Nicole Pham-Nguyen
Varun Kausika
Krish Engineer
Disha Gandhi
Khyati Jariwala

Adverse Allergic Events Project

Description of Project Goals

Description: We investigated a dataset of adverse allergic events from 2004 to mid-2017 collected by the US Federal Drug Administration (FDA) Center for Food Safety and Applied Nutrition (CFSAN) that includes approximately 90,000 reported reactions[1]. Consumers and health care practitioners reported these adverse events through the CFSAN Adverse Event Reporting System (CAERS). CAERS consists of both required and voluntary reports from industries like cosmetics, food, vitamins, and other CFSAN regulated industries.

We sought to use this dataset to identify the most significant predictors of severe outcomes, with severe outcomes including: required intervention, hospitalization, death, life-threatening conditions, disability, serious injuries, and other medical problems. Some possible significant predictors of severity of reactions or seriousness included: industry, symptoms, age, gender, and product name. We used methods such as Naive Bayes, Logistic Regression, kNN, and Random Forest to find the best possible model to predict seriousness and find significant predictors.

Importance of the Problem: Allergies are the 6th leading cause of chronic illness in the United States, at an annual cost of $18 billion[2]. In addition, allergies and the severity of reactions have become increasingly common, particularly food allergies, leading to what many researchers call "the food allergy epidemic[3]. Identification of significant predictors of the seriousness of an allergic reaction would assist healthcare providers in the efficiency of diagnoses and treatments, leading to happier patients, providers, and a lower annual cost. For example, if a patient has a combination of predictors (symptoms, age, consumption of a product in a certain industry) that are strongly correlated with seriousness, then providers can have a plan in action to start treatment earlier than usual, leading to minimized preventable deaths.

Exploratory Analysis (EDA)

The dataset contains about 90,000 rows, or reports, of adverse medical events. Each row contains details such as date, product name, industry of product, age, gender, outcomes, and symptoms. There are 45,685 unique product entries across 41 industries. Figure 1 contains a breakdown of the top industries with the most reported cases of adverse events, where the Vit/Min/Prot/Unconv Diet (Human/Animal) industry has the most reports. A further look into product names led to the discovery of many ‘redacted’ entries, which is essentially, a sealed product name due to confidentiality or legal reasons prior to publication.

Data Processing: Since these entries comprised 7% of our dataset, we kept these entries in both the exploratory analysis and model construction to avoid as little information loss as possible. However, there were a few entries that listed N/A for certain categories, such as age, date range, and gender. For age and date range, we assigned the N/A values to the mean of the category across the dataset. For gender, we randomly assigned the N/A values to "Male" or "Female." We filtered the number of symptoms from the thousands to only the top 300, as some symptoms were only expressed once or twice out of the 90,000+ entries.

Symptoms: Using the list of symptoms presented for each entry, we generated a word cloud to visualize the most common symptoms among all reports (Figure 2). The top 5 overall symptoms across all reports are diarrhoea, vomiting, nausea, abdominal pain, and malaise. We then examined the top 15 industries that had the most frequently occurring symptoms and found the most common symptoms for each of them (Figure 3). The top 15 industries and their corresponding top 3 symptoms are included in Table 1. Each of the top 3 symptoms in the industries corresponded to one of the top 5 overall symptoms listed above. However, we found that the cosmetics industry was unique in that none of the top 3 symptoms were included in the top 5 overall symptoms.

Outcomes: To dig deeper into the top 15 industries with frequently occurring symptoms, we found the most common outcomes for each of them (Figure 4). The top 15 industries and their corresponding top 3 outcomes are shown in Table 2. Non-serious injuries/illness was found to be the most common outcome that was included in all of the industries except for the Vit/Min/Prot/Unconv Diet (Human/Animal) industry. In order to demarcate seriousness of an adverse medical event, we split the outcomes into two categories: non-serious and serious. We based our categories of "serious" vs. "non-serious" off of utility. Table 3 displays the outcomes and their associated classification of non-serious or serious. Using this classification, we colored all industries in Table 2 that included at least one serious outcome red. These industries include Vit/Min/Prot/Unconv Diet (Human/Animal), Cosmetics, Dietary Conv Food/Meal Replacements, and Baby Food Prod.

Gender: When dissecting the number of reports by gender, we noticed that there were twice as many female entries as there were male. We found that there were also twice as many females categorized as serious compared to males (Figure 5). Using a bar plot, we visualized the fraction of serious to non-serious cases grouped by gender to identify whether or not gender seemed to influence seriousness (Figure 6). The fraction of the serious to non-serious cases did not appear to be significantly impacted by either gender. Thus, on the basis of the data processing done, we expected both symptoms and industry, particularly the vitamins and the cosmetics industries, to be the significant predictors of the seriousness of an adverse allergic reaction.

Solution and Insights

Feature Selection: In order to include symptoms as a feature in our models, we utilized a one-hot encoding method. The other features selected in the model included: date range (the entered date minus the reported date), gender, industry, and product role. Since this was an extremely comprehensive dataset with many possible predictors, these five were chosen due to their direct relevance to the question we posed, as opposed to non-relevant variables such as the industry code and the entry identification number.

Models: Since we had a classification problem, the classifiers we used in order to come up with the best model to accurately predict the seriousness of an adverse medical event were: Naive Bayes, kNN, Logistic Regression, and Random Forest. Prior to running the models and finding their associated accuracies, we established a baseline accuracy of 76.62% by taking the number of outcomes classified as serious over the total number of cases. The baseline accuracy told us that we would be correct 76.62% of the time if we predicted that an event was serious. In order for a model to be considered a good fit, it must have attained an accuracy higher than the baseline. Additionally, we wanted the model to have a high recall score, since greater recall means less false negatives; hence less wrongly classified serious events as non-serious.

Naive Bayes: Upon running the model, the top predictors for the Naive Bayes were found to be nine industries and the "overdose" symptom (Table 4). The recall obtained was 84%, which meant that the model correctly identified 84% of all serious events as serious. In addition, we received a training accuracy of 73.42% and a test accuracy of 73.05%. These accuracies were lower than the baseline accuracy, showing that Naive Bayes was a poor fit for predicting the seriousness of adverse medical events.

kNN: Using a model learnt by kNN with a selection of 15 nearest neighbors resulted in an accuracy of 82.45% on the training set and 81.15% on the test set, which were both fairly higher than the baseline accuracy and Naive Bayes. Thus, kNN was found to be a much better model than that of the Naive Bayes, with an improvement of around ~6% than the baseline accuracy. Additionally, we obtained a recall of 96%, which meant that kNN did better than Naive Bayes.

Logistic Regression: Upon running the model, logistic regression resulted in a training accuracy of 84.20%, while the test set had an accuracy of 83.90%. The significant predictors in this model were mainly found to be based off of the specific industry and a handful of symptoms (Figure 7). The model had a recall of 92%, which was a 1% increase from the recall of kNN.

Random Forest: Using cross-validation with random forest, we found that using 100 trees and a depth of 20 was sufficient for our model. The significant predictors in this model were mainly found to be based off of the specific industry, age, and date (Figure 8). The training and testing accuracy were 87.89% and 84.72%, respectively. Additionally, the model resulted in a recall of 95%, which is the second highest recall obtained compared to the other models tested.

Best Model Analysis: Out of all models learnt by the classifiers, the random forest model performed the best, with a test set accuracy of 84.72% and a recall of 95%. Since random forest classifiers are used for classification problems, it worked well with our problem as our goal was to classify adverse medical events as serious or non-serious. The top 5 significant predictors were the vitamin industry, age, nuts industry, vegetable industry, and date range (Figure 8).

Upon comparison of the top 5 significant predictors of both random forest and logistic regression, we saw that they were extremely different, with logistic regression assigning symptoms with more importance while random forest assigned industry with utmost importance. Thus, our intuition during the EDA that symptoms and industries would be the most important predictors, was evidently true. In addition, the random forest model did include both the vitamins and cosmetics industries in the top 6 most important predictors, as we initially predicted. Upon reflection, it makes sense since these industries have a history of provoking the immune system more than others[4]. Also, it was interesting to note that gender did not appear to play a significant role in either of our best models. This may be due to the ratios intragender being pretty close as observed during EDA (Figure 6).

