of energy, they will not be able to maintain a healthy weight and perform necessary daily functions. Additionally, without the proper amount of protein and other nutrients, their body will not function as it should, leading to further limitations and damage to their health [1].

Obesity is identified using body mass index (BMI), calculated by dividing the individual's weight in kilograms by the square of their height in meters. High body mass index is correlated with high body fat. Obesity is considered present when the individual's weight is higher than considered healthy based on their height. It is typically caused by unhealthy eating patterns and/or low physical activity. Obesity increases the risk for heart disease through increased blood pressure and cholesterol, Type 2 diabetes, sleep disorders such as asthma and sleep apnea, joint issues such as osteoarthritis and musculoskeletal discomfort, gallstones/gallbladder disease, many cancers, and premature death. In 2019 in the United States alone, obesity-related medical care costs for preventive, diagnostic, and treatment services were estimated to be $173 billion [2].

The purpose of this study is to investigate the relationship between a country's food supply and its undernourishment/obesity prevalence to gain insight on the dietary composition of the population to be used to provide potential recommendations. Insights may lead to solutions such as providing proper education on dietary causes of these conditions for improved coaching in more developed regions and for influencing food aid and funding provided to underdeveloped countries. The overall benefit of these efforts could lead to a healthier population and lowered health care costs. Previous studies have been performed to determined dietary impacts on diabetes diagnoses and food supply in general has been used in studying undernourishment and malnutrition, but the relationship between food supply and prevalence of undernourishment and obesity in different countries has not been studied in this manner.

DATASET

The dataset for food supply and undernourishment data used for this project was extracted from the Food and Agriculture Organization of the United Nations data suite (FAOSTAT). Food supply was extracted from the FAO Suite of Food Balances [3] and is presented as food supply (kcal/capita/day).

The prevalence of undernourishment is computed by the FAO by comparing the probability distribution of habitual daily dietary energy consumption with the minimum dietary energy requirement of the reference population. This data was extracted from the FAO Suite of Food Security Indicators [4].

Prevalence of obesity data was extracted from the World Health Organization (WHO). This data is a WHO estimate of the prevalence of obesity among adults indicated by a BMI greater than or equal to 30 [5].

The original dataset contains 96 different food items. For this project, these food items were categorized into 15 broader categories. For example, potatoes were categorized as roots & starches. Additionally, the food supply data for each food category was converted into a percentage of total kcal/capita/day for each country.

EXPLORATORY DATA ANALYSIS

Undernourishment Classification

The mean prevalence of undernourishment is 10.5% and the median prevalence is 5.2%. This spread between the mean and median indicates that half of the countries have a prevalence lower than 5.2%, but there are some countries with remarkably high prevalence of undernourishment pulling the mean higher to 10.5%. This is exemplified in the histogram in Figure 1. The 50% of countries with lower than the median prevalence of undernourishment are condensed in the 2.5% to 5% range, while the upper 50% is spread between 5% and 51%.

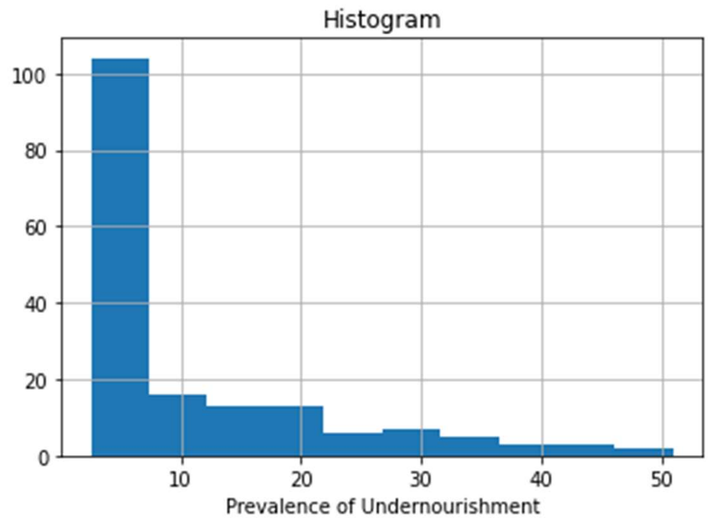

Figure 1. Histogram for Prevalence of Undernourishment

Countries with the highest prevalence of undernourishment tend to be on the continent of Africa. The country with the maximum prevalence of undernourishment is Madagascar at 51%, and the country with the minimum prevalence of undernourishment is tied by many countries at 2.5% (Table 1).

Country	Location	Prev. of Undernourishment (%)
Madagascar	East Africa	51.0
Central African Republic	Central Africa	48.7
Lesotho	South Africa	46.0
Democratic People's Republic of Korea	East Asia	45.5
Haiti	Caribbean	45.0
Zimbabwe	South Africa	38.4
Liberia	West Africa	38.4
Guinea Bissau	West Africa	37.9
Democratic Republic of Congo	Central Africa	35.3
Yemen	Middle East	34.5

Table 1. Top 10 Highest Countries in Prevalence of Undernourishment

The correlation of each of the food categories with prevalence of undernourishment was calculated (Figure 2). High percentages of roots & starches, grains, and beans in the food supply were positively correlated with higher prevalence of undernourishment. This is logical, as these food items are dense enough to provide calories for survival, but do not provide many nutrients besides carbohydrates.

