WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ
POLITECHNIKI RZESZOWSKIEJ

Wnioskowanie w warunkach niepewności

Patryk Motyka
Inżynieria i analiza danych

Rzeszów, 2022

Spis treści

1. Dane

Dane wykorzystywane w projekcie pochodzą z pliku Statlog (Heart), który zawiera 270 wierszy z danymi, na które składa się 14 kolumn.

1.Wiek
2. Płeć
3. Rodzaj bólu klatki (4 możliwości)
4. Spoczynkowe ciśnienie krwi
5. Cholesterol w surowicy w mg/dl
6. Cukier we krwi na czczo >120mg/dl
7. Spoczynkowe wyniki elektrokardiografii (wartości: 0,1,2)
8. Maksymalny zanotowany puls
9. Dławica piersiowa wywołana wysiłkiem fizycznym
10. Oldpeak czyli obniżenie odcinka ST wywołane wysiłkiem fizycznym w stosunku do odpoczynku
11. Nachylenie odcinka ST dla wysiłki fizycznego
12. Liczba głównych naczyń (0-3) zabarwionych podczas fluoroskopii
13. Talasemia: 3 = normalny; 6 = naprawiona wada; 7 = wada odwracalna
14. Obecność choroby serca-(1) brak choroby serca-(2)

*Talasemia czyli niedokrwistość tarczowatokrwinkowa ilościowe zaburzenia syntezy hemoglobiny, spowodowane wrodzonym defektem biosyntezy łańcuchów globiny.
**Odcinek ST – w terminologii medycznej określenie fragmentu zapisu elektrokardiograficznego odpowiadającego początkowej fazie repolaryzacji mięśnia komór serca.

	age	sex	Chest	Press	Cholest	Sugar	Electro	MaxRate	Angina	oldpeak	ST	Vessels	thal	class
1	70	1	4	130	322	0	2	109	0	1	2	3	3	2
2	67	0	3	115	564	0	2	160	0	1	2	0	7	1
3	57	1	2	124	261	0	0	141	0	1	1	0	7	2
4	64	1	4	128	263	0	0	105	1	1	2	1	7	1
5	74	0	2	120	269	0	2	121	1	1	1	1	3	1
6	65	1	4	120	177	0	0	140	0	1	1	0	7	1

1-1Przykładowe rekordy z pliku z danymi

Źródło danych: https://archive.ics.uci.edu/ml/datasets/statlog+(heart)

2. Teoria

Sieć bayesowska służy do przedstawienia zależności pomiędzy zdarzeniami bazując na rachunku prawdopodonieństwa. Sieć jest modelowana za pomocą skierowanego grafu, w którym wierzchołki odpowiadają zdarzeniom, natomiast łuki związkom pomiędzy tymi zdarzeniami.

Przykładowe algorytmy wykorzystywane do uczenia struktury sieci bayesowskiej:

Hc korzysta z algorytmu optymalizacji „hill climbing” należącego rodziny przeszukiwania lokalnego. Jest to algorytm iteracyjny, który zaczyna się od arbitralnego rozwiązania problemu, a następnie próbuje znaleźć lepsze rozwiązanie, dokonując stopniowej zmiany rozwiązania. Jeśli zmiana przyniesie lepsze rozwiązanie, do nowego rozwiązania wprowadzana jest kolejna przyrostowa zmiana i tak dalej, aż nie będzie można znaleźć dalszych ulepszeń.

Gs korzysta z algorytmu grow-shrink.

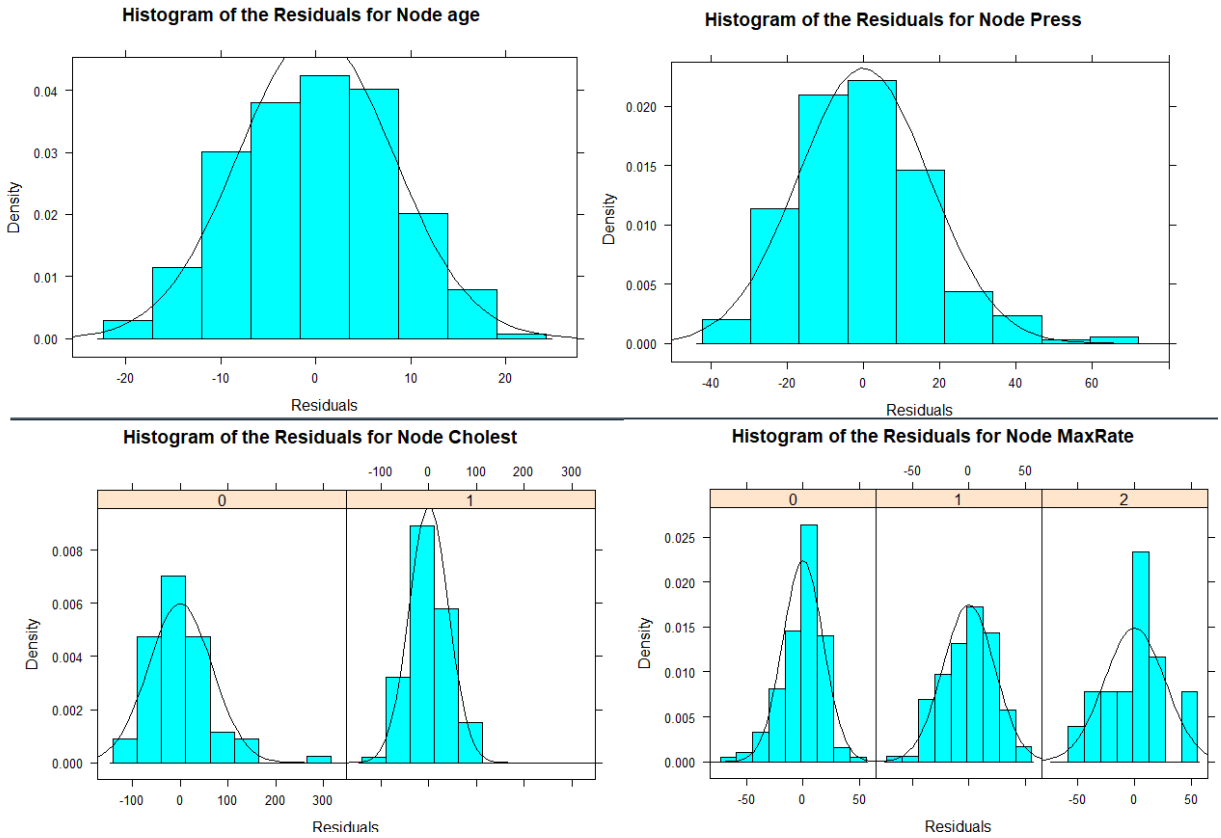
Iamb - Incremental Association Markov Blanket czyli powiązania/skojarzenia przyrostowe.

hpc hybrydowe algorytmy oparte na ograniczeniach rodziców i dzieci.

PC to prototypowy algorytm oparty na ograniczeniach do uczenia sieci bayesowskich.

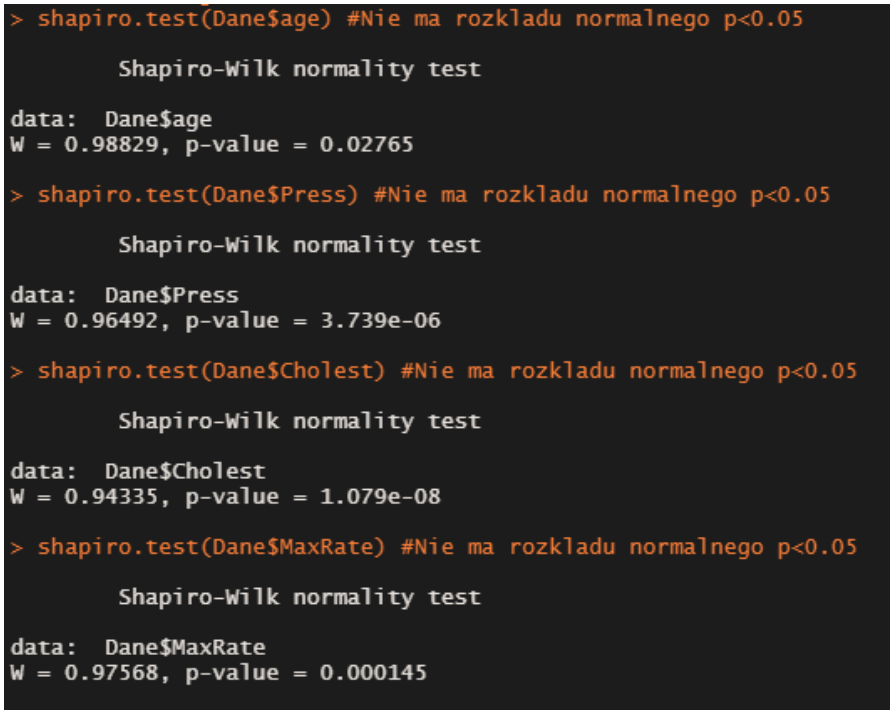
3. Działania

Po wczytaniu dane dyskretne zamieniam na czynniki, a typ danych ciągłych na numeric.
Następnie tworzę pierwszą sieć algorytmem hc.
Wykorzystuje funkcję bn.fit, która dopasowuje parametry sieci bayesowskiej w zależności od jej struktury.
Rysuję histogramy reszt dla danych ciągłych.



3-1Histogramy dla reszt danych ciągłych

W niektórych przypadkach histogramy mogą nieco przypominać rozkład normalny, więc wykonuję testy statystyczne.



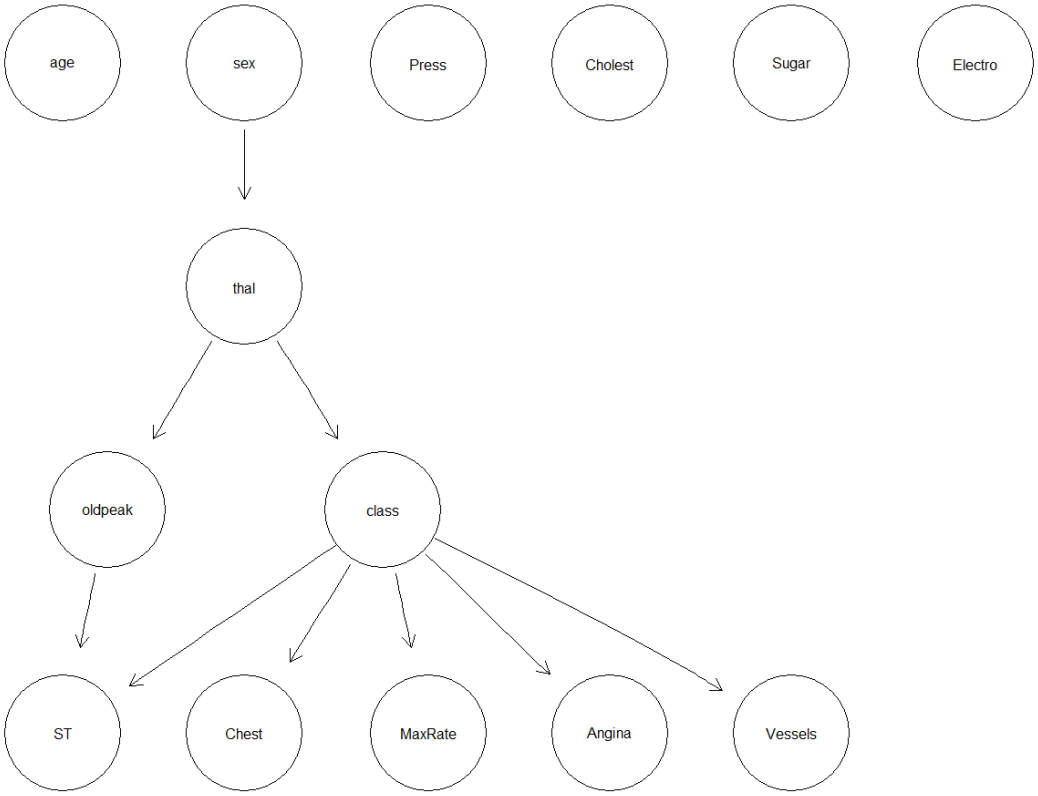
3-2 Test Shapiro-Wilka na rozkład normalny reszt

Sprawdzam przy pomocy testu Shapiro-Wilka czy występuje rozkład normalny reszt, ale nie dzieje się tak w żadnym przypadku, więc dyskretyzuję dane.

Następnie tworzę sieci w 3 wariantach: na danych przed dyskretyzacją, po dyskretyzacji oraz na danych po dyskretyzacji, ale na podstawie sieci utworzonej przed. Używam do tego algorytmów: hc, iamb, gs, pc.stable.

Nazwa algorytmu	Dane	Score
hc	Przed dyskretyzacją	-6734.621
hc	Po dyskretyzacji	-3276.803
hc	Po dyskretyzacji na sieci przed	-3477.814
iamb	Przed dyskretyzacją	-6841.408
iamb	Po dyskretyzacji	-3366.27
iamb	Po dyskretyzacji na sieci przed	-3582.758
pc.stable	Przed dyskretyzacją	-6835.526
pc.stable	Po dyskretyzacji	----------
pc.stable	Po dyskretyzacji na sieci przed	-3382.51
gs	Przed dyskretyzacją	-6921.047
gs	Po dyskretyzacji	-3404.635
gs	Po dyskretyzacji na sieci przed	-3671.392

Najlepszy wynik otrzymałem przy wykorzystaniu algorytmu hc na danych po dyskretyzacji, tak prezentuje się otrzymana sieć.



