Raport - Eksploracja Danych

Choroby serca

Patryk Zając 297110
Szymon Dośpiał, 303057

1 Eksploracyjna analiza danych **3**
1.1 Analiza pojedynczych zmiennych 3
1.1.1 Wiek 3
1.1.2 Płeć 4
1.1.3 Chest Pain Type 4
1.1.4 Ciśnienie 5
1.1.5 Cholesterol 6
1.1.6 Cukrzyca 7
1.1.7 Spoczynkowe wyniki elektrokardiograficzne 7
1.1.8 Angina indukowana wysiłkiem fizycznym 8
1.1.9 Maksymalne tętno 9
1.1.10 Oldpeak 9
1.1.11 Nachylenie szczytowego odcinka ST podczas ćwiczenia 10
1.1.12 Liczba głównych naczyń zabarwionych metodą fluoroskopii 11
1.1.13 Talasemia 12
1.1.14 Choroba serca 12
1.2 Analiza predyktorów ze zmienną celu 13
1.2.1 Wiek 13
1.2.2 Ciśnienie 14
1.2.3 Maksymalny poziom tętna 16
1.2.4 Poziom cholesterolu 17
1.2.5 Oldpeak 19
1.2.6 Płeć 20
1.2.7 Typ bólu klatki piersiowej 21
1.2.8 Cukrzyca 22
1.2.9 Spoczynkowe wyniki elektrokardiograficzne 23
1.2.10 Angina indukowana wysiłkiem fizycznym 24
1.2.11 Nachylenie szczytowego odcinka ST podczas ćwiczenia 25
1.2.12 Liczba głównych naczyń zabarwionych metodą fluoroskopii 27
1.2.13 Talasemia 28
1.3 Mapa ciepła dla parametrów numerycznych 29
1.3 Podsumowanie 29

2 Budowa modeli klasyfikujących **30**
2.1 Metoda MLP 30
2.2 Lasy losowe 34

3 Porównanie modeli **39**

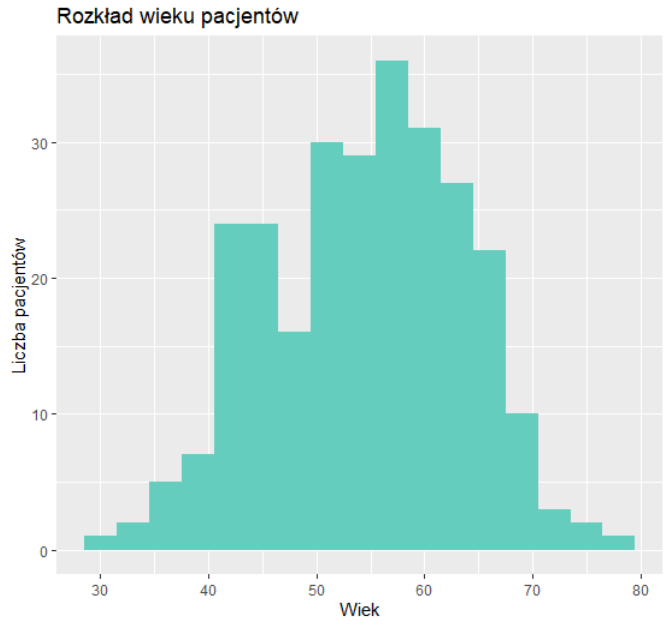
1 Eksploracyjna analiza danych

W tym akapicie przeprowadzimy eksploracyjną analizę danych.

1.1 Analiza pojedynczych zmiennych

Przeanalizujmy wykresy dla pojedynczych zmiennych. Użyjemy do tego wykresów słupkowych i histogramów.

