In [1]:
```
import pandas as pd
import numpy as np
data=pd.read_csv("heart_2020_cleaned.csv")
print(data)
```

```
       HeartDisease    BMI Smoking AlcoholDrinking Stroke  PhysicalHealth  \
0                No  16.60     Yes              No     No             3.0
1                No  20.34      No              No    Yes             0.0
2                No  26.58     Yes              No     No            20.0
3                No  24.21      No              No     No             0.0
4                No  23.71      No              No     No            28.0
...             ...    ...     ...             ...    ...             ...
319790          Yes  27.41     Yes              No     No             7.0
319791           No  29.84     Yes              No     No             0.0
319792           No  24.24      No              No     No             0.0
319793           No  32.81      No              No     No             0.0
319794           No  46.56      No              No     No             0.0

        MentalHealth DiffWalking     Sex  AgeCategory      Race Diabetic  \
0               30.0          No  Female        55-59     White      Yes
1                0.0          No  Female  80 or older     White       No
2               30.0          No    Male        65-69     White      Yes
3                0.0          No  Female        75-79     White       No
4                0.0         Yes  Female        40-44     White       No
...              ...         ...     ...          ...       ...      ...
319790           0.0         Yes    Male        60-64  Hispanic      Yes
319791           0.0          No    Male        35-39  Hispanic       No
319792           0.0          No  Female        45-49  Hispanic       No
319793           0.0          No  Female        25-29  Hispanic       No
319794           0.0          No  Female  80 or older  Hispanic       No

       PhysicalActivity  GenHealth  SleepTime Asthma KidneyDisease SkinCancer
0                   Yes  Very good        5.0    Yes            No        Yes
1                   Yes  Very good        7.0     No            No         No
2                   Yes       Fair        8.0    Yes            No         No
3                    No       Good        6.0     No            No        Yes
4                   Yes  Very good        8.0     No            No         No
...                 ...        ...        ...    ...           ...        ...
319790               No       Fair        6.0    Yes            No         No
319791              Yes  Very good        5.0    Yes            No         No
319792              Yes       Good        6.0     No            No         No
319793               No       Good       12.0     No            No         No
319794              Yes       Good        8.0     No            No         No

[319795 rows x 18 columns]
```

In [2]:
```
data.head()
```

Out[2]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	Physi
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes	
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No	
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes	
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No	
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No	

In [3]:
```
data.describe()
```

Out[3]:

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	319795.000000	319795.00000	319795.000000	319795.000000
mean	28.325399	3.37171	3.898366	7.097075
std	6.356100	7.95085	7.955235	1.436007
min	12.020000	0.00000	0.000000	1.000000
25%	24.030000	0.00000	0.000000	6.000000
50%	27.340000	0.00000	0.000000	7.000000
75%	31.420000	2.00000	3.000000	8.000000
max	94.850000	30.00000	30.000000	24.000000

In [4]:
```
data.tail()
```

Out[4]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic
319790	Yes	27.41	Yes	No	No	7.0	0.0	Yes	Male	60-64	Hispanic	Yes
319791	No	29.84	Yes	No	No	0.0	0.0	No	Male	35-39	Hispanic	No
319792	No	24.24	No	No	No	0.0	0.0	No	Female	45-49	Hispanic	No
319793	No	32.81	No	No	No	0.0	0.0	No	Female	25-29	Hispanic	No
319794	No	46.56	No	No	No	0.0	0.0	No	Female	80 or older	Hispanic	No

In [5]:
```
data.isnull().sum()
```

Out[5]:
```
HeartDisease        0
BMI                 0
Smoking             0
AlcoholDrinking     0
Stroke              0
PhysicalHealth      0
MentalHealth        0
DiffWalking         0
Sex                 0
AgeCategory         0
Race                0
Diabetic            0
PhysicalActivity    0
GenHealth           0
SleepTime           0
Asthma              0
KidneyDisease       0
SkinCancer          0
dtype: int64
```

In [6]:
```
d1=data.drop(["PhysicalHealth","Race","Diabetic","PhysicalActivity","Asthma","KidneyDisease","SkinCancer","Sex","Diffw
d1
```

