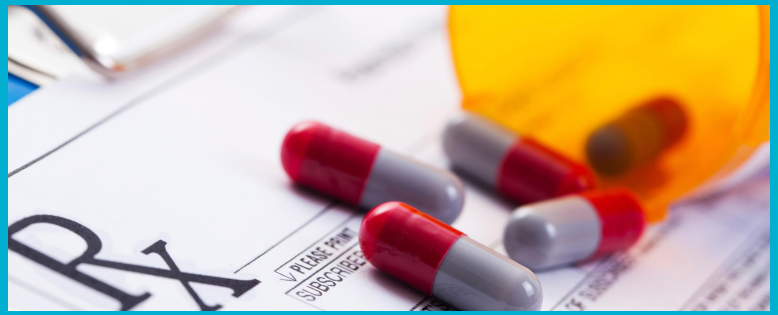
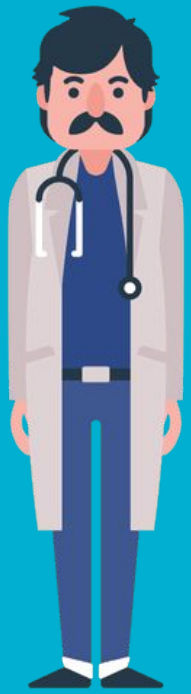
Predicting adverse drug reactions based on patient demographic information and other factors



Anthony Olvera and Annie Lydens

Project Description



For our semester project, we analyzed the FDA's drug adverse events data collection using data mining techniques in hopes of obtaining interesting findings. We focused on how outcomes were affected by different demographic characteristics. We also sought to resolve whether there are pairs of frequently occurring sets of specific patient demographic attributes. In addition, we also paired sets of common multiple reactions with certain medications which tend to occur together in the treatment process.

Questions Sought

- Given a patient's demographic information (age, sex, weight, reported location), can we predict the likelihood of a significant medical outcome?
- What are the most common set of patient characteristics and adverse drug reactions?
- Do certain medical reactions occur together frequently? If so, what are those reactions?

Data Preparation

Dataset Procurement and Preprocessing

- FDA's Adverse Event Reporting System (FAERS) Quarterly Extract datasets fetched using shell script and Chrome extension
- Dataset covers the period of 2016 Q1-2018 Q1
- Dataset was imported into a Postgresql database with 37 million+ rows
    - DB Tables: indications, outcomes, reactions, report sources, drug therapy dates, demographic info and drug info.
- Data was also downloaded as JSON files

Data Preparation (cont)

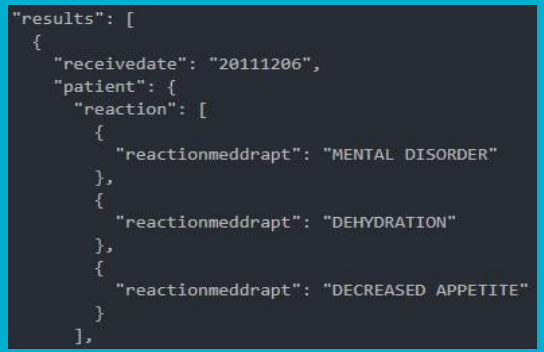
Data Cleaning and Preprocessing

- Remove duplicate adverse event cases and reports
- Missing value imputation
- Standardizing date formats and weight fields to kilograms
- Attribute subset selection by dropping unnecessary attributes
- Numerosity reduction by retaining only “Primary Suspect” drugs and primary (most representative) reactions
- Process resulted in 2 cleaned and preprocessed subsets for analysis

Data Preparation (cont)

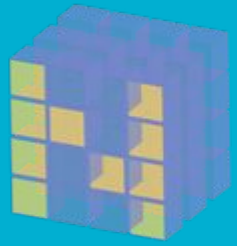
JSON Parsing

- There are a total of 788 JSON files available for download at https://open.fda.gov/tools/downloads/
- The files were downloaded and inflated as a batch using the chrome extension
- Each file contained many nested dictionaries which were flattened using a python script.
- All irrelevant information is dropped
- Data types were changed, values converted
- Outliers are eliminated
- Numeric attributes are discretized into bins

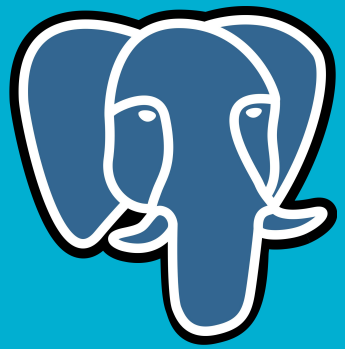


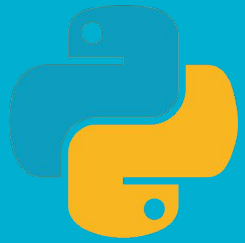
Tools Used

- Python
- Python JSON library
- Pandas
- Numpy
- SciPy
- Postgresql
- Scikit-learn
  - GaussianNB
  - BernoulliNB
  - Sklearn.model_selection
  - Sklearn.metrics



