Analyzing US County Level Food Deserts' Demographics, Food Access, and Diet Related Disease Prevelance

MORGAN L. FORD, Rensselaer Polytechnic Institute, USA

23.5 million American live in food deserts [4]. Many Americans also live with diet related health conditions. I hope to examine a potential relationship between access to grocery stores and access to healthy foods at grocery stores, the differences between urban and non-urban food deserts, the relationship between population demographics and food deserts, and the relationship between food deserts and the prevalence or mortality rates of diet-related health conditions. I hypothesis that food deserts are less likely to have a variety of healthy foods in their limited food sources compared to areas that are not food deserts. I also hypothesize that food desert communities are more likely to be at risk for diet related health issues than than areas that are not food deserts.

CCS Concepts: • **Mathematics of computing** → *Cluster analysis*; Exploratory data analysis; *Robust regression*; • **Applied computing** → *Sociology*.

Additional Key Words and Phrases: food deserts, random forest, classification, regression

ACM Reference Format:

Morgan L. Ford. 2022. Analyzing US County Level Food Deserts' Demographics, Food Access, and Diet Related Disease Prevelance. In *GOODIT '22: ACM International Conference on Information Technology for Social Good, Sept 7-9, 2022, Limassol, Cyprus.* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

According to the USDA, a food desert is defined as an area that has at least 500 people (or 33%) that live further than 1 mile from the nearest large grocery store for urban areas and 10 miles from the nearest large grocery store for non-urban areas. In addition to this criteria, a an area must also have either a poverty rate $\geq$ 20% or a median family income $\leq$ 80% of the median family income in urban areas and a median family income of 80% of the statewide median family income in non-urban areas [5]. 23.5 million American live in food deserts [4]. Many Americans also live with diet related health conditions, such as diabetes, heart disease, obesity, and colon, breast, and uterine cancers.[3]. I hope to examine a potential relationship between access to grocery stores and access to healthy foods at grocery stores, the differences between urban and non-urban food deserts, the relationship between population demographics and food deserts, and the relationship between food deserts and the prevalence or mortality rates of diet-related health conditions. I hypothesis that food deserts are less likely to have a variety of healthy foods in their limited food sources compared to areas that are not food deserts. I also hypothesize that food desert communities are more likely to be at risk for diet related health issues than than areas that are not food deserts.

2 DATA DESCRIPTION AND EXPLORATORY DATA ANALYTICS

2.1 Data Description

The three data sources were pulled from various interactive maps found online for ease of having population based data.

Manuscript submitted to ACM

The first data source is the USDA's Food Environment Atlas [**?**]. This interactive map shows statistic on food environment indicators and provides a spatial overview of communities access to healthy food. This data set contained more than 280 variables with supplemental information about population and demographic information alongside the food environment data. It was split into 4 main sets: access, assistance, restaurants, and stores.

The second data source is the USDA's Food Access Research Atlas [**?**].This interactive map shows statistics on food access indicators, especially for low income and other census tracts. This data set contains nearly 150 variables, with data about populations and food access. Some of the most important variables in this data set relate to the distance to grocery stores and other food suppliers for the general and subsets of the population.

The third data source is Global Health Data Exchange's (GHDx) US Health Map. This interactive map shows statistics on health trends in the US by county. The data is comprised on many different data sets. The data set most utilized is the county health rankings data, which has data about a variety of factors including life expectancy and more. I also used data on diabetes prevalence, cardiovascular disease mortality rates, and various diet-related cancer mortality rates.

2.2 EDA

First, I examined the food environment and food access data. Figure 1 shows a histogram of the number of different types of food stores per thousand people in a county. Figure 2 shows a scatter plot of the population beyond 1 mile from the nearest supermarket to the poverty rate for census tracts in New York state.

