Food Desert Data Mining Project

IST 707
Winter 2022

Miranda Braman

Kaitlyn Keebaugh

Matthew Pergolski

Victor Yamaykin

Table of Contents

Introduction

Food deserts are indicators of the increasing wealth gap between blue-collar and white-collar versions of the United States of America. Some areas have three or more grocery stores in a five-mile radius; others have one or none. There is a responsibility to understand what areas are at risk of becoming food deserts in the future.

The Healthy Food Financing Initiative (HFFI) defines food deserts as low-income and low-access census tract areas. Low-income means a poverty rate of more than 20% or below 80% of the statewide median family income. Low access means at least 500 people, and at least 33% of the population lives more than 1 mile from a large grocery store (10 miles, in the case of rural areas).[i]

One of the problems with these definitions is that a “grocery store” can include convenience and dollar stores. They are also vulnerable to maintenance issues; for example, the Family Dollar in York, Alabama, was forced to close temporarily after inspectors found a thousand rodents at one of its largest warehouses. In a National Public Radio (NPR) report, the mayor of York, Willie Lake, explained that dollar stores are “a double-edged sword because grocery stores have trouble competing with their low prices.”[ii] It is increasingly clear that the reliance on cheap and convenient options is not the most sustainable route moving forward.

This analysis sets out to answer one main question in terms of economic and racial justice:

- Can classification help to identify food deserts and learn more about the surrounding communities?

The Data

The United States Department of Agriculture uses data from the census every five years to build their Food Access Research Atlas. There are labeled data for 2006 but not for 2010, 2015, or 2019. The main goal is to examine the critical characteristics of counties labeled as food deserts.

For example, in Alabama, York is in Sumter County, which is labeled as not having a food desert. However, it is a low-income tract, and 3 out of 4 census tracts have low access to a grocery store by vehicle.

One of the limitations of the data set is missing nutritional information in relation to the nearest grocery store. Even places labeled as not an official ‘food desert’ can suffer from a lack of fresh produce and meat options.

With the data sets for 2006 and 2019 combined, there are over 160 attributes. In exploratory data analysis, ten points were chosen, including the target for food desert.

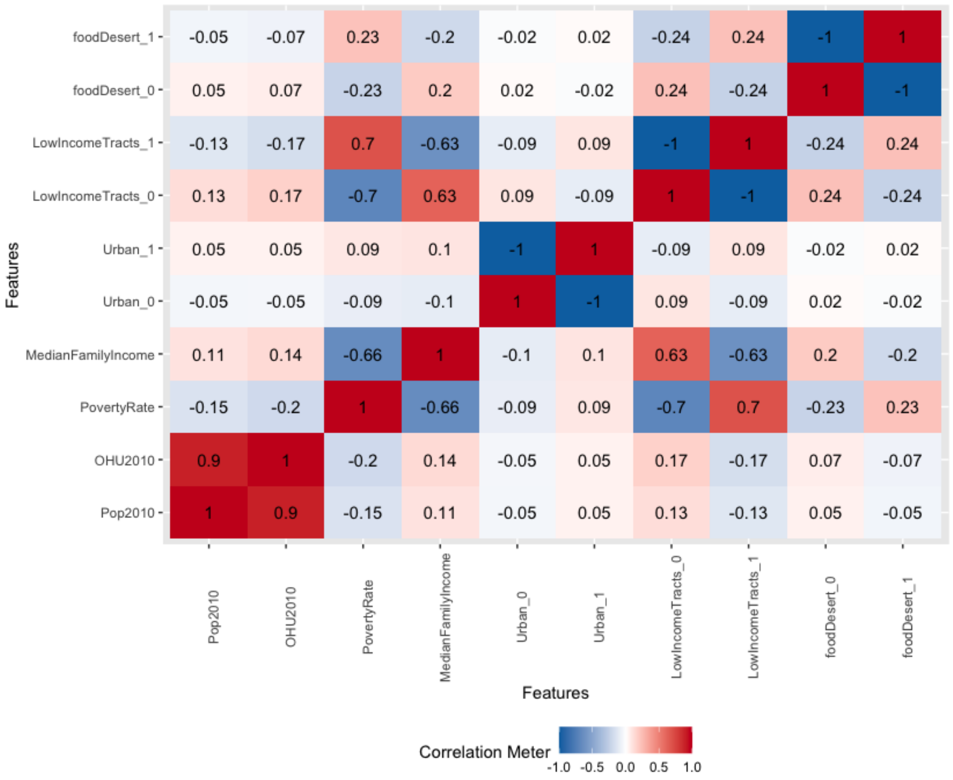
- Census Tract ID
- State

- County
- Population in 2010
- Occupied Housing Units (OHU) in 2010
- Urban Area Indicator
- Low-Income Indicator
- Poverty Rate
- Median Family Income
- Food Desert Indicator

