MACHINE LEARNING AND VIRGINIA'S FOOD DESERTS

ML4VA CATEGORY: GOVERNMENT

Paige Hipes
aph3gx@virginia.edu

Mariam Guirguis
mg6qb@virginia.edu

Avni Garg
ag9wse@virginia.edu

October 7, 2021

1 Motivation

The United States Department of Agriculture (USDA) defines food deserts as low-income areas that have limited access to nutritious and healthy foods. Although the specific mileage varies definition by definition, food deserts are often quantified as low income urban areas further than 1 mile from a large grocery store or supermarket along with low income rural areas further than 10 miles away from a large grocery store or supermarket. According to a report from The Food Trust in 2015, over 1.7 million Virginians, including over 480,000 children, live in lower income communities considered food deserts, and this limited access poses several risks to public health. Those living in food deserts are more likely to develop diabetes, obesity, and cardiovascular disease. Alongside chronic illness positively correlated with living in a food desert, medical bills following disease may impact those living in food deserts financially as well. In order to address Virginia's food desert problem, it is important to understand the situation at hand; how different features affect a location's status as a food desert; and how to predict food desert risk in a certain location. A Virginia without food deserts is an improvement in public health, specifically working towards equity, as the majority of food deserts are home to minority populations.

2 Dataset

We have chosen a dataset from Kaggle.com:

`https://www.kaggle.com/tcrammond/food-access-and-food-deserts?select=food_access_variable_lookup.csv.`

This dataset is comprised of all US Census Bureau defined land "tracts" in the United States with feature information related to demographics of the area and levels of food insecurity. The primary scope of this project is to predict a location's risk level of becoming a food desert based on 10 features, which will be decided through the research process. From our dataset, we have the census tract number which is the primary basis of location, a binary flag that identifies food deserts, median family income for the census tract, and population count, and distance count for grocery stores among other variables. Using this data, we will perform a combination of supervised and unsupervised machine learning methods to further understand the food desert problem and create a prediction model for the likelihood of a person living in a food desert.

Task modeling. We approach the overall task as a regression problem but will use classification throughout the research. For every tract of land, we will predict its share of population that does not have access to a supermarket within 1 mile for urban areas and 10 miles for rural.

Construct train and test data. Because we want to focus on Virginia and generalize onto food deserts within the state, we will subset the full Kaggle dataset of about 72,000 rows to the 1,900 that represent Virginia tracts of land. Additionally, we will have to perform some feature engineering, as the original dataset contains over 100 features. Narrowing down features will allow for greater generalization and prevent our model from overfitting.

3 Related work

Due to the seriousness surrounding the issue of Food Deserts, the U.S. Department of Agriculture conducted a one year study to assess the extent of areas with limited access to affordable and nutritious food and identify characteristics and causes of such areas. From the study, it was determined that existing data and research is insufficient to conclusively determine whether areas with limited access to grocery stores are classified as having inadequate access. However, it was determined that urban core areas with limited food access are characterized by higher levels of racial segregation and greater income inequality. In addition, in small towns and rural areas with limited food access, the most defining characteristic is the lack of transportation infrastructure. Although there was insufficient evidence to confidently correlate food deserts with different factors (such as ethnicity), the analysis still shed light on the issue. The research team used kernel density function to define low-income areas: the household income had to be below 200 percentage of the Federal poverty threshold and for a given population, at least 40 percent of the total population had to meet the poverty criterion. A systematic search criterion was used for each geographic area, which was represented by a grid, which resembles a type of kernel density function. The kernel density function serves to test each grid within the context of adjacent grids for meeting the low-income area criteria. Additional basic and statistical analysis was conducted on the data to attain the results.

Looking at a different example, food deserts were located and evaluated to predict causes of food deserts. For this research, hierarchical clustering was used. This Montreal study was conducted in 2007 with the goal of identifying food deserts in Montreal. After conducting their analysis, they came to the conclusion that proximity (distance to the nearest supermarket), diversity (number of supermarkets within a distance of less than 1000 meters), and variety (in terms of food and prices) are important in identifying food deserts. The research group performed computation of a hierarchical cluster analysis to classify and characterize census tracts (CT) in different groups of CTs with similar levels of the 3 measures aforementioned. The objective of the hierarchical cluster analysis is to locate food deserts and also categorize all CTs in terms of deprivation and accessibility. Given the fact that we have similar data to the research paper, we will also attempt hierarchical clustering to further understand food deserts and the variables that effect it.

4 Intended Experiments

For the purpose of this project, we will be using a classification algorithm (as mentioned above) to classify whether or not an area qualifies as a food desert based on other features of the data set such as income and distance from a grocery store. The official definition of a food desert is being low income and living 1 mile ,for urban, and 10 miles, for rural, away from a grocery store. A particular classifier has not been chosen as we would like to test the various classifiers to determine the appropriate approach for our data set. We plan to test the accuracy of the classifier by checking against the official definition of a food desert. This will help us determine the accuracy of our dataset and aid in determining the importance and relationship between the different variables in our dataset and their effect on food deserts. We are aiming to see if there are any particular factors correlated with food deserts such as a particular group being prone to living in food deserts or whether a certain part of the state is more food insecure than the others. We also want to use regression to predict the share of the populations in the region that are in a food desert. We will be doing feature selection to ensure that we are making a model that has relevant information to make correct classifications about areas being food deserts or not. Additionally, we may run a K-Means clustering algorithm to see whether food deserts are concentrated in one part of the state or if they are spread across the commonwealth. To measure our accuracy for classification of food deserts, we will be using an F1 score. To measure our regression accuracy, we will be using multi fold cross validation to get an RMSE score.

5 Team Contributions:

All of us sat and came up with the topic. We then divided the paragraphs and research. Paige did the motivation and dataset paragraph, Mariam did the Dataset paragraph. Avni did the intended experiments paragraph. We all checked each other's work and met again to go over it and make edits before submission.

References

[1] Apparicio, P., Cloutier, M.-S., amp; Shearmur, R. (2007, February 12). The case of Montréal's missing food deserts: Evaluation of accessibility to Food Supermarkets. International Journal of Health Geographics. Retrieved October 7, 2021, from https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-6-4.

[2] ver Ploeg, M., Breneman, V., Farrigan, T., Hamrick, K., Hopkins, D., Kaufman, P., Lin, B.-H., Nord, M., Smith, T. A., Williams, R., Kinnison, K., Olander, C., Singh, A., amp; Tuckermanty, E. (1970, January 1). Access to

affordable and nutritious food: Measuring and understanding food deserts and their consequences: Report to Congress. AgEcon Search. Retrieved October 7, 2021, from https://ageconsearch.umn.edu/record/292130/.

[3] Waldoks, R., Lang, B., amp; Treering, D. (2015). Food for Every Child: The Need for Healthy Food Financing in Virginia. The Food Trust. Retrieved October 6, 2021, from

http://thefoodtrust.org/uploads/media_items/virginia-mappingfinal-lowres.original.pdf.