Figures and Tables

Figure 1: Bar plot of most commonly reported industries with adverse medical events

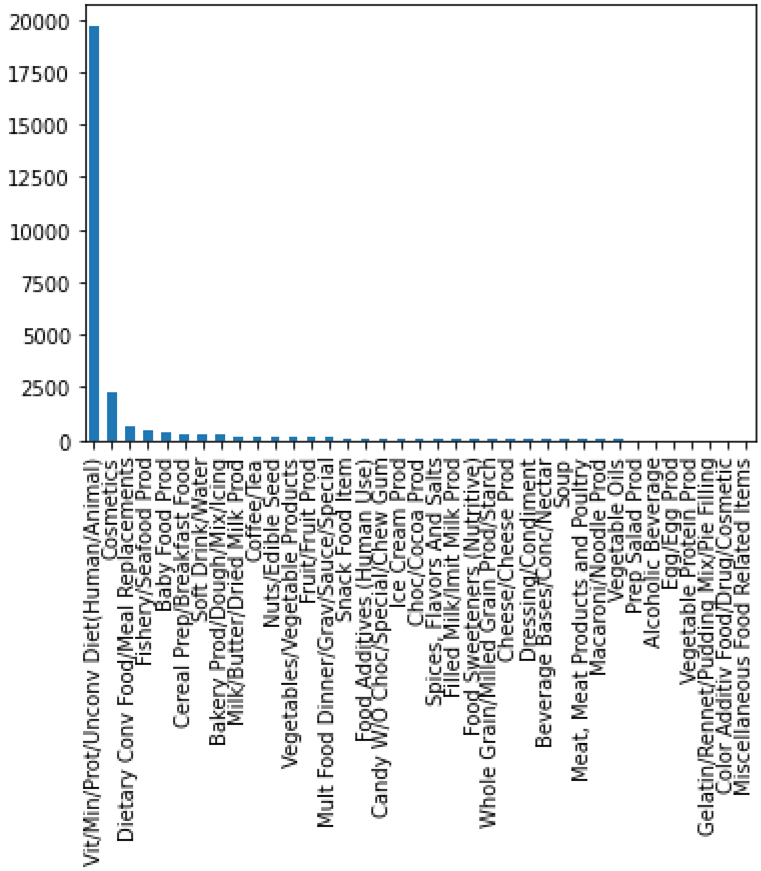


Figure 2: Word cloud displaying most common symptoms

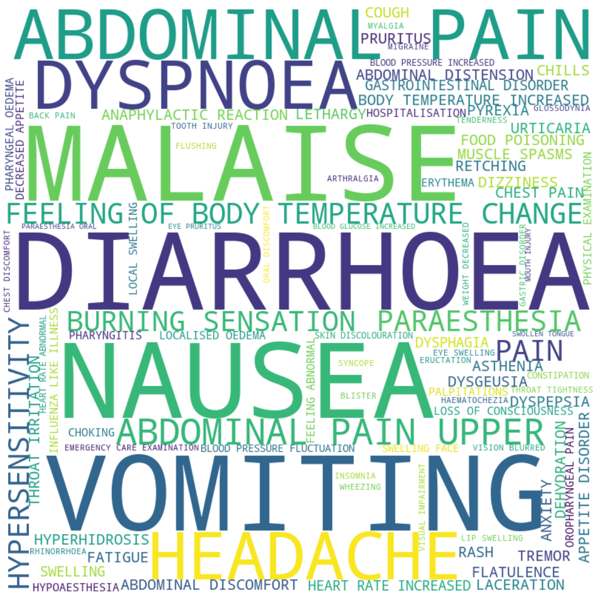
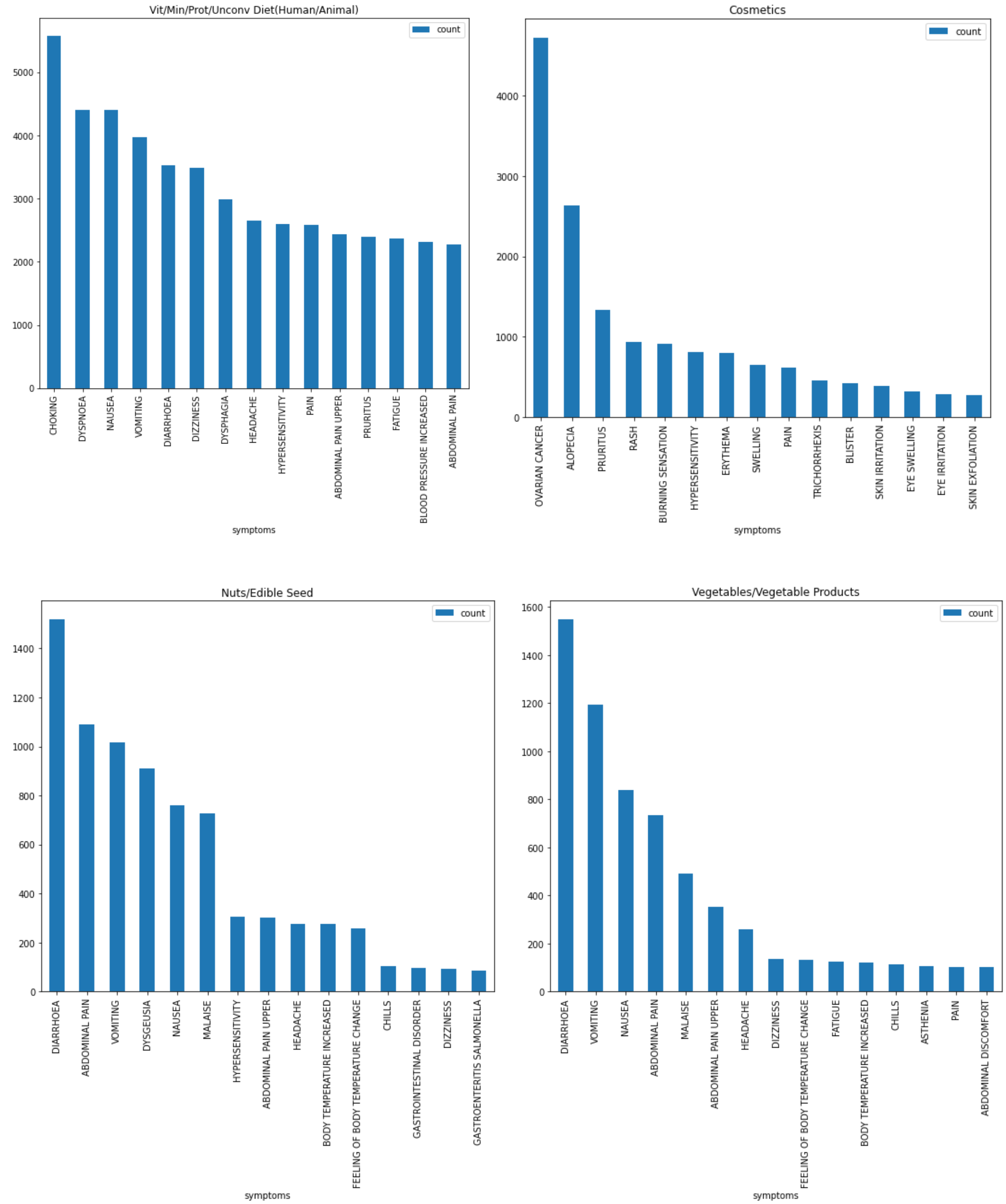
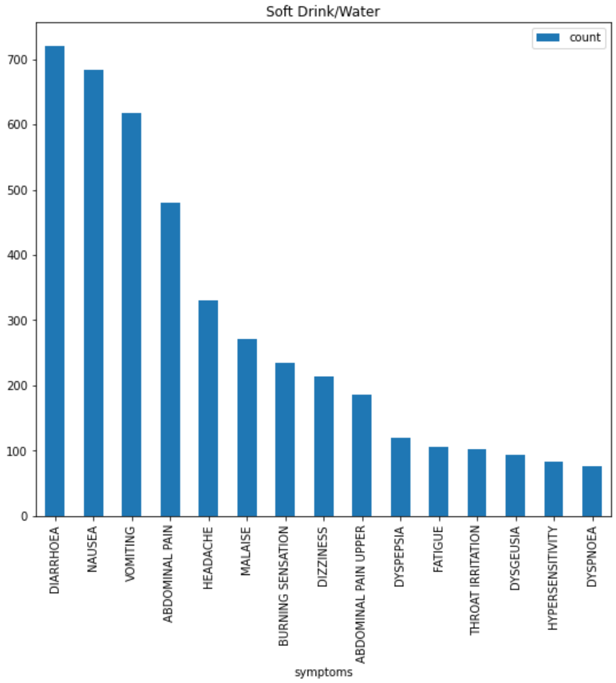
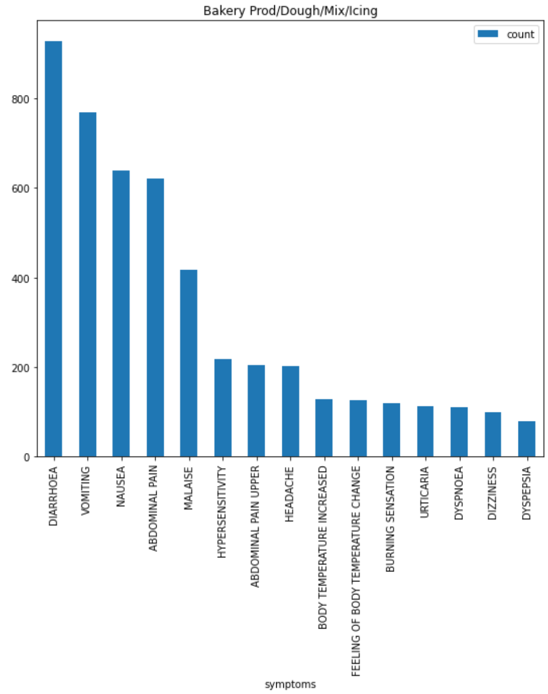


Figure 3: Bar plots depicting the 15 most common symptoms for the top 15 industries that frequently result in symptoms