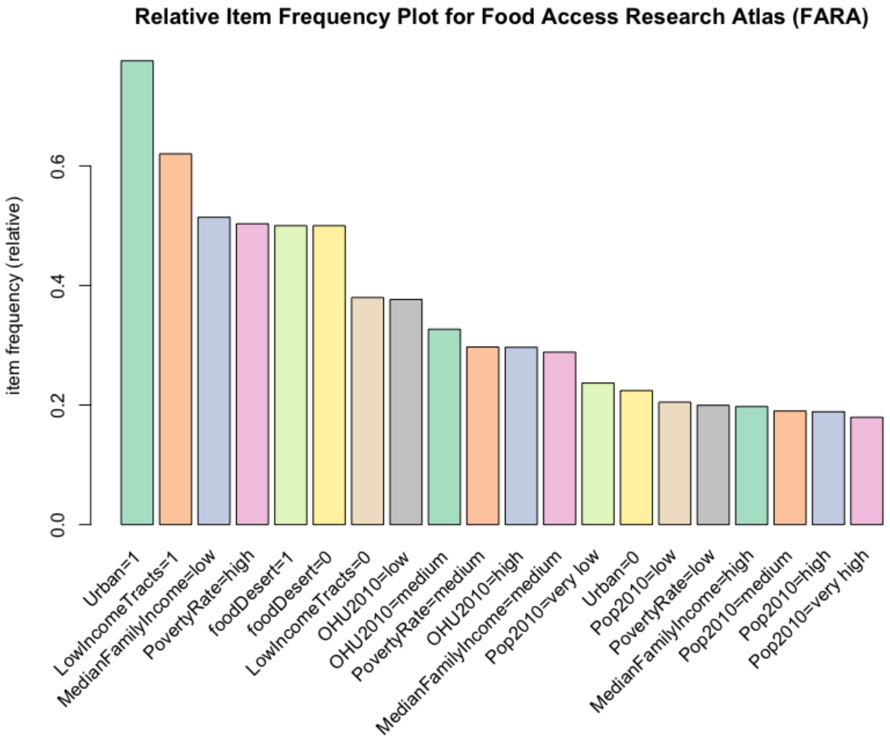
There were several methods of cleaning the data for each classification algorithm. These will be explained during each of the model breakdowns.

Exploratory Data Analysis

The correlation matrix indicates there are strong links between food deserts and low-income areas along with high poverty rate.



The relative item frequency plot demonstrates that food deserts occur in urban areas and where the median family income is low (i.e., $45k and below).



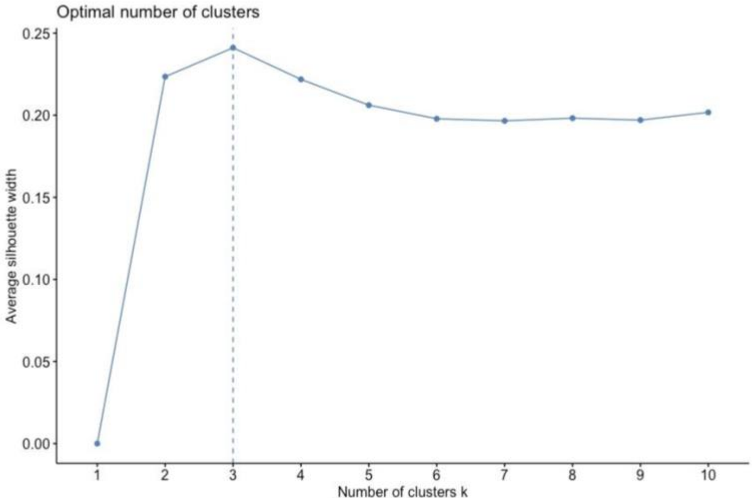
Models

Clustering

Continuing the exploratory analysis section, the unsupervised learning method of ‘clustering’ was chosen to determine groupings/subsets within the ‘Food Desert’ data set. Ideally, various states would be grouped to determine possible similarities, meaningful structure, generative features, etc.

To begin the clustering process, the data set needed to be transformed in terms of aggregating observations by the ‘State’ variable. State data was then converted from an essential column/variable to a row name of the data frame. Non-numeric variables were then omitted from the analysis (i.e., ‘State,’ ‘CensusTract,’ and ‘County’).

An additional analysis was conducted to determine the optimal number of clusters to use for the data. Overall, the ‘Average Silhouette Method’ and its respective R/RStudio functions calculated the ‘optimal’ number of clusters (i.e., 3) via the ‘fviz_nbclust’ function from the ‘factoextra’ package.

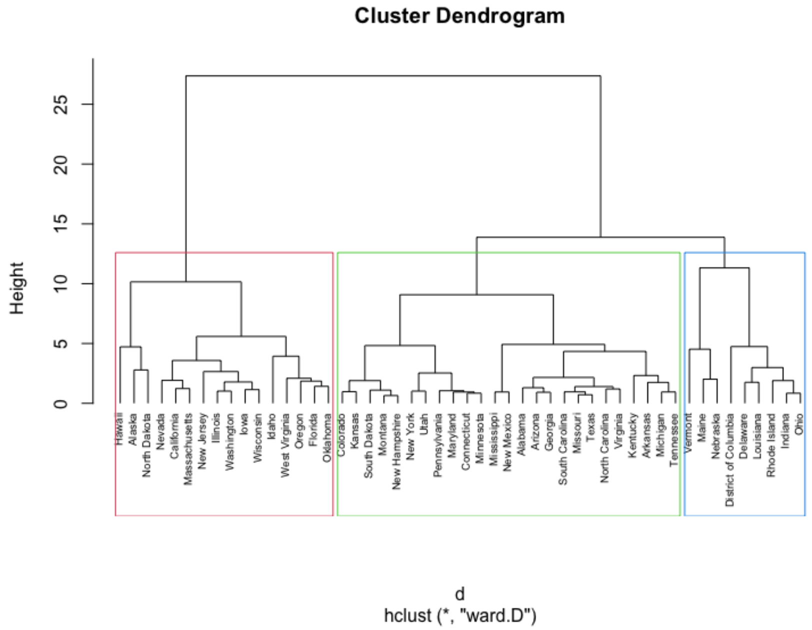


Following this munging/preliminary step in the process, various models and functions were prepped for use within the analysis section of the overall ‘clustering’ topic; for the hierarchical aspect, ‘agnes’ and ‘hclust’ algorithms were chosen. Before running the respective functions, a general method was determined before use. Available options for these methods of determining distance are listed as follows:

- Average
- Single
- Complete
- Ward

To determine which method would serve best for the analysis, a custom function was developed to generate, compute, and compare agglomerative coefficients (which conveys overall quality and fit for the clustering algorithm). In ranking from best to worst, the ‘ward’ method indicated ~0.88, followed by ‘complete,’ ‘average,’ and ‘single, which generated ~0.85, ~0.73, and ~0.60, respectively.

