Project Proposal

This repo is Data Since Bootcamp requirement.

Who is of people at risk for a stroke?

According to some research and studies, stroke is one of the leading causes of death worldwide with 11%. So that we can control some of the risk factors for stroke. I will use this project to find out the risks that put patients at risk of stroke based on certain characteristics.

Dataset Description:

To achieve the purpose of the project, the stroke prediction dataset that can be found on Kaggle will be used.

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

This dataset has the following attributes:

[id, gender, age, hypertension, heart_disease, heart_disease

ever_married, work_type, Residence_type, avg_glucose_level, bmi, ,

smoking_status, stroke]

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
2	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
3	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
5	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
6	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
7	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
8	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
9	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
10	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
11	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
13	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
14	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1

Tools:

Some tools will be used such as:

pandas

numpay

sklearn

matplotlib

ToDo :

Data will be explored with EDA phases. Then build a classification model will be used to fit the data.

NOTE : some features might be increased or changed and the model as well.