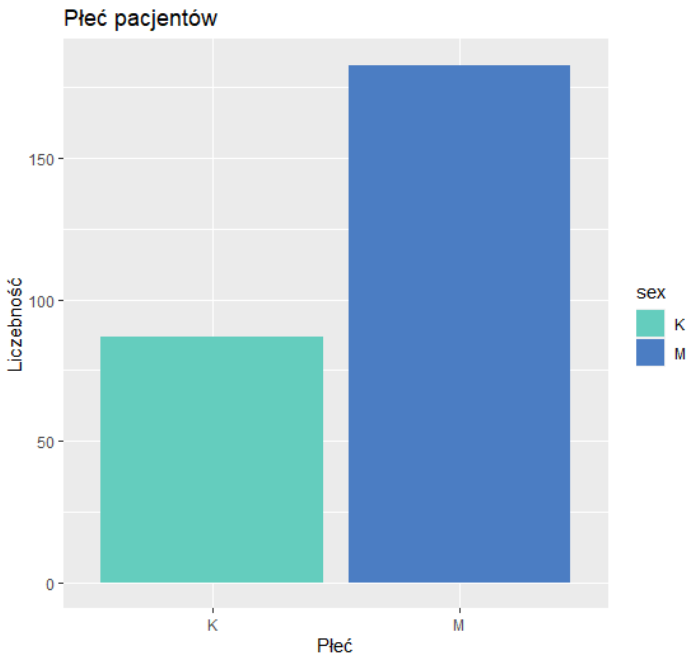
1.1.1 Wiek



Powyższy wykres pokazuje że:

- mamy zarówno mało obserwacji poniżej 40 lat jak i powyżej 70
- minimalny wiek to 29 lat, a maksymalny to 77 lat

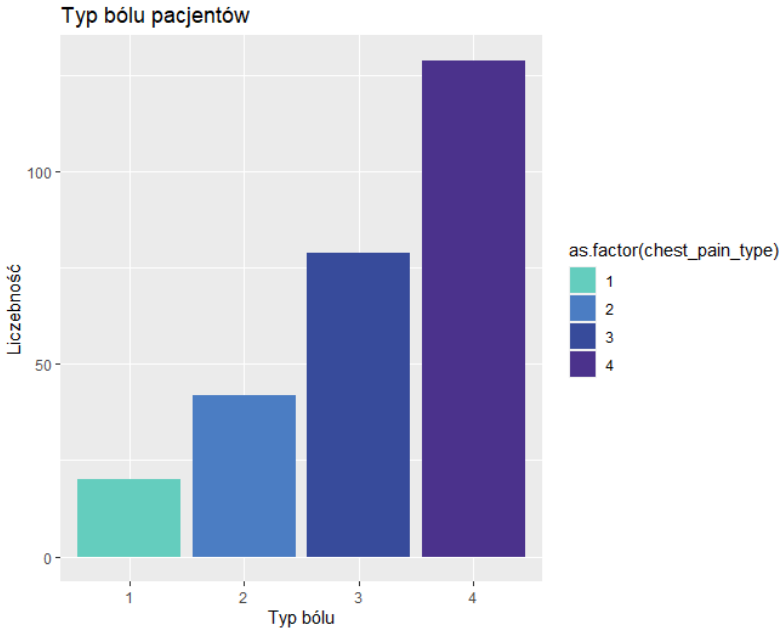
1.1.2 Płeć



Powyższy wykres pokazuje że:

- mamy dwie unikalne wartości: kobieta i mężczyzna
- w bazie danych jest zdecydowanie więcej mężczyzn, około 2 razy więcej niż kobiet
- ta baza danych jest bardziej skierowana ku mężczyznom

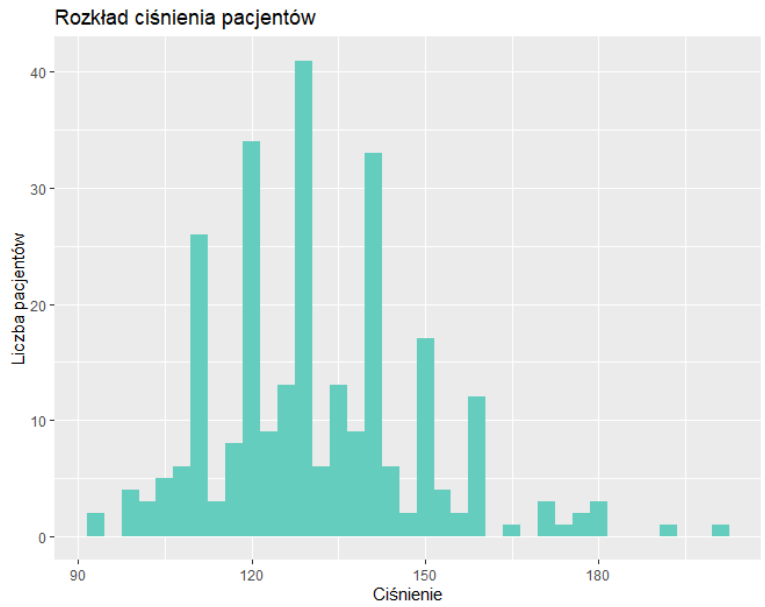
1.1.3 Chest Pain Type



Powyższy wykres pokazuje że:

- mamy cztery unikalne wartości bólu: 1, 2, 3 i 4
- najwięcej jest osób z bólem równym 4
- w bazie mamy rosnącą ilość osób względem typu bólu

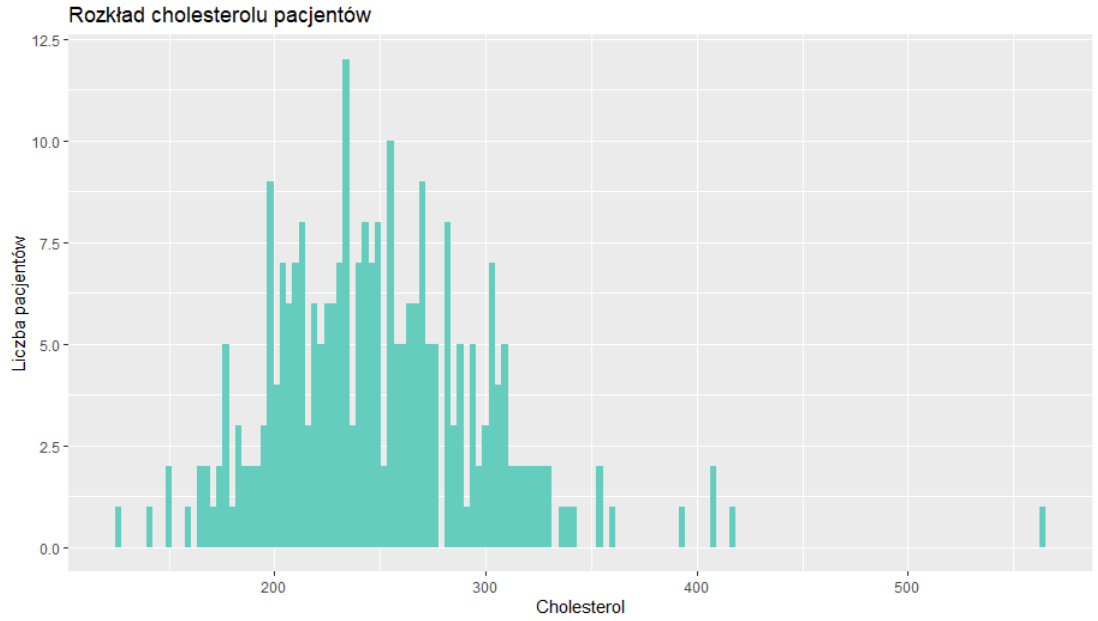
1.1.4 Ciśnienie



Powyższy wykres pokazuje że:

- zakres badanego spoczynkowego ciśnienia jest od 94 do 200
- najwięcej pacjentów ma ciśnienie w zakresie 120 a 140
- ciśnienie powyżej 165 występuje u małej ilości pacjentów

