**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. http://www.rfi.fr/en/science-and-technology/20200227-france-s-pasteur-institute-develop-coronavirus-vaccine-candidate

The first dataset we are using was compiled by analysts at the Wall Street Journal for an article on school vaccination rates. https://github.com/WSJ/measles-data

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; other 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 1, we will look at the relationships between vaccination rates and where outbreaks of measles and other contagious diseases are most likely to occur. https://www.kaggle.com/pitt/contagious-diseases

## 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 plan to also use the KNN method on this data, 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 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, marplot lib, 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.

**Anticipated Completion Dates**

Video Call 3/28 picked a data set
Set up GitHub 4/7
Clean school data 4/12
Run decision tree on school data 4/12
Clean/combine school location data 4/12
Run KNN on school data 4/12
Video Call 4/13 finalize proposal
Clean/combine MMR (measles, mumps, rubella) data 4/18
Run decision tree on MMR data 4/18
Run KNN on MMR data 4/18
Combine all our data 4/18
Run decision tree on combined data sets, using school vaccination rates as a feature and the actual rates of disease as a target this time. 4/18
Run KNN on combined data sets. 4/18
Video Call 4/19 Combine results; gather questions for 4/21 class
Create group presentation (PowerPoint slides) 4/20
Write group final report 4/20
Write final individual report 4/20
Write GUI Code 4/22
Check and double check GUI code 4/24
Video call to finalize everything 4/26
Upload all deliverables 4/27