**Heart Stroke Prediction System**

**By**

**Chandra Mani Sharma & Kapil Tomar**

**Research Scholars**

**CRDT, IIT Delhi**

**Index**

1. **Problem Statement**
2. **Significance of Problem to Various Stakeholders: Businesses, Government and People**
3. **Overview of AI based Heart Stroke Prediction System**
4. **Dataset & Key Visual Insights on It**
5. **Training Machine Learning Model**
6. **Web Interface**
7. **Deployment**
8. **Discussion & Future Scope**

**Problem Statement**

Good health is the desire of everyone. If we get to know about the critical disease on time then definitely we can cure the diseases. Unfortunately we know the diseases when there is no sufficient time to cure it. Technology can help us to know about the diseases on time so that necessary steps can be taken to cure it.

Here, the objective is to create a machine learning based web application to predict and screen the possibility of heart stroke for a person by taking into account various heath inputs. The data set used for this problem has been taken from Kaggle (  
<https://www.kaggle.com/fedesoriano/stroke-prediction-dataset> )

Based on the various inputs provided by individuals we are trying to predict about heart stroke. If there is any possibility of heart stroke then we can take necessary steps. Thus we can save the life of many people in the society.

**Significance of Problem to Various Stakeholders: Businesses, Government and People**

**1. As a business how can we turn this into Profit making or brand making service/product**

Heart Stroke is a very common problem nowadays. As a business, we can use technology to develop a service/product to predict stroke possibilities accurately. We can advertise about the product and make people aware about it. People in urban and rural areas can use our product to know about the possibilities of heart stroke based on the inputs.

We have to tie ups with various path labs, doctors, pharmaceutical companies and yoga centres. If there is any chance of heart stroke then we can suggest various solutions such as check-ups, doctor consultations, medicines and exercises. When these people will visit various suggested path labs, doctors, yoga centres with our brand name then definitely our brand value will increase and we will also earn the profit through coloration with such health care institutions.

**2. As an NGO how the impact will help around the people**

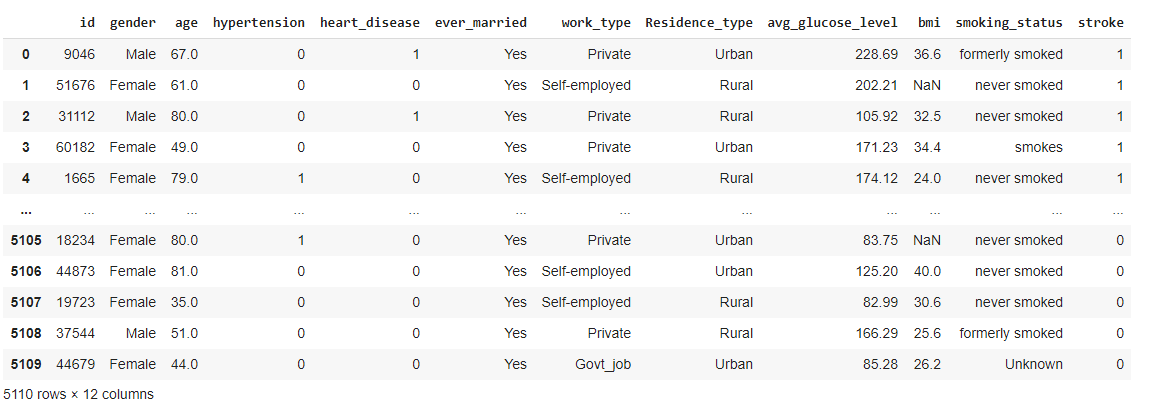
NGO always play a great role to make an impact on the society without concentrating on its own profit. NGO can organize various health camps for the community and provide proper training to reduce the possibilities of heart stroke. There are various NGOs which are running various yoga classes and provide tips to improve their health. They can guide the society about the various services/products which are available around us for the wellbeing of the community. NGO can help the society to know about the government initiatives also to improve the health of citizens. They can also share benefits of our product to the society so that maximum people can use and get the benefit.

**3. As a government representative how it will effect the image of the government in social cues**  
Government always focuses on social issues. We are trying to make a positive impact on individual health in the society. So that everyone can have better health. Such initiatives are possible with good policies and support of the government. As a government representative we are working with government goals. Everyone will get to know that the government is thinking and worried about the health of everyone in the county. It will also make a positive image of government in the society.