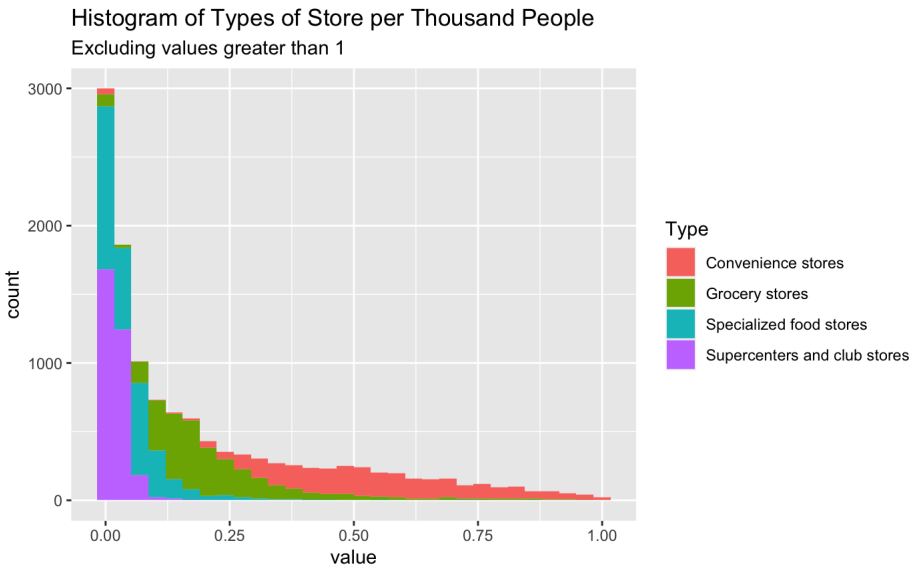

Fig. 1. A histogram of different kinds of store per thousand people.

The distribution of the data from various variables in the food access data can be seen in Figure 3. Besides the Urban factor, all the other variables are pretty positively skewed.

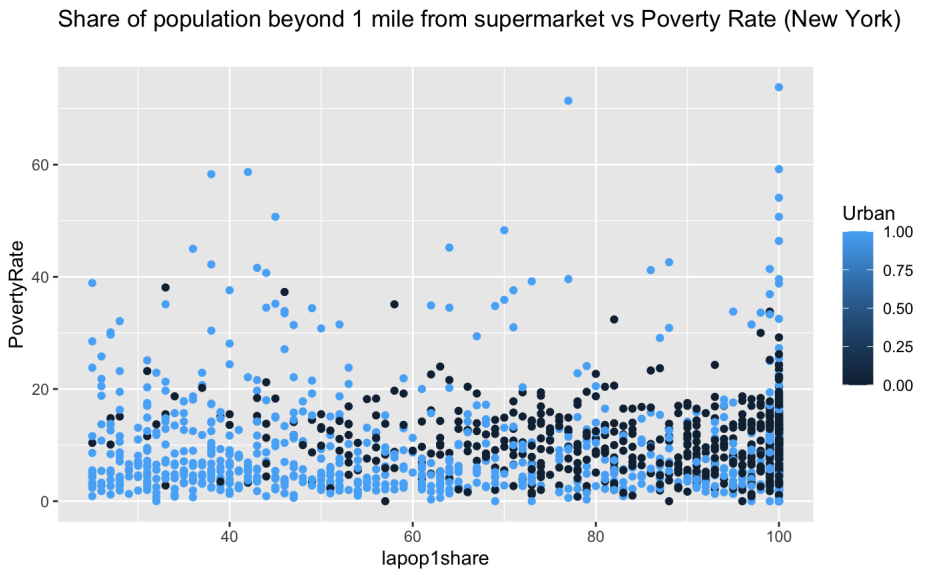

Fig. 2. A plot showing the poverty rate versus the share of population living further than 1 mile to the nearest grocery store.

3 ANALYSIS

Almost all of the data that contained numbers needed to be transformed from characters to numerics. I created a data frame called "state_dict" to help convert state names that were formed as abbreviations (such as AL) to their full names. So when the data was stored with County as one column and State as another, they all became County, State. For the data I had where there was data for individual tracts in a county, such as the data from the Food Access Research Atlas, the data from all of the tracts in a county were combined using the mean into one data point for the entire county.

