# Data Science Bootcamp

## Project Proposal

**11 clinical features for predicting stroke events**

Done by: Eman Ahmed Alzhrani

supervised by: Ms. Mariam Elmasry


Introduction

## What is Stroke as a medical condition? Causes? Who are those at risk?

•A stroke is a medical emergency called "brain attack". occurs when part of the brain loses its blood supply and stops working. This causes the part of the body that the injured brain controls to stop working. As stroke is an emergency medical condition thus prompt treatment is crucial. Early action can reduce brain damage and other complications.

•Causes of strokes include ischemia (loss of blood supply) or hemorrhage (bleeding) in the brain.

•People at risk for stroke include those who have high blood pressure, high cholesterol, diabetes, and those who smoke. People with heart rhythm disturbances, especially atrial fibrillation are also at risk.

# ❑ In my Project

•**I am focusing on** understanding the reasons that may cause stroke to people and find out if we can successfully detect stroke depends on specific features using machine learning tools and techniques.

•**By the end of this project, we should be able to predict if a person is more likely to have stroke depends on some features such as:** age, gender, body mass, smoking, blood pressure, heart disease, marriage status, working type, residence type..etc


Questions & Needs

# Questions that should be answered by this project:

1. Are older people more likely to have stroke?
2. Do smokers tend to have stroke in the future more than non-smokers?
3. As a female wanting to know if I am more likely to have stroke than males?
4. will living in a city affect the potentials to have stroke more than living in suburbs?
5. Will your health condition "body mass, hypertension, heart disease, glucose level" put you at risk of having stroke?


Data Description

# Stroke Prediction Dataset

## 11 clinical features for predicting stroke events

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

**Brief:**

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

# Attribute Information:

Target variable "the dependent variable"

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

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

## Data and Attribute info.

The data contains **5110** observations with **12** attributes.


Tools and Modeling

# Tools and Modeling

Since My target variable is a classification type of data, the best models and algorithms for classification are going to be:

- Logistic Regression
- Naïve Bayes
- Stochastic Gradient Descent
- K-Nearest Neighbours
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)

“

❑ Tools and libraries that are more likely to be used are:

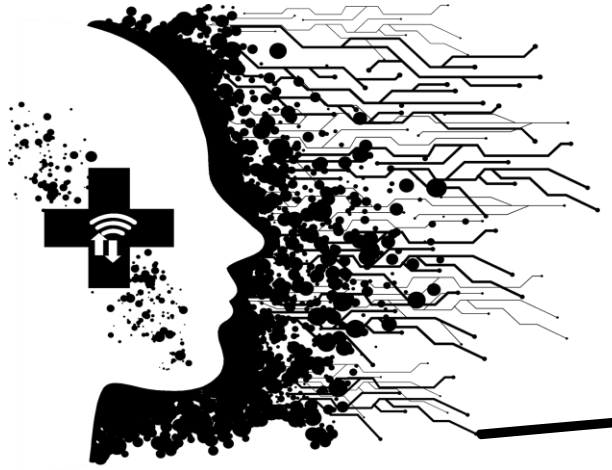
*Python 3* ***environment*** *on Jupyter notebook*

*Using python programming language*

***Libraries:***

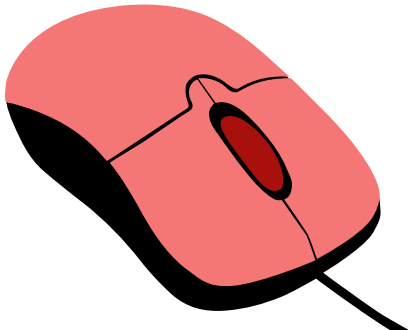
| | |
|---|---|
| numpy | seaborn |
| pandas | Plotly |
| matplotlib | sklearn |

”



# Thank You

I wish you enjoyed my work ☺



Eman Ahmed Alzhrani