**Key insights from the data**

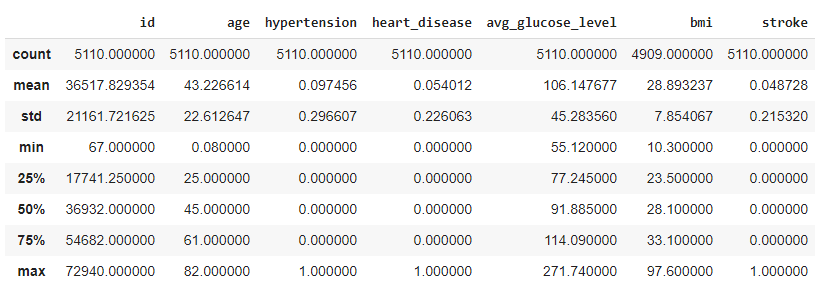
* There are 5110 Rows and 12 columns in the data set.
* In data set seven are numerical values and 5 are non-numerical,
* Total missing values in the data set is 201.
* In bmi column 201 out of 5110 values are missing.
* Mode technique can be applied in categorical missing values and mean technique can be applied to the numerical values.
* In dataset min age=0.8, max age=82 and mean age=43.22
* Out of 5110 – Female =2994, Male-2115 and other=1
* In data set 249 have heart stroke and 4861 does not have heart stroke
* Urban Population = 2595 and Rural Population = 2514
* In Rural 114 people have strokes and in Urban 135 people have strokes.

1. **There are 5110 Rows and 12 columns in the data set.**

****

* It is clear from above graph that we have 5110 Row and 12 columns in our data set.
* Starting index and ending index is also mentioned in the graph.

1. **Statistical analysis on numerical values.**



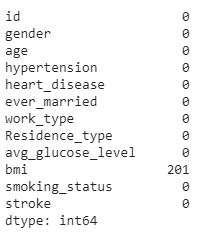
Statistical analysis provides 8 basics information about the data set on numerical values.

* **Count** – Showing the total values available in each column.
* **Mean** - Showing the minimum values of each column.
* **Std** - Showing the standard deviation o each column.
* **25%** - 25% values are less than the mentioned values in each column.
* **50%** - 50% values are less than the mentioned values in each column.
* **75%** - 75% values are less than the mentioned values in each column.
* **Max** - Showing the maximum values of each column.

1. **Counting total missing values in the data set – 201**

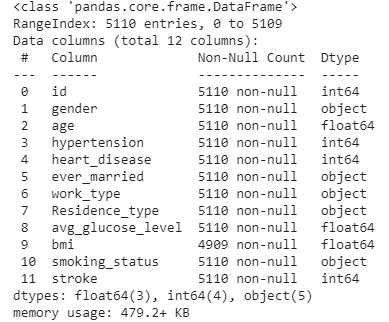
* Total 201 values are missing in the data set.

1. **Counting missing values in each column in data set:**



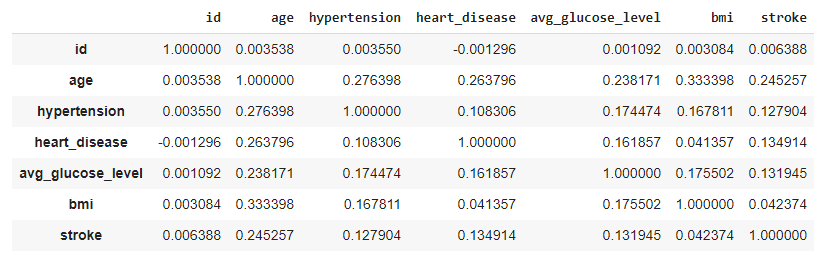
* This graph is showing missing values column wise.
* In bmi column 201 out of 5110, values are missing.

1. **Concise summary about the data:**



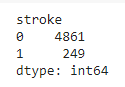
* Showing each column name in the data set with data types and available values in each column.
* Starting and ending index of each entry.
* Total number of columns available in the data set.

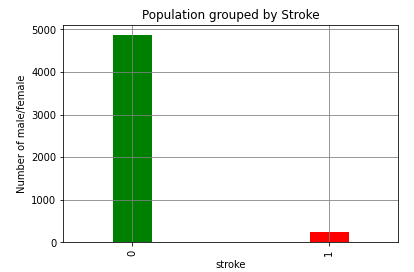
1. **Finding correlation in data set**

****

* This graph is showing the correlation between each entry.
* According to this data set Stroke is highly correlated with age and very less correlated to bmi.

