Data Analysis Project

On

Stroke Prediction Data set



Introduction

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.

Do you know, 80% Heart strokes are preventable?, yes they are. In this notebook, we attempted to visualize some key indicators that lead to heart strokes. Here data is sampled from a wide range of age groups, gender, habits and health related issues. Most of the visualizations are self expilantory and try to stick to simple visualization .

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  
\*Note: "Unknown" in smoking\_status means that the information is unavailable for this patient

Context

1.load data

2.study the features

3.fill the null and missing values

4.perform various data analysis

5.find out the relationships between various features

6. clean the data set for future work

Questions and Assumptions

1.show records of patients experienced stroke but had missing values in 'bmi' attribute

# 2. fill the missing values of bmi using median

# 3.Categorise BMI into "Underweight", "Normal Weight", "Overweight" and "Obese"

# 4.calculate the total number of stroke experienced by each bmi\_category?

# 5.print the total count of obese,normal weight, overweight and underweight category?

# 6. calculate the stroke occurence in gender category and show it in bar graph?

# 7. find a) total number of person below age 30 with and without stroke?

# b)total number of person above 60 got a stroke and not have a stroke?

# c) plot pie chart on obove

# 8. a)How many marreied persons get stroke and plot in bargraph?

# 9. a) show the count of Rural residence having hypertension and plot in graph

# 10.a)How many persons below age 40 getting a stroke?

# 11. how many patients are below 10 years old?

# 12.how many patients are above 80 years old?

# 13.how many persons having smoking habit have a stroke?

# 14.find out the female patients with high level of glucose and obese?

# 15. find persons having heart-disease,hypertention with private worktype and had stroke?a)Male

# 16. show the relation between stroke and age bracket(0-80)

# 17.show the graphs for occurence of stroke per

# a)Age Bracket & Gender

# b)stroke occurence per BMI Category and gender

# c) stroke occurence per smoking status

# 18. Replace the ever\_married column values with'yes' should

# be 1 and 'no' should be 0

# 19.converting all categorical values into numerical values

# 20.find the correlation matrix of features?

# 

Future Scope

The final cleaned dataset will be used for mechine learning to build a stroke prediction model with maximum accuracy.