Out[6]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	MentalHealth	SleepTime
0	No	16.60	Yes	No	No	30.0	5.0
1	No	20.34	No	No	Yes	0.0	7.0
2	No	26.58	Yes	No	No	30.0	8.0
3	No	24.21	No	No	No	0.0	6.0
4	No	23.71	No	No	No	0.0	8.0
...	...	...	...	...	...	...	...
319790	Yes	27.41	Yes	No	No	0.0	6.0
319791	No	29.84	Yes	No	No	0.0	5.0
319792	No	24.24	No	No	No	0.0	6.0
319793	No	32.81	No	No	No	0.0	12.0
319794	No	46.56	No	No	No	0.0	8.0

319795 rows × 7 columns

In [7]:
```
d1["HeartDisease"]=d1["HeartDisease"].map({"Yes":1,"No":0})
```

In [8]:
```
d1
```

Out[8]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	MentalHealth	SleepTime
0	0	16.60	Yes	No	No	30.0	5.0
1	0	20.34	No	No	Yes	0.0	7.0
2	0	26.58	Yes	No	No	30.0	8.0
3	0	24.21	No	No	No	0.0	6.0
4	0	23.71	No	No	No	0.0	8.0
...	...	...	...	...	...	...	...
319790	1	27.41	Yes	No	No	0.0	6.0
319791	0	29.84	Yes	No	No	0.0	5.0
319792	0	24.24	No	No	No	0.0	6.0
319793	0	32.81	No	No	No	0.0	12.0
319794	0	46.56	No	No	No	0.0	8.0

319795 rows × 7 columns

In [9]:
```
d1.groupby(["HeartDisease"]).count()
```

Out[9]:

	BMI	Smoking	AlcoholDrinking	Stroke	MentalHealth	SleepTime
HeartDisease						
0	292422	292422	292422	292422	292422	292422
1	27373	27373	27373	27373	27373	27373

In [10]:
```
d2=pd.get_dummies(d1,dtype=int)
```

In [11]:
```
d2
```

Out[11]:

	HeartDisease	BMI	MentalHealth	SleepTime	Smoking_No	Smoking_Yes	AlcoholDrinking_No	AlcoholDrinking_Yes	Stroke_No	Stroke_Y
0	0	16.60	30.0	5.0	0	1	1	0	1	
1	0	20.34	0.0	7.0	1	0	1	0	0	
2	0	26.58	30.0	8.0	0	1	1	0	1	
3	0	24.21	0.0	6.0	1	0	1	0	1	
4	0	23.71	0.0	8.0	1	0	1	0	1	
...	...	...	...	...	...	...	...	...	...	
319790	1	27.41	0.0	6.0	0	1	1	0	1	
319791	0	29.84	0.0	5.0	0	1	1	0	1	
319792	0	24.24	0.0	6.0	1	0	1	0	1	
319793	0	32.81	0.0	12.0	1	0	1	0	1	
319794	0	46.56	0.0	8.0	1	0	1	0	1	

319795 rows × 10 columns

In [12]:
```
y=d2['HeartDisease']
x=d2.drop("HeartDisease",axis=1)
```

In [13]:
```
y
```

Out[13]:
```
0         0
1         0
2         0
3         0
4         0
         ..
319790    1
319791    0
319792    0
319793    0
319794    0
Name: HeartDisease, Length: 319795, dtype: int64
```

In [14]:
```
x
```

Out[14]:

	BMI	MentalHealth	SleepTime	Smoking_No	Smoking_Yes	AlcoholDrinking_No	AlcoholDrinking_Yes	Stroke_No	Stroke_Yes
0	16.60	30.0	5.0	0	1	1	0	1	0
1	20.34	0.0	7.0	1	0	1	0	0	1
2	26.58	30.0	8.0	0	1	1	0	1	0
3	24.21	0.0	6.0	1	0	1	0	1	0
4	23.71	0.0	8.0	1	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...
319790	27.41	0.0	6.0	0	1	1	0	1	0
319791	29.84	0.0	5.0	0	1	1	0	1	0
319792	24.24	0.0	6.0	1	0	1	0	1	0
319793	32.81	0.0	12.0	1	0	1	0	1	0
319794	46.56	0.0	8.0	1	0	1	0	1	0

