Prediction of Outbreak Sizes from

Foodborne Illnesses

# Problem Statement

This project will be exploring foodborne illnesses. Hopefully, by predicting the size of the spread of the illness, hospitals and news sources can better prepare and release information about the foodborne illnesses to the public. We converted this problem into a data science problem by focusing on predicting the size of the outbreak based on the genus species, location of preparation, and food vehicle. Using supervised learning, the model will use and train based on the dataset from the CDC and predict the outbreak size.

# Solution

The results of this project are definitely something learn from and can be improved on. Since the model is used to predict the possible size of the illness outbreak instead of a binary result, the need for a large amount of data is important. This is something that we would look for more in the future. However, from these results, we see that the level of hospitalizations, the genus species, and the location of preparation are significant features to help predict the size of the outbreaks.

### Data used

The data used in this project is a dataset from the CDC about foodborne outbreaks in the US from 1998 to 2017.

### Approach and assumptions

We approached the dataset with an idea of what was expected to be cleaned, which features would be necessary to help predict the results, and which aspect of the problem we would like to see predicted from the model. Using Darwin to clean the data, we noticed that there were features that were dropped that we had expected to remove, for example, “contaminated ingredient” since there were so few data records for that feature and “serotype or genotype.” This use of Darwin was very convenient and helpful in our process of cleaning the dataset.

### Application of Darwin

Darwin was used to help with the creation of our prediction model. Instead of creating and testing the model from scratch, Darwin provided a helpful set of functions to more easily create the model. For example, Darwin chooses the best type of classification method for the dataset instead of the programmers deciding which is best. This is helpful because one of the difficult parts of creating prediction models is choosing the classification method best suited for the data.

Some aspects that we would like to see in Darwin in the future would be some more graphical tools that we could use at the end of the project to better display our results. While this is not crucial, it would be a great addition to this toolset.

# Team Engagement

The team divided the roles based on the individuals’ abilities. This way, everyone was best utilized for the project. Those who were more comfortable coding, programmed the majority of the model, and those who were more comfortable writing the essay, focused on writing. However, we believed the best way to complete the project was to collaborate and work together. This way, everyone would participate and be able to contribute to the project as well as understand how the model performed and the use of Darwin. There were some challenges being able to meet up despite all of our busy schedules, but we met in smaller groups when needed.

# General Challenges

For this project, the team ran into some problems with the research and data cleaning. We previously had several ideas for the project that was not about foodborne illnesses that we would have preferred to research, but it was extremely difficult finding datasets that had relevant information to base the model from. With this dataset, there’s also a problem cleaning the data. For the location feature, the dataset allows for multiple types of location of preparation, so it is difficult to clean and classify using that data feature even though it is useful information. With more time, we could possibly create another function to better clean up that data feature to make it more relevant for the model. Moreover, the project was using the dataset to predict the possible size of the outbreak. We would need a massive amount of data in order to accurately predict the size of the outbreak since foodborne illnesses can be affected by a variety of factors, such as local population size, the number of customers served, and other influences.

There were also some problems with the Darwin server that hindered our progress on the project slightly. At some point, we could not create and train the model because the Darwin server had faced some problems and shut down, so the project could not move past that point. However, we tried to code as much as we could without running and testing the program code, so that when the servers came back up, we were not too behind. The max unique values parameter for the data cleaning portion of the project also posed a very slight issue. While it was very convenient to use and help parse down on the data records used, it meant that we couldn’t actually keep the data feature if we wanted to if there were too few records. It may have been helpful if there was a function to help input or estimate what the data record could have been, however, most of our features used were categorical data points, so that function may not have been as helpful.

# Next Steps

Through the experience of Darwin and the results of this project, we hope to continue improving our model and work on a better method of predicting the size of the outbreak. At this point, we would like to be able to improve the model to include more information, such as how long the food was served before identifying the origin of the food illness, probable size of people served or exposed to the bacteria, and the city or a more exact location, so we would need further additional data and research. This project would also could be improved if there was an expert in this field to suggest better identifiers or features to help the model better identify the level of impact of the foodborne outbreaks. Currently, we are only working based on the limited information from the dataset from the CDC, so to have help from an expert or at least someone more knowledgeable about foodborne illnesses or in this field would be very helpful. Another area of improvement would be to alter the model so it can predict the number of hospitalizations and deaths instead of just illnesses. Predicting solely the number of illnesses does not help hospitals as much to prepare for number of possible hospitalizations since there could be a large number of symptoms but not actual hospitalizations. There is a lot more that this project could do for this topic, and a range of other areas for exploration.