CIND 820

Capstone Project Abstract

Name: Ibrahim Sayed Ahmed

Student Number: 501120722

Supervisor: …

Date: September 25, 2022

## Theme:

Classification and Predictive Analytics.

## Objective:

The purpose of this project is to apply Classification and Predictive Analysis on the data:

* Identify the most important factors for stroke prediction.
* Visualize the results
* Propose a predictive analytics approach for stroke prediction.

## Introduction:

## Stroke, also known as brain attack, happens when blood flow to the brain is blocked, preventing it from getting oxygen and nutrients from it and causing the death of brain cells within minutes. According to the World Health Organization (WHO), it is the second cause of death worldwide after ischemic heart disease. Stroke victims can experience paralysis, impaired speed, or loss of vision. While some of the Stroke risk factors cannot be modified, such as family history of cerebrovascular diseases, age, gender and race, others can and are estimated to account for 60% -80% of stroke risk in the general population. Therefore, predicting stroke outcome for new cases can be determining for them to be treated early enough and avoid disabling and mortal consequences. The purpose of this research is to create an accurate model for predicting Stroke outcome with Data Science and Machine Learning (ML) based on previous data and individual characteristics. Providing useful information for the medical staff to deploy the needed treatment and decrease risks and consequences.

## Research Questions and Methods:

**Research Question:**

* What are the attributes related to stroke?
* What is the correlation between pre-existing health conditions and stroke risks?
* How to predict stroke risk?

The plan is to use data collected to identify the most important factors for stroke prediction.

The machine learning methods, techniques to be applied to address these questions are as follows:

* **What are the attributes related to stroke?**

Using Exploratory Data Analysis

* **Implementation of supervised ML classification Algorithms**

Building several models and compare which will be the best, Like Decision Tree Classifier (DT), Random Forest Classifier (RF), K-Nearest Neighbour (KNN)

## Tools

Language to be used is Python and suitable libraries of Python.

## Datasets

The dataset to be used for this project are Stroke Prediction Dataset.

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

The dataset has about 5k records. There are 12 columns: -id - gender - age - hypertension -heart\_disease - ever married - work\_type - Residence\_type - avg\_glucose\_level - bmi -smoking\_status – stroke.

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