319795 rows × 9 columns

In [16]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.30,random_state=32)
```

In [17]:
```
x_test
```

Out[17]:

	BMI	MentalHealth	SleepTime	Smoking_No	Smoking_Yes	AlcoholDrinking_No	AlcoholDrinking_Yes	Stroke_No	Stroke_Yes
147608	31.38	0.0	7.0	1	0	1	0	1	0
25159	29.95	0.0	6.0	0	1	1	0	1	0
302982	20.34	0.0	5.0	1	0	1	0	1	0
1625	21.95	0.0	4.0	0	1	1	0	1	0
88586	32.98	2.0	6.0	1	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...
122822	59.45	0.0	7.0	1	0	1	0	1	0
187720	28.89	0.0	6.0	0	1	1	0	1	0
14244	22.71	0.0	8.0	0	1	0	1	1	0
305527	30.27	0.0	6.0	1	0	1	0	1	0
314537	20.22	0.0	8.0	0	1	1	0	1	0

95939 rows × 9 columns

In [18]:
```
x_train
```

Out[18]:

	BMI	MentalHealth	SleepTime	Smoking_No	Smoking_Yes	AlcoholDrinking_No	AlcoholDrinking_Yes	Stroke_No	Stroke_Yes
269005	25.75	4.0	7.0	1	0	1	0	1	0
134271	24.39	0.0	8.0	0	1	1	0	1	0
261945	26.39	0.0	8.0	1	0	1	0	1	0
36370	29.53	0.0	8.0	1	0	1	0	1	0
137035	23.06	2.0	8.0	0	1	1	0	1	0
...	...	...	...	...	...	...	...	...	...
216135	21.14	2.0	10.0	0	1	1	0	1	0
282558	30.43	0.0	7.0	1	0	1	0	1	0
75062	36.34	10.0	8.0	1	0	1	0	1	0
130949	23.03	0.0	6.0	0	1	1	0	1	0
10967	27.99	0.0	7.0	1	0	1	0	1	0

223856 rows × 9 columns

In [19]:
```
y_test
```

Out[19]:
```
147608    0
25159     1
302982    0
1625      0
88586     0
         ..
122822    0
187720    1
14244     0
305527    0
314537    0
Name: HeartDisease, Length: 95939, dtype: int64
```

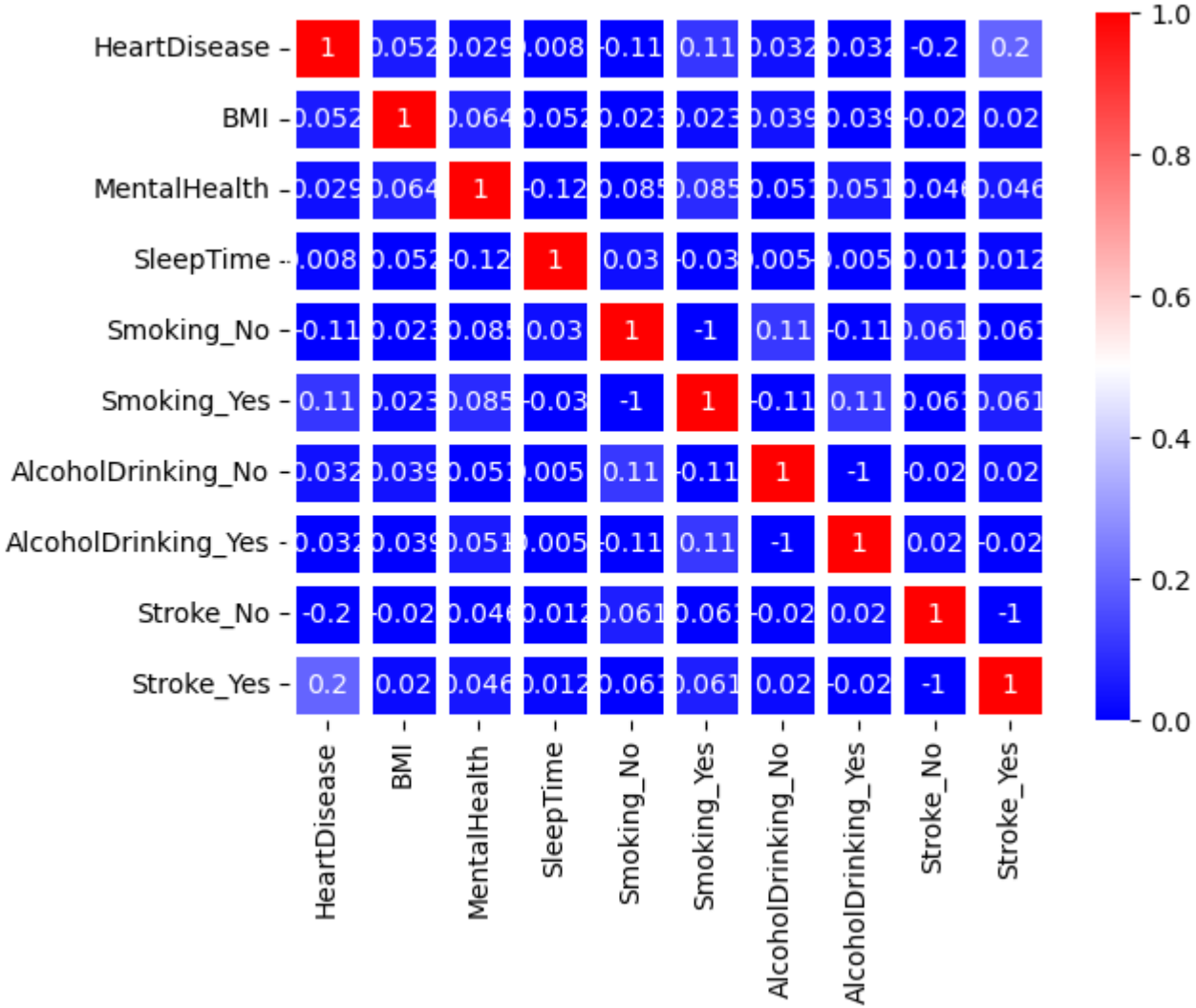
In [20]:
```
core=d2.corr()
core
```

Out[20]:

	HeartDisease	BMI	MentalHealth	SleepTime	Smoking_No	Smoking_Yes	AlcoholDrinking_No	AlcoholDrinking_Yes	Stro
