In [55]:

```
import pandas as pd
data= pd.read_csv("/Users/zahiramohammed/Desktop/DAPM_original.csv")
```

In [56]:

```
data.head() #to check if the dataset is loaded
```

Out[56]:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

In [57]:

```
data.info() #print the information of the dataframe
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   gender               100000 non-null  object 
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64  
 3   heart_disease        100000 non-null  int64  
 4   smoking_history      100000 non-null  object 
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64  
 8   diabetes             100000 non-null  int64  
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
```

In [58]:

```
data.isna().sum() #checking missing values
```

Out[58]:

```
gender                 0
age                    0
hypertension           0
heart_disease          0
smoking_history        0
bmi                    0
HbA1c_level            0
blood_glucose_level    0
diabetes               0
dtype: int64
```

In [59]:

```
data.drop_duplicates(inplace=True) #returns dataframe with duplicates removed
```

In [62]:

```
#shape of the dataframe
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 96146 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   gender               96146 non-null  object 
 1   age                  96146 non-null  float64
 2   hypertension         96146 non-null  int64  
 3   heart_disease        96146 non-null  int64  
 4   smoking_history      96146 non-null  object 
 5   bmi                  96146 non-null  float64
 6   HbA1c_level          96146 non-null  float64
 7   blood_glucose_level  96146 non-null  int64  
 8   diabetes             96146 non-null  int64  
dtypes: float64(3), int64(4), object(2)
memory usage: 7.3+ MB
```

In [63]:

```
#mapping the categorical variables into numericals
data['smoking_history']=data['smoking_history'].map({'never':0,'current':1,
                                                    'former':-2,'ever':2,'not current':-1})
data['gender']=data['gender'].map({'Female':1,'Male':0})
```

In [66]:

```
data
```

Out[66]:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	1.0	80.0	0	1	0.0	25.19	6.6	140	0
1	1.0	54.0	0	0	NaN	27.32	6.6	80	0
2	0.0	28.0	0	0	0.0	27.32	5.7	158	0
3	1.0	36.0	0	0	1.0	23.45	5.0	155	0
4	0.0	76.0	1	1	1.0	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...	...
99994	1.0	36.0	0	0	NaN	24.60	4.8	145	0
99996	1.0	2.0	0	0	NaN	17.37	6.5	100	0
99997	0.0	66.0	0	0	-2.0	27.83	5.7	155	0
99998	1.0	24.0	0	0	0.0	35.42	4.0	100	0
99999	1.0	57.0	0	0	1.0	22.43	6.6	90	0

96146 rows × 9 columns

In [67]:

```
data.dropna(inplace=True)#to drop null values
```

In [68]:

```
data
```

Out[68]:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	1.0	80.0	0	1	0.0	25.19	6.6	140	0
2	0.0	28.0	0	0	0.0	27.32	5.7	158	0
3	1.0	36.0	0	0	1.0	23.45	5.0	155	0
4	0.0	76.0	1	1	1.0	20.14	4.8	155	0
5	1.0	20.0	0	0	0.0	27.32	6.6	85	0
...	...	...	...	...	...	...	...	...	...
99992	1.0	26.0	0	0	0.0	34.34	6.5	160	0
99993	1.0	40.0	0	0	0.0	40.69	3.5	155	0
99997	0.0	66.0	0	0	-2.0	27.83	5.7	155	0
99998	1.0	24.0	0	0	0.0	35.42	4.0	100	0
99999	1.0	57.0	0	0	1.0	22.43	6.6	90	0

63247 rows × 9 columns

In [74]:

```
#since the dataset is huge, new dataset(data_new) is created with 500 random samples to perform further analysis
data_new=data.sample(500)
```

In [79]:

```
#saving the new dataset
data_new.to_csv("/Users/zahiramohammed/Desktop/DAPM_dataset_new.csv")
```

In [80]:

```
data_new.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 500 entries, 43383 to 23438
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   gender               500 non-null    float64
 1   age                  500 non-null    float64
 2   hypertension         500 non-null    int64  
 3   heart_disease        500 non-null    int64  
 4   smoking_history      500 non-null    float64
 5   bmi                  500 non-null    float64
 6   HbA1c_level          500 non-null    float64
 7   blood_glucose_level  500 non-null    int64  
 8   diabetes             500 non-null    int64  
dtypes: float64(5), int64(4)
memory usage: 39.1 KB
```

In [78]:

```
data_new["diabetes"].value_counts()
```

Out[78]:

```
diabetes
0    450
1     50
Name: count, dtype: int64
```

In [ ]: