NORTHEASTERN UNIVERSITY
CS6220 DATA MINING TECHNIQUES

PROJECT PROPOSAL
OCTOBER 2018

SRIJIT RAVISHANKAR
SURUPA TUSHAR CHATTERJEE

1. Background

Over the past decade there have been numerous diseases that have been detected and reported to have caused outbreaks that led to many deaths. Many of these diseases have been identified to have origins from food that a person eats. The **Center for Disease Control and Prevention (CDC)** estimates roughly 1 in 6 Americans (48 million people) get sick, 128,000 are hospitalized, and 3,000 die of foodborne diseases each year. A foodborne disease outbreak occurs when two or more people get the same illness from the same contaminated food or drink. While most foodborne illnesses are not part of a recognized outbreak, outbreaks provide important information on how germs spread, which foods cause illness, and how to prevent infection.

Here we use the dataset that contains data on foodborne disease outbreaks reported to CDC from 1998 through 2016. Data fields include year, state (outbreaks occurring in more than one state are listed as "multistate"), location where the food was prepared, reported food vehicle and contaminated ingredient, etiology (the pathogen, toxin, or chemical that caused the illnesses), status (whether the etiology was confirmed or suspected), total illnesses, hospitalizations, and fatalities. In many outbreak investigations, a specific food vehicle is not identified; for these outbreaks, the food vehicle variable values are blank.

We aim to apply data mining techniques that can help discover interesting patterns in this dataset and apply algorithms to predict the patterns of foodborne illness outbreaks, identify the key ingredients/sources in food that lead to infections, specific regions/areas across U.S where the risk of foodborne diseases is high/low, where the patient consumed the food(Restaurants/Home etc.), specific period of the year when a particular disease is more prominent and spreads, kind of pathogens or germs that leads to a disease or specific toxin/chemical that contributes to the disease, method of the food preparation that may lead to the outbreak.

We plan to explore the above-mentioned dataset and use additional datasets such as a dataset that would contain features pertaining to what time of the year a particular food pathogen may be more active and more prone to grow and spread, food materials that contain specific toxic materials and quantities of those toxins that can lead to infections/diseases and so on.

2. Goal and Outline

The goal of our project is to apply supervised machine learning techniques to build predictive models to:

1. Predict whether a case is suspected, confirmed or no food borne illness.
2. Number of people that may be affected with food borne illnesses.

We also intend to address the following questions:

1. Whether a particular food can cause foodborne disease?
2. Which pathogens, toxins or chemicals in the food lead to diseases?
3. Food Combinations and their potential infection source(pathogen/toxin/chemical).
4. Deem a food to be fatal on the basis of number of hospitalizations and fatalities?

3. Preparation and Implementation

The first step towards addressing the problem statements is to process and clean the dataset obtained. After this step, we need to normalize the features and also engineer new features as necessary. The next step is to create an appropriate model for each of the problem statements. Once this is accomplished, we can evaluate the model using appropriate accuracy metrics.

3.1. Data Preprocessing

3.1.1. Missing Data

The major preprocessing step is to handle the missing data. For this dataset, the missing data for continuous variables are going be populated by using mean of their respective columns and the missing data for categorical variables will be populated using **K-Nearest Neighbors (KNN) algorithm**.

3.1.2. Feature Normalization

Normalization for categorical and continuous variables have to be handled in different ways. For continuous/ numerical variables, the method of **Mean Normalization** will be employed. Categorical variables will be normalized using **One Hot Encoding Technique**.

3.1.3 Dimensionality Reduction and Feature Engineering

Based on the data and its behavior, a few new features can be derived from the existing features to add more value to the model. This feature set can then be reduced using various techniques like **Correlation Analysis**, **Feature Importance plots**, etc. This helps us to arrive at the most important/ influential feature set that can be fed to the model.

3.2. Predictive Modelling

To predict whether a case is suspected/confirmed/no food borne illness is a classification problem. Based on the data and its patterns, we can apply various classification algorithms like **Logistic Regression, Decision trees, Random forests, KNN**, etc. Based on the resulting accuracies, two ways can be adopted to better the model – **Parameter tuning** by estimating the optimal regularization parameters based on accuracies or having an **Ensemble of classifiers** that together predict the target class.

Predicting the number of people to fall ill due to food borne illness can be achieved by employing one or more of the following Machine learning algorithms: **Multiple linear regression, Lasso or Ridge regression, Gradient Descent, Support Vector Regression**, etc. Based on the data, Gradient Descent based Ridge regression seems to be a good choice.

3.3. Model Evaluation

For problem statement (1) of predicting whether a case is suspected/confirmed/no food borne illness, evaluation metrics like **Sensitivity, Specificity and Youden's index** can be used. **Receiver Operating Characteristic (ROC) curves** can be plotted, and **Area Under the Curve (AUC) values** can be determined to evaluate the model performance. Various models can be drafted, and their accuracies can be compared to arrive at the best model.

$$\textbf{Sensitivity} = \frac{TP}{TP+FN}$$

$$\textbf{Specificity} = \frac{TN}{TN+FP}$$

Youden's index = Sensitivity - Specificity -1

Where, TP - True Positives, TN - True Negatives, FP - False Positives, FN - False Negatives.

For problem statement (2) of predicting the number of people to fall ill due to food borne illness, metrics like **RMSE (Root Mean Squared Error)** and $\mathbf{R^2}$ can be used to evaluate model performance. Here again, various models can be drafted, and their performance metrics can be compared to arrive at the best model.

$$\textbf{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)}{T}}$$

Where, $\hat{y}_t$ is the predicted value, y_t is the ground truth target variable value, T is the total number of data points in the dataset

4. References

https://www.kaggle.com/cdc/foodborne-diseases/home
https://web.uri.edu/foodsafety/cause-and-prevention-of-foodborne-illness/