High percentages of dairy, vegetables, meat, sugars & sweeteners, beverages, seafood, and fats & oils were negatively correlated with higher prevalence of undernourishment. This is logical, as dairy, vegetables, meat, and seafood provide protein and many healthy nutrients. Fats & oils, beverages, and sugars & sweeteners do not provide many nutrients, but are calorie-dense and support weight gain.

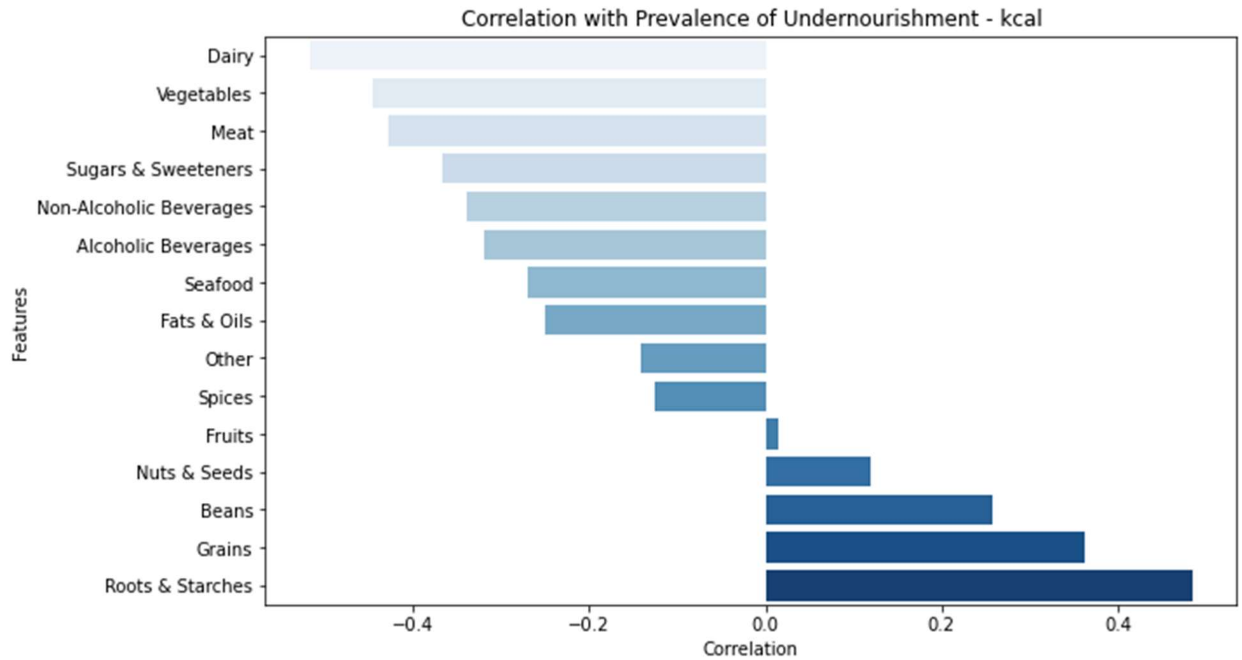

Figure 2. Food Supply Correlation with Prevalence of Undernourishment

An additional column was added to identify if the country's undernourishment prevalence was greater than the median. If the prevalence of undernourishment for a country is above 5%, it is considered to have population undernourishment presence above the median and is represented with a 1. Otherwise, it is not considered to have population undernourishment presence below the median and is represented with a 0.

Four of the highly correlated food categories were plotted against the prevalence of undernourishment classes of below and above the median (0 and 1, respectively). The roots & starches food supply upper confidence interval is much higher, and the distribution is much wider for countries classified above the prevalence of undernourishment median. On the other hand, the medians are higher, and distributions are wider for dairy, vegetables, and meat food supply in countries below the median for prevalence of undernourishment. This further indicates that these variables likely have an impact on classifying prevalence of undernourishment.

