Group 3: Grace Cote, Sara Edgar, Michael Good

UPENN-PHI-DATA-PT-09-2020-U-C

Final Project Proposal

**Am I Dying?**

*Topic Overview*

Nearly 80 million people use WebMD every month to research diseases, self-diagnose, and connect with community programs. We wanted to take a swing at making our own diagnostic model using a simplified dataset

.

*Project Overview*

The ***Project***takes a set of symptoms and evaluates what disease the patient is experiencing. First we will categorize each patient based upon their symptoms and then we will evaluate the accuracy of the model with a testing dataset. We will evaluate not only the binary accuracy of the model: “Did the model correctly categorize the patients?” but also how significant a deviation was made: “The patient had allergies and was diagnosed with the common cold vs the patient had tuberculosis and was diagnosed with the common cold.”

We plan on using SQL, jupyter notebook, plotly, scikit and heroku to launch the project.

*Data Source*

<https://www.kaggle.com/kaushil268/disease-prediction-using-machine-learning>

The dataset being used for the project provides over 130 symptoms with binary data and a diagnosis of over 40 diseases to go along with each patient. The data is used to teach physicians-in-training what symptoms intersect to become what disease. While it already comes separated as training and test, the two files will be combined in order to create unique train and test datasets.