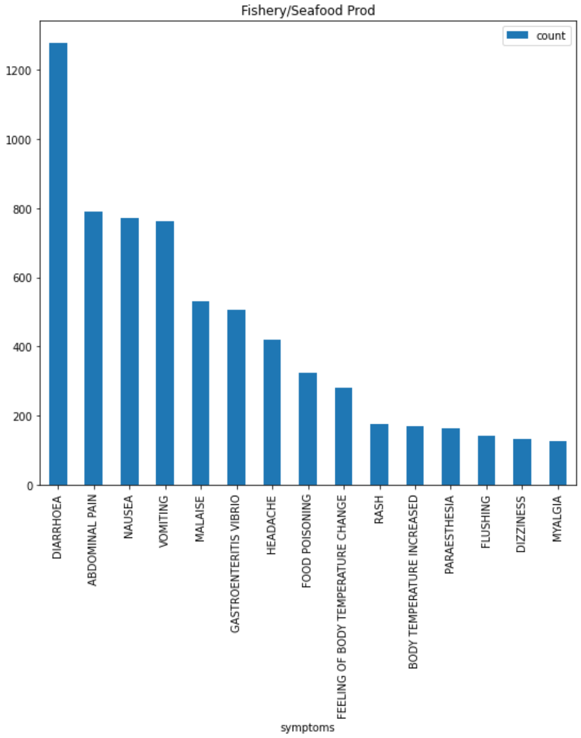




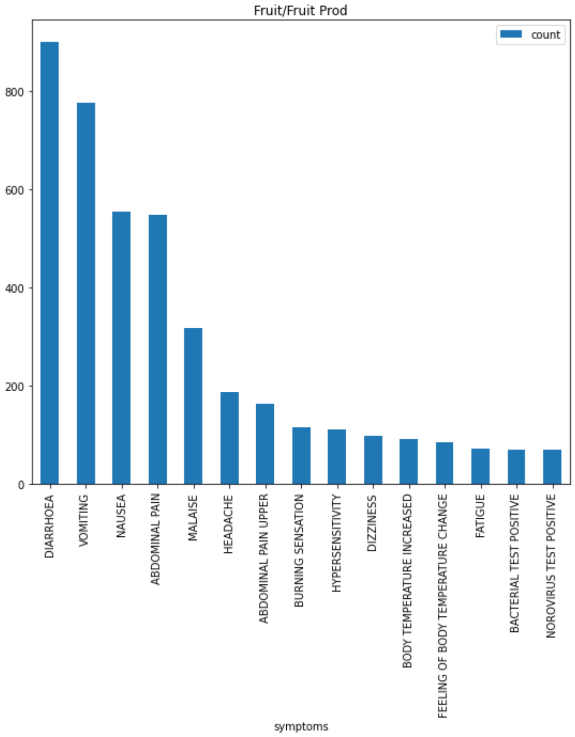
Soft Drink/Water



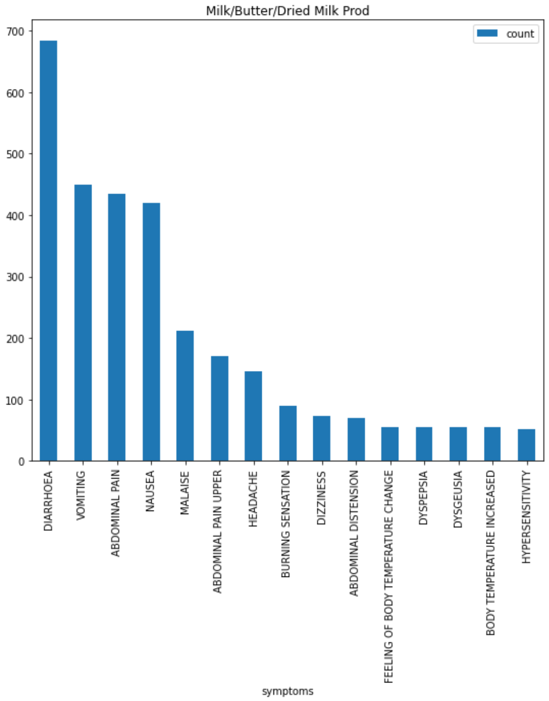
Bakery Prod/Dough/Mix/Icing

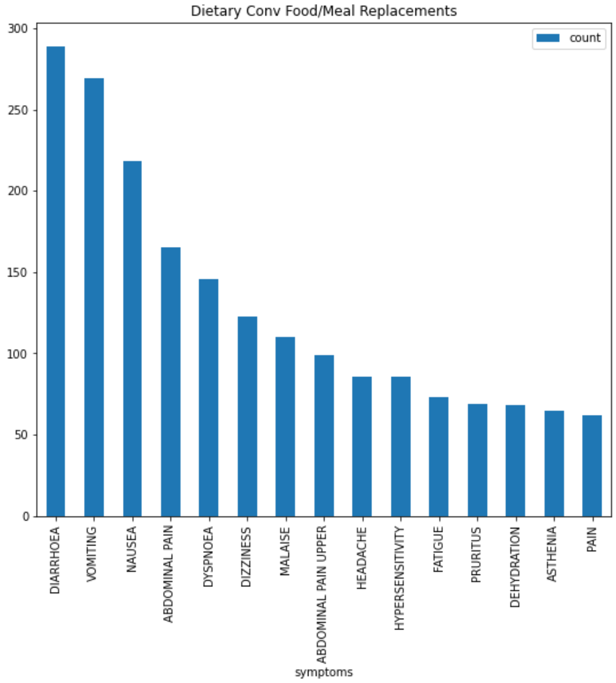


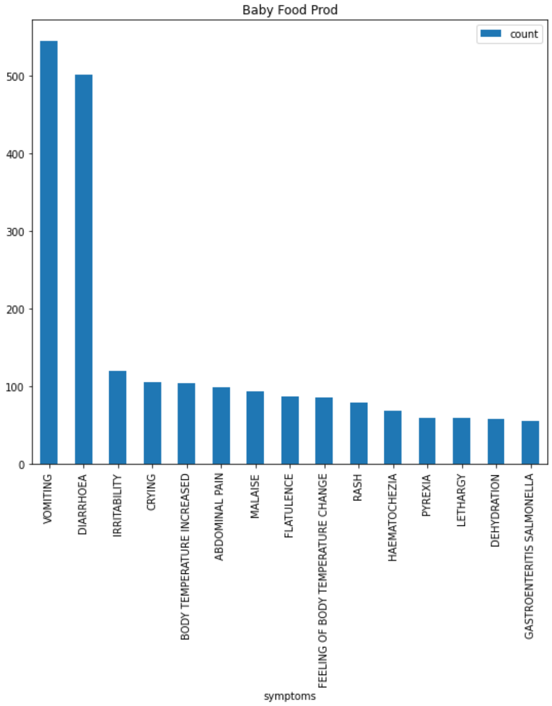
Fishery/Seafood Prod

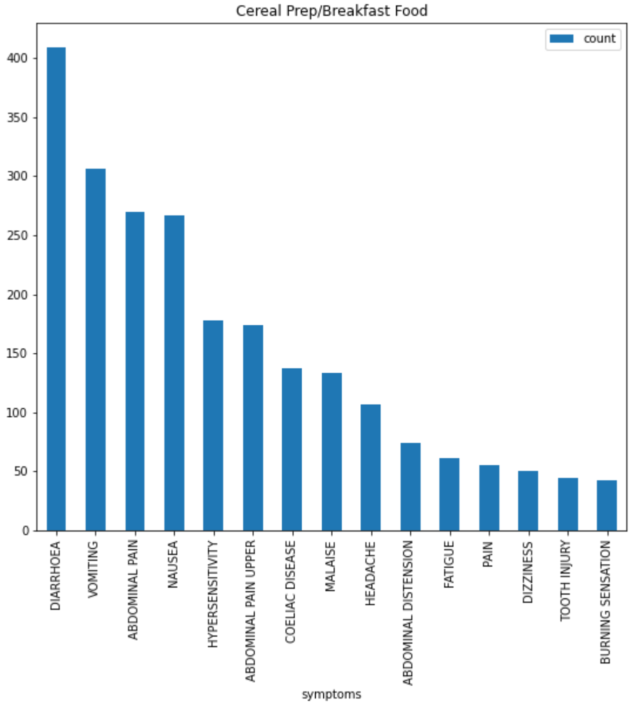


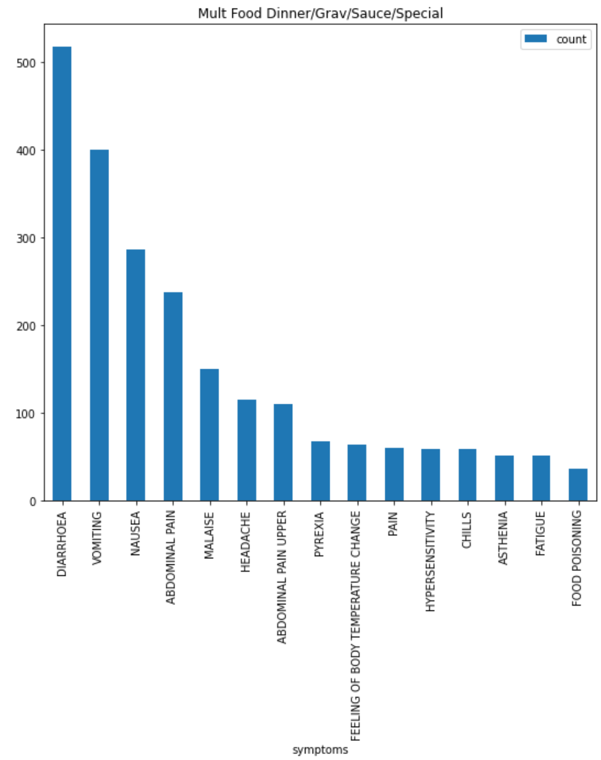
Fruit/Fruit Prod

