HAP_DataProcessing

February 26, 2022

[23]:
```
import numpy as np
import pandas as pd
```

[24]:
```
#Reading Datasets
va_df = pd.read_table("C:\BU\portfolio\HeartAttackPrediction\dataset\processed.
 ↪va.data", header = None, sep = ",")
va_df = va_df.apply(pd.to_numeric, errors='coerce')
va_df.columns = ["age" ,"sex" ,"cp" ,"trestbps", "chol" ,"fbs" ,"restecg"␣
 ↪,"thalach", "exang", "oldpeak", "slope", "ca", "thal" ,"num"]

cl_df = pd.read_table("C:\BU\portfolio\HeartAttackPrediction\dataset\processed.
 ↪cleveland.data", header = None, sep = ",")
cl_df = cl_df.apply(pd.to_numeric, errors='coerce')
cl_df.columns = ["age" ,"sex" ,"cp" ,"trestbps", "chol" ,"fbs" ,"restecg"␣
 ↪,"thalach", "exang", "oldpeak", "slope", "ca", "thal" ,"num"]

sz_df = pd.read_table("C:\BU\portfolio\HeartAttackPrediction\dataset\processed.
 ↪switzerland.data", header = None, sep = ",")
sz_df = sz_df.apply(pd.to_numeric, errors='coerce')
sz_df.columns = ["age" ,"sex" ,"cp" ,"trestbps", "chol" ,"fbs" ,"restecg"␣
 ↪,"thalach", "exang", "oldpeak", "slope", "ca", "thal" ,"num"]

hg_df = pd.read_table("C:\BU\portfolio\HeartAttackPrediction\dataset\processed.
 ↪hungarian.data", header = None, sep = ",")
hg_df = cl_df.apply(pd.to_numeric, errors='coerce')
hg_df.columns = ["age" ,"sex" ,"cp" ,"trestbps", "chol" ,"fbs" ,"restecg"␣
 ↪,"thalach", "exang", "oldpeak", "slope", "ca", "thal" ,"num"]

#Combining Datasets
datasets = [va_df, cl_df, sz_df, hg_df]
df0 = pd.concat(datasets)
df0 = df0.sample(frac=1).reset_index(drop=True)# Randomize Data rows
df0.sample(10)
df0['num'] = df0['num'].mask(df0['num'] > 0, 1)
hap_df = df0.copy(deep = "True")
hap_df.sample(10)
```

[24]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
231	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	1.0	0.0	3.0	0
830	57.0	1.0	4.0	110.0	335.0	0.0	0.0	143.0	1.0	3.0	2.0	1.0	7.0	1
574	58.0	1.0	4.0	128.0	216.0	0.0	2.0	131.0	1.0	2.2	2.0	3.0	7.0	1
884	61.0	0.0	4.0	145.0	307.0	0.0	2.0	146.0	1.0	1.0	2.0	0.0	7.0	1
492	67.0	1.0	4.0	125.0	254.0	1.0	0.0	163.0	0.0	0.2	2.0	2.0	7.0	1
473	57.0	1.0	4.0	140.0	0.0	0.0	0.0	120.0	1.0	2.0	2.0	NaN	6.0	1
159	74.0	1.0	3.0	NaN	0.0	0.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	0
566	70.0	1.0	3.0	160.0	269.0	0.0	0.0	112.0	1.0	2.9	2.0	1.0	7.0	1
12	55.0	1.0	2.0	130.0	262.0	0.0	0.0	155.0	0.0	0.0	1.0	0.0	3.0	0
530	61.0	1.0	3.0	200.0	0.0	NaN	1.0	70.0	0.0	0.0	NaN	NaN	3.0	1

[25]:
```
hap_df.head()
```

[25]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	44.0	1.0	4.0	120.0	169.0	0.0	0.0	144.0	1.0	2.8	3.0	0.0	6.0	1
1	53.0	1.0	4.0	123.0	282.0	0.0	0.0	95.0	1.0	2.0	2.0	2.0	7.0	1
2	52.0	1.0	1.0	118.0	186.0	0.0	2.0	190.0	0.0	0.0	2.0	0.0	6.0	0
3	68.0	1.0	3.0	118.0	277.0	0.0	0.0	151.0	0.0	1.0	1.0	1.0	7.0	0
4	57.0	1.0	4.0	140.0	192.0	0.0	0.0	148.0	0.0	0.4	2.0	0.0	6.0	0

[26]:
```
hap_df.isnull().sum()
```

[26]:
```
age          0
sex          0
cp           0
trestbps    58
chol         7
```

```
fbs          82
restecg       1
thalach      54
exang        54
oldpeak      62
slope       119
ca          324
thal        222
num           0
dtype: int64
```

[27]:
```
hap_df = hap_df.dropna()
```

[28]:
```
hap_df.isnull().sum()
```

[28]:
```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
num         0
dtype: int64
```

[29]:
```
display (hap_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 595 entries, 0 to 927
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       595 non-null    float64
 1   sex       595 non-null    float64
 2   cp        595 non-null    float64
 3   trestbps  595 non-null    float64
 4   chol      595 non-null    float64
 5   fbs       595 non-null    float64
 6   restecg   595 non-null    float64
 7   thalach   595 non-null    float64
 8   exang     595 non-null    float64
 9   oldpeak   595 non-null    float64
```

```
 10  slope     595 non-null    float64
 11  ca        595 non-null    float64
 12  thal      595 non-null    float64
 13  num       595 non-null    int64
dtypes: float64(13), int64(1)
memory usage: 69.7 KB
```

None

[31]:
```
hap_df.to_csv("C:\BU\portfolio\HeartAttackPrediction\dataset/heart_attack.csv",␣
↪index=False)
```