AIT 664

G01394251

Rakam sai shivani

Part-1 Project Overview Data Acquisition Report

Brain Strokes Prediction.

My area of focus for the project is brain strokes. Strokes are a leading cause of disability and mortality worldwide, emphasizing the importance of early diagnosis and treatment. Common risk factors include hypertension, smoking, diabetes, and a family history of strokes. Timely medical intervention, such as clot-busting medications or surgical procedures, can significantly improve outcomes for stroke patients. Aging is a significant risk factor for strokes, with the risk doubling each decade after the age of 55.

There are a wide range of data source available through various websites, and I choose do my research on Brain stroke prediction from Kaggle.

Analysis Focus:

The primary focus of this analysis is to predict the occurrence of brain strokes in individuals based on their demographic and health-related attributes. The goal is to develop a robust predictive model that aids in identifying individuals at risk of brain strokes based on their attributes.

Requirements and Objectives:

A. **Outcome Expectations:**

- Develop a predictive model that can accurately classify individuals into stroke or non-stroke categories. Achieve a high level of model accuracy, precision, and recall in identifying stroke cases.
- Provide actionable insights to assist medical practitioners and policymakers in stroke prevention and early intervention.
- Provide insights into on how smoking contributes to stroke risk.

B. **Data Inputs:**

- Extract and preprocess relevant features from the dataset, including age, gender, hypertension status, heart disease status, marital status, work type, residence type, average glucose level, smoking statis and body mass index (BMI).

C. **Key Questions:**

- What are the main factors associated with the likelihood of having a brain stroke?
- How do age and gender influence brain stroke risk?
- Are there any significant relationships between hypertension, heart disease, and brain stroke?
- Can we build a predictive model to identify individuals at higher risk of brain stroke?

D. **Population Variables:**

- The population of interest comprises individuals represented in the dataset.
- Key variables of interest include those that are both statistically significant and clinically relevant to stroke prediction.

E. **Data Types:**

- The dataset includes both numerical (e.g., age, average glucose level, BMI) and categorical (e.g., gender, hypertension status) variables.
- Data preprocessing will involve handling missing values, encoding categorical variables, and normalizing numerical data.

Hypothesis:

Based on limited evidence and prior knowledge, we hypothesize that certain demographic and health-related factors, such as age, hypertension status, and average glucose level, are significant predictors of stroke occurrence. Specifically, we anticipate that:

- Advanced age is positively associated with an increased risk of stroke.
- Individuals with a history of hypertension are more likely to experience strokes.
- Elevated average glucose levels are correlated with a higher likelihood of stroke.
- Individuals with smoking history are also associated with higher chances of stroke.

This initial hypothesis will serve as a guiding framework for our investigation into the dataset. We will explore the data to determine if there is substantial evidence to support or refute these hypotheses. Additionally, we will consider alternative hypotheses to ensure a comprehensive and rigorous analysis. The results of our analysis will help validate or refine these hypotheses and contribute to a better understanding of stroke prediction and risk factors.

Collecting data and information from a variety of sources:

A. The authors of the paper "Early Stroke Prediction Methods for Stroke Prevention" stress the value of noninvasive methods in healthcare, especially for spotting strokes in their early stages. They emphasise that for stroke patients, early detection can literally save their lives. The paper focuses on creating affordable noninvasive approaches for early stroke diagnosis in India, where stroke cases are on the rise, while previous literature frequently relies on expensive MRI and CT scan pictures for diagnosis. The paper suggests time series-based methods to assess processed EEG data and forecast strokes, such as LSTM, biLSTM, GRU, and FFNN. The research findings show encouraging results, with GRU achieving the maximum accuracy of 95.6%, thereby helping doctors identify strokes earlier and maybe saving lives.[1]

B. When anything prevents blood flow to a portion of the brain or when a blood artery in the brain bursts, a stroke, also known as a brain attack, happens.
The brain either ages or suffers harm in both scenarios. A stroke may result in permanent brain damage, chronic disability, or even fatality.