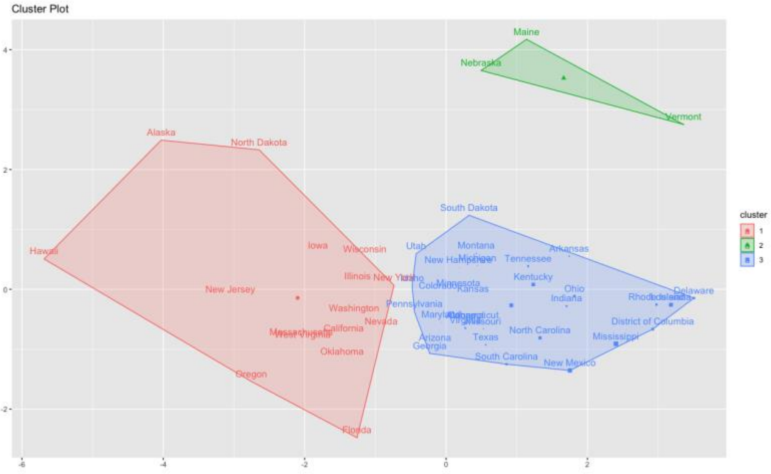
Since ‘ward’ yielded the highest agglomerative coefficient value, this method was chosen for the ‘agnes’ and ‘hclust’ algorithms. By keeping 3 clusters in mind, the following dendrogram was generated (with state information shown near the bottom of the figure):



Lastly, the ‘K-Means’ algorithm was also chosen as a viable function for the ‘clustering’ topic. The three (3) overall clusters consisted of 16, 3, and 31 observations (which equate to 50 – the total number of US states). The ‘Sum of Squares’ values by cluster equate to the following:

- ~75.7
- ~11.5
- ~108.9

A visualization also accompanies the algorithm via the ‘fviz_cluster’ function of the same ‘factoextra’ package introduced earlier in this section.



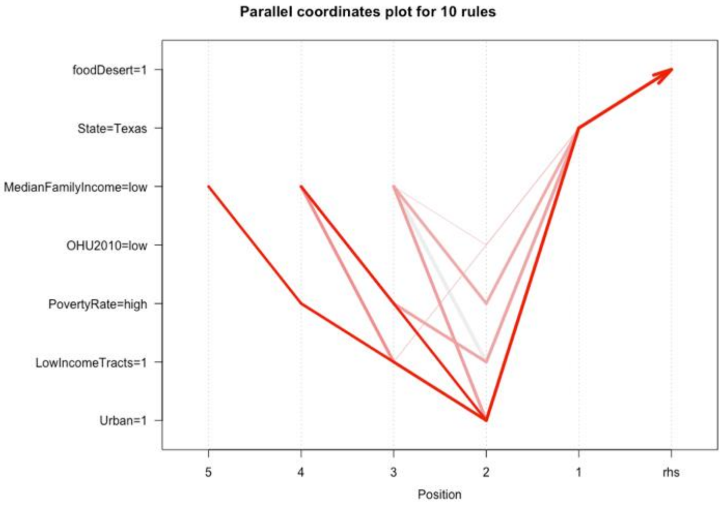
As shown in the Cluster Plot, three (3) distinctive clusters contain US state information. The larger the geometric point, the higher the mean poverty rate for the state. These subsets of states may lead to valuable classification in the latter stages of the report when supervised learning methods are introduced to the audience. At this point in the analysis, it may be viable to preliminarily confirm clusters with higher poverty rates obtain positive correlation with food deserts.

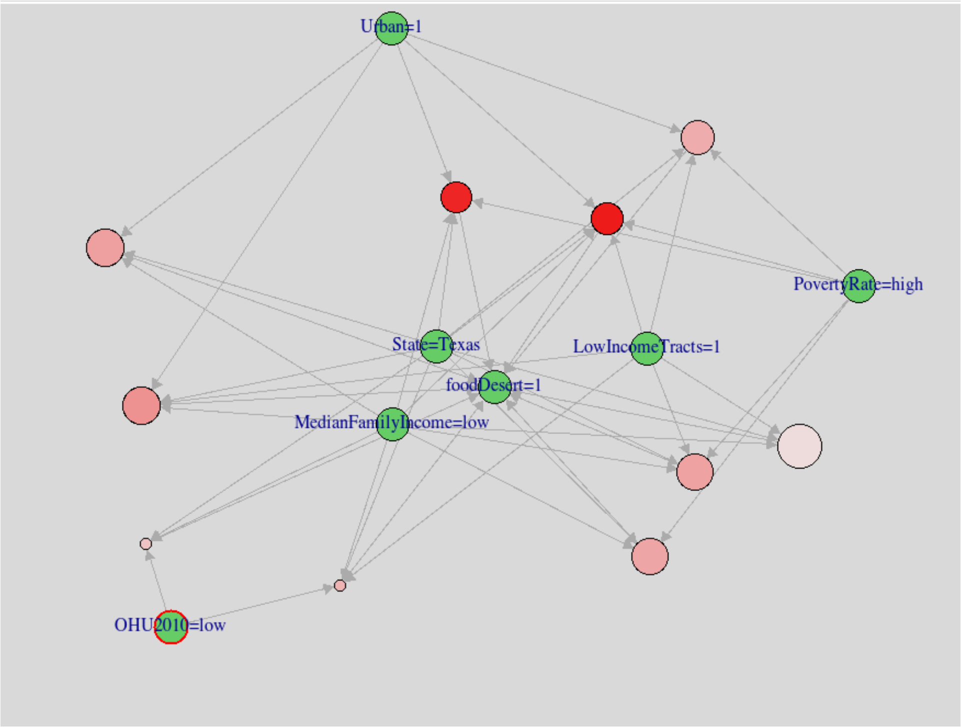
Association Rule Mining (ARM)