3-3 Graficzne przedstawienie sieci z najlepszą wartością score

Wykorzystuję funkcję bn.fit dla tej sieci, następnie compile(as.grain).
Dla tej sieci obliczam prawdopodobieństwa.

Prawdopodobieństwo wartości „normalna” dla talasemii, gdy wartość dla płci to „0” (prawdopodobnie oznacza to kobietę):

Prawdopodobieństwo nachylenia ST o wartości „2” dla class „2”, czyli braku choroby serca i oldpeak(obniżenia odcinka ST po wysiłku w stosunku do odpoczynku) o wartości „0” :

$$P(ST = 2 \mid class = 2, oldpeak = 0) = \frac{P(ST = 2 \cap class = 2 \cap oldpeak = 0)}{P(class = 2 \cap oldpeak = 0)} = \frac{4/270}{24/270} = 0.166667$$

```
>   querygrain(war_21,nodes="ST")$ST
ST
        1         2         3
0.8333333 0.1666667 0.0000000
>
```

3-4Obliczenie prawdopodobieństwa przy pomocy querygrain

Prawdopodobieństwo całkowite dla wartości thal=3.

$$P(thal = 3) = P(thal = 3\,|sex = 0) * P(sex = 0) + P(thal = 3\,|sex = 1) * P(sex = 1)$$
$$= 0.85057 * 03222 + 0.42623 * 0.67778 = 0.56296$$

Prawdopodobieństwo całkowite dla wartości thal=6.

$$P(thal = 6) = P(thal = 6|sex = 0) * P(sex = 0) + P(thal = 6\,|sex = 1) * P(sex = 1)$$
$$= 0 + 0.07650 * 0.67778 = 0.05185$$

Prawdopodobieństwo całkowite dla wartości thal=7.

$$P(thal = 7) = P(thal = 7\,|sex = 0) * P(sex = 0) + P(thal = 7\,|sex = 1) * P(sex = 1)$$
$$= 0.14942 * 0.32222 + 0.49726 * 0.67778 = \ 0.38518$$

```
Conditional probability table:

    sex
thal          0          1
   3 0.85057471 0.42622951
   6 0.00000000 0.07650273
   7 0.14942529 0.49726776
> bn_hc_siecDys$sex

  Parameters of node sex (multinomial distribution)

Conditional probability table:
         0         1
0.3222222 0.6777778
>
> querygrain(jun,nodes="thal")$thal
thal
         3          6          7
0.56296296 0.05185185 0.38518519
>
```

3-5Wartości prawdopodobieństwa z funkcji bn.fit oraz obliczona przy pomocy querygrain dla thal

Prawdopodobieństwo całkowite występowania choroby serca (class=1).

$$P(class = 1) = P(class = 1|thal = 3) * P(thal = 3) +$$
$$(class = 1|thal = 6) * P(thal = 6) + (class = 1|thal = 7) * P(thal = 7)$$
$$= 0.78289 * 0.56296 + 0.42857 * 0.05185 + 0.24038 * 0.38518 = 0.55555$$