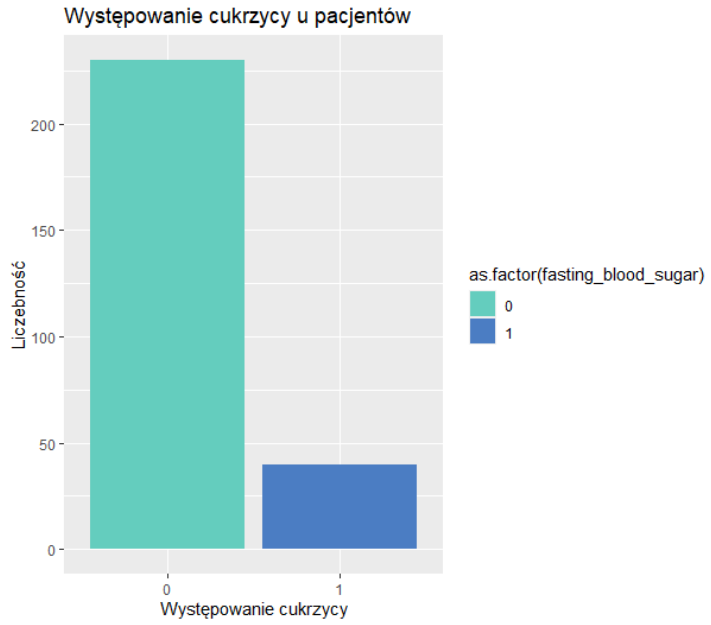
1.1.5 Cholesterol



Powyższy wykres pokazuje że:

- zakres badanego cholesterolu jest od 126 do 564
- zdecydowana większość pacjentów posiada cholesterol w zakresie 200 do 300 (Cholesterol całkowity u zdrowego człowieka nie powinien przekraczać 200 mg/dl)
- mamy też kilka obserwacji z zdecydowanie za wysokim cholesterolem – powyżej 350 mg/dl
- baza danych jest zbudowana głównie z osób które mają za wysoki poziom cholesterolu

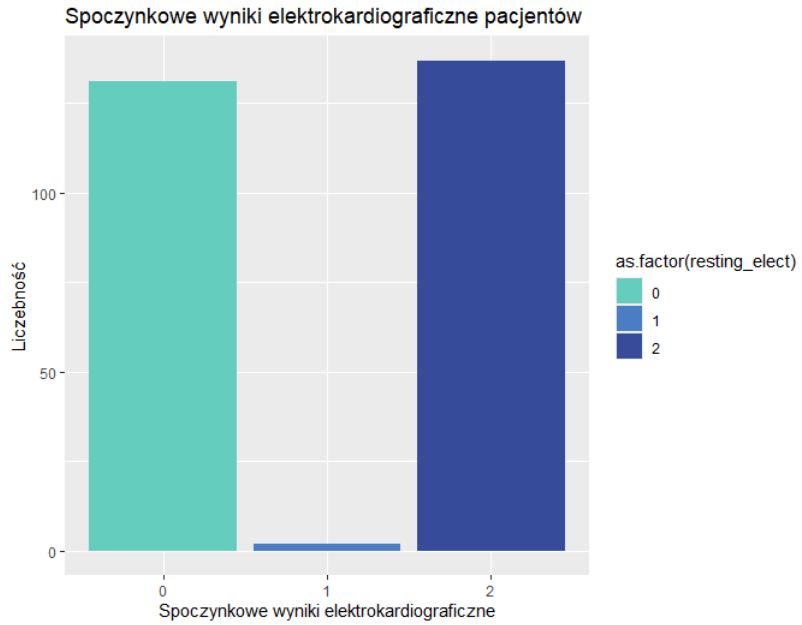
1.1.6 Cukrzyca



Powyższy wykres pokazuje że:

- mamy dwie unikalne wartości: 0 – osoba zdrowa, 1 – osoba chora
- baza danych składa się głównie z osób niechorujących na cukrzycę
- osoby chorujące stanowią zdecydowaną mniejszość w obserwacjach, jest ich mniej niż 50

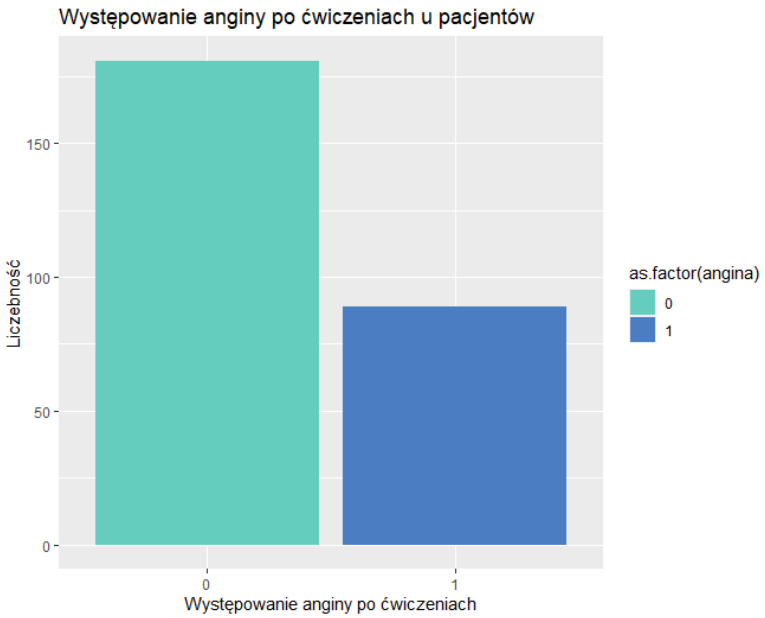
1.1.7 Spoczynkowe wyniki elektrokardiograficzne



Powyższy wykres pokazuje że:

- mamy trzy unikalne wartości: 0, 1, 2
- większość pacjentów ma wynik badania równy 0 lub 2
- mamy zaledwie kilka obserwacji gdzie wynik badania jest równy 1
- brakuje obserwacji z wynikiem badania równym 1, żeby bardziej zbalansować te kategorię

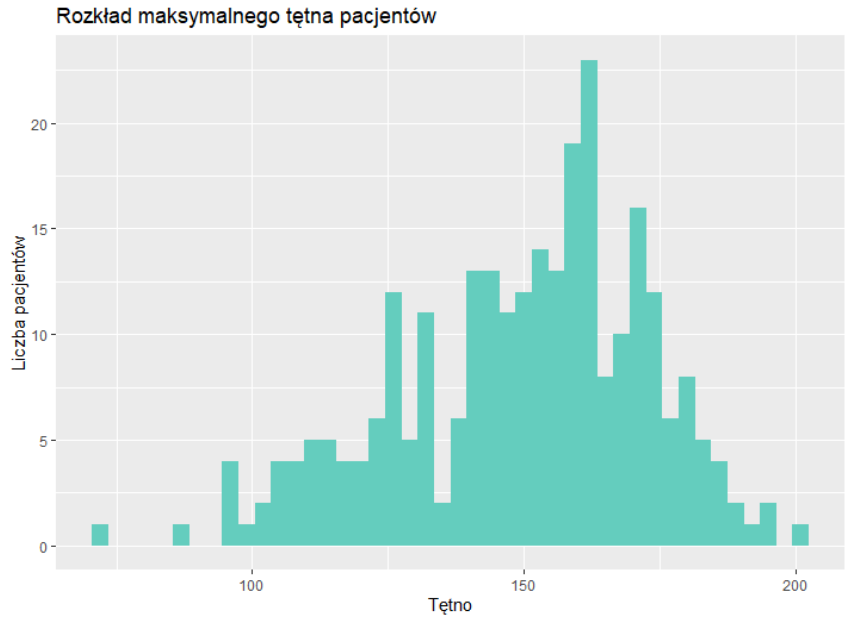
1.1.8 Angina indukowana wysiłkiem fizycznym