NumPy



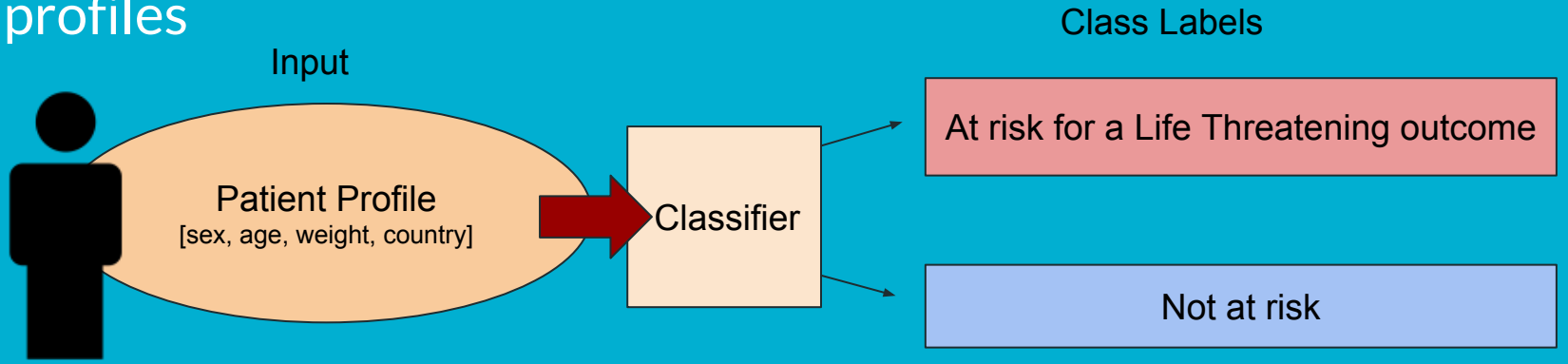


matplotlib

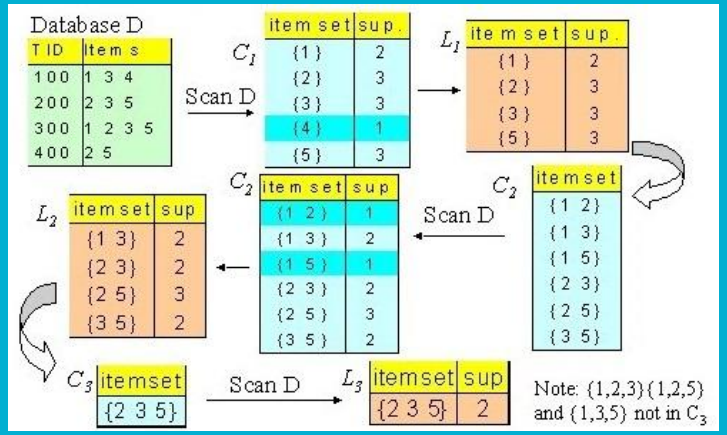
Data Mining Techniques Applied

Classification using Naive Bayes

Classified patient into one of two classes: “life-threatening reaction” and “non-life threatening reaction” based on their demographic profiles



Techniques Applied (cont)



Results from Apriori Analysis

- First, a set of each unique value is generated
- Next, the table is scanned once for each item-set of size k and each which meets the minimum support kept in a list . This continues until no more item-sets meet the minimum support.
- Lists were generated for all item-sets meeting a minimum support thresholds of 20%, 10%, 7%, 5%, 3%, 1%
- Most sets were not interesting such as {“male”, “elderly”}
- Most interesting sets were discovered with minimum supports at or below 5%
- Support and confidence is calculated for the sets we find interesting.

Knowledge Gained

Naive Bayes Classification

- Older men most at risk
- Younger women least at risk
- Changing even a single attribute in a patient test vector affected the estimated probabilities
- Significant differences between reporting countries

Estimated Probability of A Life Threatening Reaction Based on Demographic Attributes

	Male, 60, 80kg	Female, 60, 80kg	Male, 30, 80kg	Female, 30, 80kg
JP	13.36%	8.32%	10.60%	6.52%
GB	11.54%	7.14%	9.12%	5.58%
US	10.83%	6.67%	8.54%	5.21%
CH	13.90%	8.69%	11.05%	6.82%
CN	12.38%	7.68%	9.80%	6.02%
CA	12.85%	7.99%	10.19%	6.26%
DE	11.95%	7.40%	9.45%	5.79%
IT	15.13%	9.50%	12.06%	7.47%

Knowledge Gained (cont)

Results from Apriori Analysis

- Most common single drug reaction: drug ineffective with a support of 5.468%
- Most common drug in use in the data set is “FORTEO” , followed by “CRESTOR” and “LUNESTA”
- Most frequent pair of patient characteristic and reaction was drug ineffective and female : DRUG INEFFECTIVE => female(0.03497, 0.59910)
- Most frequent pair of drug and reaction : LUNESTA => DRUG INEFFECTIVE(0.02126, 0.47654)
- Most frequent pair of two reactions : INSOMNIA => DRUG INEFFECTIVE(0.01166, 0.46942)
- What do Lunesta and Insomnia have in common?

Knowledge Applications

- Bring awareness to common co-occurring adverse reactions.
  - Could help physicians support patients during treatment

- Allow medical staff to understand demographic factors that may contribute to adverse reactions in patients.

- Help identify groups of patients as higher risk for life threatening reactions.