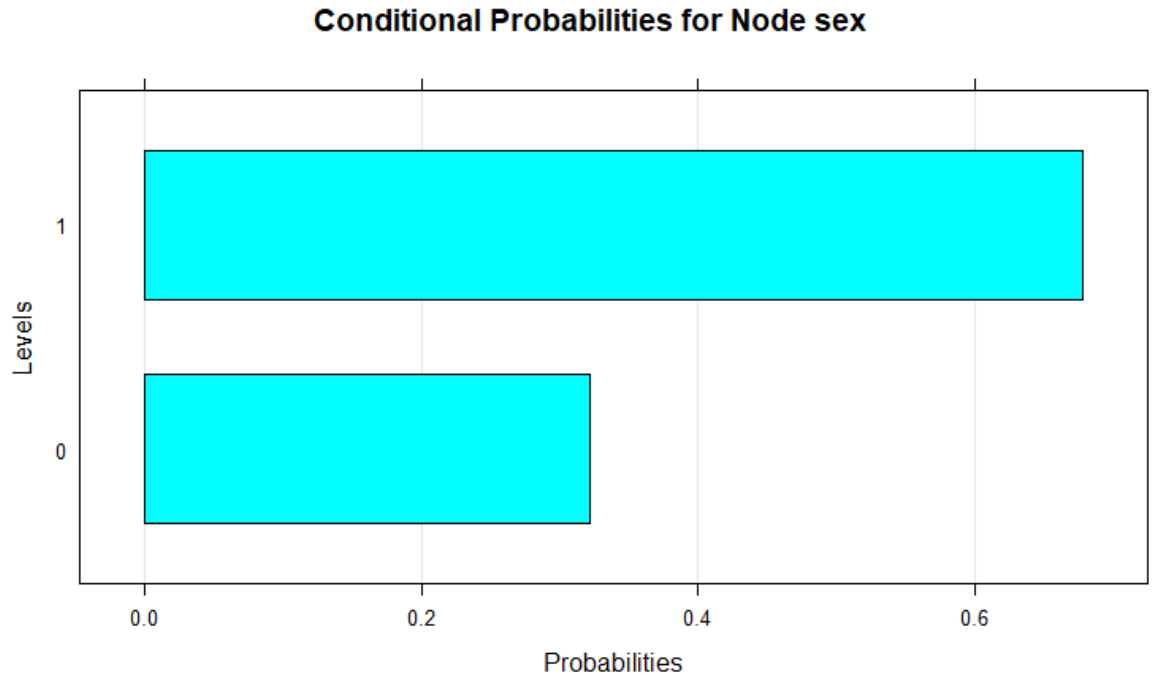
Prawdopodobieństwo całkowite braku występowania choroby serca (class=2).

$$P(class = 2) = P(class = 2|thal = 3) * P(thal = 3) +$$
$$(class = 2|thal = 6) * P(thal = 6) + (class = 2|thal = 7) * P(thal = 7)$$
$$= 0.21710 * 0.56296 + 0.57142 * 0.05185 + 0.75961 * 0.38518 = 0.44443$$

```
> bn_hc_siecDys$class

  Parameters of node class (multinomial distribution)

Conditional probability table:

     thal
class         3         6         7
    1 0.7828947 0.4285714 0.2403846
    2 0.2171053 0.5714286 0.7596154
>
> querygrain(jun,nodes="class")$class
class
        1         2
0.5555556 0.4444444
>
```

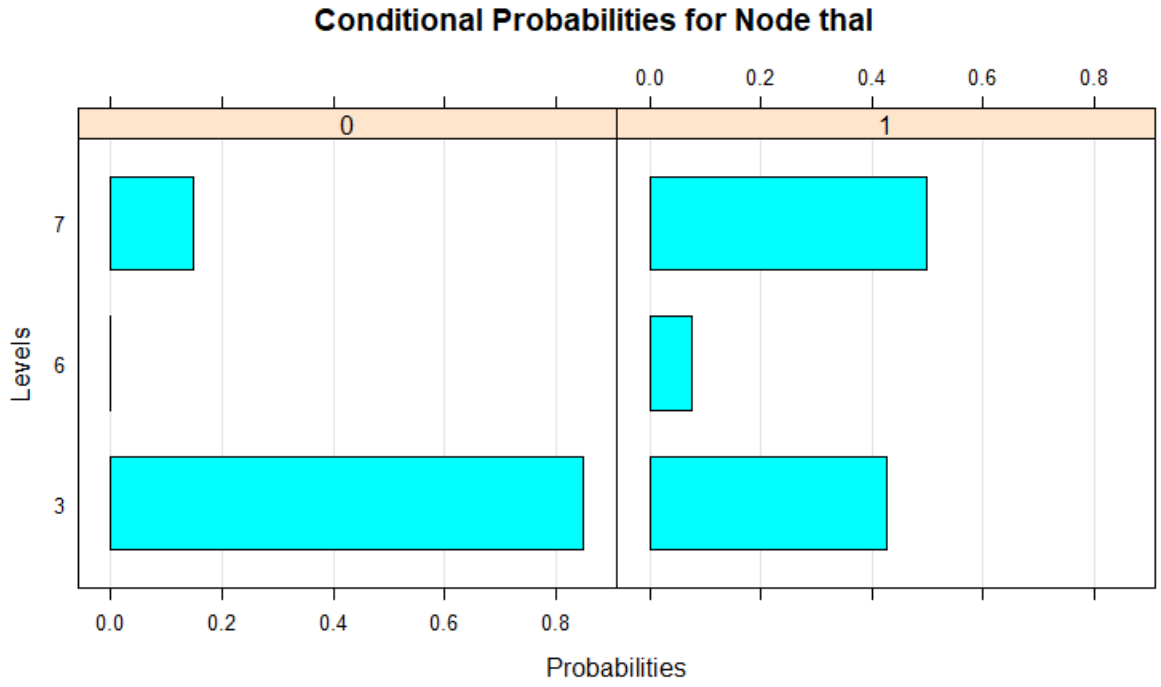
3-6 Wartości prawdopodobieństwa z funkcji bn.fit oraz obliczona przy pomocy querygrain dla class