HeartDisease	1.000000	0.051803	0.028591	0.008327	-0.107764	0.107764	0.032080	-0.032080	-0.
BMI	0.051803	1.000000	0.064131	-0.051822	-0.023118	0.023118	0.038816	-0.038816	-0.
MentalHealth	0.028591	0.064131	1.000000	-0.119717	-0.085157	0.085157	-0.051282	0.051282	-0.
SleepTime	0.008327	-0.051822	-0.119717	1.000000	0.030336	-0.030336	0.005065	-0.005065	-0.
Smoking_No	-0.107764	-0.023118	-0.085157	0.030336	1.000000	-1.000000	0.111768	-0.111768	0.
Smoking_Yes	0.107764	0.023118	0.085157	-0.030336	-1.000000	1.000000	-0.111768	0.111768	-0.
AlcoholDrinking_No	0.032080	0.038816	-0.051282	0.005065	0.111768	-0.111768	1.000000	-1.000000	-0.
AlcoholDrinking_Yes	-0.032080	-0.038816	0.051282	-0.005065	-0.111768	0.111768	-1.000000	1.000000	0.
Stroke_No	-0.196835	-0.019733	-0.046467	-0.011900	0.061226	-0.061226	-0.019858	0.019858	1.
Stroke_Yes	0.196835	0.019733	0.046467	0.011900	-0.061226	0.061226	0.019858	-0.019858	-1.

In [22]:
```
import seaborn as sns
sns.heatmap(core,vmax=1,vmin=0,annot=True,linewidth=5,cmap='bwr')
```

Out[22]: <Axes: >



In [23]:
```
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(n_estimators=100, max_depth=10, random_state=42)
model.fit(x_train, y_train)
```

Out[23]:
```
RandomForestClassifier(max_depth=10, random_state=42)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [25]:
```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
predictions = model.predict(x_test)
accuracy = accuracy_score(y_test,predictions)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.9135492344093643
```