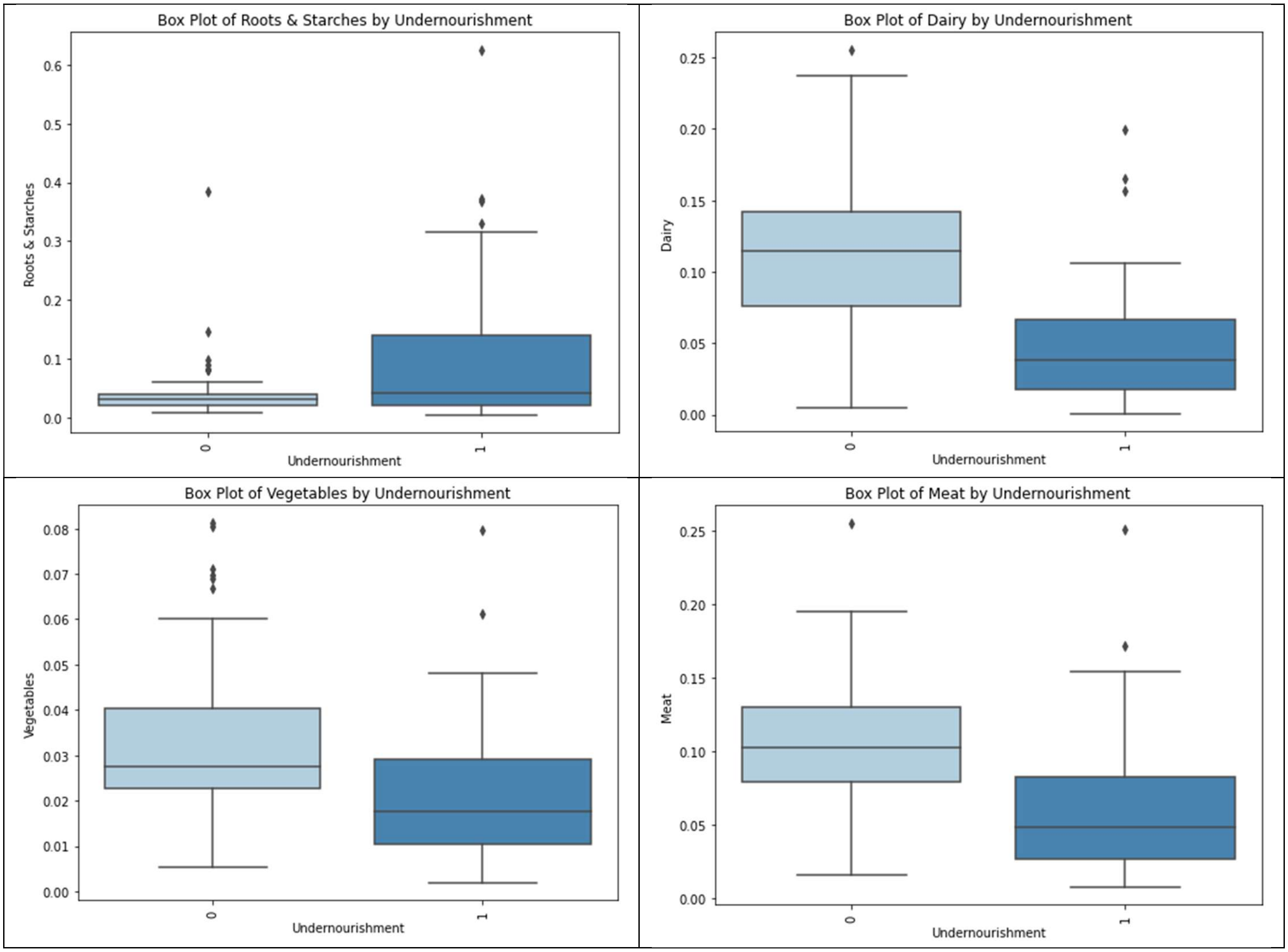

Table 2. Prevalence of Undernourishment Classification Boxplots

Obesity Classification

The mean prevalence of obesity is 19% and the median prevalence is 20.6%. The mean and median are remarkably close, indicating that there is generally an even spread between the top 50% and bottom 50%. However, looking at the histogram in Figure 3, there are many countries lying at the median and mean, and the upper 50% is more sparsely spread out than the lower 50%. The 50% of countries with lower than the median prevalence of obesity are in the 2.5% to 20% range, while the upper 50% is spread between 20% and 61%.

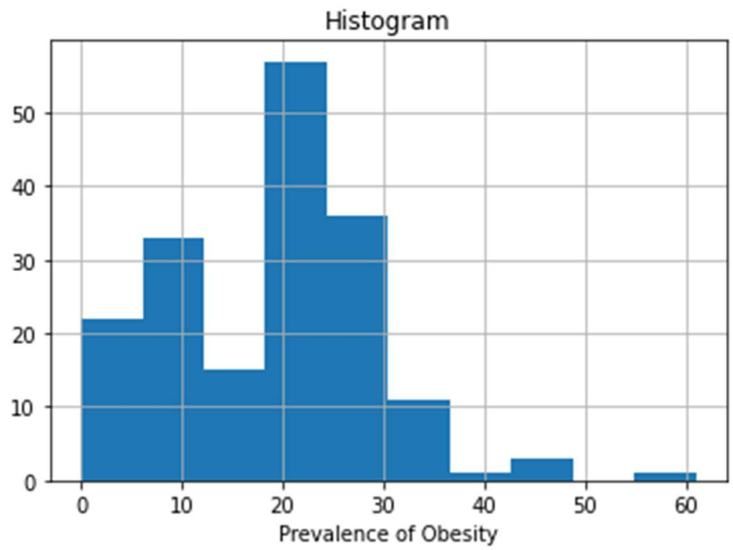

Figure 3. Histogram for Prevalence of Obesity

Countries with the highest prevalence of obesity tend to be in Oceania and the Middle East plus the United States of America. The country with the maximum prevalence of obesity is Naura at 61%, and the country with the minimum prevalence of obesity is Vietnam at 2.1%. The high presence of obesity in Naura, approximately 14% higher than the second highest country, explains the spread of the upper 50% (Table 3).