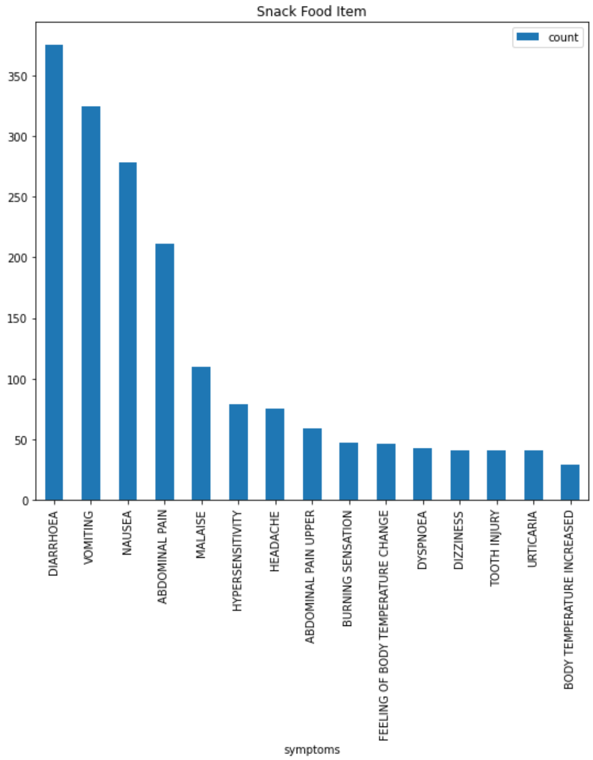

Milk/Butter/Dried Milk Prod


Dietary Conv Food/Meal Replacements


Baby Food Prod


Cereal Prep/Breakfast Food


Mult Food Dinner/Grav/Sauce/Special


Snack Food Item

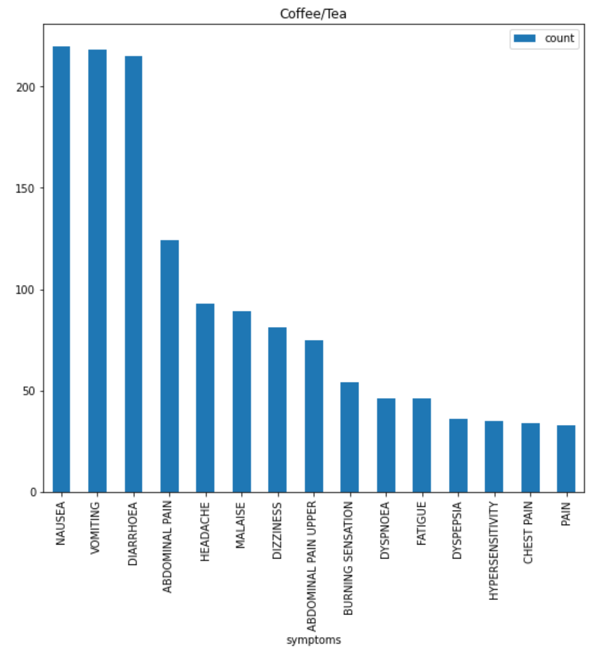
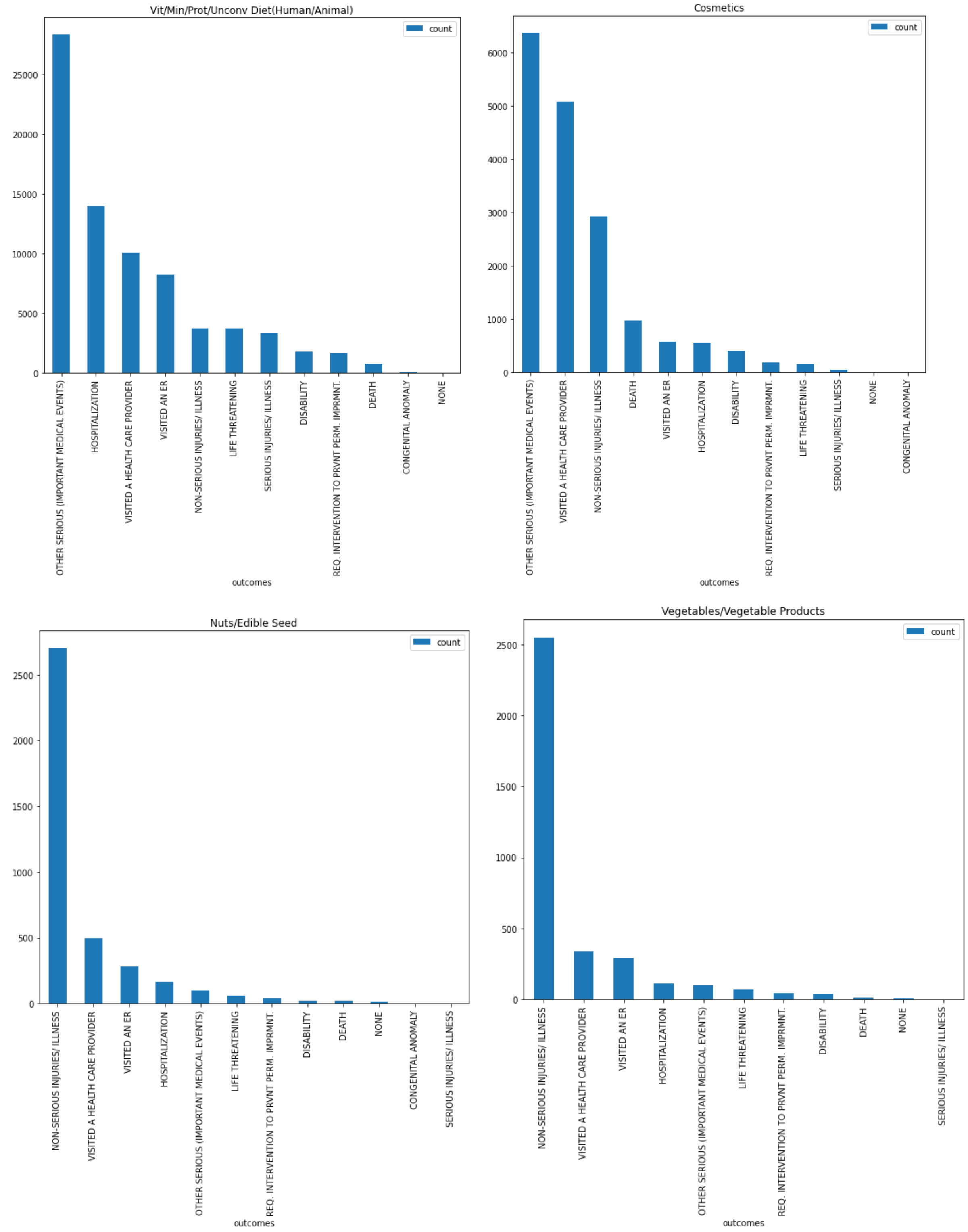
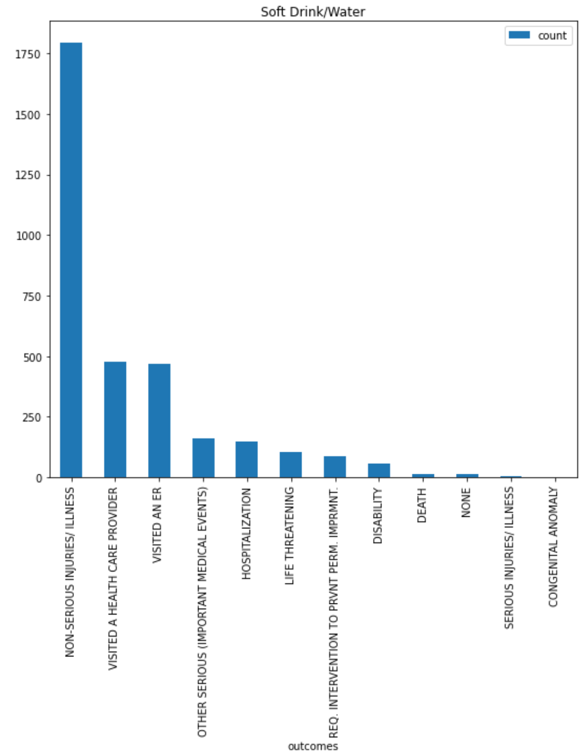
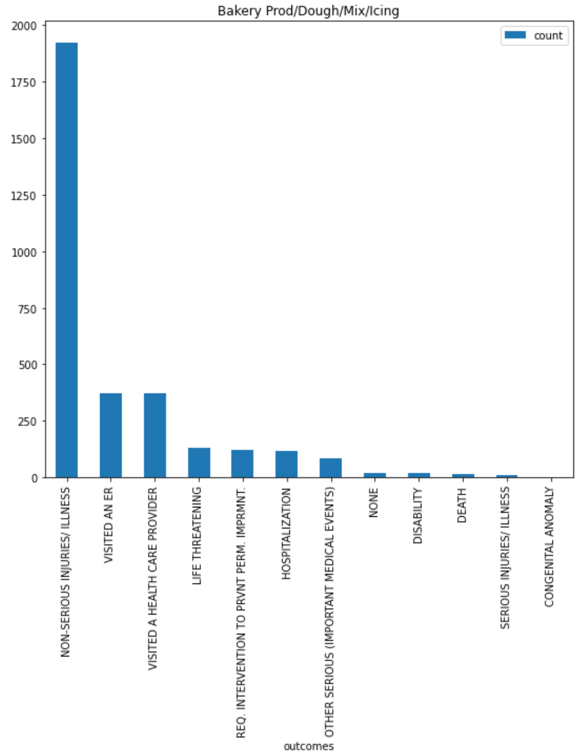
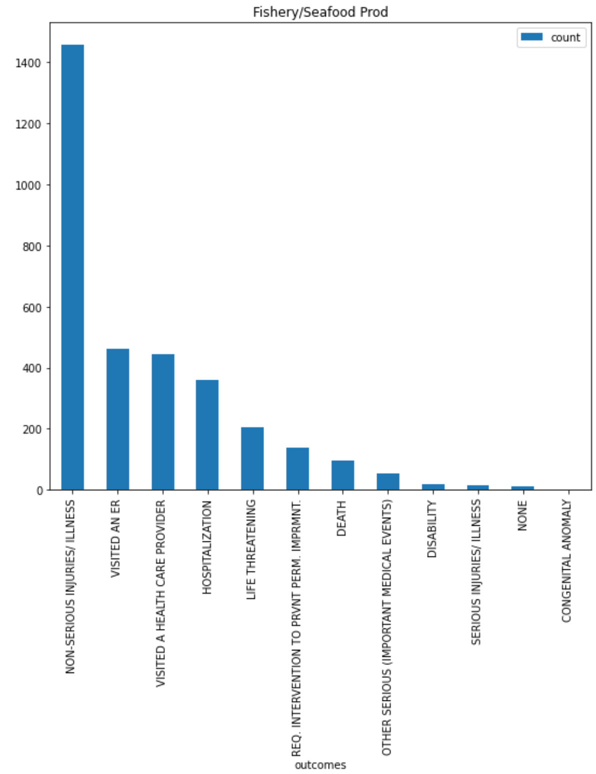