In [26]:
```
predictions
```

Out[26]:
```
array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

In [27]:
```
results=pd.DataFrame(columns=["original","predicted"])
```

In [28]:
```
results["original"]=y_test
```

In [31]:
```
results['predicted']=predictions
```

```
results
```

In [32]:
```
results
```

Out[32]:

	original	predicted
147608	0	0
25159	1	0
302982	0	0
1625	0	0
88586	0	0
...	...	...
122822	0	0
187720	1	0
14244	0	0
305527	0	0
314537	0	0

95939 rows × 2 columns

In [33]:
```
results.head(10)
```

Out[33]:

	original	predicted
147608	0	0
25159	1	0
302982	0	0
1625	0	0
88586	0	0
163850	0	0
68742	0	0
236971	0	0
216191	0	0
152376	0	0

In [34]:
```
results=results.reset_index()
results
```

Out[34]:

	index	original	predicted
0	147608	0	0
1	25159	1	0
2	302982	0	0
3	1625	0	0
4	88586	0	0
...	...	...	...
95934	122822	0	0
95935	187720	1	0
95936	14244	0	0
95937	305527	0	0
95938	314537	0	0

95939 rows × 3 columns

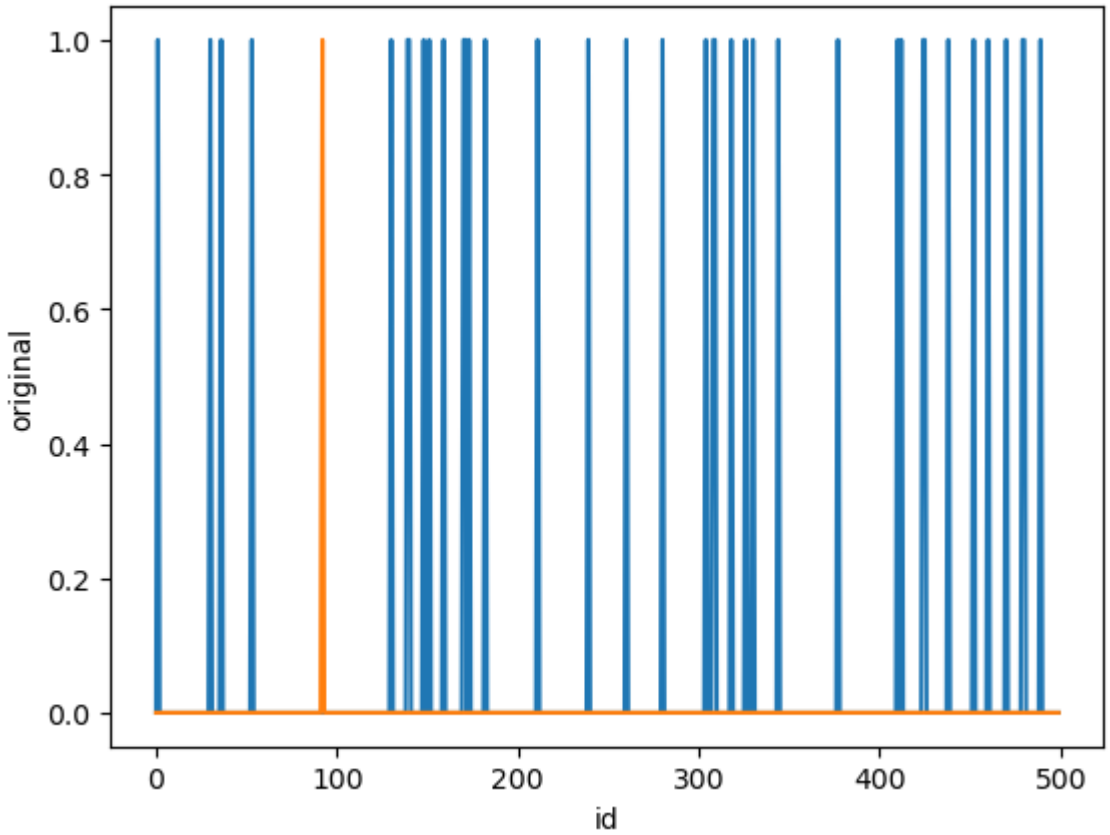
In [35]:
```
results['id']=results.index
results.head(10)
```

Out[35]:

	index	original	predicted	id
0	147608	0	0	0
1	25159	1	0	1
2	302982	0	0	2
3	1625	0	0	3
4	88586	0	0	4
5	163850	0	0	5
6	68742	0	0	6
7	236971	0	0	7
8	216191	0	0	8
9	152376	0	0	9

In [37]:
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.lineplot(x="id",y="original",data=results.head(500))
sns.lineplot(x="id",y="predicted",data=results.head(500))
plt.plot()
```

Out[37]: []



In [ ]: