Ahmed Mohamed, Byron Dickerson, Clara Comstock, Julia Goyco, Rick Djeuhon

Final Project Question Proposal

4/7/2019

After exploring many datasets, and evaluating possible solutions and relevance to societal

need, we decided to dive into the topic of health, specifically disease outbreaks. We

narrowed down the general topic, and we chose a topic dealing with California immunization

rates, in which we found a large dataset dealing with California Pertussis cases and rates from

2010 to 2014. The dataset includes number of students, school name, school county, the

number of PBEs, PMEs, number of students vaccinated and if the school was private or public.

The datasets are provided to Kaggle by the California Department of Public Health (CDPH).

Upon inspection on the dataset, we found it interesting to examine what role vaccination rates in every county played in infant Pertussis cases, as well as how the reported exemption for a given school affected the outbreak.

Our question is: can we predict which counties in California are at a higher risk of

Pertussis outbreaks based on previous years’ vaccination records? This question would be

of societal need as the Pertussis cases and rates over the years lead to infant hospitalization and

unfortunately, some deaths. In addition, if we can build a predictive model that uses previous

years to predict the upcoming years risk, then early intervention and an increase in vaccination rates can help combat the issue and reduce outbreak cases. Furthermore, the model will be tested and validated on the 2014 year to help further improve the model and set possible modifications for future improvements of the model.

Work Cited

[1] https://www.kaggle.com/broach/california-kindergarten-immunization-rates#InfantData.csv