To use the ‘apriori’ algorithm, some of the attributes needed to be discretized; in particular, the data for population, occupied housing units, poverty rate, and median family income was put in categories such as ‘low,’ ‘medium,’ and ‘high.’ Another issue was the imbalanced data set. The ‘over’ sampling method helped to retrieve a similar number of observations for food desert classification. The rules were sorted by support, lift, and confidence. The best results were achieved with support = 0.25, confidence = 0.7, and minimum length = 3.

Below is an example of a rule with high confidence and lift. It provides information that a food desert is 1.6 times more likely to be in an area with a high poverty rate and low median family income (possibly in an urban census tract). The parallel coordinates plot does a better job of indicating the strength of the link between the attributes and the presence of a food desert. However, the network diagram points out that the level of occupied housing units has a weaker link to the presence of a food desert.

Left-hand side	Right-hand side	Support	Confidence	Coverage	Lift	Count
State=Texas, Urban=1, LowIncomeTracts=1, PovertyRate=high, MedianFamilyIncome=low	foodDesert =1	0.039	0.82	0.047	1.6	5203

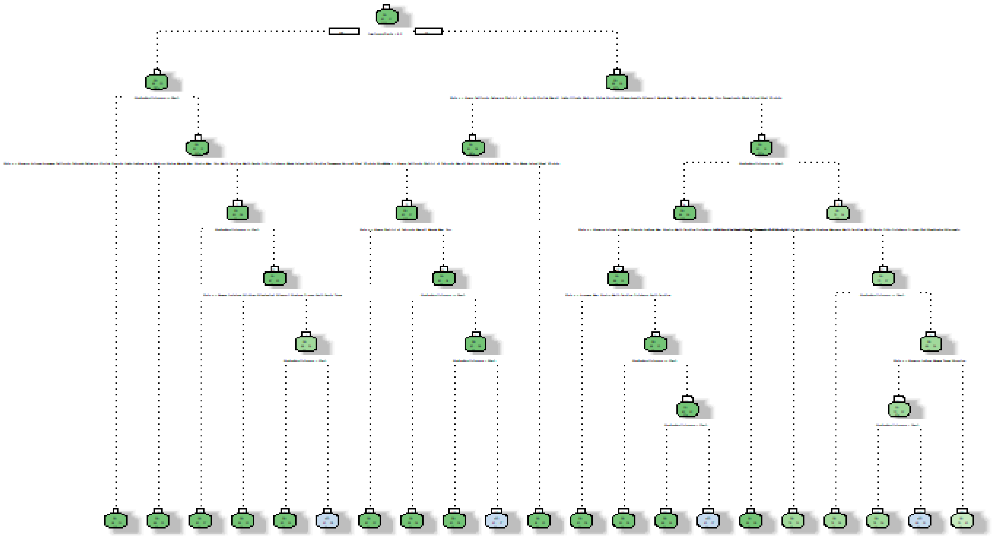




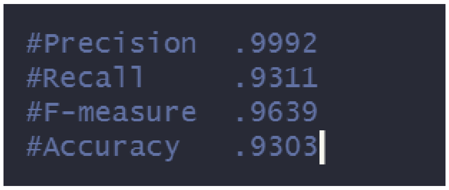
Decision Trees

Decision trees are considered ‘supervised’ learning methods since they build a model using a training data set and subsequently predict outcomes based on a separate testing data set. These decision trees are the classification type because they decide whether an area is a food desert or not. With each decision tree, the functions ‘rsq.rpart,’ ‘plotcp,’ and ‘printcp’ were used to help determine the ideal size of the tree and ‘cp’ value. Since this data set is so extensive, only some variables were chosen for splitting into training/testing sets. They include ‘Median Family Income,’ ‘Low Income Tract,’ and ‘States.’ The data was then shuffled and split 80/20 into training and testing sets, respectively.

Without scaling down the variables and the function, this tree became overly complex and started overfitting. The first tree took about 5-6 minutes to create due to the large number of records – 14,000 in total. The accuracy was very high (93.3%), but this model had to be tested a through additional attempts for validation. To try and prevent the overfitting on subsequent attempts, a set seed of 341 was used, and a max-depth control function was conducted to ensure the most extended tree branch was only seven (7) nodes long. This produced great accuracy, precision, and recall. The following decision trees are smaller and more specific, which would provide a better assumption for the concise data prediction.