The first merged data set combined the Food Access Research Atlas and the Food Environment Atlas. Figure 4 show the heatmap of this data frame. It was created by joining the data from each of these sources. It contains columns with the percent of tracts in a county that are food deserts, the percent of tracts in a county that are urban, and the number of various types stores and restaurants per thousand people in a county.

This data was then also split into two separate data frames, one with counties with at least 50% Urban tracts and one with less. When the number of food deserts in a county needed to converted into a factor to classify a county as being a food desert or not, the threshold for classifying it as being a food desert was 33%. When training and testing sets were created, 70% was always used.

The next merged data set combined the Food Access Research Atlas's LA1and10 and County information with the Health Rankings Data Subrankings. The Quartile data was used for classification instead of the ranks in order for there to be fewer possible classes.

The third merged data set combined the Food Access Research Atlas's LA1and10 and County information with data from various sources about certain diet related diseases prevalence in the population or mortality rates. These diseases include: Obesity, Diabetes, Cardiovascular Disease, and Colon, Uterine, and Breast Cancer. This data took extensive cleaning since it came from many different sources, much of this needed to be done straight in Excel. Again, the County

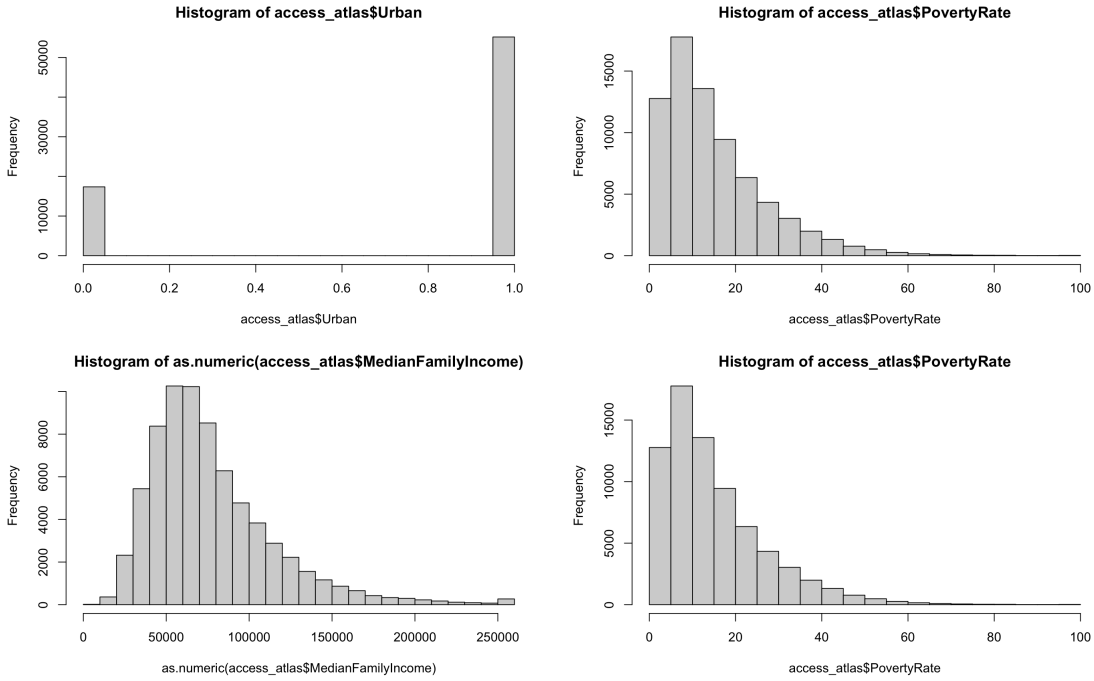

Fig. 3. Histograms of various factors in the food acess data.

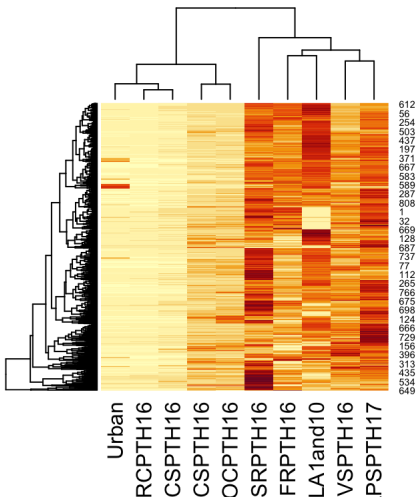

Fig. 4. A heat map of the joined Food Access Research Atlas data and the Food Environment Atlas data.

names needed to be unified. In addition, for the cancer related data, the error needed to be separated from the data points.

