In [2]:

```
import pandas as pd
import numpy as np
```

In [3]:

```
df = pd.read_csv("C:/Users/ameya/OneDrive/Desktop/DSBDAL/healthcare-dataset-stroke-data.
```

In [4]:

```
df.head()
```

Out[4]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_ty
0	9046	Male	67.0	0	1	Yes	Private	Urb
1	51676	Female	61.0	0	0	Yes	Self-employed	Ru
2	31112	Male	80.0	0	1	Yes	Private	Ru
3	60182	Female	49.0	0	0	Yes	Private	Urb
4	1665	Female	79.0	1	0	Yes	Self-employed	Ru

In [5]:

```
df.tail()
```

Out[5]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence
5105	18234	Female	80.0	1	0	Yes	Private	
5106	44873	Female	81.0	0	0	Yes	Self-employed	
5107	19723	Female	35.0	0	0	Yes	Self-employed	
5108	37544	Male	51.0	0	0	Yes	Private	
5109	44679	Female	44.0	0	0	Yes	Govt_job	

In [6]:

```
df.describe()
```

Out[6]:

	id	age	hypertension	heart_disease	avg_glucose_level	b
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.0000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.8932
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.8540
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.3000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.5000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.1000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.1000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.6000

In [7]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64  
 4   heart_disease      5110 non-null   int64  
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

In [8]:

```
df.dtypes
```

Out[8]:

```
id                   int64
gender              object
age                float64
hypertension         int64
heart_disease        int64
ever_married        object
work_type           object
Residence_type      object
avg_glucose_level  float64
bmi                float64
smoking_status      object
stroke               int64
dtype: object
```

In [9]:

```
df.isnull().sum()
```

Out[9]:

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

In [10]:

```
df['bmi'].fillna(df['bmi'].mean,inplace=True)
```

In [11]:

```
df.isnull().sum()
```

Out[11]:

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
stroke               0
dtype: int64
```

In [12]:

```
df.shape
```

Out[12]:

```
(5110, 12)
```

In [13]:

```
df['avg_glucose_level'] = df['avg_glucose_level'].astype(int)
```

In [14]:

```
df.dtypes
```

Out[14]:

```
id                   int64
gender              object
age                float64
hypertension         int64
heart_disease        int64
ever_married        object
work_type           object
Residence_type      object
avg_glucose_level    int32
bmi                 object
smoking_status      object
stroke               int64
dtype: object
```

In [20]:

```
df.dtypes
```

Out[20]:

```
id                    int64
gender               object
age                 float64
hypertension          int64
heart_disease         int64
ever_married         object
work_type            object
Residence_type       object
avg_glucose_level     int32
bmi                  object
smoking_status       object
stroke                int64
dtype: object
```

In [21]:

```
col = ["hypertension","heart_disease"];
```

Z-score normalization

In [23]:

```
from sklearn.preprocessing import StandardScaler
scalar = StandardScaler()
df_scaled = scalar.fit_transform(df[col].to_numpy())
df_scaled = pd.DataFrame(df_scaled,columns=col)
```

In [24]:

```
df_scaled
```

Out[24]:

	hypertension	heart_disease
0	-0.328602	4.185032
1	-0.328602	-0.238947
2	-0.328602	4.185032
3	-0.328602	-0.238947
4	3.043196	-0.238947
...	...	...
5105	3.043196	-0.238947
5106	-0.328602	-0.238947
5107	-0.328602	-0.238947
5108	-0.328602	-0.238947
5109	-0.328602	-0.238947

5110 rows × 2 columns

In [ ]:

```

```