Country	Location	Prev. of Obesity (%)
Nauru	Oceania	61.0
Samoa	Oceania	47.3
Kiribati	Oceania	46.0
Federal States of Micronesia	Oceania	45.8
Kuwait	Middle East	37.9
United States of America	North America	36.2
Jordan	Middle East	35.5
Saudi Arabia	Middle East	35.4
Qatar	Middle East	35.1
Libya	North Africa	32.5

Table 3. Top 10 Highest Countries in Prevalence of Obesity

The correlation of each of the food categories with prevalence of obesity was calculated (Figure 4). The results were almost the opposite as the correlation values for prevalence of undernourishment. It is not a surprise that the variables negatively correlated with undernourishment are positively correlated with prevalence of obesity. High percentages of sugars & sweeteners, meat, fats & oils, dairy, fruits, and beverages in the food supply were positively correlated with higher prevalence of obesity. This is logical, as these food items are calorie-dense, and/or provide healthy nutrients such as proteins and fats. It is sensical that the highest correlated food is sugars & sweeteners. Sugars & sweeteners can be highly addictive, and they tend to spike blood sugar and insulin levels, which cause the body to store energy (calories and fat).

High percentages of grains, beans, roots & starches, and nuts & seeds were negatively correlated with higher prevalence of obesity. These foods are not calorie-dense and do not provide many nutrients other than carbohydrates and fiber. It is interesting that nuts & seeds are negatively correlated with obesity, as they are extremely high in calories and fat.

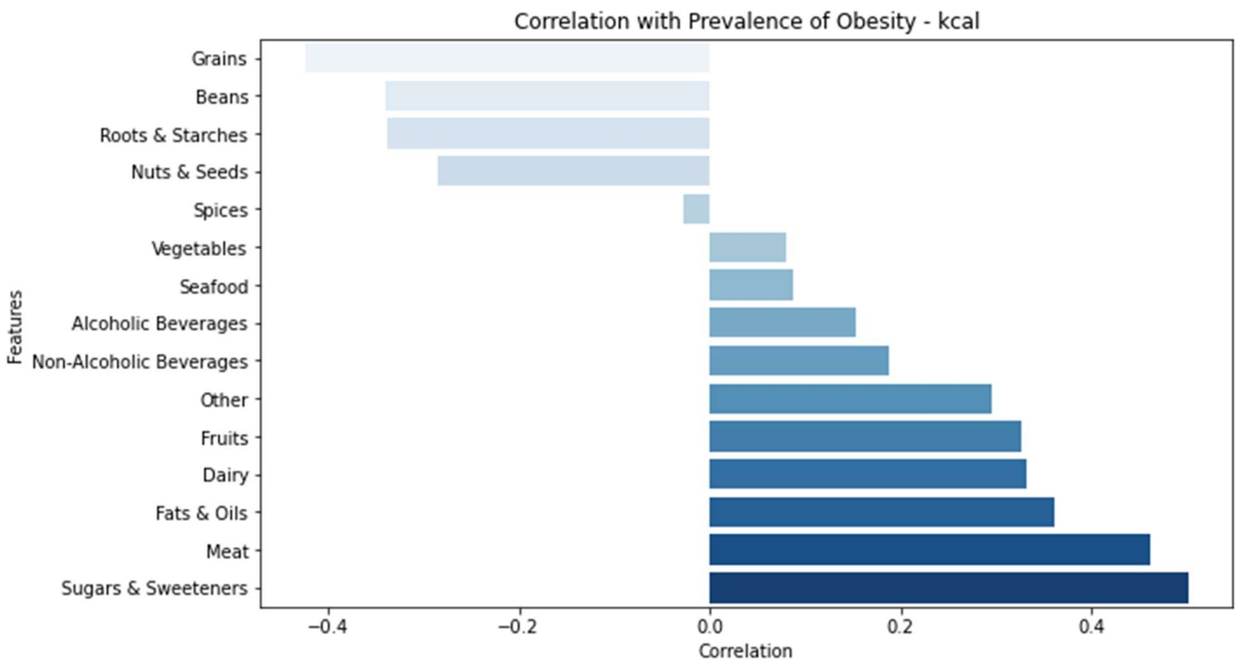

Figure 4. Food Supply Correlation with Prevalence of Obesity

An additional column was added to identify if the country's obesity prevalence was greater than the median. If the prevalence of undernourishment for a country is above 20%, it is considered to have population obesity presence greater than the median and is represented with a 1. Otherwise, it is considered to have population obesity presence below the median and is represented with a 0.

Four of the highly correlated food categories were plotted against the prevalence of obesity classes of below and above the median (0 and 1, respectively). The medians are higher for sugars & sweeteners, meat, and fats & oils food supply in countries above the median for prevalence of obesity. The median was higher for grains food supply in countries below the median for prevalence of obesity. This further indicates that these variables may have an impact on classifying prevalence of obesity.