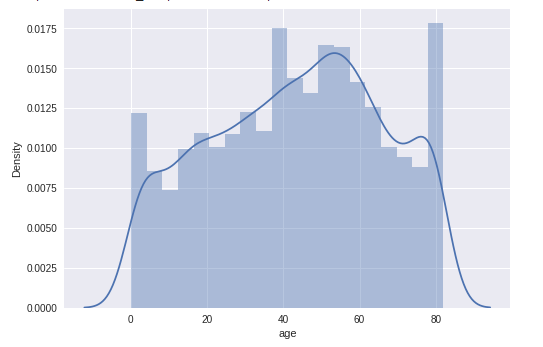
1. **Total population based on the stroke -**





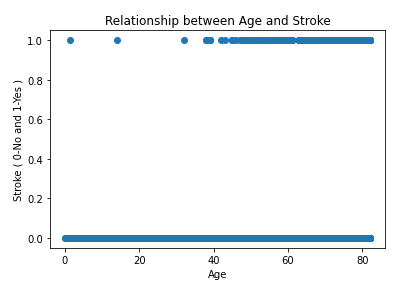
* Total number of people in the data set is 5110.
* Out of 5110 only 249 gave the stroke and 4861 did not have the stroke.
* Only 4.8% people have the stroke. It means data is unbalanced.

1. **Variation in Age parameter in data set**

****

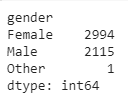
* This graph is showing the age density in the data set.

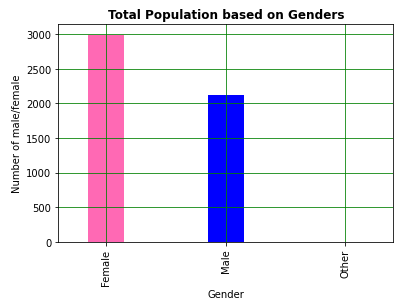
1. **Relationship between age and stroke**

****

* This scatter graph is showing the relationship between age and the stroke.
* It shows that people who have the stroke are mostly 40+ in age and very few less than 40 in age.

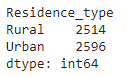
1. **Total Population based on Genders.**

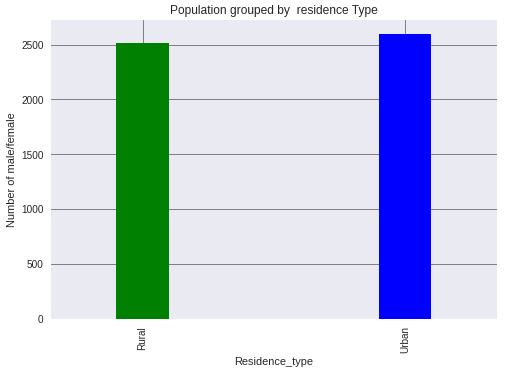




* This graph is showing the details gender wise. That how many Male, Female and other are available in the data set.
* Female=2994, Male=2115 & other=1

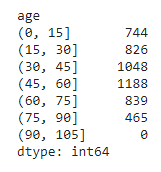
1. **Total population according to Residence Type –**

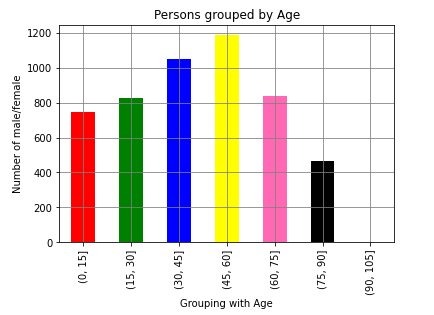
****

****

* This graph is showing the population living in Rural and Urban areas.
* 2514 People are living in Rural.
* 2595 People are living in Urban.

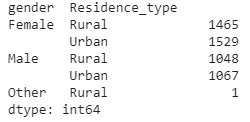
1. **Total number of male/female grouped by age**

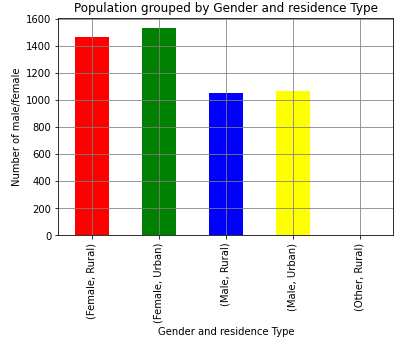




* In this graph total population is divided into a particular age group.
* In data set major population is belonging to 45-60 age group which is equal to 1188.

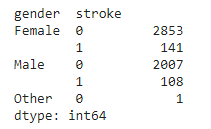
1. **Population Grouped with Gender and Residence Type**

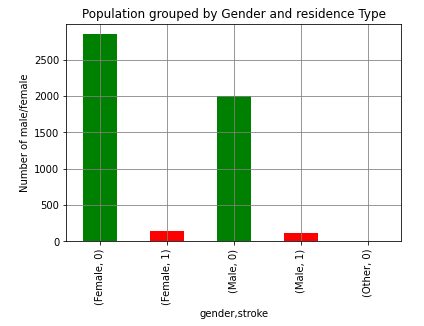
****



* This graph is showing the relationship between gender and residence type.
* Females, who are living in the Urban, are maximum in the data set.
* Males, who are living in the Rural, are minimum in the data set.

