**TITLE: PREDICTING STROKE IN POPULATION**

**Introduction**

World Health Organization cites stroke-related diseases as are the second highest cause of death third highest cause of disability worldwide [1]

Stroke is a very challenging disease that occurs when a normal blood flow to the brain is interrupted. This is caused either by a blood clot formation in one of the arteries that supplies blood to brain (ischemic stroke) or by a blood leak in the brain by artery (Hemorrhagic stroke). Complications such as paralysis or loss of muscle movement may occur as a result. Difficulties in talking or swallowing are considered stroke complications as well.

Many studies and projects have been conducted around the world by health care professionals, academics and data scientists in order to understand the rout-causes and the medical aspects of the disease. Data Mining methods can be used to detect the occurrence of stroke and benefit healthcare providers to make effective decisions to enhance patient health.

In this project, our main goal is to understand how Machine Learning tools can be applied to reliably predict stroke occurrences in population by targeting key modifiable and non-modifiable stroke factors.

We will start by exploring and highlighting works of various analysts and understanding developed Deep Learning Models used to predict stroke based on medical history and behaviors of patients. We will also try to draw comparisons between those models.

In the next literature review section, we are going to explore just a few most frequently used models by researchers and present the summary of their findings.

**Literature Review**

There have been many articles, projects and conferences conducted on the application of machine-learning algorithms in the study of stroke type classification and stroke predictions. Most notable models focus on Deep-learning neural networks (DNN), Decision Tree, Naïve Bayes, K Nearest Neighbors (KNN), just to name a few.

In one of project carried out by Kunder Akash Mahesh and his colleagues [2], he tries to develop and validate 3 deep learning algorithms of Decision Tree, Naïve Bayes and Artificial Neural Network. His team used a dataset with attributes that are the closest to my projects, notably, the gender of patient, age, hypertension, heart disease, ever married, work type, average glucose level and smoking status. The author underscores clear advantage of the Decision Tree as a model that can handle high dimensional data, a supervised model with a high accuracy, stability and easy for interpretation. Overall findings by his team is that all models could identify patients with stroke at an acceptable accuracy.

The Operation Research Center at MIT Sloan School of Management Department conducted a study aiming at building a stroke predictive model that is not just accurate but also highly interpretable. He presents the model in the form of *sparse decision list*, which consists of if…then…statements. His model called Bayesian List Machine (BLM) provides a new type of balance between accuracy, interpretability and computation. The author compares it to a typical Bayesian Decision Tree. He underscores the BLM main advantage as its construction of the pre-mined rules in the first step that leads to a massive reducing the computation power in the second step where the rules are fully ordered. What I found very compelling is how he deals with the imbalanced dataset taken from the University members where few members had stroke. He applies the class weighting - rule mining algorithm- separately to each class to avoid the rejection of the under-represented class by the minimum support threshold.

In the work conducted by Songhee Choen and co-authors[4], Deep Neural Network(DNN) with scaled Principal Component Analysis(PCA) was used to estimate the stroke occurrence. They were able to predict stroke using indirect and limited data, such as medical service use history and health behaviors. The authors suggest that this approach can reduce future medical costs and facilitate diagnostic. As they derived correlation coefficients between various patient variables and stroke. These did not however clearly reveal the relationship between principal components and stroke.

Medical data is often presented in semi-structured or unstructured form in hospital information systems. Elham Sedghi from the University of Victoria implemented a novel classification method where a technique from Natural Language Processing (NLP) was used [5]. He describes a procedure to process the raw data, transform it into clean, well-structured data that can be effectively used by Data Mining (DM) learning algorithms. Utilizing the sampling method, they were able to balance the dataset and to distinguish migraine from stroke with a high sensitivity and specificity of about 80% and 75% respectively.

**Dataset**

Patient data was obtained from Kaggle open-source forum [6]. The data are used to predict the likelihood of a patient being diagnosed as having stroke based on factors various factors listed below:

*Attributs information:*

* id: unique identifier
* gender: male, female, other
* age: age of the patient
* hypertension: 0 if the patient does not have hypertension, 1 if he/she has one
* heart disease: 0 if the patient doesn’t have heart disease ,1 if he/she has one
* ever married: Yes or No
* work type: children, Government (Gov) job, never worked, Private, Self- employed
* residence type: Rural or Urban
* average(avg) glucose\_level: average glucose level in blood.
* bmi: body mass index
* smoking status: formerly smoked, never smoked, smokes, unknown.
* stroke: 1 if the patient had stroke or 0 if not(target)

Table

Description automatically generated

**Approach**

**Defining a problem**

**Prediction of the outcome**

**Defining data Requirement**

**Training/Validate Models**

**Loading Data**

**Exploratory Data Analytics -Cleaning**

**Reviewing data**

1.World Health Organization. Bulletin of World Health Organization. Editorials, Stroke: a global response is needed. https://www.who.int/bulletin/volumes/94/9/16-181636/en/

2.Prediction od Stroke Using Machine Learning. Dept.of Computer Science & Engineering CMRIT \, Bangalore Karnataka, India. June 4, 2020. <https://www.researchgate.net/publication/342437236_Prediction_of_Stroke_Using_Machine_Learning>

3.An Interpretable Stroke Prediction Model Using Rules and Bayesian Analysis. MIT Sloan School of Management. Authors: Benjamin Letham, Cynthis Rudin, Tyler H. McCormick, David Madigan. <https://dspace.mit.edu/bitstream/handle/1721.1/82148/Rudin_SWP_5040-13.pdf>?sequence=1&isAllowed=y

4. International Journal of Environmental Research and Public Health. Article: The use of Deep Learning to Predict Stroke Patient Mortality. Authors: Songhee Cheon,Jungyoon Kim, Jihye, Li

5.A Novel Stroke Prediction Model based on Clinical Natural Language Processing (NLP) and Data Mining. By Elham Sedghi. Msc. University of Victoria, 2017.

6. Source: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>

By Alexandre Kabanda