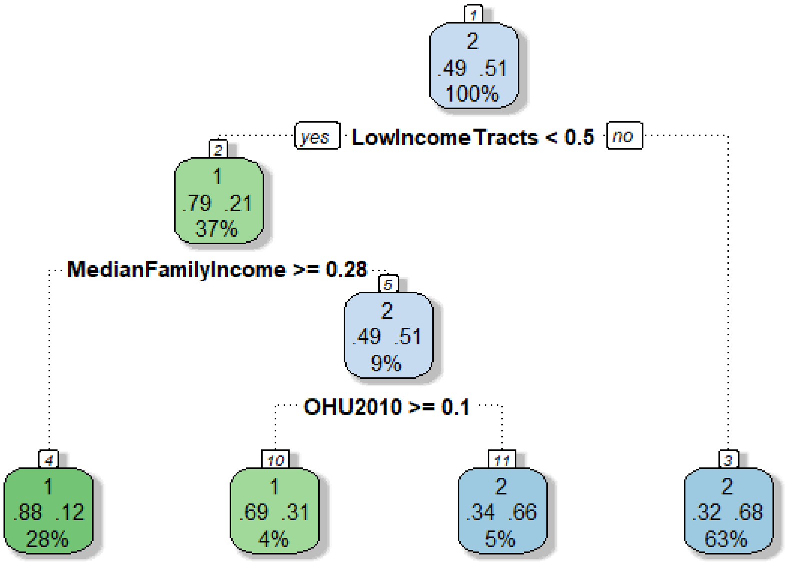


Rattle 2022-Feb-17 15:54:53 kkeeb

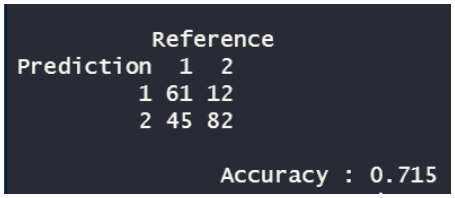


```
#Precision   .9992
#Recall      .9311
#F-measure   .9639
#Accuracy    .9303
```

The second and third decision trees used a much smaller data set containing only 1,000 records and fewer columns. The variables used to determine a food desert in this case were ‘Median Family Income,’ ‘OHU 2010,’ ‘Poverty Rate,’ ‘Low-Income Tracts,’ and ‘Urban.’ Moreover, ‘Median Family Income,’ ‘OHU2010,’ and ‘Poverty Rate’ were normalized so that all the variables would have the same scale from 0-1. This step ensures that the variables have relatively the same weights, whereas if they were not normalized, they would skew the results for many of the prediction methods. The second tree had parameters of cp = 0.013 and a maximum depth of 4. This decision tree produced an accuracy of 71%.

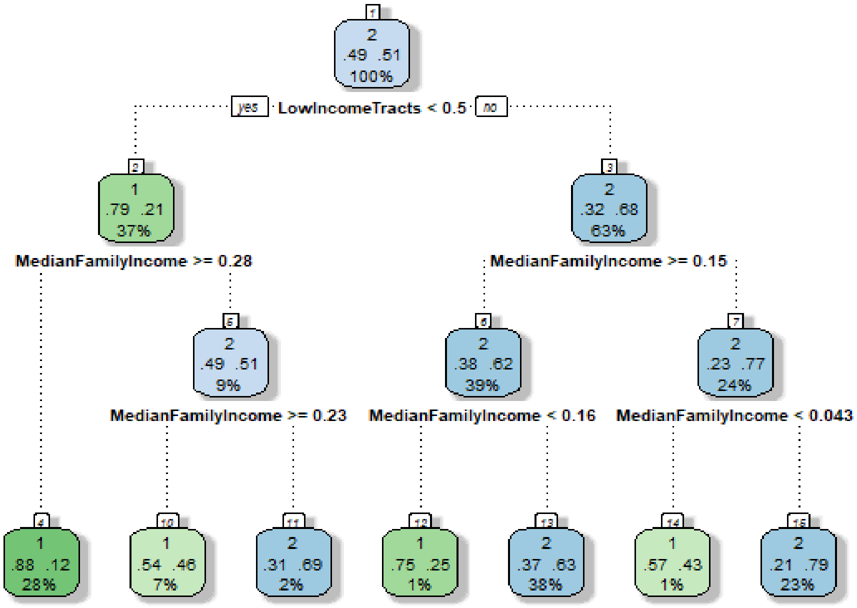


Rattle 2022-Mar-15 19:09:24 kkeeb

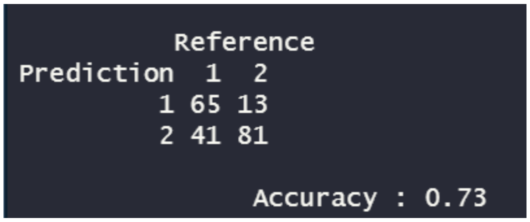


```
          Reference
Prediction  1  2
         1 61 12
         2 45 82

               Accuracy : 0.715
```

Finally, the third decision tree was even smaller. It focused on the variables ‘Low-Income Tracts’ and ‘Median Family Income.’ The parameters were cp = 0.005 and maximum depth of 4. This tree produced an accuracy of 73% – meaning the first decision tree was most likely overfitting due of the number of data/observations.