Powyższy wykres pokazuje że:

- mamy dwie unikalne wartości: 0 – niewystąpienie anginy, 1 – wystąpienie anginy
- pacjentów u których nie wystąpiła angina jest około 2 razy więcej

1.1.9 Maksymalne tętno

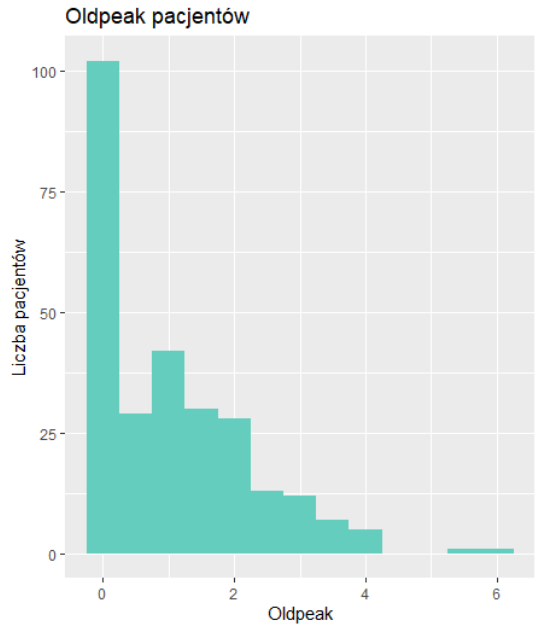


Powyższy wykres pokazuje że:

- zakres badanego maksymalnego tętna jest od 71 do 202
- zdecydowana większość pacjentów posiada maksymalne tętno w zakresie 125 do 175
- mamy też kilka obserwacji z tętnem poniżej 100 i jedną z tętnem 202

1.1.10 Oldpeak

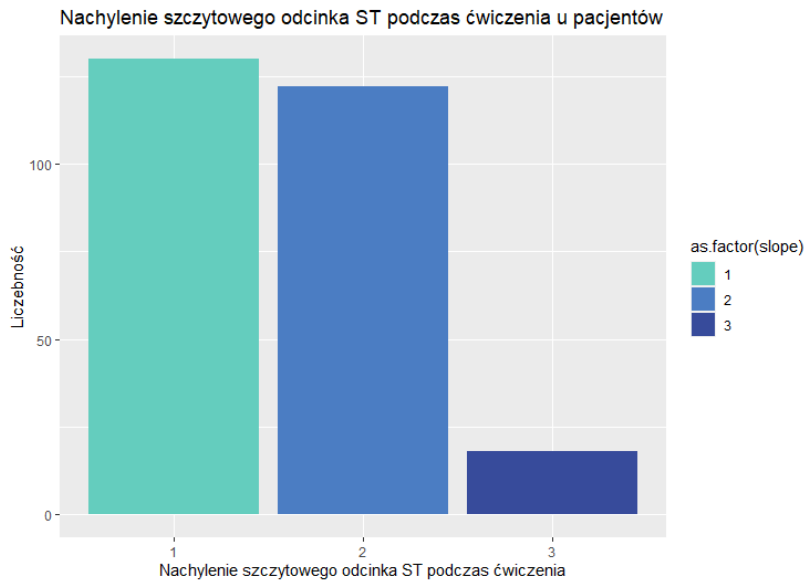
Oldpeak to obniżenie odcinka ST wywołane wysiłkiem fizycznym w stosunku do odpoczynku



Powyższy wykres pokazuje że:

- zakres obniżenia odcinka ST jest od 0 do 6.2
- większość pacjentów ma wartość oldpeak w zakresie od 0 do 2
- wartości powyżej 3 są rzadziej spotykane wśród badanych

