Data Collection Report

Brain Stroke Prediction - Lightweight and Explainable Approach

By- Pranav Pant B21CS088

Introduction and Problem Statement:

My problem statement is to develop a stroke prediction system which would be lightweight, easy to run as well as explainable, which means that we would be able to infer as to which of the given feature is responsible for the stroke prediction.
For the given problem statement, we need a dataset which contains the required features for our problem. For e.g. **BMI** is a very important indicator which heavily can influence if the given person can suffer from stroke or not. Hence, we need a dataset which covers the features and also has many samples to train our Machine Learning Model.

Dataset Selection:

For my problem, I need a dataset which is first of all, **publicly available** for use with required licenses and also has the relevant features required. Searching on internet, I came across these given datasets.

Stroke Prediction Dataset (Kaggle):

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

This dataset was created by McKinsey & Company after testing on real patients for predictive modeling and better healthcare.

This dataset contains **5,110 patient records** with **12 features**, including key medical and demographic indicators such as **age, hypertension, heart disease, BMI, and smoking status**.

FEDESORIANO · UPDATED 4 YEARS AGO

Stroke Prediction Dataset

11 clinical features for predicting stroke events

Dataset is publicly available on Kaggle.

Here are all the features in the dataset in detail:

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

These factors can be said as strong predictors of brain stroke, as confirmed by medical researchers. The dataset is clean (as it contains less missing values) and small in size (around 300KB), making it a good choice for our problem statement.

Drawbacks of given dataset:

The dataset is real and taken from real patients with good no. of features. The only drawback of this dataset is that there is a big **class imbalance** in the dataset.

It means that there are less samples for Stroke class(1) as compared to No Stroke(0). This can significantly impact our model.

---*Report Ends*--