**The Predictive Power of Food Consumption**

1. **Problem Statement**

Countries struggling with poverty are faced with insurmountable health challenges. In fact, “The rapid rise in [non-communicable diseases] NCDs is predicted to impede poverty reduction initiatives in low-income countries, particularly by increasing household costs associated with health care” (World Health Organization, 2018). Impoverished nations necessitate health resources and restructuring more than ever as the problem is only expected to grow. The aim of this research was to address a possible correlation between food consumption rates and non-communicable diseases due to the fact that the cause of these diseases has been tied to diet. If a link could be drawn between consumption rates of particular foods and NCD rates, then consumption trends could be used to both predict rates of disease and allocate health care resources.

This research was designed to explore the intersection of food consumption and health outcomes in developing countries. This research analyzed foods divided into 32 categories: rice (in all forms except flour), bread (fresh bread and special bread), cheese (cheese and curd), and pasta products to name a few. Consumption rates from each category were gathered from the World Bank Group (2010). The data for nutritional value was then gathered for each type of item in the category and the median of the individual food items was used to represent the category (U.S. Department of Agriculture, 2020). Finally, the health outcomes for each developing country were gathered from the World Health Organization (2018). The health outcomes and consumption rate datasets were 2010 data, but the nutritional data was 2020 data. This was a constraint on the accuracy of the results attained. Finally, the population rates were obtained in order to calculate per capita metrics (Data Hub, 2018).

This research aimed to achieve a predictive model for health outcomes in developing nations based on their current consumption rates with the purpose of providing health and government officials the information necessary to focus resources and maximize care.

1. **Hypothesis**

**H0:** Food consumption has no impact on health outcomes in developing nations, **H1:** in nations with lower consumption of high-fat and high-caloric foods, there are more favorable health outcomes, **H2:** nations with greater consumption of high-fat and high-caloric foods result in better health outcomes.

1. **Summary of the Data Analysis Process**

The data first needed to be analyzed to see if it met the assumptions for modeling: normalcy, linear association, homogeneity, residual normalcy, and various statistical thresholds for these metrics. The data met the basic assumptions necessary for modeling.

Once the data met the assumptions, the data was split into two categories: training and validation. The training dataset was used to develop the model and the validation dataset was used to evaluate the model. The first step was to create a binary variable for the health outcomes that coded a one for the upper quartile of the data, and a zero for the remainder. This would allow the model to target food consumption rates that were correlated with much higher rates of NCDs. The next step was analyzing the data for missing values. This was an important step because the logistic modeling procedure in SAS eliminates observations with missing data from the model and could drastically affect the results. In this particular case, there were no missing values.

Dealing with food categories could have caused a high dimensionality and/or quasi-complete separation. In order to avoid these phenomena that could have negatively affected the results of the model, the categorical variable was smoothed into a continuous variable using the smooth weight of evidence technique (*SAS* *Predictive Modeling*, 2020).

A particular statistical method called best subsets selection was implemented to select the best model for each number of variables. Due to a significantly better modeling result, the food category and consumption rate were chosen as the best predictors for NCDs. This fact was confirmed using model fit statistics.

The models for the interaction of these variables as predictors for each NCD were developed. Then the validation datasets were prepared in the same way that the training datasets were prepared in order to produce the best results. The models were used to score the validation datasets and were evaluated on their ability to predict higher rates of NCDs. This study used the Receiver-Operator Curve to measure the ability of the model to classify the rates of each NCD (Narkhede, 2018).

1. **Outline of the Findings**

The findings showed a significant relationship between food categories, consumption rates, nutritional values, and rates of NCDs. The highest predictor for diabetes was preserved milk and milk products, eggs and egg-based products, and preserved vegetables. On the other hand, foods least associated with diabetes were seafood, beer, and pork. One consideration is that many of the developing countries have a state religion that forbids the consumption of alcohol and pork. Additionally, seafood is not as available in regions farther inland, possibly explaining its presence at the bottom of the list.

The highest predictors for malignant neoplasms were flour and cereal, preserved milk and milk products, bread, and other meats. The lowest predictors for malignant neoplasms were beer, frozen fruit, and pork.

