**What problem did you select and why did you select it?**

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>

**What database/dataset will you use? Does it need to be cleaned?**

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 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 data to look at the relationships between vaccination rates and where outbreaks of measles and other contagious diseases may be most likely to occur. <https://www.kaggle.com/pitt/contagious-diseases>

**What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?**

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.

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.

**What softwares will you use to implement the network? Why?**

**[I’M NOT EXACTLY SURE WHAT THIS MEANS – DO YOU GUYS GET THIS?]**

**What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?**

**How will you judge the performance of your results? What metrics will you use?**

We’ll use both the entropy and gini models, and look at the classification report to check our accuracy.

**Provide a rough schedule for completing the project.**

Kristin – run decision tree on school data

Russell – run KNN on schools data

Ben – run decision tree on measles, mumps, rubella data

Combine all our data and run decision tree/KNN on data using school vaccination rates as a feature and the actual rates of disease as a target this time.