Rattle 2022-Mar-15 19:27:04 kkeeb

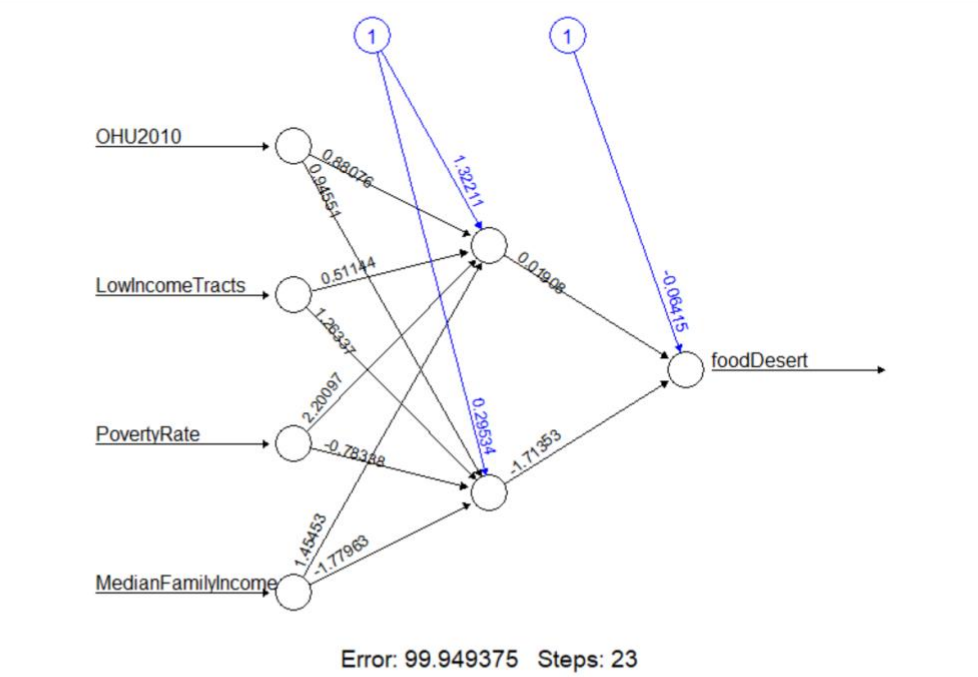


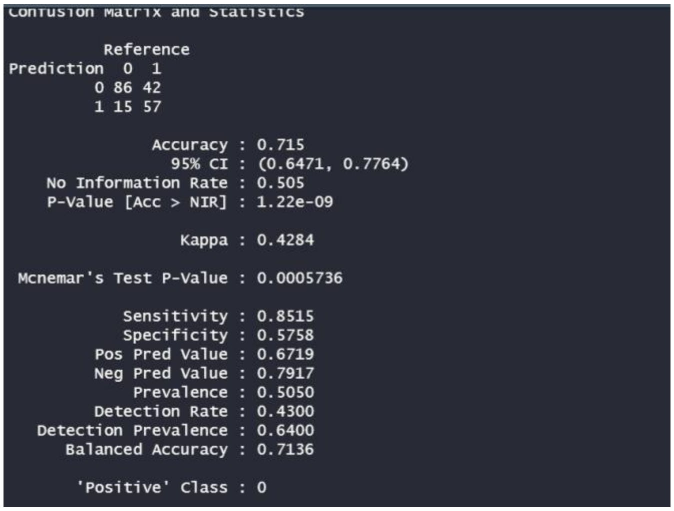
```
          Reference
Prediction  1  2
         1 65 13
         2 41 81

               Accuracy : 0.73
```

Neural Networks

Neural networks are machine learning models where the machine learns to perform a task by analyzing training examples. The weights and coefficients are initially random values while the network is trained. Neural networks can adapt to changing input, so the network generates the best possible result without redesigning the output criteria. The 'neuron' collects and classifies information according to the neural network functions. Hidden layers then fine-tune the input weights until the margin of error is minimal.

The function used for this neural network is 'sigmoid' because it calculates output values between 0 and 1. The variables were narrowed down to 'OHU2010,' 'Median Family Income,' 'Poverty Rate,' and 'Low-Income Tract.' Subsequently, the neural network received the training data as input and rendered it through different hidden layers to pinpoint the best accuracy value. The overall outcome variable contained many numbers between 0 and 1, so the function 'ifelse' was referenced to round the numbers from above 0.5 to 1 and below 0.5 to 0, respectively. In summary, the accuracy for the neural network predicting a food desert was 71.5%.

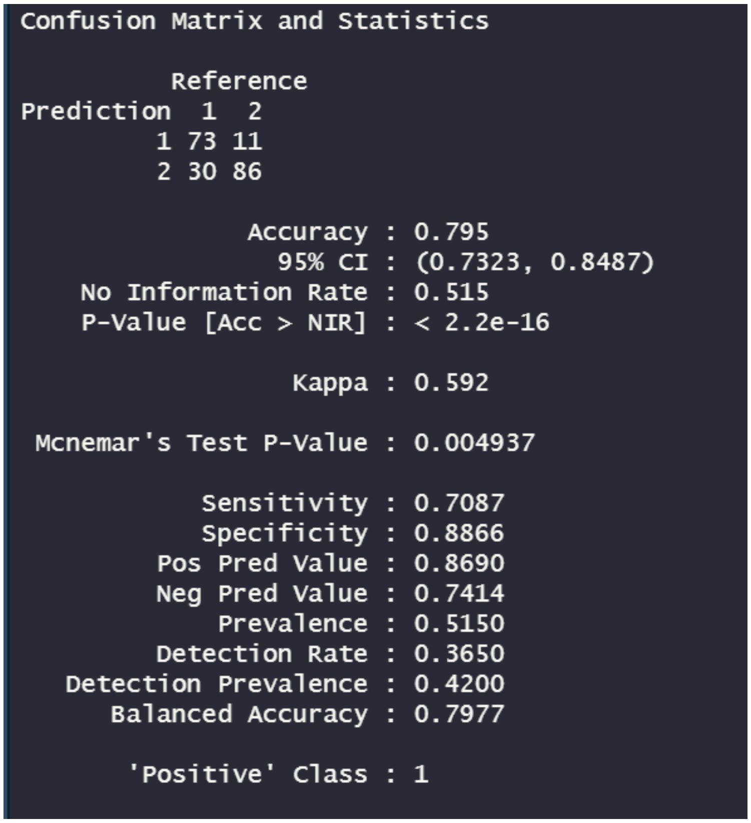




```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 86 42
         1 15 57

               Accuracy : 0.715
                 95% CI : (0.6471, 0.7764)
    No Information Rate : 0.505
    P-Value [Acc > NIR] : 1.22e-09

                  Kappa : 0.4284

 Mcnemar's Test P-Value : 0.0005736

            Sensitivity : 0.8515
            Specificity : 0.5758
         Pos Pred Value : 0.6719
         Neg Pred Value : 0.7917
             Prevalence : 0.5050
         Detection Rate : 0.4300
   Detection Prevalence : 0.6400
      Balanced Accuracy : 0.7136

       'Positive' Class : 0
```

