Final Project Proposal – Group 4

Benjamin Lee, Kristin Levine, Russell Moncrief

DATS 6103-10: Introduction to Data Mining

Due: April 14, 2020

Problem and Datasets

With all the news about coronavirus and the search for a possible vaccine, we thought it would be interesting to consider the challenges of convincing an entire population to embrace a vaccine. The MMR vaccine for measles, mumps, and rubella has been around for years, and yet there are still many areas around the country where at least 95% of school children are not vaccinated as recommended by the CDC.

This seems especially relevant because scientists in France are using the current MMR vaccine as the backbone of their vaccine candidate.

The first dataset we are using was compiled by analysts at the Wall Street Journal for an article on school vaccination rates.

The WSJ did some cleaning of the data, but we will probably need to do some more, as the data comes from different states. Some states provided MMR vaccine rates; others provided only the overall vaccination rate. There are certain schools on this list where there is no vaccination data provided. We are going to use methods from this class to try to predict vaccination rates at schools. Perhaps we can also find more data from states that are not represented in the WSJ files.

The second dataset comes from the University of Pittsburgh and contains outbreaks of measles and other contagious diseases. We will first need to combine the data for measles, mumps, and rubella, as those are the diseases the MMR vaccine prevents. Using the data from part one, we will look at the relationships between vaccination rates and where outbreaks of measles and other contagious diseases are most likely to occur.

Algorithms, Methods and Software

For the schools' dataset, we will run a decision tree. We'll use whether the vaccination rate at a specific school is over 95% as the target, and then create a couple of different features. For example, was a school public or private? Was the average city vaccination rate above or below 95%? Was the average state vaccination rate above or below 95%? Were the exemptions at the school above or below a certain level? We will be compiling and running our code using PyCharm software. Decision tree algorithms will be displayed using Graphviz.

We plan to also use the K-nearest neighbors algorithm on our datasets, since it's known that vaccine refusers tend to be near each other. The latitude and longitude data are given in the separate state files, so we'll have to combine that data and clean it in order to make it work. We plan to construct a world map to show the distribution of MMR rates across the United States for the purpose of visualization. If we have time, we will attempt to apply Naïve Bayes to our datasets as well.

We are planning to use similar methods on the University of Pittsburgh data, (decision tree/KNN), but adding in the school vaccination rates as a FEATURE this time. We may also use random forest on both of these datasets. We'll use both the entropy and Gini models and look at the classification report and confusion matrix to check our accuracy.

To process our data, we will use many different packages including sci-pi, sci-kit learn, matplotlib and pandas to name a few. We are currently planning to write most of the code ourselves, but if we look up code from GitHub, Kaggle, or a similar place we will be sure to reference our sources.

Reference Materials

Although a majority of our reference material is derived from the PowerPoint slides provided for our class lectures, we have used a variety of sources including Towards Data Science, Wikipedia and Stack Overflow.

Anticipated Completion Dates

- March 28: Video call, picked a data set
- April 7: Set up GitHub
- April 12: Preprocess and clean school data
- April 12: Clean and combine school location data
- April 12: Run decision tree and KNN on school data
- April 13: Video call, finalize proposal
- April 18: Clean and combine MMR (measles, mumps, rubella) data
- April 18: Run decision tree and KNN on MMR data
- April 18: Create map visualization
- April 18: Run Naïve Bayes on MMR data
- April 18: Combine all our data
- April 18: Run decision tree and KNN on combined data sets, using school vaccination rates as a feature and the actual rates of disease as a target this time.
- April 19: Video call, combine results and gather questions for class
- April 20: Create group presentation (PowerPoint slides)
- April 20: Write group final report and individual final reports
- April 22: Write GUI Code
- April 24: Check (and double-check!) GUI code
- April 26: Video call to finalize everything
- April 27: Upload all deliverables