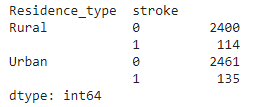
1. **Population Grouped with Gender, and Stroke**

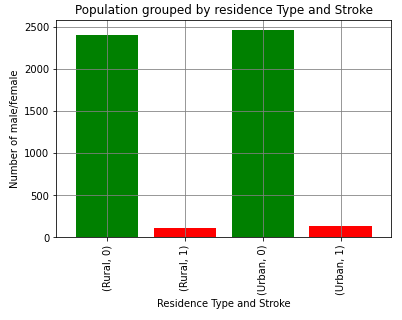
****

****

* This graph is showing the relation between Gender and Stroke.
* 148 Females and 108 males are having the stroke.

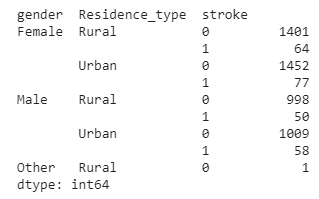
1. **Population Grouped with Residence Type and Stroke**

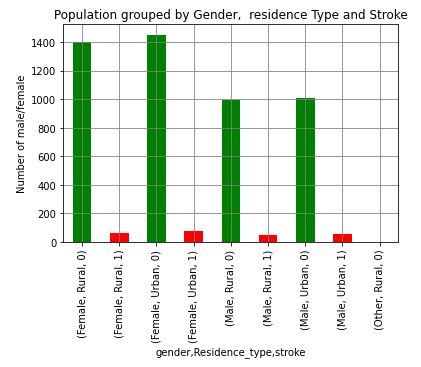
****

****

* This graph is showing the relation between Residence Type and Stroke.
* Maximum number of people who got the stroke are living in Urban areas which is 135 n numbers.

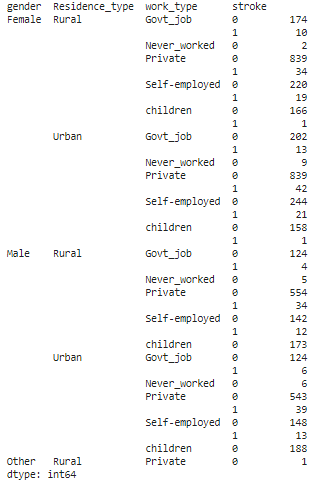
1. **Population Grouped with Gender, Residence Type and Stroke**

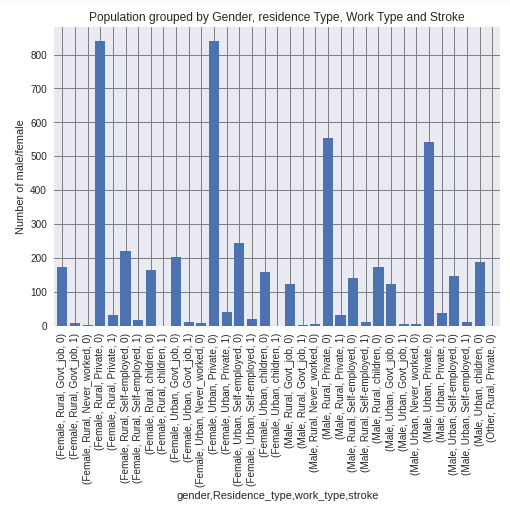




* This graph is showing the relation between Gender, Residence Type and Stroke.
* Maximum - 77 Females living in Urban got the heart stroke.
* Minimum - 58 Males living in Urban got the heart stroke.

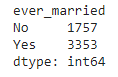
1. **Population Grouped with Gender, Residence Type, Work Type and Stroke**

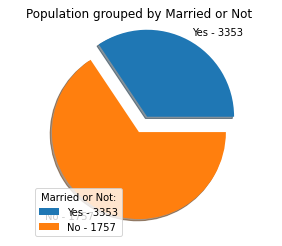




* This graph is showing the relation between Gender, Residence Type, Work Types and Stroke.
* **Females** who are living in **Urban** and working in **Private** organization are the maximum (**42**) in numbers who got the **stroke**.
* **Females** who are living in **Rural** and working in **Government** organization are the minimum (**4**) in numbers who get the stroke except the children worktype.
* Work type is mentioned children. It shown wrong entries in data set.