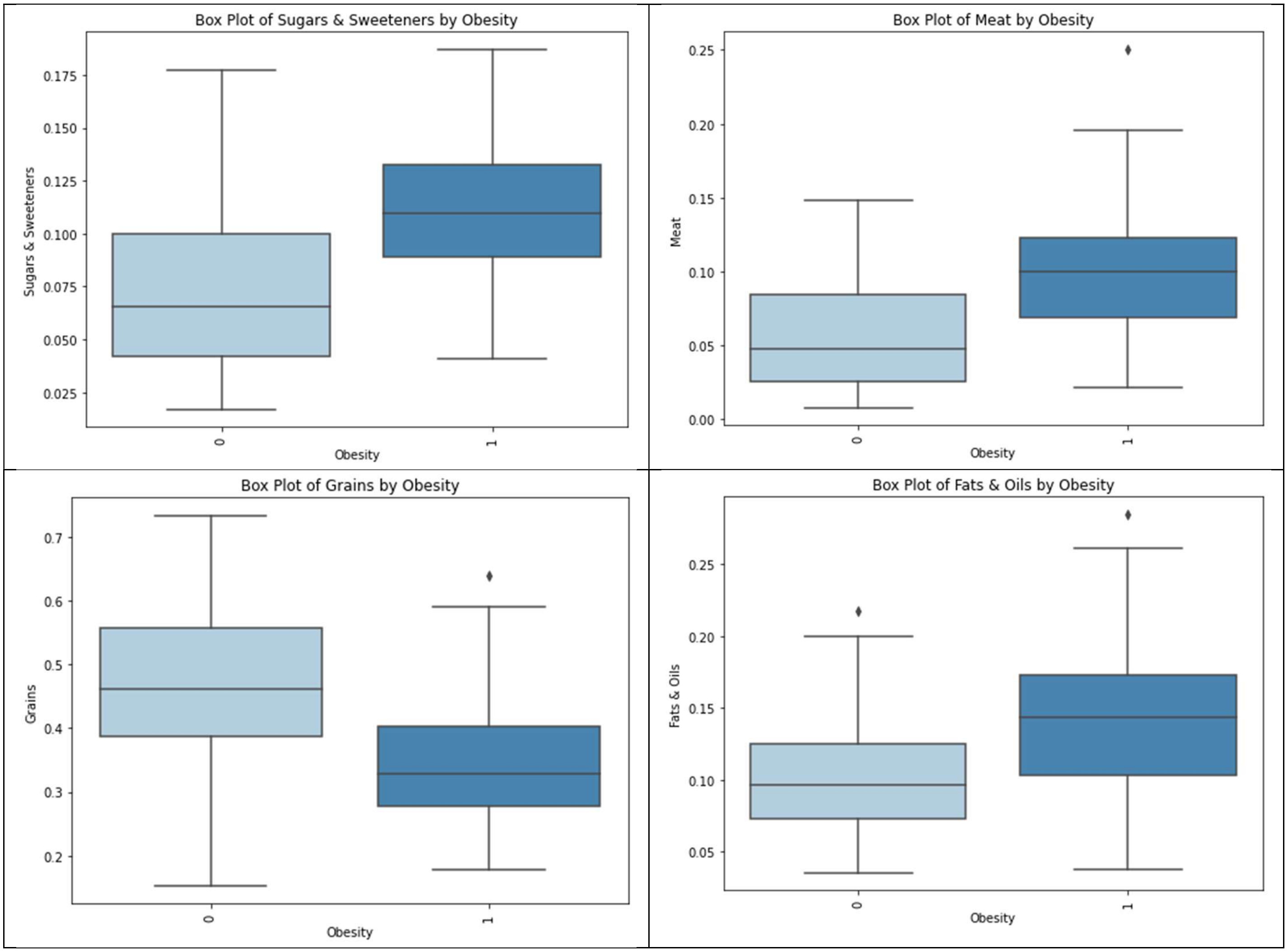

Table 3. Prevalence of Obesity Classification Boxplots

METHODOLOGY

Feature Selection

Feature selection was performed through determining the chi-squared scores of the features and ranking them by their score. Based on these scores, the dairy, roots & starches, grains, and meat food supply features have the highest rankings and scores (Table 4). This aligns with the findings in the EDA section. Based on their low rankings and scores, only the features in bold were used in modeling (dairy, roots & starches, grains, meat, alcoholic beverages, fats & oils, beans).

Feature	Score
Dairy	**1.87**
Roots & Starches	**1.30**
Grains	**1.26**
Meat	**0.86**

Alcoholic Beverages	**0.41**
Fats & Oils	**0.35**
Beans	**0.23**
Vegetables	0.19
Sugars & Sweeteners	0.11
Fruits	0.07
Non-Alcoholic Beverages	0.06
Seafood	0.05
Spices	0.02
Other	0.01
Nuts & Seeds	0.00

Table 4. Feature scoring for Undernourishment Classification

The same process was followed for prevalence of obesity. Based on these scores, the roots & starches, grains, and dairy food supply features have the highest rankings and scores (Table 5). This also aligns with the findings in the EDA section. Based on their low rankings and scores, only the features in bold were used in modeling (roots & starches, grains, dairy, meat, fats & oils, sugars & sweeteners).