4 MODEL DEVELOPMENT AND APPLICATION OF MODELS

A support vector machine model was model to classify if a county that was a food desert or not, using the variable that flagged if a tract that flagged at low access for 1 mile for urban areas and 10 miles for non-urban areas, and the following demographic variables : total population, total housing units, poverty rate, median family income, total low income population, total child population, total senior population, total white population, total Black population, total Asian population, total native Hawaiian or other pacific islander population, total native American and Alaska native population, total multiracial population, total Hispanic population, total housing units without a vehicle, and total housing units receiving SNAP benefits. A Gaussian Radial Basis kernal function was formed and can be approximated as :

$$K(x, x^{'}) = exp(-\frac{||x - x^{'}||^2}{2(0.203)^2})$$

Using 732 support vectors, it performed with 74.55% accuracy. Figure 5 shows the confusion matrix.

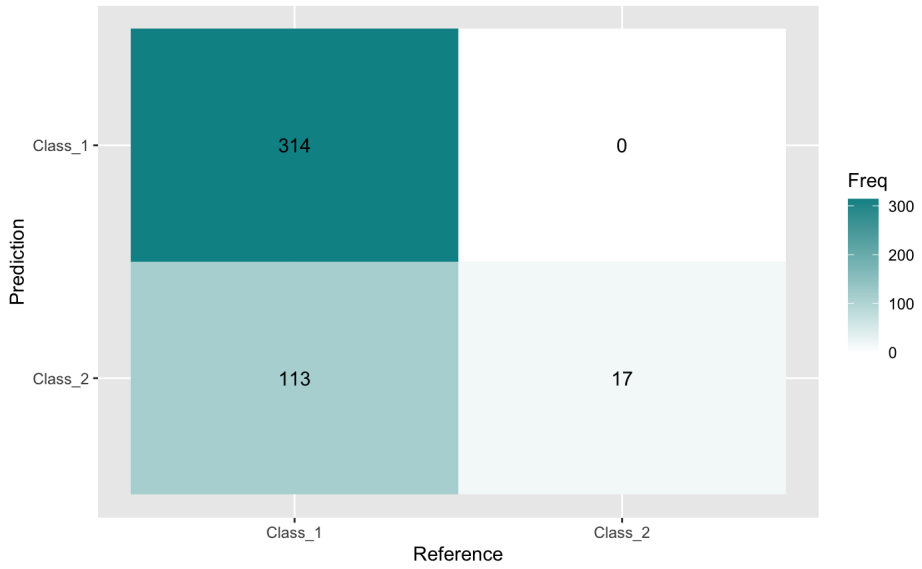

Fig. 5. The confusion matrix for the svm created to classify food deserts based on demographic information.

A linear regression model was created to examine feature importance and to examine the difference between urban and non-urban areas. While these aren't good models by any means, it is interesting to see the results. Table 1 shows the coefficients of said linear models.

Examining this model, in urban areas food deserts have significantly more convenience stores than grocery stores, and even more so for super-centers and club stores. In non-urban areas, there are a lot of super-centers and club stores. The urban linear model is likely overfit due to the lack of data. The non-urban linear model performs with residual standard error of 0.3614 and multiple R-squared of 0.04345. This model may also be overfit, but does perform with relatively low error.

The next goal was to see if it was possible to correctly classify a county as having 33% or more food desert tracts based on the food environment. A random forest classifier was made with 50 trees and 3 variables tried at each split. It

Coefficient (per thousand, 2016)	Urban Estimate	Non-Urban Estimate
Grocery Stores	-15.8089	0.13203
Super-centers and Club Stores	-45.6754	3.03135
Convenience Stores	6.9483	0.10744
Specialty Stores	NA	0.27167
SNAP Authorized Stores	NA	-0.17213
WIC Authorized Stores	NA	0.13954
Fast Food Restaurants	NA	-0.02529
Food Service Restaurants	NA	-0.02453

Table 1. The coefficents for the linear regression model for the number of food deserts in a county compared between urban and non-urban areas.

performed with 97.3% accuracy. Figure 6 shows the error per trees, confusion matrix, factor importance by accuracy, and factor importance by gini plots.

