https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7872834/

In this paper, researchers want to use machine learning methods to predict foodborne disease pathogens (a total of four types) and investigate what predictors play significant roles using datasets from the National Foodborne Disease Surveillance Reporting System. They utilized four tree-based algorithms: a single decision tree, random forest, gradient boost decision tree, and adaptive boost decision tree.

During feature selection, researchers included predictors such as patient age, gender, time of illness, and home address recorded as three variables (province, city, and district): symptom, diagnosis, food name, and food type name. In particular, variables symptom and diagnosis are binary variables (diarrhea, vomiting, fever, etc.). Variable food names are preprocessed using a neural network called word2vec to convert texts to vector form. Variable food type names are one-hot coded into 23 food categories.

The gradient boost decision tree method has the highest accuracy, with a macro-averaged F1 score of nearly 69%. Moreover, they revealed that predictors such as latitude, longitude, sick time, and patient age affect predictions most (the importance value is calculated using Gini importance).

https://www.nature.com/articles/s41598-021-00766-w

In this paper, researchers developed a model called BERTweet to perform two tasks: the first one is to classify whether a given tweet was associated with a foodborne illness, and the other one is to extract critical information such as food type, symptoms, and location information in the given tweet that was predicted associating with a foodborne illness.

Researchers trained the model using a dataset manually labeled by professionals, where the model architecture was similar to the original BERT model. The f1-score for classifying unreported foodborne illnesses was 87%, outperforming all previous state-of-the-art models. The f1-score obtained for entity extraction was around 61%

After training the model, researchers applied the model to the U.S. Tweets only to analyze trends of foodborne illnesses across time. They discovered spikes in January to February, March to May, September to October, and December. Moreover, the top 20 food entities identified by the model in tweets in the U.S. were similar to the actual foodborne outbreaks, such as pizza, chicken, milk, cheese, and salad.

In the end, they even performed a case study by analyzing Tweets related to “lettuce” foodborne illnesses in 2018. The model successfully captured the two spikes of Tweets occurring from April to June and November to December.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7664469/

The model is GSARIMA - review for next time.

https://medinform.jmir.org/2021/8/e29433

In this paper, researchers want to predict foodborne diseases in various regions using a spatial-temporal model. This model was based on LSTMs and used an encoder-decoder structure to predict regional foodborne illnesses. Furthermore, to increase robustness, researchers introduced a spatial-temporal attention mechanism. Ultimately, they evaluated the model's external features, such as holidays and temperature.

In particular, the model framework consists of five parts. The first part is the data generation module, where all the processing for temporal sequence and multiple spatial graphs is done. The second part is the multigraph fusion module, where numerous spatial correlations are considered. The third part is the encoder-decoder module that uses LSTM networks to model temporal dependence and spatial dependence of foodborne diseases using the edge LSTM and the node LSTM, respectively, in the encoder. The node LSTM is used in the decoder to predict foodborne disease risk in each region. The fourth part is the spatial-temporal attention module, which assigns temporal importance values to timesteps and spatial importance values to adjacent edges of nodes. The last part is the external feature embedding module that considers the encoder's exterior features, such as holidays and temperature.

The performance of this model outperforms all the state-of-the-art models in predicting foodborne diseases in China. Moreover, this model is validated using a case study where the model results match the real-world situation.

https://www.nature.com/articles/s41746-018-0045-1

Researchers in this paper built a model called FINDER, which can identify restaurants likely to be considered unsafe during inspection. FINDER applies machine learning to Google search and location logs to infer which restaurants have major food safety violations. To increase the robustness of the model, FINDER leverages a collection of signals beyond the query string itself, such as the aggregated clicks on those results and the content of the opened web pages.

Of all the restaurants identified by FINDER, 52.3% were deemed unsafe upon inspection, which is outstanding compared to all the other models.

New Summaries:

https://academic.oup.com/aje/article/190/10/2188/6239824?searchresult=1

In this paper, researchers proposed several ways for generating hypotheses that can significantly ease subsequent time, effort, and resources needed and increase the likelihood of rapid and conclusive prediction of the contaminated food vehicle. In detail, the researchers presented a framework for hypothesis generation focusing on three primary sources of information: 1. known sources of the pathogen causing illness; 2. person, place, and time characteristics of cases associated with the outbreak; and 3. case exposure assessment.

https://www.pnas.org/doi/10.1073/pnas.2115714119

In this paper, researchers created a model called SOURCE to replicate how risks of opioid misuse will evolve. SOURCE is a dynamic, continuous-time differential equation simulation model that tracks the US noninstitutionalized opioid-using population whose age is more significant than twelve. It closely replicates the historical trajectory of the opioid crisis from 1999 to 2020. In particular, across all 15 time series used in model estimation, the average r-squared for simulated values is 75.6%.

https://www.sciencedirect.com/science/article/pii/S0749379720303998

This paper summarizes a meeting discussion regarding the ongoing opioid systems modeling work. Experts in this meeting discussed key national data sources, data needs, and data considerations for developing opioid modeling. In summary, the main issues raised revolve around definitional inconsistencies, a general lack of data, and problems with overestimation, underestimation, and potential overlap between data sources. Some potential solutions presented by the experts include maintaining close collaboration among modeling teams, enhancing data collection to better-fit modeling needs, and focusing on bridging the most crucial information gaps.

https://academic.oup.com/epirev/article/43/1/147/6427243?login=true

In this paper, researchers searched and screened 1398 articles about simulation models of opioid use and overdose. Among them, 88 studies are qualified and further analyzed. The most frequent types of models among those 88 papers are compartmental(36%), Markov(20%), system dynamics(16%), and agent-based models(16%).

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7664469/

In this paper, researchers present a model for forecasting non-typhoidal salmonellosis outbreaks. This model is based on the fitted values of the time series modeled by GSARIMA. It is validated by analyzing the case of Salmonella enterica serovar Enteritidis in Sydney, Australia.

In the case study, the final covariates selected include two lags, yt-1 and yt-2; three predictors mean maximal temperature, tell 3 pm relative humidity, and weekly demand for eggs to predict the number of reports of Salmonella enterica serovar Enteritidis cases. The final r-squared value for this model is 0.88.