Predicting whether a patient is likely to get stroke or not

Tariq Almutiri

Abstract

The goal of this project is to use classification model to predict which patients are close to have stroke and compute the accuracy of this model. These outcomes will help to focus on potential patients and avoid them being in stroke critical case. I will work on dataset that published on Kaggle.com with URL:

“https://www.kaggle.com/fedesoriano/stroke-prediction-dataset”

Data Description

The dataset contains 5110 unique records with 12 attributes for each, collecting from 2995 females and 2115 males. The last column contains ‘1’ if the patient had stroke and ‘0’ if he or she hadn’t. A few attributes like age, marriage, and hypertension will be consider in prediction model. The dataset will be divided in three parts:

- 20% for testing.
- 60% for training.
- 20% for validation.

##Tools

- pandas.
- numpy.
- matplotlib.
- seaborn
- sklearn.

##MVP goal:

- I will use pandas and numpy to extract the dataset and clean it. Then matplotlib and seaborn to plot and visualize data in exploratory data analysis phase. In step of applying models on dataset I will use four or five classification algorithms from sklearn like:
    - K-Nearest Neighbors,
    - Decision Tree,
    - Naive Bayes,
    - Support Vector Machines,
    - and Logistic Regression.

  models will apply on training and validation datasets. Merging validation dataset into training dataset is the next step. At the end, models’ performance will be measured by cross-validation on training and testing datasets to pick which model is best for this kind of data.