KNN

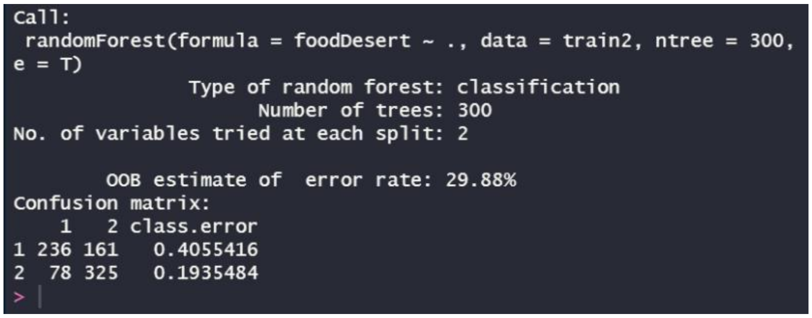
K Nearest Neighbor (KNN) is another machine learning classification method used in the analysis. The ‘training’ and ‘validation’ error rates are two parameters needed to access different K-values. The error rate initially decreases and reaches a minimum; after the minimum point, it increases with an increasing ‘K’ value and conveys a positive correlation. The testing and training data sets had to be adjusted for KNN. These new datasets only included the variables ‘Food Desert,’ ‘Median Family Income,’ ‘OHU 2010,’ ‘Poverty Rate,’ ‘Low-Income Tracts,’ and ‘Urban.’ The KNN algorithm was called with ‘K’ set to 5, 10, 15, 27, 28, and 29. Ultimately, K = 28 produced the highest accuracy of 79.5%.



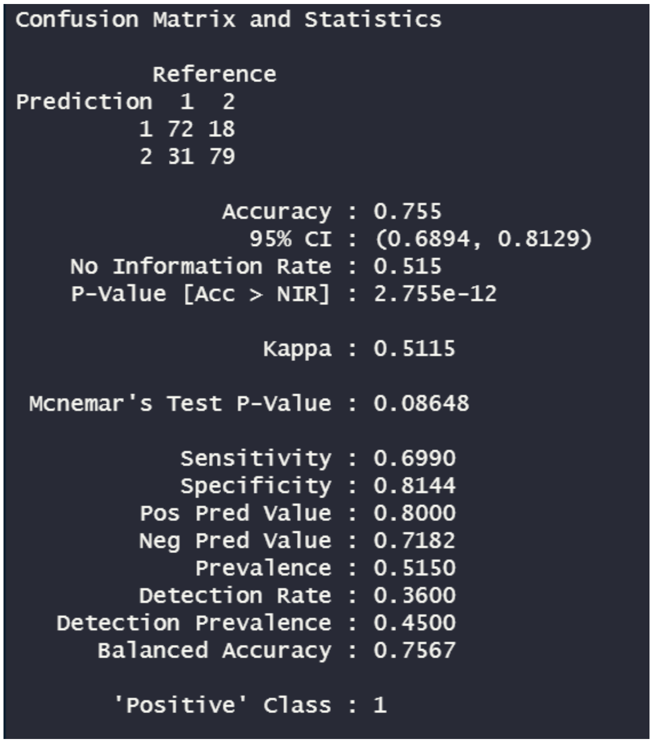
```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1 73 11
         2 30 86

               Accuracy : 0.795
                 95% CI : (0.7323, 0.8487)
    No Information Rate : 0.515
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.592

 Mcnemar's Test P-Value : 0.004937

            Sensitivity : 0.7087
            Specificity : 0.8866
         Pos Pred Value : 0.8690
         Neg Pred Value : 0.7414
             Prevalence : 0.5150
         Detection Rate : 0.3650
   Detection Prevalence : 0.4200
      Balanced Accuracy : 0.7977

       'Positive' Class : 1
```

