**BDPP PROJECT INITIAL REPORT**

**SHANKARANARAYANAN BR**

For this project,I am using the “Healthcare Dataset Stroke Data” to predict if the person will have the heart stroke or not. This is a classification problem. We will be using the test2v, and train2v dataset which are given. This dataset was download from the github project(https://github.com/surajdidwania/Healthcare-Stroke-Patients-Data/tree/master/healthcare-dataset-stroke-data) website. The dataset consist of the following.

Both the files train and test consist of the following columns

|  |  |
| --- | --- |
| **Columns** | **Description** |
| ID | Patient ID |
| Gender | Gender of the Patient |
| Age | Age of the Patient |
| Hypertension | 1-Incase suffering from hypertension, 0-No hypertension |
| Heart Disease | 1. Incase of suffering from heart disease,0- No heart disease |
| Ever\_married | Yes/No |
| Work\_type | Type of Occupation |
| Residence Type | Area of residence(Rural, Urban) |
| Avg\_glucose\_level | Average\_glucose\_level(measured after meal) |
| BMI | Body mass index |
| Smoking\_status | Patient’s smoke status |
| Stroke | 0-no stroke, 1-suffered Stroke |

There are 43400 entries and 12 attributes as mentioned above in the train data, and the test data has 18601 entries.

This project was interesting to me as it I could use different attributes(Column values) of the data, which provided and make classifications. This dataset also has considers different physical health condition such as hypertension, area of residence etc. Also heart conditions have been the most common causes of death among all age groups according to the WHO