Class: DSC630 T302 2231-Fall2022

Course: Predictive Analytics

Author: Anjani Bonda

Date: Sep-10 2022

**Project Milestone2**

**(Prediction of Stroke)**

**Introduction:**

This is something personal topic to me as I have seen a lot of people in my family and friends experienced a stroke. A stroke, sometimes called a brain attack, occurs when something blocks blood supply to part of the brain or when a blood vessel in the brain bursts. In either case, parts of the brain become damaged or die. A stroke can cause lasting brain damage, long-term disability, or even death. The brain controls our movements, stores our memories, and is the source of our thoughts, emotions, and language. The brain also controls many functions of the body, like breathing and digestion. To work properly, your brain needs oxygen. Your arteries deliver oxygen-rich blood to all parts of your brain. If something happens to block the flow of blood, brain cells start to die within minutes**,** because they can’t get oxygen. This causes a stroke.

Common Symptoms:

* Sudden numbness or weakness in the face, arm, or leg, especially on one side of the body.
* Sudden confusion, trouble speaking, or difficulty understanding speech.
* Sudden trouble seeing in one or both eyes.
* Sudden trouble walking, dizziness, loss of balance, or lack of coordination.
* Sudden severe headache with no known cause.

Facts & Stats per CDC:

* In 2020, 1 in 6 deaths from cardiovascular disease was due to stroke.
* Every 40 seconds, someone in the United States has a stroke. Every 3.5 minutes, someone dies of stroke.2
* Every year, more than 795,000 people in the United States have a stroke. About 610,000 of these are first or new strokes.
* About 185,000 strokes—nearly 1 in 4—are in people who have had a previous stroke.2
* About 87% of all strokes are [ischemic strokes](https://www.cdc.gov/stroke/types_of_stroke.htm), in which blood flow to the brain is blocked.
* Stroke-related costs in the United States came to nearly $53 billion between 2017 and 2018. This total includes the cost of health care services, medicines to treat stroke, and missed days of work.
* Stroke is a leading cause of serious long-term disability.2 Stroke reduces mobility in more than half of stroke survivors age 65 and older.
* Stroke is a leading cause of death for Americans, but the risk of having a stroke varies with race and ethnicity.
* Risk of having a first stroke is nearly twice as high for Blacks as for Whites, and Blacks have the highest rate of death due to stroke.
* Though stroke death rates have declined for decades among all race/ethnicities, Hispanics have seen an increase in death rates since 2013.

Some of the risk factors that can possibly increase one’s risk of stroke include:

* Age
* Sex
* Race/Ethnicity
* High Blood Pressure
* High Cholesterol
* Heart Disease
* Diabetes
* Unhealthy Diet/Physical Inactivity
* Obesity
* Tobacco/Alcohol Use

**Problem Statement/Project Proposal:**

How can we predict if a person is likely to experience a stroke?

What’s the likeliness of a stroke based on his/her available medical history?

The project proposal therefore is “Stroke Prediction” using common risk factors and applying predictive modeling techniques.

**Dataset Used:**

The dataset used for this project is sourced from Kaggle.com. This dataset is a popular one from Kaggle with about 12 attributes and 5110 observations. Here is the link for the data source being used.

[**https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download**](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download)

**Attribute Information**

1) id: unique identifier  
2) gender: "Male", "Female" or "Other"  
3) age: age of the patient  
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension  
5) heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease  
6) ever\_married: "No" or "Yes"  
7) work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"  
8) Residence\_type: "Rural" or "Urban"  
9) avg\_glucose\_level: average glucose level in blood  
10) bmi: body mass index  
11) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*  
12) stroke: 1 if the patient had a stroke or 0 if not  
\*Note: "Unknown" in smoking\_status means that the information is unavailable for this patient

**Models to be used/Approach:**

This project will follow CRISP-DM model for data exploration/understanding, modeling, and evaluation. Python programming language will be used for data load, exploration, fitting the models and testing the accuracy along with some of the machine learning libraries that might be required.

Firstly, EDA will be performed to better understand the data and compute descriptive statistics followed by visualizations representing the dataset.

Secondly, various machine learning algorithms will be employed to examine and choose the best possible prediction model. Some of them include Logistic Regression, K Nearest Neighbor (KNN), Decision Tree, Naïve Bayes, Support Vector Machine (SVM) and Random Forest Classifier.

**Evaluation of results:**

Based on the models used and carefully examining them, I expect to curate an effective and most accurate and sensible working model (with train and test datasets) to predict the likeliness of stroke in patients based on the available medical history of patients from the given dataset (sourced).

**Risks:**

Following the potential risks involved in this project:

* This project assumes that the dataset sourced is valid and accurate.
* There is always a chance of data loss due to incompleteness or extremities/anomalies.
* The number of observations in the dataset used is low (5110) for this study.
* This dataset does not include all the risk factors of a stroke like diabetes, alcohol consumption, physical activity etc., hence the model cannot be fully accurate.

**Contingency Plan:**

If for some reason, the project cannot be completed with the available data or shows no sign of predictive potential, then I would take a portion of data or consider a new similar dataset altogether to come up with a possible predictive outcome/model.

**Learning:**

I anticipate and hope to learn the application of predictive modeling techniques on a given dataset and come up with the best possible accurate and effective model/outcome. Learn to come up with all the relevant information to be able to present to the management/teams to be able to understand and make informed decisions. Also, to learn and improve the approach and model deployment based on the inputs/review from peers.

**References:**

1. Centers for Disease Control and Prevention. [Underlying Cause of Death, 1999–2018](http://wonder.cdc.gov/ucd-icd10.html). CDC WONDER Online Database. Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020.
2. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. [Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Associationexternal icon](https://www.ahajournals.org/doi/10.1161/CIR.0000000000001052). *Circulation*. 2022;145(8):e153–e639.
3. Jackson G, Chari K. [National Hospital Care Survey Demonstration Projects: Stroke Inpatient Hospitalizationsexternal icon](https://pubmed.ncbi.nlm.nih.gov/32510306/). *Natl Health Stat Report.* 2019 Nov;(132):1-11.
4. Fang J, Keenan NL, Ayala C, Dai S, Merritt R, Denny CH. [Awareness of stroke warning symptoms—13 states and the District of Columbia, 2005](http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5718a2.htm). *MMWR.*2008;57(18):481–5.
5. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download