Feature	Score
Roots & Starches	**1.91**
Grains	**1.50**
Dairy	**1.21**
Meat	**0.77**
Fats & Oils	**0.68**
Sugars & Sweeteners	**0.49**
Fruits	0.26
Beans	0.24
Alcoholic Beverages	0.18
Nuts & Seeds	0.12
Other	0.03
Non-Alcoholic Beverages	0.02
Vegetables	0.01
Spices	0.00
Seafood	0.00

Table 5. Feature scoring for Obesity Classification

The differences between the features selected for prevalence of undernourishment and obesity are that undernourishment will utilize alcoholic beverages and beans in addition to common features, and obesity will utilize sugars & sweeteners in addition to common features. Note that in addition to the scoring process used above, different combinations of features were evaluated to achieve the best classifications.

Modeling

The following classification models were fitted using the training data and evaluated using the testing data. Parameter tuning was performed using a grid-search technique on the training data with a 5-fold cross validation splitting strategy and scored using prediction accuracy. The models tuned and trained are listed below in Table 6 and Table 7.

Logistic Regression utilizes the logit/sigmoid function to create a boundary between two classes. It is a linear model that performs well when there is a linear relationship between the features and the target. Regularization adds a penalty to the cost function to help prevent overfitting. The regularization parameter, C, is used to control the regularization penalty. Smaller values of C increase the regularization penalty effect.

The K-Nearest Neighbors (KNN) model classifies datapoints based on its similarity to the datapoints nearby. The parameter n-neighbors is used to specify how many neighboring datapoints will influence the classification of the datapoint being predicted. Larger numbers of neighbors may not capture the data well enough, while smaller numbers of neighbors may lead to overfitting.

The Decision Tree model performs classification through splitting data into leaves based on node criteria. The criteria parameter indicates how the quality of the split will be measured, the max depth parameter indicates the maximum number of expansions allowed, and the minimum samples per split indicates the minimum number of samples allowed per leaf.

The Random Forest model is an ensemble method consisting of many decision trees. Each of the decision trees are constructed individually on a random subset of the data. The results from the individual trees are compiled as votes to determine the ultimate classification of the data points. Random forests are known to be robust due to their cross-validation-like technique and excel at handling high-dimensional data.

Gradient Boosting is also an ensemble learning method that works by training tree models sequentially. With each iteration, the model works to correct deficiencies in the previous model. Through this method, the weak tree learners are combined to create a stronger learner. Like AdaBoost, this model includes the number of estimators and learning rate parameters. Additionally, it includes a maximum depth parameter to be applied to each individual tree in the ensemble.

The hyperparameters used for each model in both Undernourishment and Obesity Classification are presented in Table 6 and Table 7, respectively.

Model	Tuned Parameters
Logistic Regression	regularization parameter, C = 10
K-Nearest Neighbors (KNN)	n-neighbors = 2
Decision Tree	criteria = entropy maximum depth = None minimum samples per split = 4
Random Forest	n-estimators = 100 maximum depth = None minimum samples per split = 4
Gradient Boosting	learning rate = 0.2 maximum depth = None n-estimators = 200

Table 6. Modeling parameters used for Undernourishment Classification

Model	Tuned Parameters
Logistic Regression	regularization parameter, C = 100
K-Nearest Neighbors (KNN)	n-neighbors = 8
Decision Tree	criteria = entropy maximum depth = None minimum samples per split = 2
Random Forest	n-estimators = 200 maximum depth = None minimum samples per split = 5
Gradient Boosting	learning rate = 0.2 maximum depth = None n-estimators = 300

Table 7. Modeling parameters used for Obesity Classification

After tuning and fitting the model to the training data, the target classes of the test data were predicted and compared to the actual test data targets. Metrics reported include accuracy, precision, recall, and F1-score. Accuracy is defined as the proportion of correctly assigned instances out of total instances. Precision is defined as the ability of the model to correctly identify true positives. Recall is defined as the ability of the model to correctly identify all positive instances. The F1-score is the mean of precision and recall. Note that precision, recall, and F1-scores reported are macro averages of both variables to ensure that the model performs well for all classes, regardless of sample size.

RESULTS

Undernourishment Classification

The coefficients determined by the Logistic Regression model are presented in Table 8. It is logical that the grains, dairy, roots & starches, and meat food supply coefficients are among the top highest absolute value, as they were ranked highly in feature selection. Likewise, it is not surprising that beans, fats & oils, and alcoholic beverage food supply coefficients had lower absolute values and lower impact on the model, since they were scored lower in feature selection.

Feature	Coefficient
Grains	6.4663
Roots & Starches	5.2789
Beans	1.8496
Alcoholic Beverages	-1.5630
Fats & Oils	-1.7669
Meat	-3.7438
Dairy	-6.1823

Table 8. Undernourishment Classification Logistic Regression Coefficients

The Logistic Regression model performed the worst among the models (accuracy = 80%). The lower performance of this model suggests that the data may not have a linear relationship with the prevalence of undernourishment classes.