Ischemic stroke:
Ischemic strokes are the most common type. When blood clots or other substances obstruct the brain's blood arteries, an ischemic stroke happens.
The accumulation of fatty deposits known as plaque in the blood vessels can also result in blockages.

Hemorrhagic stroke:
An artery in the brain ruptures or bleeds blood, resulting in a hemorrhagic stroke. The pressure from the blood leak causes brain cells to get damaged. Examples of diseases that can result in a hemorrhagic stroke include high blood pressure and aneurysms, which are balloon-like bulges in an artery that have the potential to stretch and explode.
Expectations following a stroke:
After a stroke, you can make significant progress toward regaining your independence. However, several issues might still exist:

- On one side of the body, a person may have weakness, paralysis, or both.
- Problems with memory, judgment, learning, awareness, and attention.
- difficulties understanding or speaking.
- difficulty with emotion regulation or expression.
- Strange or numb sensations.
- Hand and foot pain that gets worse with activity and variations in temperature.
- difficulty swallowing and chewing.
- issues controlling one's bowels and bladder.
- Depression.[2]

C. Cigarette smoking poses a well-established risk for all types of strokes, with a persistent prevalence despite awareness of its vascular dangers. Around one in five U.S. adults are regular smokers, often beginning in their teens. Notably, there exists a significant dose-response relationship between smoking and stroke risk, a fact that's sometimes underrecognized. This article underscores the scientific evidence linking smoking to stroke and highlights the associated costs, both for individuals and society at large.[3]

D. This dataset is designed for electromagnetic-based stroke classification. It consists of simulated signals generated in the context of electromagnetic imaging. The dataset includes information related to brain injury localization and stroke classification using graph-based approaches. The primary purpose of this dataset is to facilitate the testing and classification of stroke types. Stroke is categorized into two primary types: intracranial haemorrhage (ICH) and ischemic stroke (IS). ICH is caused by the rupture of a blood vessel, while IS results from a clot that restricts blood flow to brain tissues. Differentiating between these stroke types is crucial for timely and appropriate treatment, as the treatment approach can vary significantly between the two types. Accurate classification of stroke types is therefore a critical issue in a microwave system.[5]

From the available information and sets I choose brain stroke prediction dataset from Kaggle to do the further analysis.
Evaluation of Data Sources for Validity and Quality:

- The dataset provides clear definitions for each attribute, including gender, age, hypertension, heart disease, marital status, work type, residence type, glucose level, BMI, smoking status, and stroke occurrence. This characteristic is well-established.
- All attributes in the dataset are quantifiable and measurable. For instance, age is measured in years, glucose level in blood is measurable, and BMI (body mass index) is a measurable indicator of body composition.
- The dataset contains attributes with consistent units of measurement. Age is in years, glucose level is likely in milligrams per deciliter (mg/dL), and BMI is typically measured in kg/m^2. This consistency ensures that data is unitized. The dataset doesn't explicitly state whether numerical attributes have been scaled or standardized. Further details on data normalization techniques would be helpful to assess this characteristic fully.
- The dataset doesn't provide specific details about the source of data, the methods of data collection, or the data collection process's quality control. Therefore, the data quality aspect is not fully addressed. Quality assurance, data cleaning, and validation procedures should be documented to ensure data quality.

Overall, from my focus area the following need to be addressed, Potential data structure issues could include missing data, as mentioned earlier. It's important to handle missing values appropriately, either by imputation or removal. The "smoking_status" variable contains an "Unknown" category, which may need special handling during analysis or preprocessing. Outliers in numerical variables, such as age, average glucose level, and BMI, may need to be addressed.

References:

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9017592/
[2]https://www.cdc.gov/stroke/about.htm#:~:text=A%20stroke%2C%20sometimes%20called%20a,term%20disability%2C%20or%20even%20death.
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928253/
[4] https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset
[5] https://ieee-dataport.org/documents/dataset-brain-injury-localization-and-stroke-classification-electromagnetic-imaging-using