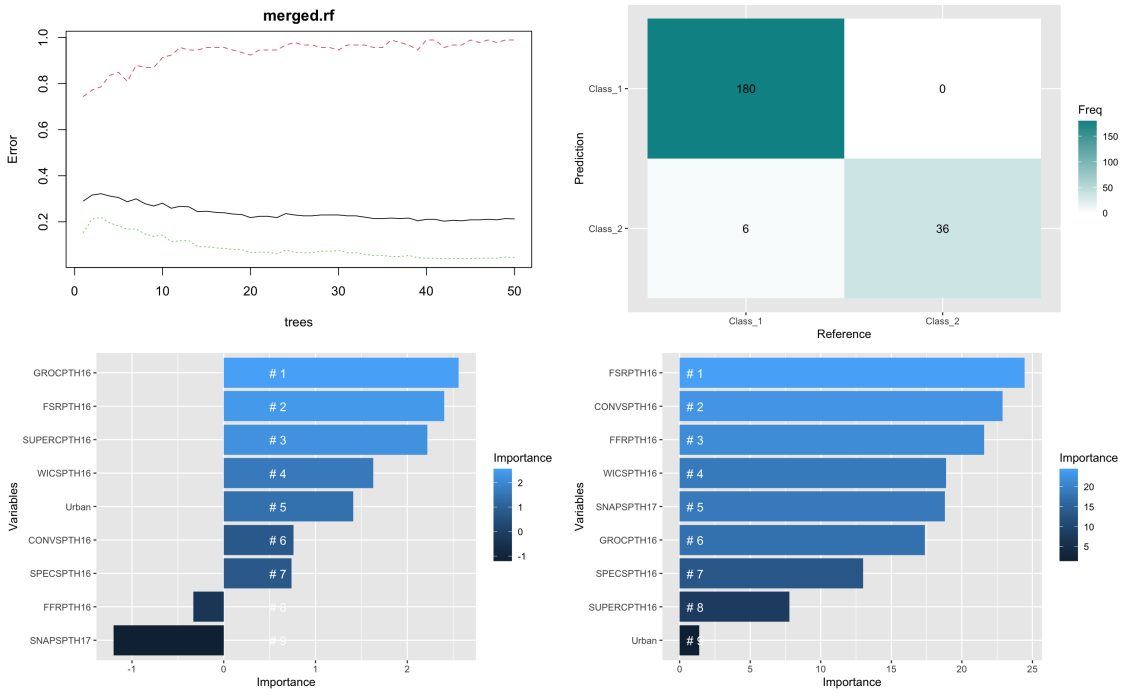

Fig. 6. The plots for the random forest model to classify a county having food deserts.

This model correctly classifies food deserts from the food environment pretty well. It is also interesting to examine that the most important factors for accuracy the number of grocery stores, food service restaurants, and convenience stores.

Next quality of life factors were examined. Two knn models were created, one examining the length of life based on food deserts in a county and the other examining quality of life. The factor were created to classify which quartile

of the counties it's rank was. The length of life knn model performed with 72.3% accuracy and the quality of life knn model performed with 94.15% accuracy. The confusion matrices are seen in Figure 7.