Interestingly, the Random Forest, Decision Tree, and Gradient Boosting models all performed with the same accuracy (83%), precision, recall, and therefore F1-Score. It is interesting that these models all performed the same, but this is likely due to the small sample size of the test set. There are limited ways the data points can be classified and there is not much variability for the different models to work with. It is especially interesting that the ensemble methods (Random Forest and Gradient Boosting) performed the same as the non-ensemble Decision Tree model. The Random Forest is essentially multiple decision trees, so it typically performs better. Again, this is likely due to the small sample size.

The KNN model performed the best of all the classification models on prediction prevalence of undernourishment (accuracy = 86%). The KNN model compares test points with surrounding points to make classifications. Since it compares the test points with the entire training set, it is logical that it would perform well with a low sample size if the datapoints do not have high variability.

The results of all models used to classify prevalence of undernourishment are presented in Table 9.

Model	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors (KNN)	0.86	0.90	0.82	0.84
Random Forest	0.83	0.82	0.82	0.82
Decision Tree	0.83	0.82	0.82	0.82
Gradient Boosting	0.83	0.82	0.82	0.82
Logistic Regression	0.80	0.80	0.81	0.80

Table 9. Undernourishment Classification Model Results

All the models performed with a recall score of around 82%. Even in the KNN model, where the precision score was 90%, the recall score was much lower at 82%. This indicates that the model is missing a high number of positive instances and is conservative with predicting these cases (prevalence of undernourishment is greater than the median). The confusion matrix for the KNN model in Figure 5 shows that all cases of prevalence of undernourishment less than the median were classified correctly (21/21). However, of the 14 positive cases of prevalence of undernourishment above the median, five were incorrectly classified as less than the median.

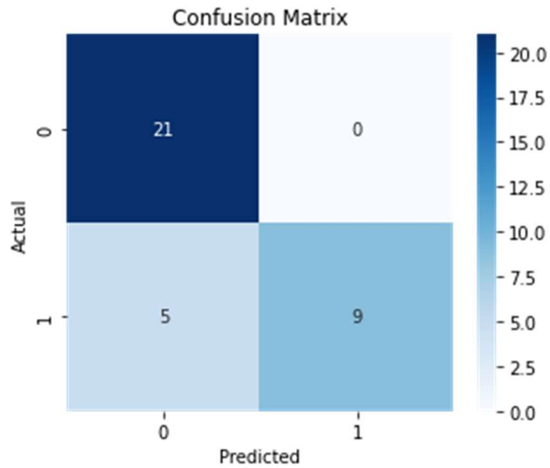

Figure 5. Confusion Matrix of KNN Model for Undernourishment Classification

Obesity Classification

The coefficients determined by the Logistic Regression model are presented in Table 10. It is logical that the sugars & sweeteners, fats & oils, roots & starches, and grains food supply coefficients are among the top highest absolute value, as they were ranked highly in feature selection. It is interesting that the meat and dairy food supply coefficients were smaller, having less impact on the model, as they were ranked more highly than fats & oils and sugars & sweeteners in feature selection.

Feature	Coefficient
Sugars & Sweeteners	7.8835
Fats & Oils	7.6854
Dairy	4.3843
Meat	1.1270
Roots & Starches	-8.6438
Grains	-11.5474

Table 10. Obesity Classification Logistic Regression Coefficients

The Logistic Regression and Decision Tree classification models performed the worst in prevalence of obesity classification (accuracy = 69%). The lower performance of the Logistic Regression model suggests that the data may not have a linear relationship with the prevalence of obesity classes. Likewise, the lower performance of the Decision Tree model suggests that the data is highly nonlinear.

The KNN and Gadient Boosting models performed moderately with an accuracy of 72%. It is interesting that the Gradient Boosting model performed at the same level as the KNN model, since typically ensemble methods are more powerful than non-ensemble methods. Like the classification of prevalence of undernourishment, this is likely due to the small sample size of the test set. As stated previously, there are limited ways the data points can be classified and there is not much variability for the different models to work with.

The ensemble Random Forest model performed the best of all the models (accuracy = 75%). It is logical that the Random Forest model performed quite better than the Decision Tree model, since the Random Forest combines multiple Decision Tree models to create one stronger model.

It is interesting that the accuracy, precision, and recall were equivalent within each model. This indicates that the model can balance precision-recall tradeoffs. This is also an indication that the dataset itself is balanced, allowing for all metrics to be equivalent in classification.

The results of all models used to classify prevalence of undernourishment are presented in Table 11.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.75	0.75	0.75	0.75
Gradient Boosting	0.72	0.72	0.72	0.72
K-Nearest Neighbors (KNN)	0.72	0.72	0.72	0.72
Logistic Regression	0.69	0.69	0.69	0.69
Decision Tree	0.69	0.69	0.69	0.69

Table 11. Obesity Classification Model Results

The confusion matrix for the Random Forest model in Figure 6 shows that 12 of 16 cases of prevalence of obesity less than the median and 15 of 20 positive cases of prevalence of obesity above the median were classified correctly. This is an example of the even precision and recall scores obtained by the models.

