**Stroke Prediction Program**

Group 4: Atabay Kadiroglu, Bingxuan Hu, Qiaoyi Tan, Wenyi Wang, Xin Bian, Zixin Ye

**I. Problem Statement**

Stroke (Brain attack) occurs when the blood flow to a person’s brain is blocked by a blood clot or bursts. In either case, part of the brain cannot get the oxygen it needs, and blood cells die, causing lasting brain damage, long-term disability or even death. According to the World Health Organization (WHO), stroke is the second leading cause of death and major cause of disability worldwide.

In the United States, on average someone has a stroke every 40 seconds, and someone dies of a stroke every 4 mins. Clearly, stroke is a major health problem in the U.S, but it poses a big financial problem as well. Stroke related costs were almost $103 billion in 2016 in the US, and with the aging population and double the risk of stroke in each decade after the age of 55 years, these costs are projected to increase to $240.67 billion by 2030, an increase of 120%.

Analyzing the available data set, our group intends to create a model to predict the likelihood of someone having a stroke, as well as a program for our patients to input their information and get an predicted probability of having a stroke. We believe our analysis will help our patients take early actions such as making healthy choices or controlling any health conditions they may have. We also believe that our analysis will assist in lowering the national health expenditures on stroke.

**II. Data**

Kaggle offers an original Stroke Prediction Dataset available for public access at the following web link: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/discussion/229886>.

Each observation in the dataset provides the relevant health information of an anonymous individual, such as gender, age, hypertension, heart disease, marital status, work type, residence type, glucose level, bmi smoking status and occurance of stroke.

The dataset contains 5,110 observations with 12 attributes. Unknown or N/A imply that the information is either unavailable or not applicable.

For this analysis, we will use occurrence of stroke as the dependent variable. Occurrence of stroke is implied by 1 for an individual who has a stroke, 0 for an individual who does not have a stroke.

**III. Data Analysis**

We intend to compile a program that can predict the likelihood of an individual having a stroke based on his or her health conditions. The program consists of two parts: 1) data analysis and 2) user interface.

**3.1 Data analysis**

Our analysis will use historical data to create a regression or classification model that can predict the likelihood of an individual having a stroke.

As mentioned, we have ten variables out of twelve attributes, which are: gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose, bmi and smoke status. Our goal is to analyze whether and how each of them affects stroke incidence using either the logistic regression model or the decision tree model.

Specifically, we will firstly make an analysis of the overall data sample. We would determine all the age patterns and their gender proportion in both the stroke and non-stroke group.

Then we are going to determine how the following attributes would affect the stroke and non-stroke groups and to visualize the results.

* Different Age Groups
* Different BMI Groups
* Gender
* Jobs & Residence Type
* Smoking status

**3.2 User Interface**

For demonstration purposes, the program would have an interface that allows users to input their information and return the probability.

We would give a prediction of whether the patient would get a stroke based on his/her input information about his/her health status or whether this person is in a high-risk or low-risk group.

**Conclusion**

Stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths.In addition strokes cause a greater range of disabilities than any other condition. However, it is also a medical emergency that can be treated by getting the right treatment as soon as possible. Getting warning and acting ahead is the best way to prevent a stroke

The output of our analysis can determine all the factors that affect the stroke incidence rate. Our user interface can help users to be alerted by any early symptoms and prevent them from getting a stroke.

**Reference**

<https://www.cdc.gov/stroke/facts.htm>

<https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death#:~:text=Since%202000%2C%20the%20largest%20increase,6%25%20of%20total%20deaths%20respectively>.

<https://www.stroke.org/en/about-stroke/stroke-symptoms>

<https://www.jns-journal.com/article/S0022-510X(19)32408-6/fulltext>

<https://www.ahajournals.org/doi/pdf/10.1161/STR.0b013e31829734f2>