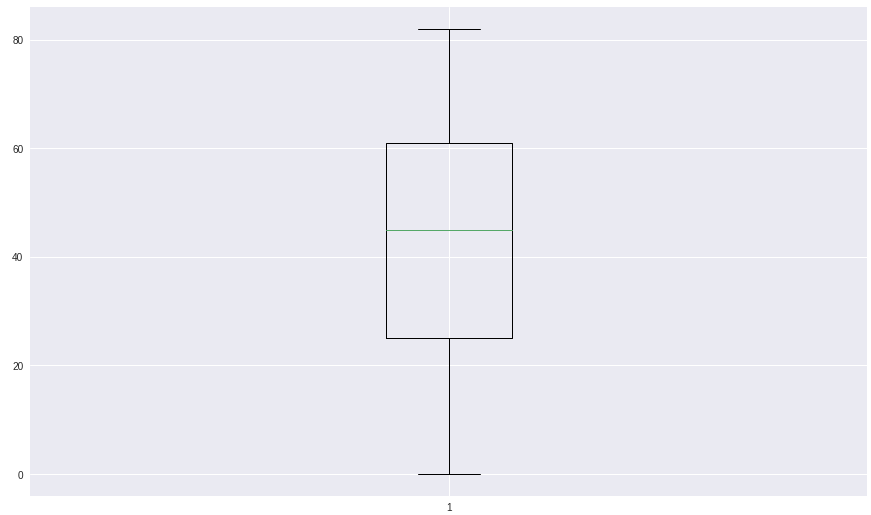
1. **Population based on Married or not -**

****

****

* This pie chart shows that 3353 are married and 1757 are not married.

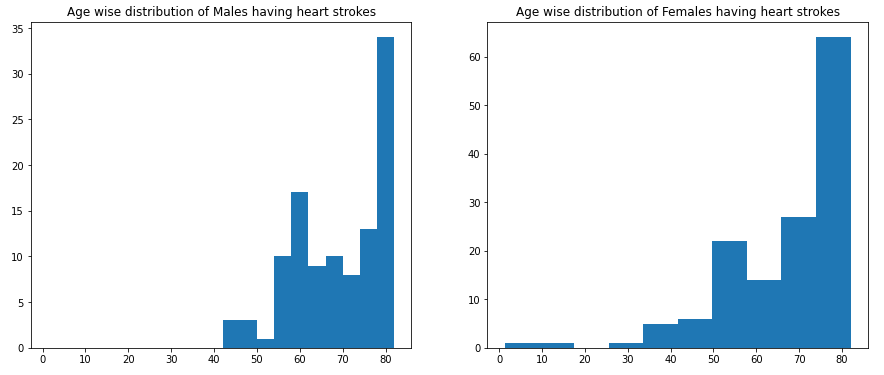
1. **Boxplot on the age of the population –**

****

This Box Plot tells the following information.

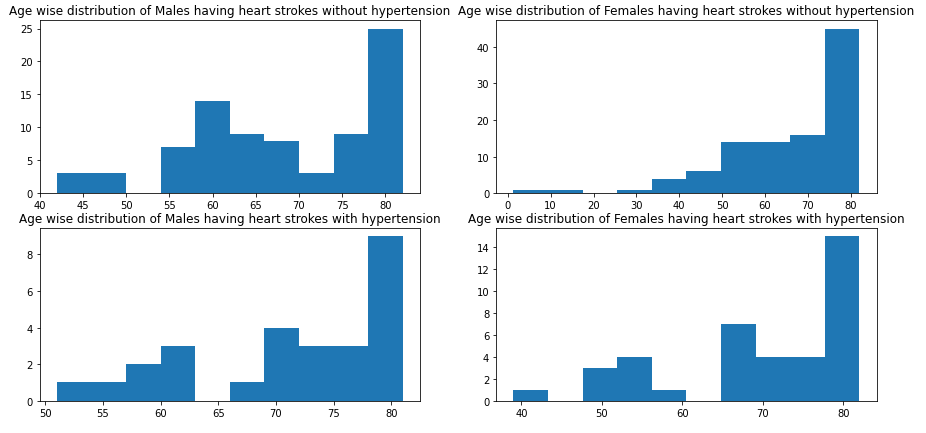
* Mean age of the population = 43.22
* Minimum age of the population = 0.08
* Maximum age of the population = 82.0
* Maximus population is between age of 25 and 53

