Predicting Health

Mining patient data to gain invaluable insights

Group 2

Anthony Le

Description

For my data mining project, I aim to build a predictive model using a dataset containing healthcare data. There are a variety of interesting questions that can be answered using this data, from trying to predict patient outcomes based on health indicators of a patient and their demographic to helping patients navigate learning about how costly treatment may be for their condition or how long they may expect to be hospitalized.

Prior Work

Field of healthcare analytics has always been pretty vast, and with the boom in machine learning, there has been more extensive looks into building predictive models that help healthcare providers offer more robust care for patients or help providers identify how to provide better care for patients, helping identify various other indicators of health that aren't so obvious that can change the health outcome for a patient

- https://news.vumc.org/2024/03/05/ge-healthcare-vanderbilt-publish-data-on-ai-models-predicting-patient-response-to-immunotherapy/
- https://aws.amazon.com/blogs/publicsector/one-small-team-created-cloud-based-predictive-modeling-solution-improve-healthcare-services-uk/
- https://healthitanalytics.com/news/improving-social-determinates-of-health-with-predictive-analytics

Datasets:

There were two datasets that were of interest to me

- SyntheticMass
    - 1,000,000 records
        - Can help train a model to be more robust with more data to look at
- Healthcare Dataset
    - 10,000 records
        - Can be useful to try and look at something small scale first

Given the nature of HIPAA, both dataset are completely synthetic - none of the entries in these dataset belongs to any real patient. While it is synthetic, it still provides an avenue to be able to explore healthcare analysis and patient records

I haven’t downloaded the SyntheticMass dataset yet but I have downloaded and done some exploration into the second dataset

Proposed Work

There is plenty of work to be done with the dataset I chose, ranging from:

- Trim the overall structure of the dataset
  - SyntheticMass zip file is broken down further into 12 output folders that contains various data formats (JSON, CSV, etc.)
    - Looking at using just CSV, need to explore what data sets are worth keeping and look to combine/consolidate the files into one
- Identify missing or null values and develop an approach for handling them (removal, insertion of “default value”, etc.)
- Identify and remove any duplicate records
- Standardize the format of values (consistent capitalization, correct identify misspellings, etc.)
- Identify any outliers and develop an approach for handling them
- Convert categorical variables to numeric values
- Look into more literature on predictive models and how to build out a rudimentary one for the project

Tools that can be used that will be helpful for this proposed workflow will be:

- Python
  - Plenty of libraries to help augment the language and data workflow
- Git/GitHub
  - Helpful to help with having a non-local repository to have my work saved to that I can pull and work on anywhere

Evaluation

To evaluate my results, there are various quantitative metrics I could utilize - sklearn/scikit library offers plenty of modules with methods that can help provide numeric values to evaluate my project on: accuracy, precision, confusion matrix, etc.

Another way is to find and utilize results obtained from real healthcare data that may have been published in research papers to compare my results against and see how far off they may be. For example, if I utilize the model to predict cost of treatment for X condition or length of hospitalization for X condition, I can see what the actual average value for those outcomes is and see how it compares - although there is some discrepancy here as there is a variety of different complexities that can come into play that isn't accurately captured in a synthetic dataset.

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7239479/
    - Research published in 2020 that studies hospitalizations of cancer patients that I could look at

The concept of evaluating my project and finding ways to meaningfully evaluate it is still a bit over me, so this will still be a ongoing and rolling thought process of determining ways to evaluate my results as best I can