main

April 2, 2023

1 Experiment 8 - Modelling Medical Data with a Bayesian Network

[ ]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
```

[ ]:
```
data = pd.read_csv(r'./data.csv')
data
```

[ ]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
..	…	…	..	…	…	…	…	…	…	…	…	..	…	…
298	45	1	1	110	264	0	0	132	0	1.2	2	0	7	1
299	68	1	4	144	193	1	0	141	0	3.4	2	2	7	2
300	57	1	4	130	131	0	0	115	1	1.2	2	1	7	3
301	57	0	2	130	236	0	2	174	0	0.0	2	1	3	1
302	38	1	3	138	175	0	0	173	0	0.0	1	?	3	0

[303 rows x 14 columns]

[ ]:
```
data.columns
```

[ ]:
```
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'num'],
      dtype='object')
```

[ ]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    object 
 12  thal      303 non-null    object 
 13  num       303 non-null    int64  
dtypes: float64(1), int64(11), object(2)
memory usage: 33.3+ KB
```

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

[ ]:

	age	sex	cp	trestbps	chol	fbs	\
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	54.438944	0.679868	3.158416	131.689769	246.693069	0.148515	
std	9.038662	0.467299	0.960126	17.599748	51.776918	0.356198	
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	
50%	56.000000	1.000000	3.000000	130.000000	241.000000	0.000000	
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	

	restecg	thalach	exang	oldpeak	slope	num
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	0.990099	149.607261	0.326733	1.039604	1.600660	0.937294
std	0.994971	22.875003	0.469794	1.161075	0.616226	1.228536
min	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000

```
25%      0.000000  133.500000    0.000000    0.000000    1.000000    0.000000
50%      1.000000  153.000000    0.000000    0.800000    2.000000    0.000000
75%      2.000000  166.000000    1.000000    1.600000    2.000000    2.000000
max      2.000000  202.000000    1.000000    6.200000    3.000000    4.000000
```

[ ]:
```
columns = data.columns
```

[ ]:
```
X = data.iloc[:, :-1]
X
```

[ ]:
```
      age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  \
0      63    1   1       145   233    1        2      150      0      2.3
1      67    1   4       160   286    0        2      108      1      1.5
2      67    1   4       120   229    0        2      129      1      2.6
3      37    1   3       130   250    0        0      187      0      3.5
4      41    0   2       130   204    0        2      172      0      1.4
..    ...  ...  ..       ...   ...  ...      ...      ...    ...      ...
298    45    1   1       110   264    0        0      132      0      1.2
299    68    1   4       144   193    1        0      141      0      3.4
300    57    1   4       130   131    0        0      115      1      1.2
301    57    0   2       130   236    0        2      174      0      0.0
302    38    1   3       138   175    0        0      173      0      0.0

     slope ca thal
0        3  0    6
1        2  3    3
2        2  2    7
3        3  0    3
4        1  0    3
..     ... ..  ...
298      2  0    7
299      2  2    7
300      2  1    7
301      2  1    3
302      1  ?    3

[303 rows x 13 columns]
```

[ ]:
```
y = data['num']
y
```

[ ]:
```
0      0
1      2
2      1
3      0
4      0
      ..
```

```
298    1
299    2
300    3
301    1
302    0
Name: num, Length: 303, dtype: int64
```

[ ]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, shuffle=True,␣
↪random_state=42)
```

[ ]:
```
X_train.shape, X_test.shape
```

[ ]:
```
((227, 13), (76, 13))
```

[ ]:
```
y_train.shape, y_test.shape
```

[ ]:
```
((227,), (76,))
```

[ ]:
```

```