1. **Age wise distribution of Male/Female having Stroke.**

****

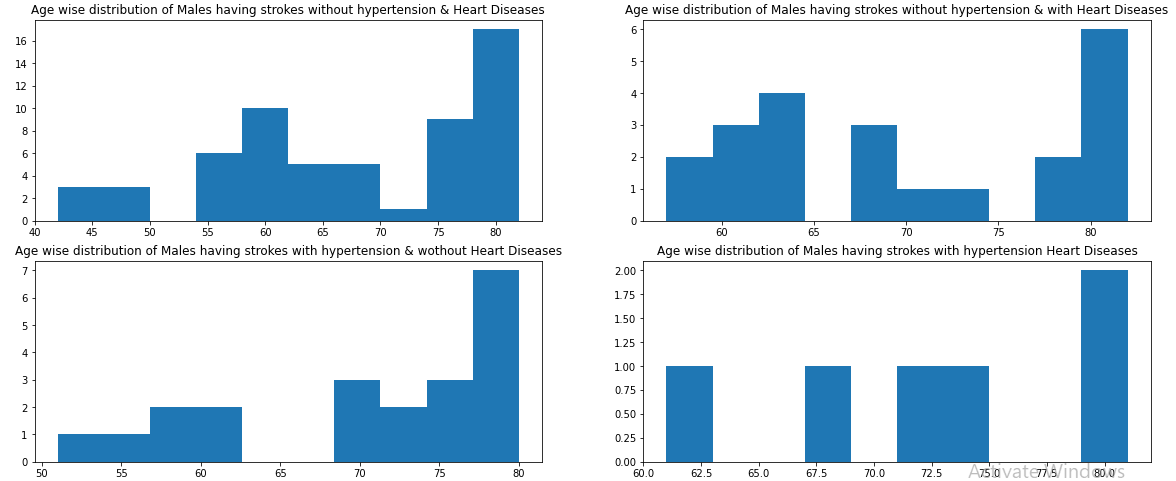
|  |  |
| --- | --- |
| * Age wise distribution of Males who are having the heart stroke. * Males less than 42 years of age did not get the heart stroke. * Maximum number of Males who got the heart stroke belongs to age group **55-60** and **75-80.** | * Age wise distribution of Females who are having the heart stroke. * Females of all age group got the heart stroke. * Maximum number of Females who got the heart stroke belongs to age group **75-80.** |

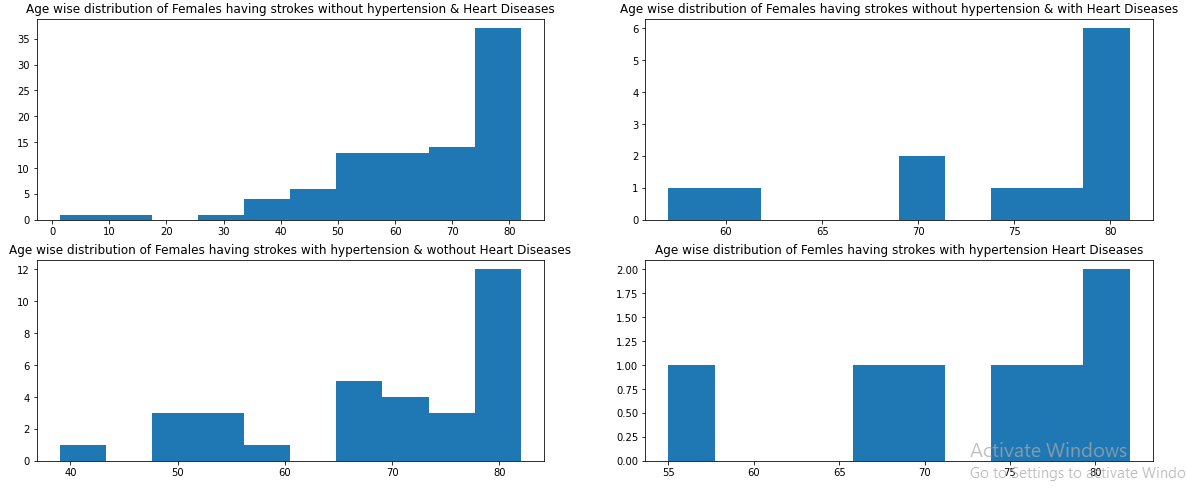
1. **Age wise distribution of Male/Female having Stroke with and without hypertension.**

****

* Four graphs are showing Age wise distribution of Male/Female having Stroke with and without hypertension.
* It’s clear from Graph [0,1] and Graph [1,1] that in same age group (77-82) more females have heart strokes either they have hypertension or not compare to men.
* Graph [0,1] reflects that more number of women have heart stroke even they do not have hypertension.
* Same is true for men if we compare Graph [0,0] and Graph [1,0].
* It means in women and men hypertension does not play a major role for having heart stroke.