Prawdopodobieństwa wyrysowane przy pomocy bn.fit.barchart



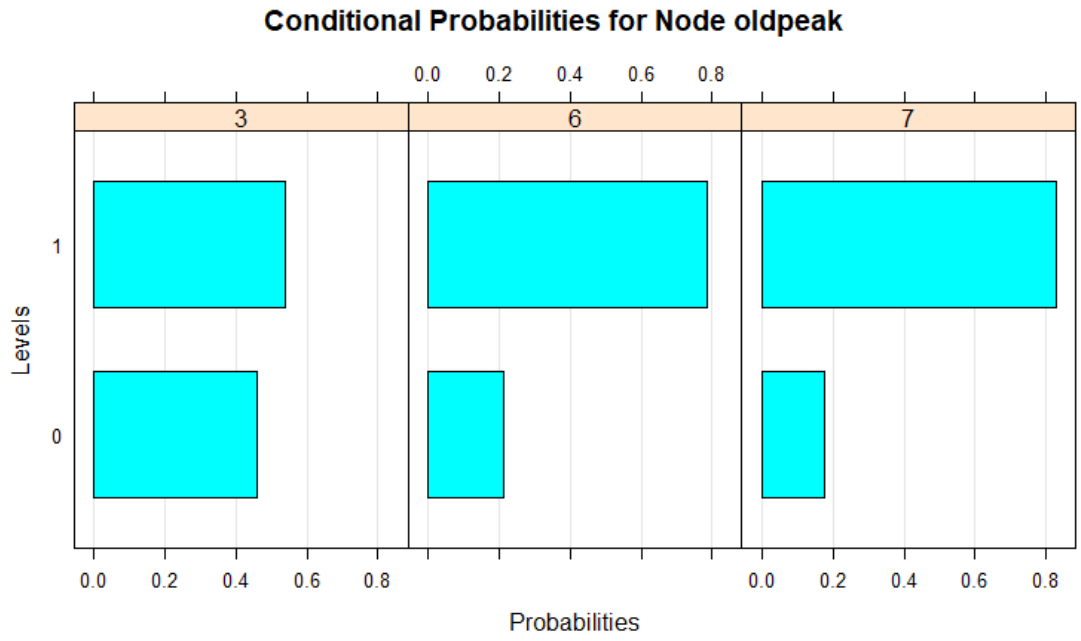
3-7 Prawdopodobieństwa dla węzła płeć

Niestety w opisie danych nie ma wyjaśnienia która płeć jest oznaczona przez 0, a która przed 1, około jedna trzecia badanych to płeć „0”.



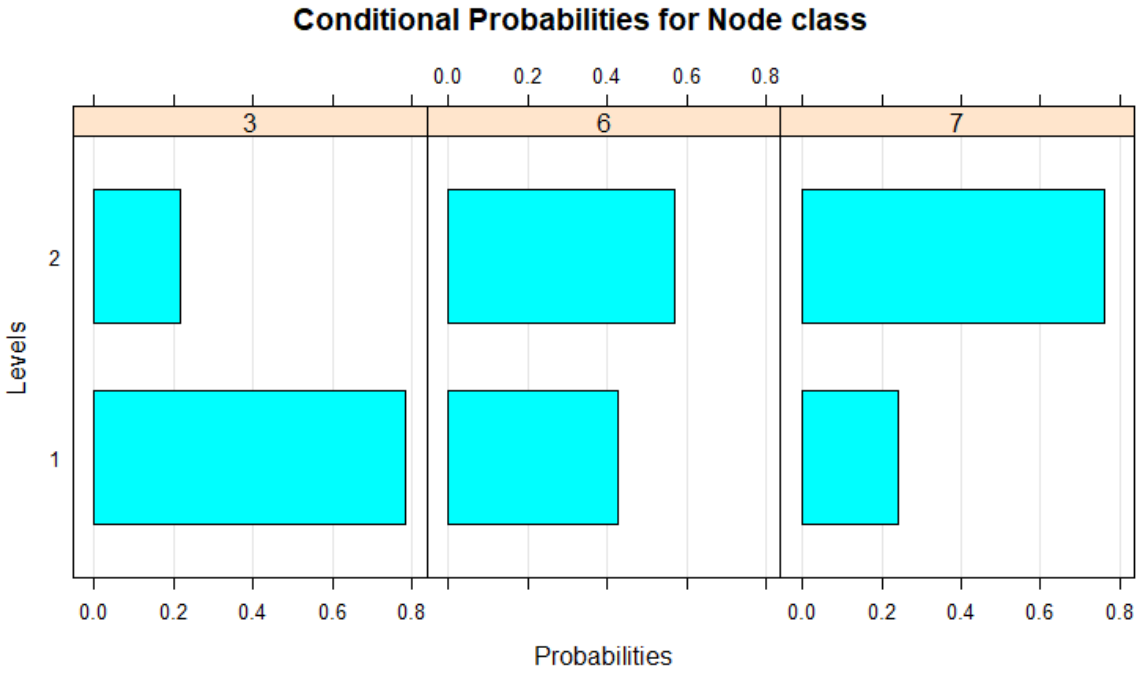
3-8 Prawdopodobieństwa dla węzła talasemia