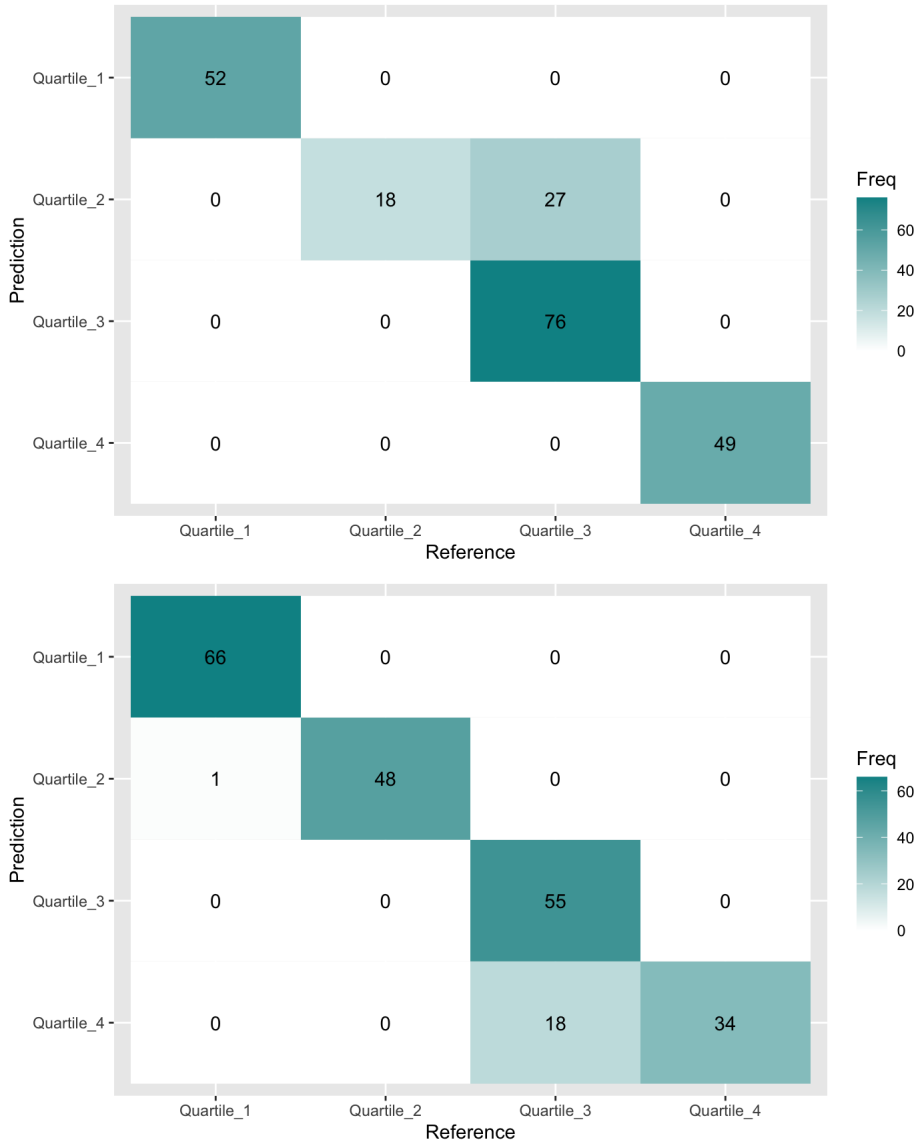

Fig. 7. The confusion matrix for knn models for classifying length and quality of life.

Next, a random forest model was created to see if it could correctly classify an area as a county with 33% or more food desert tracts based on diet related health prevalence and mortality rates. It performs rather well with 94.1% accuracy. Figure 8 shows the error, confusion matrix, accuracy, and gini plots respectively. Interestingly, the most important factors are colon cancer and cardiovascular disease.