Coffee/Tea

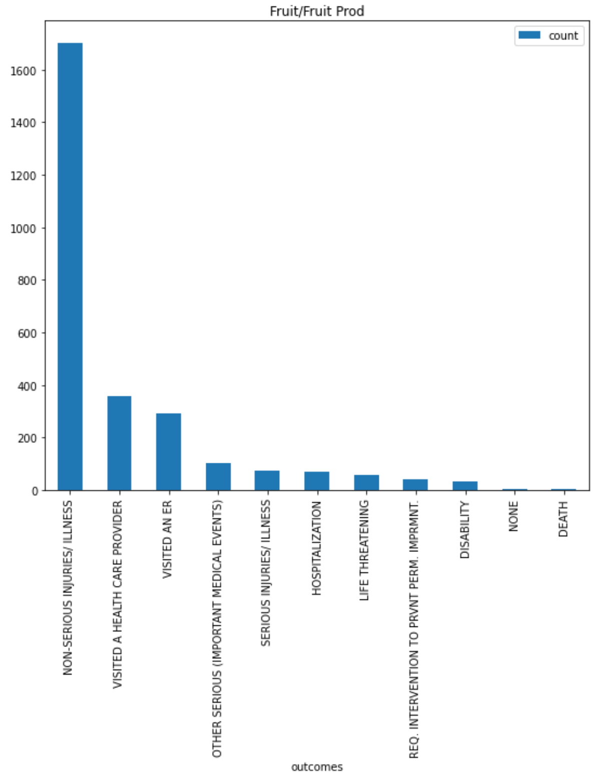
Figure 4: Bar plots depicting the 15 most common outcomes for the top 15 industries that frequently result in symptoms