Wśród grupy z płci „0” wartości dla talasemii w ponad 80% znajdowały się w normie, wśród płci „1” było to mniej niż 50%. Wartość „naprawiona wada” wśród płci „0” nie występuje, wśród „1” jest rzadka.



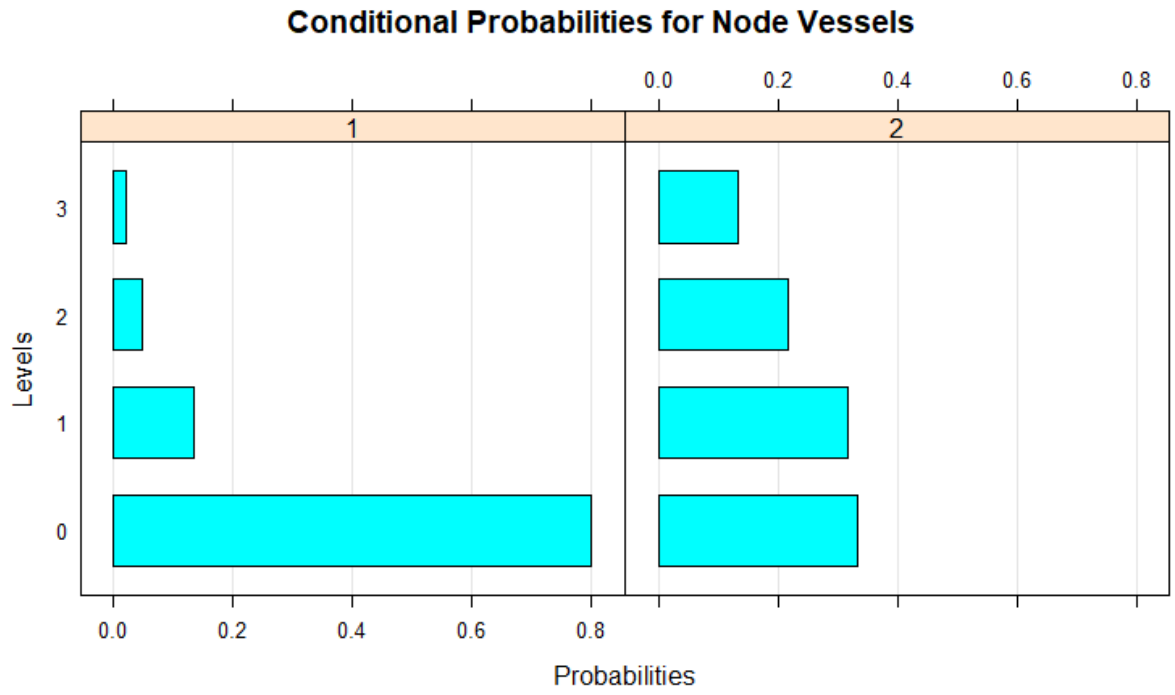
3-9 Prawdopodobieństwa dla węzła oldpeak

W przypadku wartości „normalna” w węźle dotyczącym niedokrwistości wartość dla oldpeak rozkłada się prawie równomiernie, w innych wypadkach znacznie częściej pojawia się wartość „1”, która według moich przypuszczeń (bo nie jest to podane w opisie) oznacza wystąpienie obniżenia odcinka ST wywołanego wysiłkiem fizycznym w stosunku do odpoczynku.