Random Forest

‘Random Forest’ is the final classification method used to predict food deserts. This method is generally viable because it avoids overfitting, deals with many features, and identifies important variables. Many uncorrelated trees are created using a subset of the data and are subsequently used to identify the most common output. The ‘test’ and ‘train’ data sets had to be adjusted by turning the ‘Food Desert’ variable into a factor – otherwise, the algorithm would run the model as regression instead of classification. The parameters were set at 300 trees, which gave an ‘OOB’ error rate of 29.88%, which is not optimal, but was the best this model could provide. The test data set was then initialized, and the ‘Random Forest’ algorithm generated an accuracy percentage of 75.5%.



```
Call:
 randomForest(formula = foodDesert ~ ., data = train2, ntree = 300,
e = T)
               Type of random forest: classification
                     Number of trees: 300
No. of variables tried at each split: 2

        OOB estimate of  error rate: 29.88%
Confusion matrix:
    1   2 class.error
1 236 161   0.4055416
2  78 325   0.1935484
> |
```



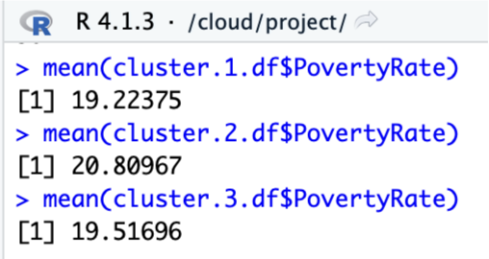
```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1 72 18
         2 31 79

               Accuracy : 0.755
                 95% CI : (0.6894, 0.8129)
    No Information Rate : 0.515
    P-Value [Acc > NIR] : 2.755e-12

                  Kappa : 0.5115

 Mcnemar's Test P-Value : 0.08648

            Sensitivity : 0.6990
            Specificity : 0.8144
         Pos Pred Value : 0.8000
         Neg Pred Value : 0.7182
             Prevalence : 0.5150
         Detection Rate : 0.3600
   Detection Prevalence : 0.4500
      Balanced Accuracy : 0.7567

       'Positive' Class : 1
```

Conclusions

With this oversimplified model to predict food deserts, there were several tradeoffs. By using the predictions from 2006 and imposing them on 2019, there were a disproportionate number of census tracts not classified as food deserts. The high variance and low bias led to overfitting in several of the models. It is extraordinary to have accuracy over 70% in machine learning for social sciences; however, the methods provided in this analysis could be used with other variables (such as percentage of ethnic population with low access to a grocery store within one mile for urban areas and 10 miles for rural).

The association rule mining provided a glimpse into the states that might have the largest concentration of food deserts. There were several rules found for Texas, Indiana, and Tennessee. These states also show up in cluster group 2 (indicated in blue on the Cluster Plot). There is high confidence and lift for these rules that demonstrate low income and high poverty rate are strong indicators of a food desert.

The clustering algorithm indicated a result of three (3) overall clusters of states based of similarities to one another via the 'ward' distance measure. When comparing mean poverty rates across these three (3) groups of states, the following was discovered:



```
> mean(cluster.1.df$PovertyRate)
[1] 19.22375
> mean(cluster.2.df$PovertyRate)
[1] 20.80967
> mean(cluster.3.df$PovertyRate)
[1] 19.51696
```

Overall, cluster group number 2 obtained the highest poverty rate of ~21%, indicating the most likely food desert zone, as further confirmed in the supervised learning section of this document. Basic 'takeaways' for the audience in reading this information is pushing for additional funding from political leaders to ensure all obtain easy access to food, housing, and safe transportation. A collective effort from both the public and private sector will be needed to improve the situation in all three clusters mentioned above.

The results of the classification algorithms confirm that the low-income and high poverty rates are strong indicators of a food desert. All the classification models produced almost the same accuracy. KNN and Random Forest gave the highest percentages of 79.5% and 75.5%, but the others were not far behind at 71.5%, 71.5%, and 73%, respectively. Other variables could be added, or a bigger dataset could be tested in the models. However, these percentages still provide a viable overview of where food deserts are ultimately located. The government could reassess the overall quality of the data to ensure fully accurate information is provided for future similar analysis.

The encouraging aspect of these models is the many parameters eligible for modification to improve further predictions in the future. When building decision trees, the number of nodes, leaves, splits, and overall size of the tree can be adjusted, for example. As seen above, the change in 'cp' and other variables aid in the increase of accuracy percentages, but also larger trees do not necessarily mean they are of higher quality. Some errors are overfitting or underfitting. Neural Networks can be manipulated with hidden nodes, changing the number of variables, or changing the weights of those variables. In this case, the lowest number of hidden nodes constructed the best accuracy percentage. KNN has additional functions available that indicate the ideal 'K' value for a data set. The 25-30 range generated almost the same confusion matrix metrics – but in the end, the optimal value was 28. 'Random Forest' parameters can also be changed, such as the number of trees to generate and the number of variables used in each tree. Put simply, building classification models assist in prediction investigation efforts because of the many features that can be changed.

Classification models are usually measured by accuracy, which is how many correct predictions were generated in the model output. A better way to determine the overall quality of a model is to first create a baseline and compare results with ongoing attempts. Other measures are 'precision' and 'accuracy,' which may be higher than just 'accuracy' alone, and this provides a better explanation of the models' inner workings in this analysis. In the next iteration of future research, there will be additional attributes related to low access and ethnicity. The data sets used in these revisions could be 2,000, 5,000, and 10,000 observations (instead of the 1,000 used in this research) which may render more accurate results. It will be wise for the government track these changes in a time-series analysis.

[i] *USDA ERS - Documentation*. (2017). Usda.gov. https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/

[ii] *A rodent infestation shut down Family Dollar stores. How one Alabama town is coping*. (2022, February 26). NPR.org. https://www.npr.org/2022/02/26/1083217868/a-rodent-infestation-shut-down-family-dollar-stores-how-one-alabama-town-is-copi