Stroke Prediction

First Deliverable

Team - Group 2

Kavya Chigurupati	**801210026**

GitHub Link

https://github.com/kavyachigurupati/StrokePrediction

Domain

The domain we chose to work in is Health, Medicine. We perform supervised learning to predict the stroke.

Problem Statement

Stroke is a medical disorder that occurs when there is damage to blood arteries. It can also happen if blood stops flowing to the brain.It is the most common cause of death and disability worldwide, according to the World Health Organization. It is the fifth-highest cause of mortality, according to the Centers for Disease Control and Prevention (CDC).

Prior identification of this medical condition could help save the lives of people by providing appropriate medical support. With the help of machine learning, we can identify the chance of occurrence of stroke. So, the goal of this project would be to predict the accuracy of a person getting a stroke based on the textual data provided by Kaggle.

Research

Academic Review

We have identified a few research papers which tried to solve stroke prediction by applying various data preprocessing techniques and machine learning models.

https://thesai.org/Downloads/Volume12No6/Paper_62-Analyzing_the_Performance_of_Stroke_Prediction.pdf https://thesai.org/Downloads/Volume9No1/Paper_63-Prediction_of_Stroke_using_Data_Mining.pdf

Scope of the project.

We are trying to predict the chance of occurrence of the stroke to a person by feeding the data to a machine learning algorithm. Initially, we would perform data collection, then apply data cleaning techniques to remove noise, outliers, duplicate values, and fix the missing values. Secondly, We analyze the data and perform preprocessing steps. We feed this data to the Machine learning model and analyze the results. Finally, construct a confusion matrix of the results.

What are we trying to achieve?

We analyzed a few states of art techniques mentioned in academic papers that solved stroke prediction. They used various data analyzing steps; and machine learning models like Naive Bayes, random forest, and decision trees. Most of the methods they followed had drawbacks related to either accuracy or could not solve the problem efficiently. Some models do not suit real-world examples as they selected specific features and did not use all the features from the dataset.

So, we are trying to build a model that solves most of the issues mentioned above. Our goal is to

1. Train a model that includes all the features and produces accurate results compared to state-of-art techniques.
2. Compute a confusion matrix for the results.

Data Source

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

We have downloaded the dataset from Kaggle. The dataset contains 5110 observations with 12 attributes to proceed with this task. For training and test, we will split the dataset to 80% for training and 20% for testing.

Future Work

1. Our future goal would be to train the model in the least time possible.

2. The major drawback is that the current model would only work on textual data and not on real-world image datasets. We would improvise the model in a way that could work on heart or brain CT scan images.