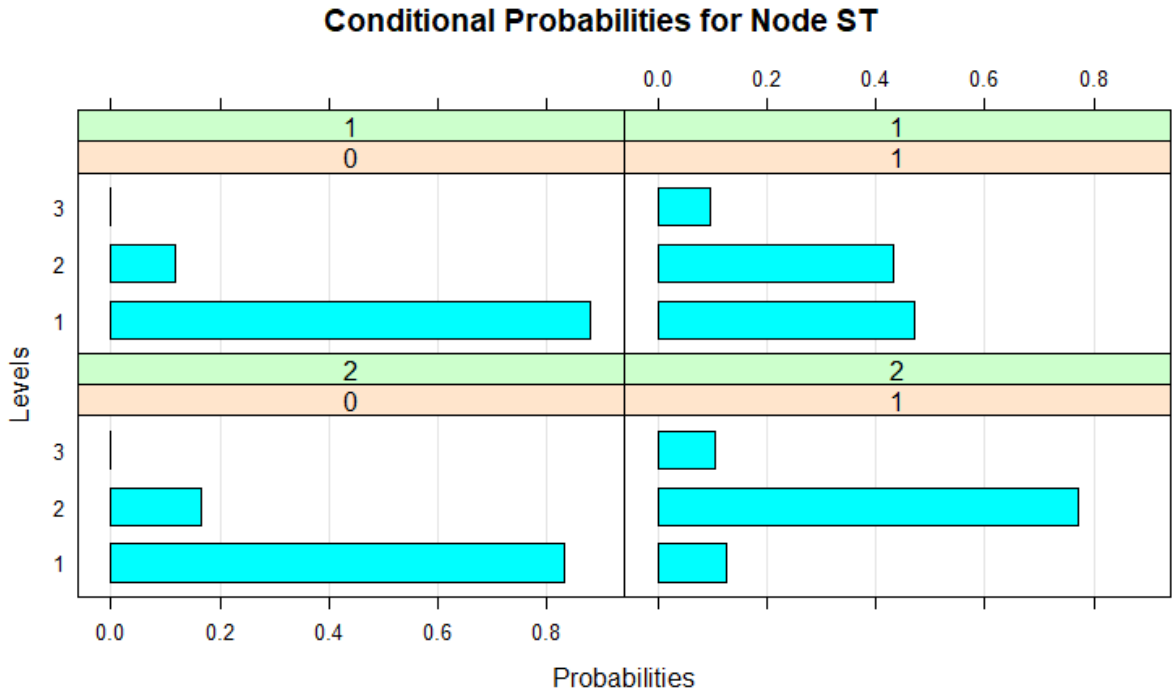
3-10 Prawdopodobieństwa dla węzła dotyczącego występowania chorób serca

Z powyższego wykresu możemy odczytać, że w przypadku braku nie niedokrwistości tarczowatokrwinkowej prawdopodobieństwo choroby serca znacząco maleje do poniżej 20%, w przypadku wartości opisanej jako „naprawiona wada” szanse na chorobę serca wynoszą ponad 50%, a w przypadku „wady odwracalnej” ponad 75%.



3-11 Prawdopodobieństwa dla węzła dotyczącego fluoroskopii

W przypadku braku choroby serca liczba głównych naczyń zabarwionych podczas fluoroskopii. W przypadku pacjentów bez chorób serca, w 80% była to liczba 0, nieco ponad 10% stanowiła liczba 1, rzadziej 2 i 3, wśród chorych rozkład jest bardziej równomierny, około 30-35% 0 oraz 1, nieco ponad 20% 2 i poniżej 20% 3.



3-12 Prawdopodobieństwa dla węzła dotyczącego odcinka ST

W przypadku kiedy oldpeak ma wartość 0, występowanie choroby serca lub jej brak nie ma znaczącego wpływu na ST, natomiast dla oldpeak 1 znacznie częściej w węźle ST pojawiają się wartości 2, w przypadku braku natomiast wartości 1 są nieco częstsze niż wartości 2.

4. Bibliografia

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(heart) dostęp 13.05.2022

[2] https://www.aaai.org/Papers/FLAIRS/2003/Flairs03-073.pdf dostęp 14.05.2022

[3] https://www.doz.pl/czytelnia/a14936-Talasemia_niedokrwistosc_tarczowatokrwinkowa__przyczyny_formy_objawy_i_leczenie dostęp 14.05.2022

[4] https://pl.wikipedia.org/wiki/Sie%C4%87_bayesowska dostęp 15.04.2022