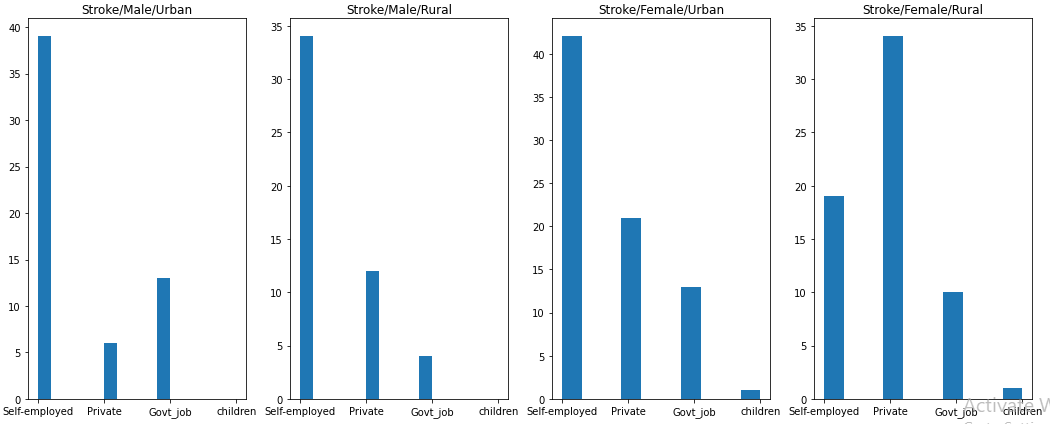
1. **Age wise distribution of heart stroke in Male/Female with and without hypertension and heart diseases.**





* These are four combinations of Age wise distribution of heart stroke in Men and women with and without hypertension & heart diseases**.**
* In men it is clearly visible that when they do not have hypertension and heart disease then heart stroke is high. Same is observed with women

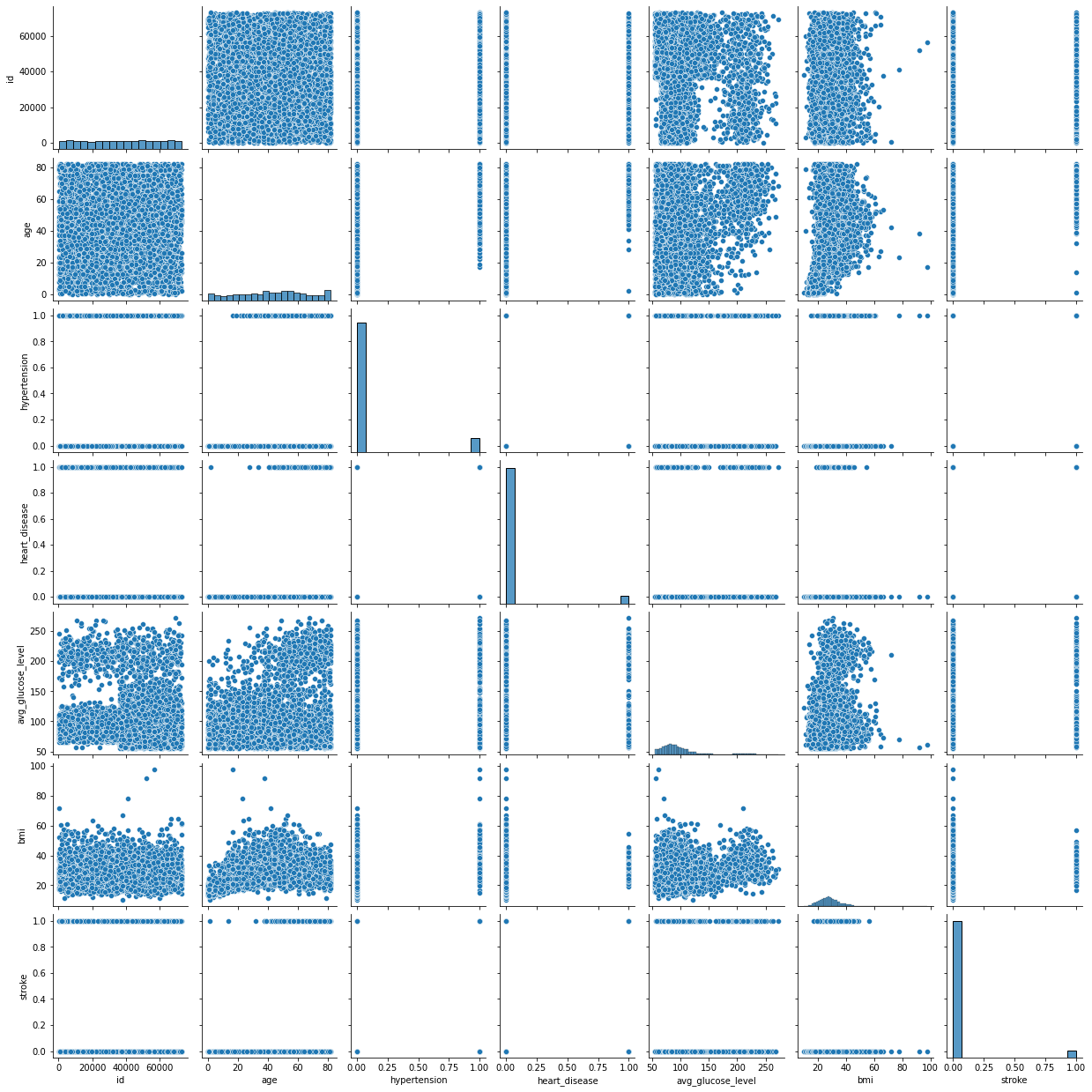
1. **Work type distribution of heart stroke in Male/Female respect to the residence type.**

