Capstone Project 3 Proposal

Adam Reichenbach
Springboard Data Science Career Track, June 2024 Cohort

For my 2nd Springboard Capstone project I will be using a dataset containing information of patients and whether or not they have suffered a stroke. The data can be obtained from *Kaggle.com*.

Context and Problem Statement

According to the World Health Organization (WHO), stroke is the 3rd leading cause of death in 2021 globally (2nd in low-income, 3rd in lower-middle-income, 1st in upper-middle income, 3rd in high-income countries).

Can we use the collected data to make a machine-learning algorithm that can classify how susceptible someone is to having a stroke based on different variables (age, sex, smoking status, heart disease status, employment type, BMI, etc.)?

Criteria for success

We need to develop a classification machine learning model that can give a very accurate prediction on whether a patient will have a stroke or not based on the other data that is included.

Solution Space

We are trying to build a model that can predict if someone will have a stroke or not, this will be a classification ML model. Key metrics of focus will be accuracy, precision, recall, F1-score, and AUC-ROC.

Some possible constraints is that the data may not include other features that could lead to the possibility of strokes such as medications that they may be taking, or family history of strokes or other medical issues, blood pressure, cholesterol levels, sleep quality. From the datacard itself, there looks to be some serious imbalance with not only a couple of the features, but also the target variable has a ton more non-strokes than strokes recorded. There are some techniques that can help overcome that last problem such as SMOTE.

Stakeholders

- Hospitals and healthcare providers since the model can be used for early intervention and helps doctors give more correct diagnoses as well as give better preventative care for their patients.
- Medical software & tech companies can use a model that they can integrate into their systems that could help HCPs and patients alike.
- Public health agencies who want to help reduce strokes to the population as a whole and help distribute resources to more at-risk people accordingly.
- Health insurers who want to reduce the amount of expensive claims they have to pay out to stroke victims and encourage more preventative care.

- Medical researchers can use the model to help uncover other patterns not seen by anyone else.

Data Source

There are 11 features (already excluding the unique patient ID column) in the data and we will build a classification model that trains the dataset to predict someone's susceptibility to having a stroke. The data used is from Kaggle and the author's data source is confidential as it is used for educational purposes only.

Data source:

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data