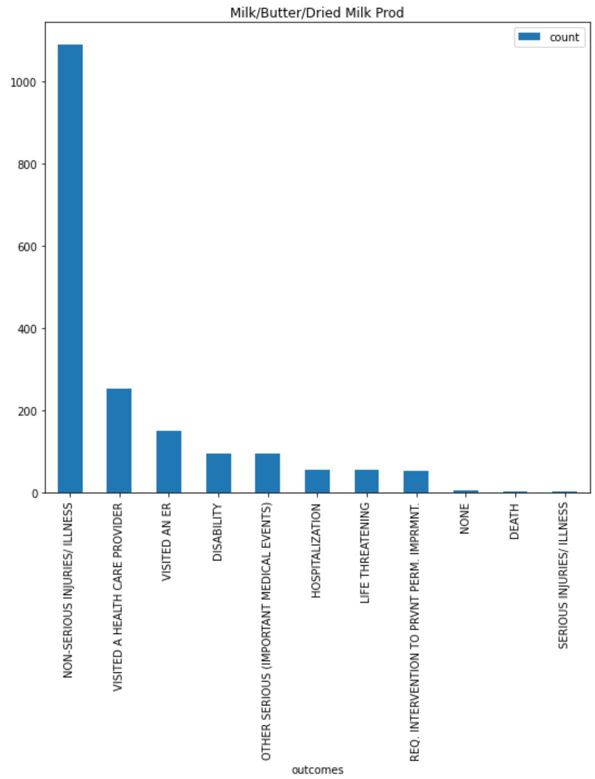


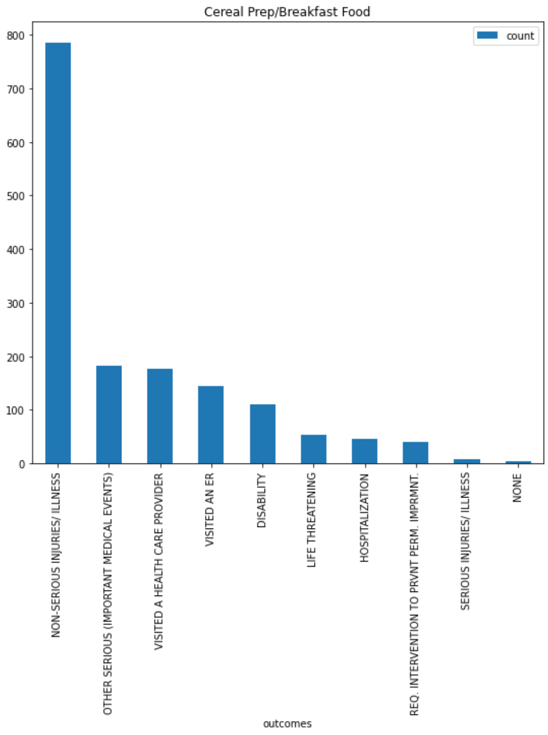

Soft Drink/Water

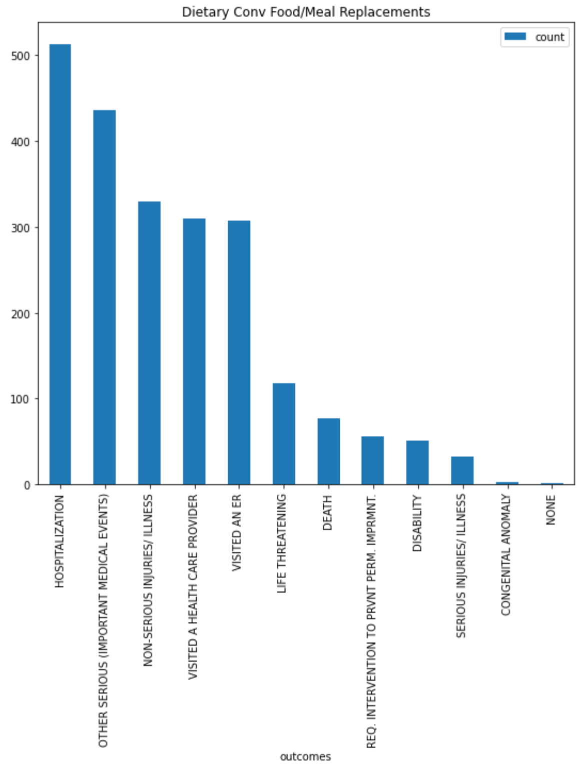

Bakery Prod/Dough/Mix/Icing

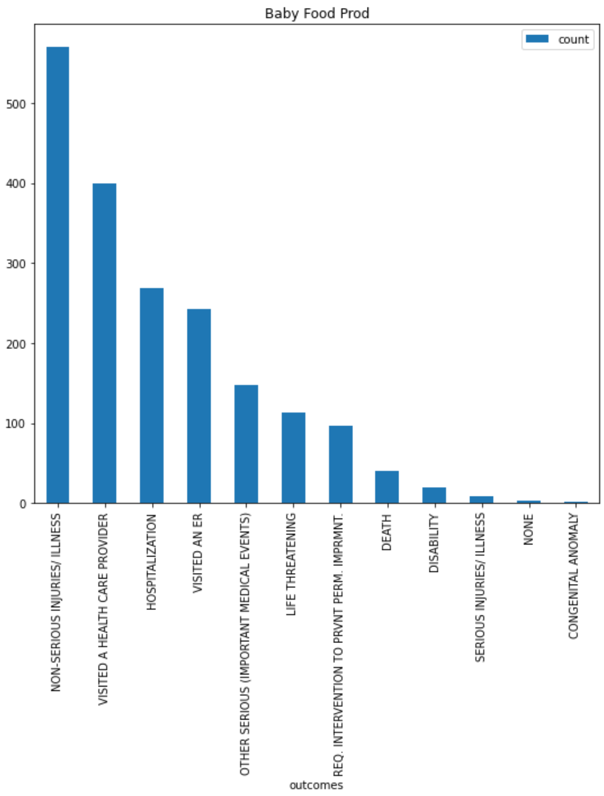

Fishery/Seafood Prod

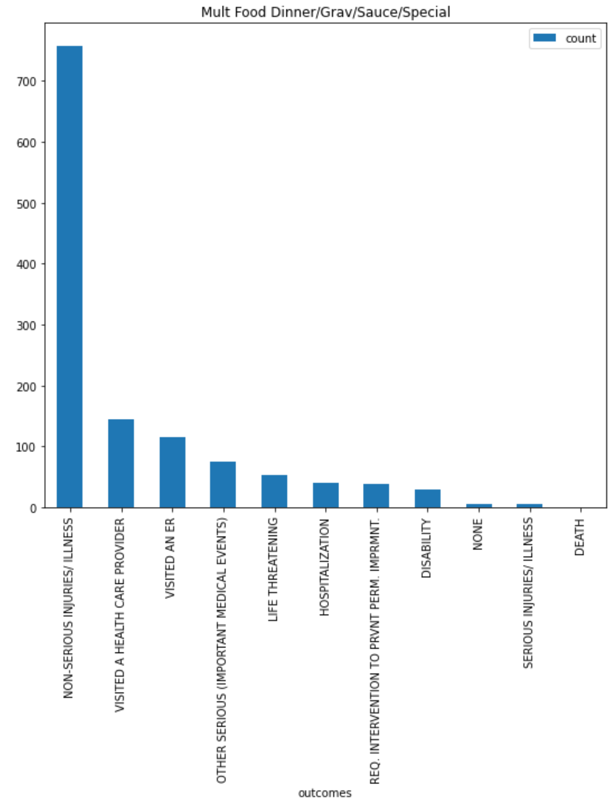

Fruit/Fruit Prod

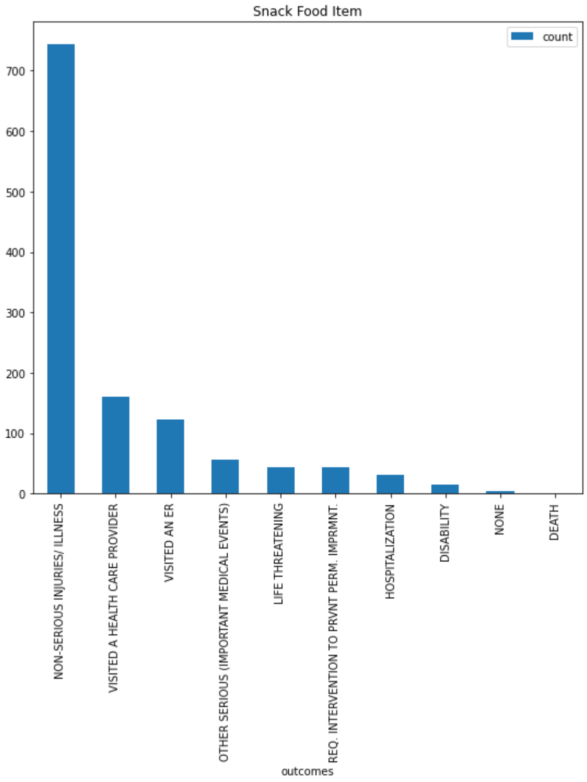

Milk/Butter/Dried Milk Prod
count
0
200
400
600
800
1000
NON-SERIOUS INJURIES/ ILLNESS
VISITED A HEALTH CARE PROVIDER
VISITED AN ER
DISABILITY
OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)
HOSPITALIZATION
LIFE THREATENING
REQ. INTERVENTION TO PRVNT PERM. IMPRMNT.
NONE
DEATH
SERIOUS INJURIES/ ILLNESS
outcomes


Cereal Prep/Breakfast Food
count
0
100
200
300
400
500
600
700
800
NON-SERIOUS INJURIES/ ILLNESS
OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)
VISITED A HEALTH CARE PROVIDER
VISITED AN ER
DISABILITY
LIFE THREATENING
HOSPITALIZATION
REQ. INTERVENTION TO PRVNT PERM. IMPRMNT.
SERIOUS INJURIES/ ILLNESS
NONE
outcomes


Dietary Conv Food/Meal Replacements
count
0
100
200
300
400
500
HOSPITALIZATION
OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)
NON-SERIOUS INJURIES/ ILLNESS
VISITED A HEALTH CARE PROVIDER
VISITED AN ER
LIFE THREATENING
DEATH
REQ. INTERVENTION TO PRVNT PERM. IMPRMNT.
DISABILITY
SERIOUS INJURIES/ ILLNESS
CONGENITAL ANOMALY
NONE
outcomes


Baby Food Prod
count
0
100
200
300
400
500
NON-SERIOUS INJURIES/ ILLNESS
VISITED A HEALTH CARE PROVIDER
HOSPITALIZATION
VISITED AN ER
OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)
LIFE THREATENING
REQ. INTERVENTION TO PRVNT PERM. IMPRMNT.
DEATH
DISABILITY
SERIOUS INJURIES/ ILLNESS
NONE
CONGENITAL ANOMALY
outcomes


Mult Food Dinner/Grav/Sauce/Special
count
0
100
200
300
400
500
600
700
NON-SERIOUS INJURIES/ ILLNESS
VISITED A HEALTH CARE PROVIDER
VISITED AN ER
OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)
LIFE THREATENING
HOSPITALIZATION
REQ. INTERVENTION TO PRVNT PERM. IMPRMNT.
DISABILITY
NONE
SERIOUS INJURIES/ ILLNESS
DEATH
outcomes


Snack Food Item
count
0
100
200
300
400
500
600
700
NON-SERIOUS INJURIES/ ILLNESS
VISITED A HEALTH CARE PROVIDER
VISITED AN ER
OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)
LIFE THREATENING
REQ. INTERVENTION TO PRVNT PERM. IMPRMNT.
HOSPITALIZATION
DISABILITY
NONE
DEATH
outcomes

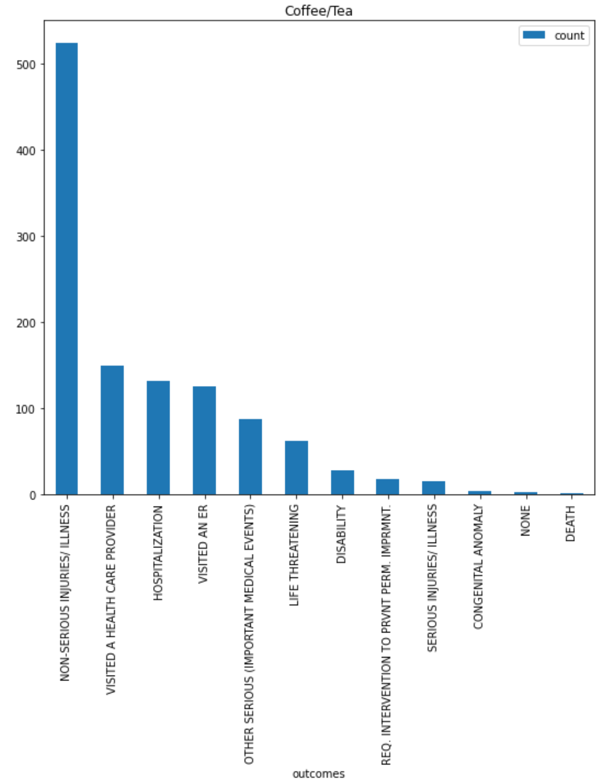

Coffee/Tea
count
0
100
200
300
400
500
NON-SERIOUS INJURIES/ ILLNESS
VISITED A HEALTH CARE PROVIDER
HOSPITALIZATION
VISITED AN ER
OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)
LIFE THREATENING
DISABILITY
REQ. INTERVENTION TO PRVNT PERM. IMPRMNT.
SERIOUS INJURIES/ ILLNESS
CONGENITAL ANOMALY
NONE
DEATH
outcomes

Figure 5: Breakdown of the number of serious events reported by gender

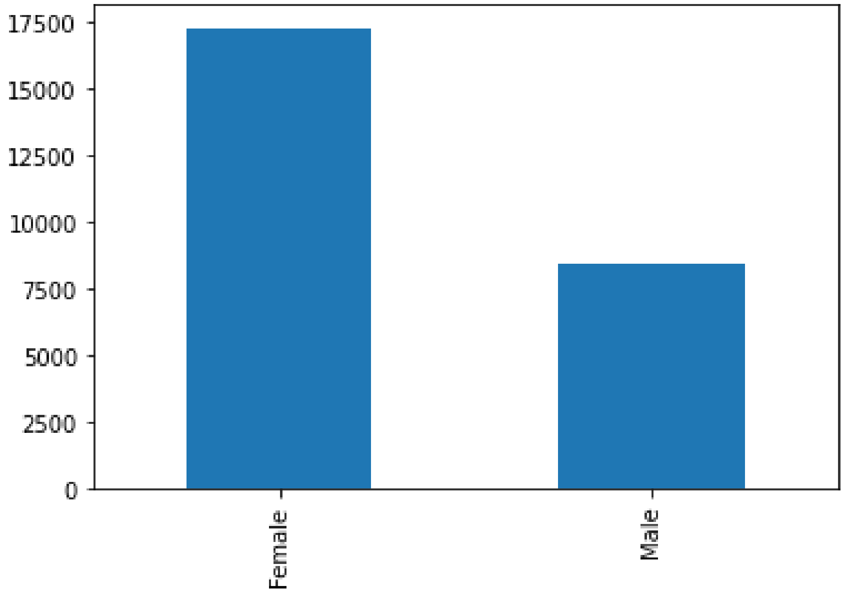


Figure 6: Comparison of the fraction of serious to non-serious cases grouped by gender prior to data cleaning