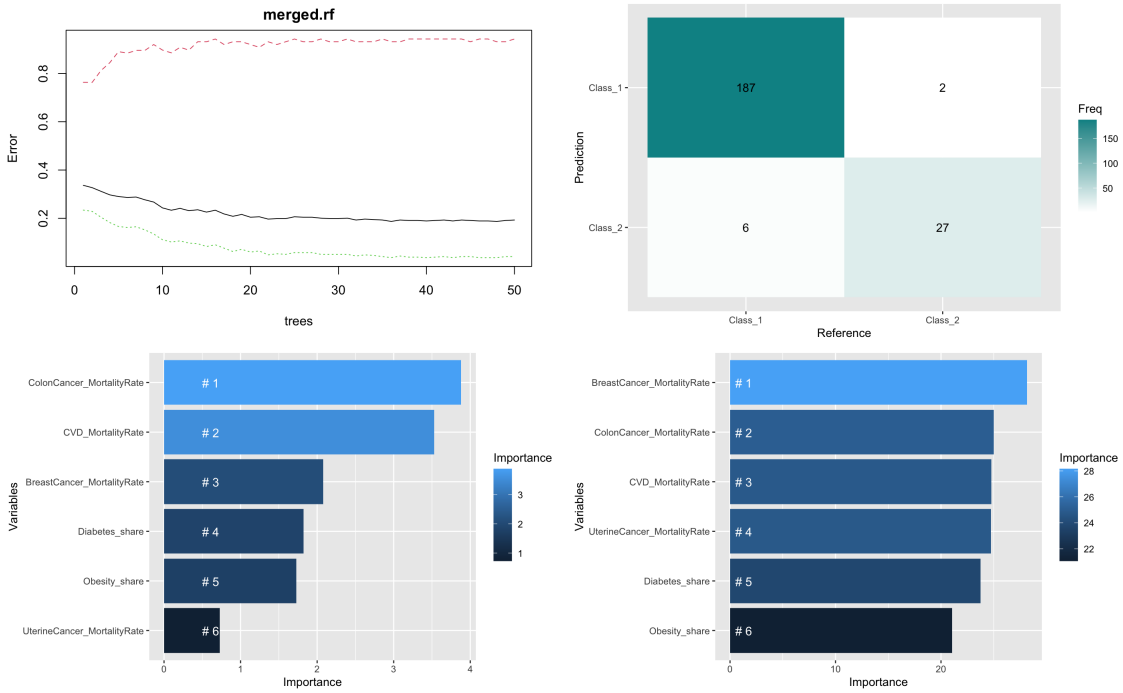

Fig. 8. The plots from the random forest model to classify a county as having food deserts based on diet-related diseases.

5 CONCLUSION AND DISCUSSION

Testing my first hypothesis, I run a welch two sample t-test to see if the mean number of grocery stores is different between areas where 33% of the tracts in a county are food deserts or not. The p-value is 0.4352, so the null hypothesis, that the difference between the means is not equal to 0, is failed to be rejected. Surprisingly, the number of grocery stores per thousand is actually higher in food desert areas. Testing my second hypothesis, I run a welch two sample t-test to see if the affects of various diet related diseases are different in food desert areas. The null hypothesis is rejected for obesity, but failed to be rejected for diabetes and cardiovascular disease. For the latter, the mean in higher in non-food desert areas. While these t-tests disprove my hypothesis, I believe that the information gained by creating these models is still useful. The colon cancer, breast cancer, and cardiovascular cancer mortality rates were found to be the most important factors while classifying food deserts based on disease rates. My models for classifying food deserts based on food access and disease rates also performed relatively well. I believe that if these models were to be fine tuned and expanded upon with larger datasets, a clearer pattern could be seen.

REFERENCES

[1])]FARS Economic Research Service (ERS). [n. d.]. Food Access Research Atlas. https://www.ers.usda.gov/data-products/food-access-research-atlas/
[2])]FEA Economic Research Service (ERS). [n. d.]. Food Environment Atlas. https://www.ers.usda.gov/data-products/food-environment-atlas/
[3] National Center for Chronic Disease Prevention and Health Promotion. 2021. Poor Nutrition. https://www.cdc.gov/chronicdisease/resources/publications/factsheets/nutrition.htm
[4] U.S. Department of Agriculture. 2009. Access to Affordable and Nutritous Food: Measuring and Understanding Food Deserts and Their Consequences - Report to Congress. *USDA, Economic Research Service* AP-036 (June 2009). http://www.ers.usda.gov/Publications/AP/AP036/

[5] Michele Ver Ploeg Paula Dutko and Tracey Farrigan. 2012. Characteristics and Influential Factors of Food Deserts. *U.S. Department of Agriculture, Economic Research Service* ERR-140 (Aug. 2012). https://www.ers.usda.gov/webdocs/publications/45014/30940_err140.pdf