Problem Statement Worksheet (Hypothesis Formation)

Can we build a machine learning model to accurately (>95%) classify 41 diseases based on the presence or absence of 132 symptoms in a presenting patient?

H

1 Context

Diagnosis continues to be one of the trickiest steps of treating patients, and can create wildly different outcomes for patients based on the interpretation and accuracy of individual physicians. Creating an accurate machine learning model that presents a ranking of the most likely diseases a patient may be suffering with can streamline the diagnosis process and create more uniform, better patient outcomes.

2 Criteria for success

I will seek to build a machine learning model to classify a set of 41 diseases based on the presence or absence of 132 diverse symptoms. The successful model should have >95% accuracy, and I will also seek to include a feature that presents the "next alternative" diagnosis to account for any inaccuracies in the model and give the patient and physician a chance to use their reasoning and experience to make final diagnosis conclusions.

3 Scope of solution space

This model will apply only to the 41 diseases included in our data set, and will only be useable when the presence or absence of each of the 132 symptoms has been accounted for in each patient's case.

4 Constraints within solution space

41 diseases is a limited number in the scope of all possible diseases, which may hinder useability by physicians.

Assessing 132 individual symptoms in all patients will be time-consuming and may turn off some patients and physicians.

5 Stakeholders to provide key insight

Myself - for the purpose of creating a Capstone Data Science project from beginning to end

My mentor - Kenneth Gil-Pasquel, who will offer feedback and direction on this project as I work through it

6 Key data sources

Data is sourced from Kaggle at:
https://www.kaggle.com/kaushil268/disease-prediction-using-machine-learning

There are distinct training and test sets. The training set has 120 instances of each diagnosis, and the test set tests for accurate classification of each disease in the training set.

H D E I P