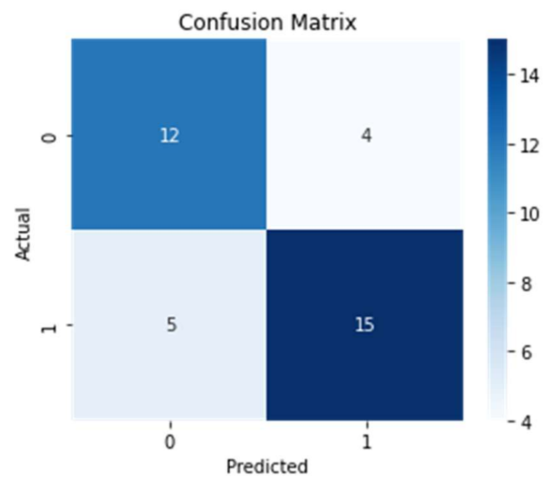

Figure 6. Confusion Matrix of Random Forest Model for Obesity Classification

CONCLUSION

The K-Nearest Neighbors model was able to predict prevalence of undernourishment above the median with 86% accuracy and the Random Forest model was able to predict prevalence of obesity above the median with 75% accuracy. The precision of the KNN model in classifying prevalence of undernourishment was much higher than the recall. This imbalance indicates that the model is missing a high number of positive instances and is conservative with predicting these cases of undernourishment prevalence above the median.

These modeling results indicate that food supply is indeed related to the prevalence of undernourishment and obesity within a country's population. It is concluded that countries in undernourished states would benefit from additional supply and consumption of more nutritionally dense foods that are negatively correlated with undernourishment, such as fats, meat, and dairy. This information can be used by humanitarian groups to focus their efforts on supplying the proper foods to these countries.

Additionally, it is concluded that countries with high prevalence of obesity would benefit from reduced consumption of foods highly correlated with obesity, such as sugars, sweeteners, fats, and oils. This information can be used to educate these populations and potentially incentives could be provided for increased production and consumption of healthier options such as vegetables, grains, beans, and roots/starches.

This study utilized food supply to quantify diet of a country and created prevalence classes using the median percentage of all countries. Further studies may investigate diverse ways of quantifying diet and/or prevalence of undernourishment and obesity to obtain a stronger model. Additionally, this study quantifies caloric intake as a percentage of the total, instead of general quantity, potentially missing the impact of generally low or high calorie food supply. This could provide further actionable insights in solving the issues of undernourishment and obesity.

REFERENCES

[1] Part II: Methods for the measurement of food deprivation and undernutrition. Keynote Paper: FAO methodology for estimating the prevalence of undernourishment. (n.d.). https://www.fao.org/3/Y4249E/y4249e06.htm.

[2] Centers for Disease Control and Prevention. (2023, September 21). Overweight & obesity. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/.

[3] Food and Agriculture Organization of the United Nations (FAO). (2023, May 17). Suite of Food Balances. FAOSTAT. https://www.fao.org/faostat/en/#data/FBS.

[4] Food and Agriculture Organization of the United Nations (FAO). (2023, August 23). Suite of Food Security Indicators Metadata. FAOSTAT. https://www.fao.org/faostat/en/#data/FS.

[5] World Health Organization (WHO). (2017, September 22). Prevalence of obesity among adults, BMI >= 30, age-standardized Estimates by WHO region. https://apps.who.int/gho/data/view.main.REGION2480A?lang=en.

[6] Over 20 million more people hungry in Africa's "Year of nutrition." Oxfam International. (2023, February 17). https://www.oxfam.org/en/press-releases/over-20-million-more-people-hungry-africas-year-nutrition#:~:text=Today%20a%20fifth%20of%20the,result%20of%20inadequate%20political%20choices.

[7] Punch, D. (2020, March 4). Below the surface: Law and the roots of obesity in the Pacific. https://www.mccabecentre.org/news-and-updates/below-the-surface-law-and-the-roots-of-obesity-in-the-pacific.html#:~:text=An%20estimated%2043%25%20of%20adults,the%20world's%20most%20obese%20countries.

APPENDIX

Undernourishment in Africa

In 2023, it was reported that one fifth of Africa's population is undernourished, including 55 million children under the age of five. These children are so malnourished that they experience stunted growth and development. It is speculated that the undernourishment crisis is a direct result of poor political decisions. Most governments on the continent only spend around 3.8% of the country's budget on agriculture. The lack of agricultural support from the government causes low food production and therefore supply to the population [6].

Obesity in Oceania

In 2020, it was reported that based on census data from 2016, an estimated 43% of adults in the Pacific Islands are obese. This number is three times the global average. Historically, the region's diet consisted mainly of green leafy vegetables, roots, and seafood. In the past few decades, much of these foods have been exported, and cheaper processed foods have been imported. The processed foods are much higher in sugar, salt, and fat than the traditional food consumed by the region, resulting in remarkably high rates of obesity [7].