****

* It is observed from the above four combinations that Female living in Urban areas who are self-employed got heart stroke.
* Women living in Rural working in Private job having the highest heart stroke.

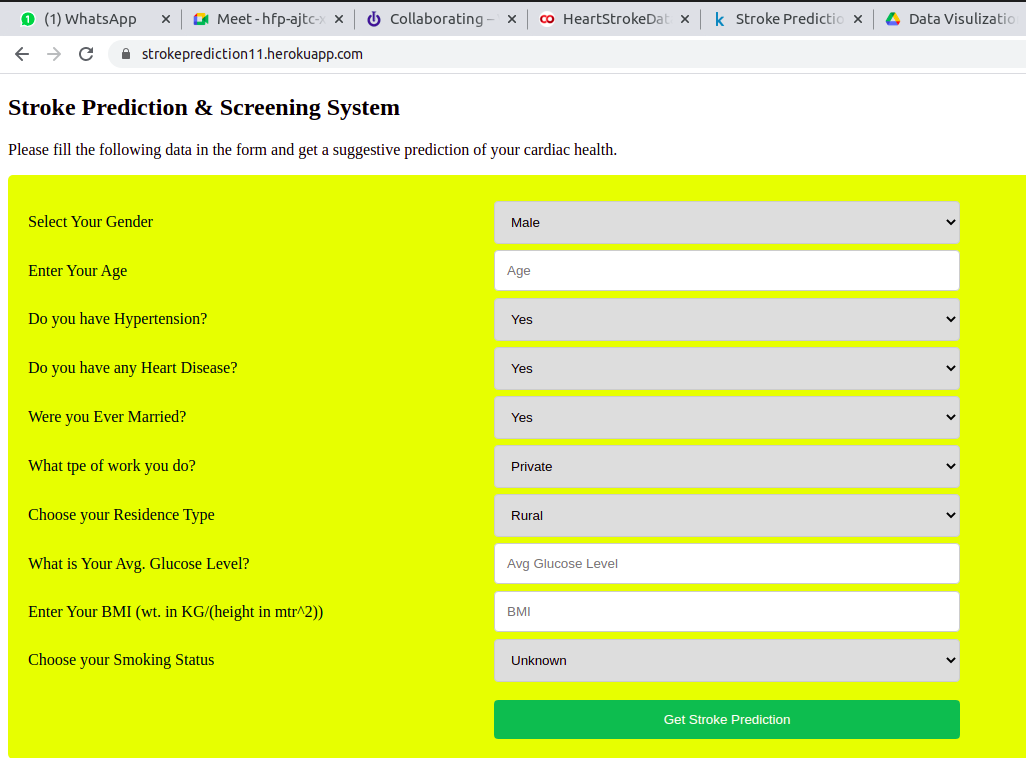
1. **Pair Plot**

Pair plot is showing the relationship between each parameter with each other.

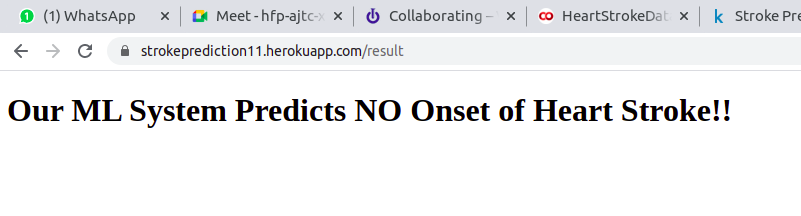
****

* Pair plot is showing the relationship between each parameter with each other.

**Web Interface**

****

*1.1 Web Interface of the prediction System*

**

*1.2 Result page of the prediction system*

**Discussion & Future Scope**

We are predicting heart stroke based on the input provided by individuals. It will help individuals to know that they have a chance of a heart stroke or not.1. We can add a few more inputs to get a more accurate result. There are some missing values in the data set. We can try to get the values in an automatic manner so that there is less change of missing values.