# DATA-390B: Model and Data Usage (Version 1)

Maha Mapara

17th November, 2020

## 1 Data sets

### 1.1 Food Research Atlas Data

The Food Research Data Atlas by the US Department of Agriculture presents a spatial overview of food access indicators for low-income and other census tracts using different measures of supermarket accessibility, it provides food access data for populations within census tracts; and offers census-tract-level data on food access that can be downloaded for community planning or research purposes.[3]

The data collection work started in 2013. The data I downloaded was last updated in May 2017. (There was a new update on November 13, 2020 and my current work does not include that update)

The data set from the Atlas has 148 variables including food access, geographical, census tract, poverty, income, and racial information. I will be focusing on the food access variables like low food access depending on income, proximity to grocery store and vehicle access.

## 1.2 Louisiana County Health Ranking Data

As my work is currently focused on Louisiana, I am using the county health rankings from the state's Department of Health. The data set contains the ranks and scores for each county in Louisiana and the underlying data details for the measures used in calculating the 2020 County Health Rankings. [2] Beside the main county rankings, there are sub ranks for variables like quality of life, length of life, clinical care, socio-economic factors and environmental factors. Then there is data that is used to calculate the rankings using information on smoking, alcohol use, physical exercise, nutrition, obesity, insurance, education, income, and poverty.

I will focus on obesity, nutrition, exercise, and environment from the ranking calculation data and the overall health and sub rankings.

## 1.3 Quick service restaurant data

I found a data set on Kaggle on fast food restaurants across the USA from Datafiniti's Business Database. [1] This data set has information on the restaurant name, city, postal code, and coordinates. I can use it to get a count of specific fast food places but the challenge is that there is no county information which makes a county-level analysis difficult.

An ideal data set would be one where I could get data about the proximity of these food places to residential areas and how many are within a particular county. For this I will be going to Datafiniti's Business Database to look at the full data set and see if there are variables that will be more useful for me.

### 1.4 Data I would like to find

I am looking for food sale/purchase data for the Louisiana counties. This will provide information the main food purchases by food categories; and help in seeing the correlation between health outcomes and overall nutritional intake.

## 2 Modeling approaches

I have decided that I do not want a binary prediction of obesity for a county but a probabilistic one. This rules out methods like the logistic regression model. Some of the approaches I am trying to decide between are:

1. Using support vector machine (SVM) to separate data into different classes by drawing hyperplanes and dividing data points. However, SVM does not perform well on large data sets like the ones I am using, so it may not be good choice.

2. Naive Bayes for a probabilistic classification outcome. As this is based on Bayes' theorem with independence assumptions between the features, it may not be useful for my analysis as the features are not necessarily independent. For example, a higher poverty rate will be highly correlated with lack of vehicular access to a supermarket.

3. Ensemble models like random forest or gradient boosting. With these I can use different training data sets (as might be the case with 4 different data sets), and get an aggregated prediction of each base model; with the final result being a prediction for the unseen data.

3. Neural networks for merging multiple data sets and predicting obesity. Neural networks can be useful here as the data sets being used are very large and in addition to using it for predictions, it can be used to detect patterns I may not

be able to discern.

## References

[1] Kaggle. Fast food restaurants across america. 2020.

[2] LA Department of Health. Louisiana county health ranking data. 2020.

[3] USDA. Food access research atlas. 2017.