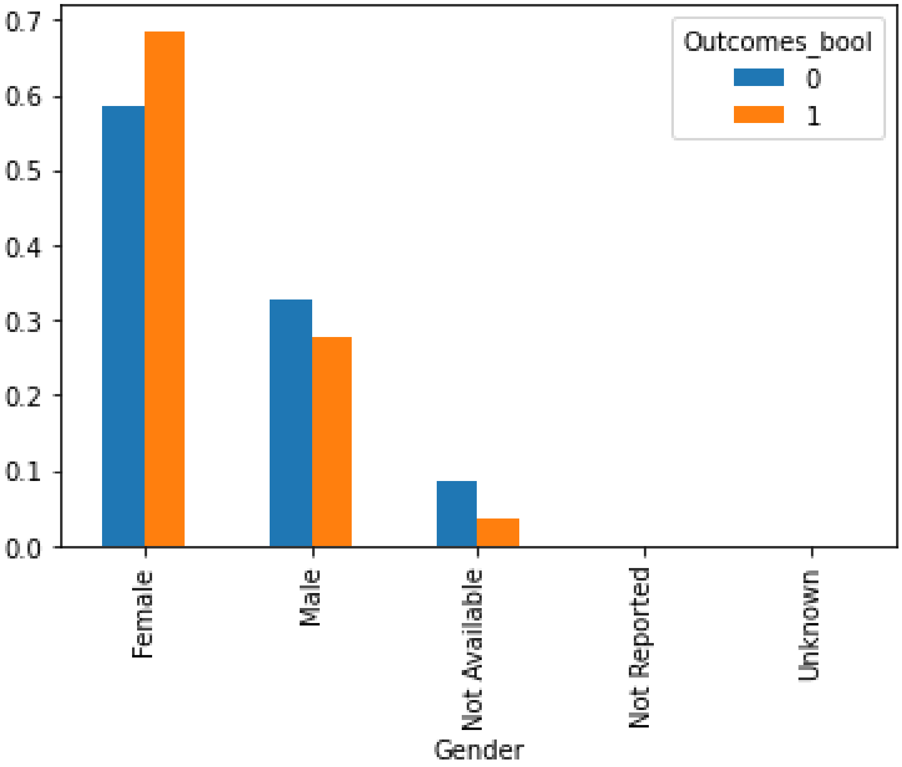


Figure 7: Feature importance bar plot from logistic regression model

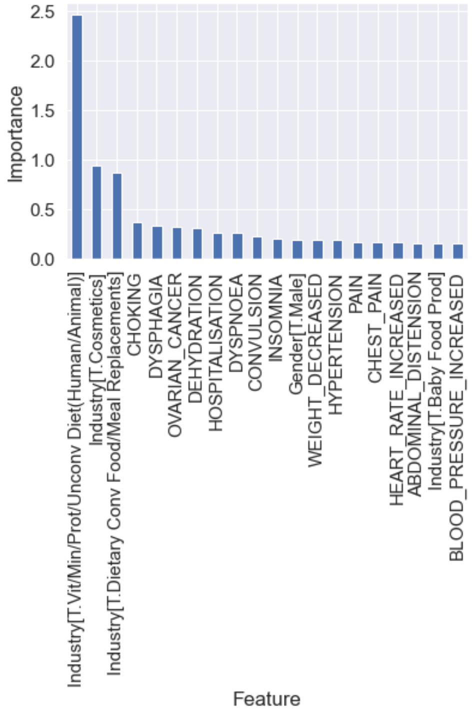


Figure 8: Feature importance bar plot from random forest model

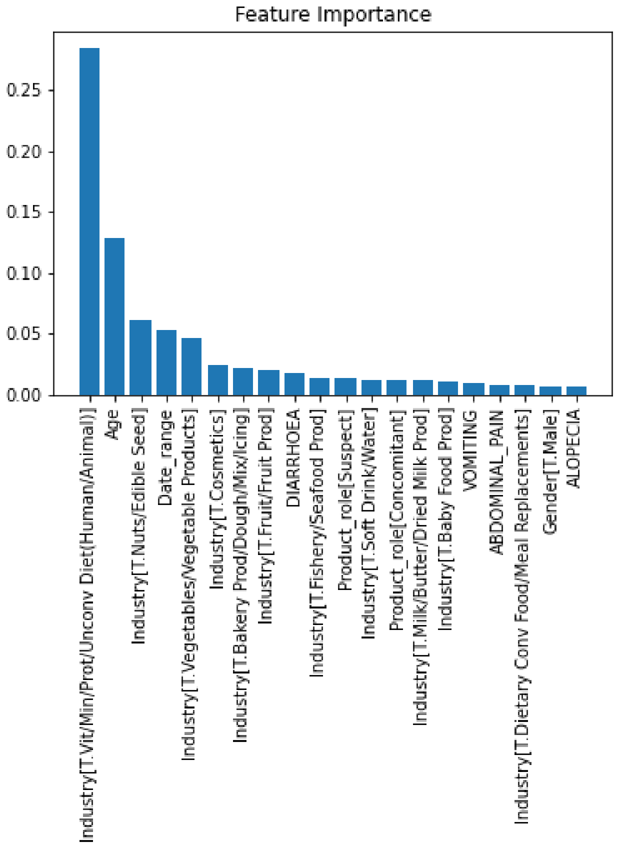


Figure 9: Confusion matrices for all models; contains the breakdown of true positives, true negatives, false positives, and false negatives

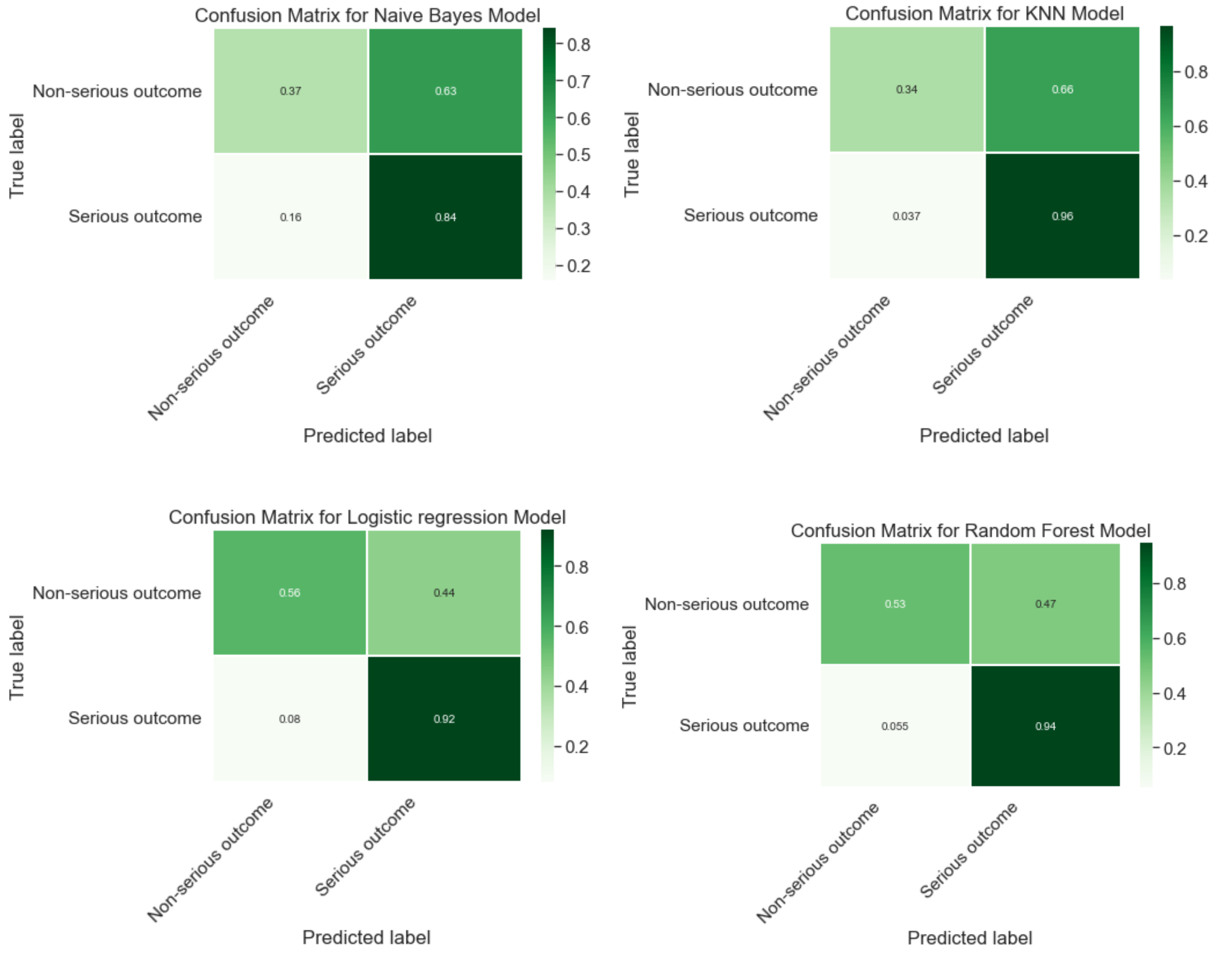
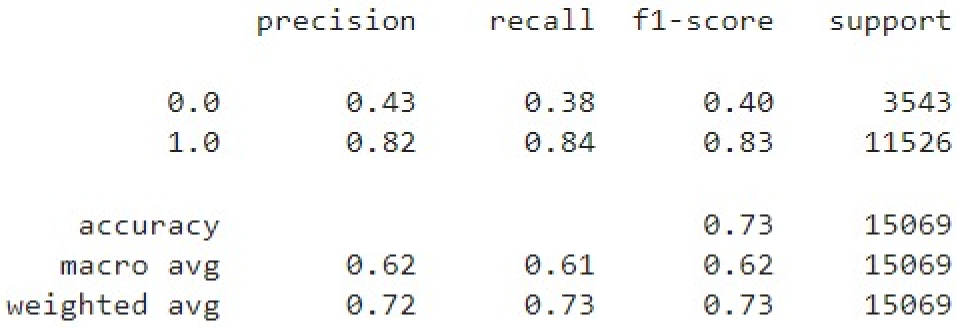