The highest predictors for chronic obstructive pulmonary disease were fresh fruit, oil and fats, and soft drinks and juice. The lowest predictors for chronic obstructive pulmonary were wine, beer, and spirits. A glass of alcohol per day has been linked to greater health overall, however lower consumption can also be due to religious reasons.

The highest predictors for cardiovascular disease were fresh vegetables, fresh seafood, and rice. The lowest predictors for cardiovascular disease were jams, eggs and egg-based products, and frozen vegetables.

Overall, these findings assert that fish, fruit, vegetables, cheese, and pasta were overwhelmingly negatively associated with NCDs. Alternatively, rice, bread, cereals and flour, as well as oils and fats have a strong positive relationship with NCDs overall. While some of these foods are more commonly consumed with each meal, another reason could be that their nutritional value is the culprit.

1. **Limitations of Techniques and Tools Used**

The data is limited by the confusion caused by using a smoothed weight of evidence to represent the food category variable. As a result of this fact, the data is difficult to interpret. However, if interpreted with this in mind, the model results can serve as an indicator for emerging nations to shift the focus of their markets to certain food groups that are not correlated with higher rates of NCDs.

Upon examination the original dataset did not depict a high degree of normalcy. However, instead of reducing the diversity and variance of the dataset, the study proceeded with the abnormal data. This could affect the results of the study and another model that adjusts for normalcy should also be considered before drawing conclusions as a way of mitigating this limitation (*SAS Predictive Modeling*, 2020).

Another limitation is the binning of the health outcome variables so that a one would represent the top quartile of the health outcome rates, and the remainder of the rates would be zero. This was selected in order for the study to target possible causes for these extremely high rates. While this method did assist in targeting the upper rates of each health outcome, but a non-binary logit should be examined as well in order to compare the model results and validate the method. The sample dataset could be unrepresentative of the population as a whole and could affect this technique exponentially.

A main limitation of this analysis is the cumulative effect of limitations within each dataset. The food consumption rates are based on the consumption as best estimated by the World Bank Group. Accurate consumption rates were not available due to government entities failing to disclose, as well as the unavailability of rates due to informal markets. This combined with the fact that the median nutrition rates for each food category may not be the most accurate representation of the food category is a major constraint.

1. **Expected Benefits of the Study and Proposed Actions**

There must be many more studies conducted in order to confirm the results of this study and take action with them. However, a significant relationship has been identified, and the particular food groups that are positively or negatively related to the target non-communicable disease should be examined for their potential harmful/beneficial effects on the population at large. Once this study is validated with further research and analysis, this study will be able to advise companies on marketing strategies for foods that are negatively correlated with NCDs. It could also be used to inform governments of which areas of the country are at highest risk for each NCD, based on the food consumption rates in that area. Governments can, in turn, use that information to re-allocate health resources in these at-risk areas and take preventative measures against the diseases.

In summary, there is clearly a significant relationship between food consumption rates, the categories of food being consumed, and the subsequent health outcomes. This relationship should be further analyzed with more accurate data that analyzes more specific food groups and more legitimate nutrition information. This would provide a better model that would assist the resourcing of developing nations. A more in-depth analysis of the specific areas within these countries that are consuming each food would be more useful for these purposes as well.

1. **Sources**

Data Hub. (2018). In *Population Figures by Country*. John Snow Labs. Retrieved July 10, 2020, from https://datahub.io/JohnSnowLabs/population-figures-by-country#readme.

Narkhede, S. (2018, June 26). Understanding AUC - ROC Curve. Medium. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

SAS Institute. (2020). *Predictive Modeling Using Logistic Regression (15.1)*. https://vle.sas.com/course/view.php?id=3472.

U.S. Department of Agriculture. (2020, April). Download FoodData Central Data. Retrieved May 08, 2020, from https://fdc.nal.usda.gov/download-datasets.html.

World Bank Group. (2010). Global Consumption Database. Retrieved May 08, 2020, from http://datatopics.worldbank.org/consumption/product/Rice.

World Health Organization. (2018, April 5). NCD Deaths by Cause and Sex - Data by Country. Retrieved May 08, 2020, from https://apps.who.int/gho/data/view.main.NCDDEATHCAUSESNUMBERv?lang=en.