Abstract

In an article published by Harvard Health in 2019, it was discussed that the major problem with trying out new medications is the lack of knowledge of the side effects that are presented with it. From a drug review dataset that includes various features such as drug rating, reviews, useful count (amount of users that found the review useful), drug name, and drug condition. Topics were modeled from the reviews and in addition to the sentiment they were compared to the usefulness. This will give insight into what types of topics (side effects) are most important to a customer when trying out a new drug.

Design

To conduct this; EDA will be conducted after text-preprocessing and sentiment analysis has been run for the review. Using the processed text we will run vectorize these results and run a topic modeling algorithm (CorEx, NMF, LSA) to see which topics are most prevalent in each review. Then we will order these results by useful count to see which topics were found to be the most useful.

Data

The data used in this project was obtained through Kaggle from a UC Irvine drug review compilation. (https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018). This dataset contains over 160000 unique inputs. Each row describes an individual's condition, review, overall rating, and usefulness of the evaluation by other people. I hope to be able to run a sentiment analysis on these reviews to get an idea of how review makes it beneficial to other patients.

Algorithms

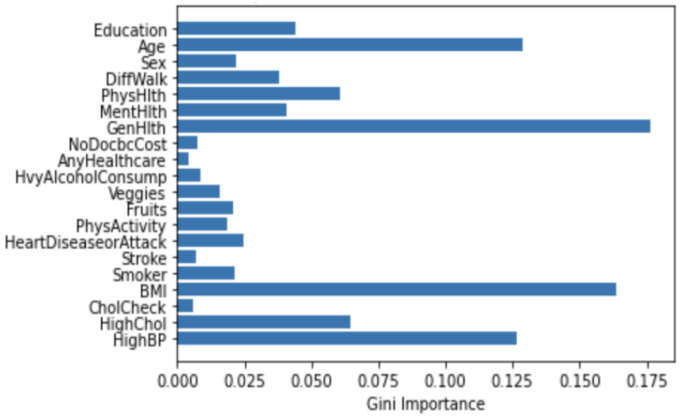
- Exploratory data analysis was done on the raw data to view the basic relationships between
- CorEx, NMF was used to do topic modeling

Tools

- Pandas, Numpy for EDA and data manipulation
- NLTK and spaCy was used for text processing
- Topic Modeling was done with sklearn and nltk's CorEx, NMF, LSA

Communication

Influence on Diabetes



Influence on Diabetes