Figure 10: Calculated precision, recall, f1-score, and support for Naive Bayes

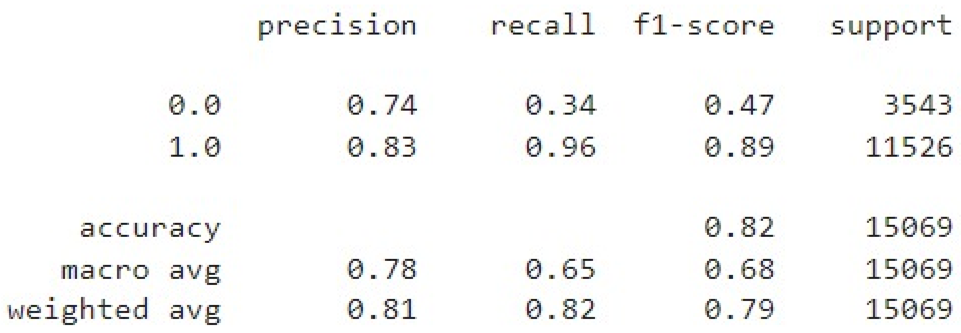


```
              precision    recall  f1-score   support

         0.0       0.43      0.38      0.40      3543
         1.0       0.82      0.84      0.83     11526

    accuracy                           0.73     15069
   macro avg       0.62      0.61      0.62     15069
weighted avg       0.72      0.73      0.73     15069
```

Figure 11: Calculated precision, recall, f1-score, and support for kNN

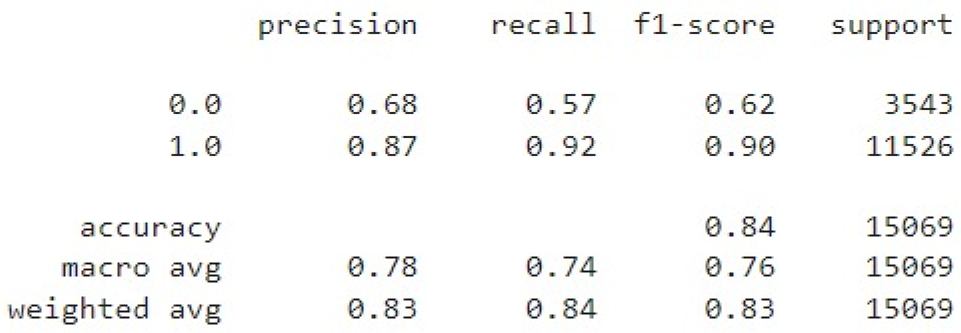


```
              precision    recall  f1-score   support

         0.0       0.74      0.34      0.47      3543
         1.0       0.83      0.96      0.89     11526

    accuracy                           0.82     15069
   macro avg       0.78      0.65      0.68     15069
weighted avg       0.81      0.82      0.79     15069
```

Figure 12: Calculated precision, recall, f1-score, and support for logistic regression

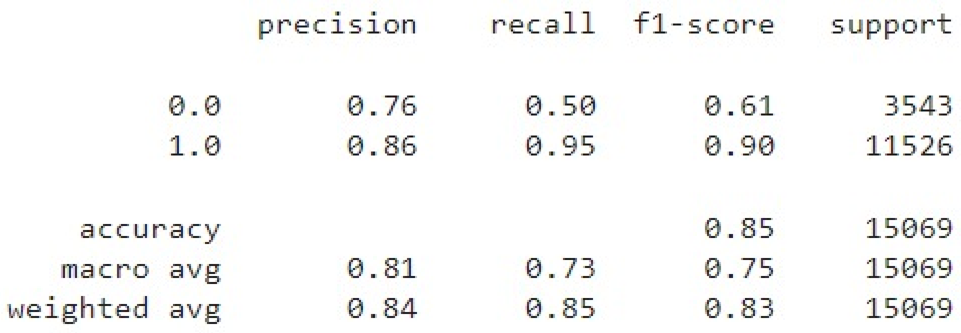


```
              precision    recall  f1-score   support

         0.0       0.68      0.57      0.62      3543
         1.0       0.87      0.92      0.90     11526

    accuracy                           0.84     15069
   macro avg       0.78      0.74      0.76     15069
weighted avg       0.83      0.84      0.83     15069
```

Figure 13: Calculated precision, recall, f1-score, and support for random forest



```
              precision    recall  f1-score   support

         0.0       0.76      0.50      0.61      3543
         1.0       0.86      0.95      0.90     11526

    accuracy                           0.85     15069
   macro avg       0.81      0.73      0.75     15069
weighted avg       0.84      0.85      0.83     15069
```

Table 1: Top 3 symptoms in each of the top 15 industries with frequently occurring symptoms

Industry	**Top 3 Symptoms**
Vit/Min/Prot/Unconv Diet (Human/Animal)	Choking, Dyspnoea, Nausea

Cosmetics	Ovarian Cancer, Alopecia, Pruritus
Nuts/Edible Seed	Diarrhoea, Abdominal Pain, Vomiting
Vegetables/Vegetable Products	Diarrhoea, Vomiting, Nausea
Soft Drink/Water	Diarrhoea, Nausea, Vomiting
Bakery Prod/Dough/Mix/Icing	Diarrhoea, Vomiting, Nausea
Fishery/Seafood Prod	Diarrhoea, Abdominal Pain, Nausea
Fruit/Fruit Prod	Diarrhoea, Vomiting, Nausea
Milk/Butter/Dried Milk Prod	Diarrhoea, Vomiting, Abdominal Pain
Dietary Conv Food/Meal Replacements	Diarrhoea, Vomiting, Nausea
Baby Food Prod	Vomiting, Diarrhoea, Irritability
Cereal Prep/Breakfast Food	Diarrhoea, Vomiting, Abdominal Pain
Mult Food Dinner/Grav/Sauce/Special	Diarrhoea, Vomiting, Nausea
Snack Food Item	Diarrhoea, Vomiting, Nausea
Coffee/Tea	Nausea, Vomiting, Diarrhoea

Table 2: Top 3 outcomes in each of the top 15 industries with frequently occurring symptoms; red coloring indicates serious events as defined by Table 3

Industry	**Top 3 Outcomes**
Vit/Min/Prot/Unconv Diet (Human/Animal)	Other Serious (Important Medical Events), Hospitalization,

	Visited a Health Care Provider
Cosmetics	Other Serious (Important Medical Events), Visited a Health Care Provider, Non-Serious Injuries/Illness
Nuts/Edible Seed	Non-Serious Injuries/Illness, Visited a Health Care Provider, Visited an ER
Vegetables/Vegetable Products	Non-Serious Injuries/Illness, Visited a Health Care Provider, Visited an ER
Soft Drink/Water	Non-Serious Injuries/Illness, Visited a Health Care Provider, Visited an ER
Bakery Prod/Dough/Mix/Icing	Non-Serious Injuries/Illness, Visited an ER, Visited a Health Care Provider
Fishery/Seafood Prod	Non-Serious Injuries/Illness, Visited an ER, Visited a Health Care Provider
Fruit/Fruit Prod	Non-Serious Injuries/Illness, Visited a Health Care Provider, Visited an ER
Milk/Butter/Dried Milk Prod	Non-Serious Injuries/Illness, Visited a Health Care Provider, Visited an ER
Dietary Conv Food/Meal Replacements	Hospitalization, Other Serious (Important Medical Events), Non-Serious Injuries/Illness
Baby Food Prod	Non-Serious Injuries/Illness, Visited a Health Care Provider, Hospitalization
Cereal Prep/Breakfast Food	Non-Serious Injuries/Illness, Other Serious (Important Medical Events), Visited a Health

	Care Provider
Mult Food Dinner/Grav/Sauce/Special	Non-Serious Injuries/Illness, Visited a Health Care Provider, Visited an ER
Snack Food Item	Non-Serious Injuries/Illness, Visited a Health Care Provider, Visited an ER
Coffee/Tea	Non-Serious Injuries/Illness, Visited a Health Care Provider, Hospitalization

Table 3: Classification of outcomes as serious or non-serious

Non-Serious Outcomes	**Serious Outcomes**
VISITED AN ER	REQ. INTERVENTION TO PRVNT PERM. IMPRMNT
VISITED A HEALTH CARE PROVIDER	HOSPITALIZATION
NON-SERIOUS INJURIES/ILLNESS	DEATH
NONE	LIFE THREATENING
	DISABILITY
	SERIOUS INJURIES/ ILLNESS
	CONGENITAL ANOMALY
	OTHER SERIOUS (IMPORTANT MEDICAL EVENTS)

Table 4: Top 10 important predictors according to Naive Bayes

Predictor	**Relative Importance**

Vegetables/Vegetable Products Industry	2.74
Nuts/Edible Seed Industry	2.73
Egg/Egg Prod Industry	2.73
Miscellaneous Food Related Items Industry	2.50
Fruit/Fruit Prod Industry	2.36
Food Service/Conveyance Industry	2.21
Bakery Prod/Dough/Mix/Icing Industry	2.07
Ice Cream Prod Industry	2.02
Cheese/Cheese Prod Industry	2.01
Overdose	2.00

References

1. Administration, F. and D. (2017, September 7). *Adverse Food Events*. Kaggle. Retrieved August 7, 2022, from https://www.kaggle.com/datasets/fda/adverse-food-events?select=README.pdf
2. Nadolpho. (2022, April 13). *Facts and stats - 50 million Americans have allergies: Acaai Patient*. ACAAI Public Website. Retrieved August 7, 2022, from https://acaai.org/allergies/allergies-101/facts-stats/
3. *Facts and Statistics - FoodAllergy.org*. (n.d.). Food Allergy Research & Education. Retrieved August 7, 2022, from https://www.foodallergy.org/resources/facts-and-statistics
4. Whipps, H. (2010, May 12). *Why are only some people allergic to some foods?* LiveScience. Retrieved August 7, 2022, from https://www.livescience.com/32581-why-are-only-some-people-allergic-to-some-foods.html