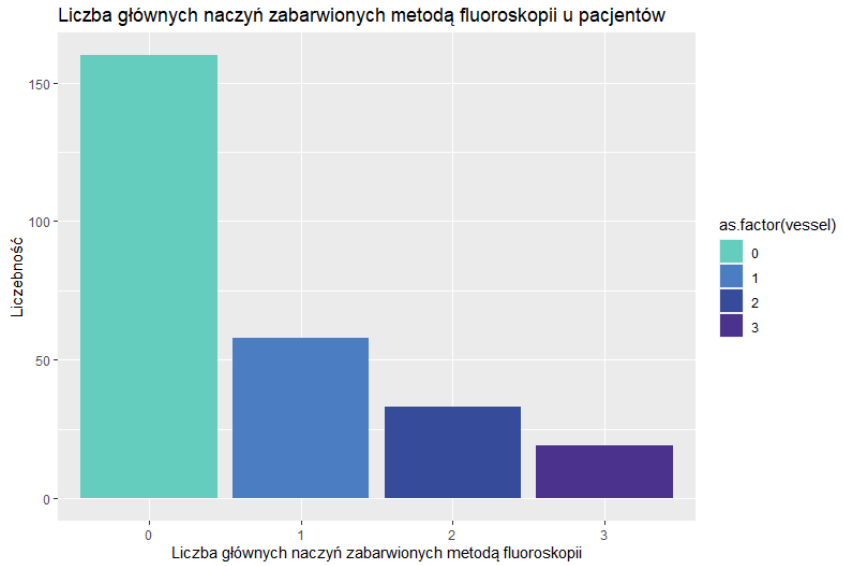
1.1.11 Nachylenie szczytowego odcinka ST podczas ćwiczenia



Powyższy wykres pokazuje że:

- mamy trzy unikalne wartości: 1, 2, 3
- najwięcej mamy pacjentów z wartością 1 i 2
- pacjenci z wartością 3 są w mniejszości

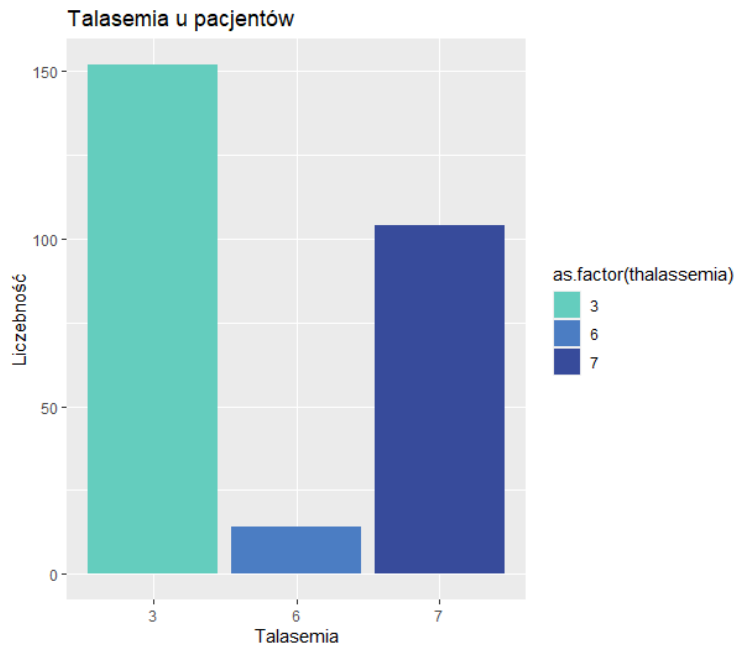
1.1.12 Liczba głównych naczyń zabarwionych metodą fluoroskopii



Powyższy wykres pokazuje że:

- mamy cztery unikalne wartości: 0, 1, 2, 3
- pacjenci z ilością naczyń równą 0 dominują w tej bazie danych
- najmniej obserwacji znajduje się przy ilości naczyń równej 3

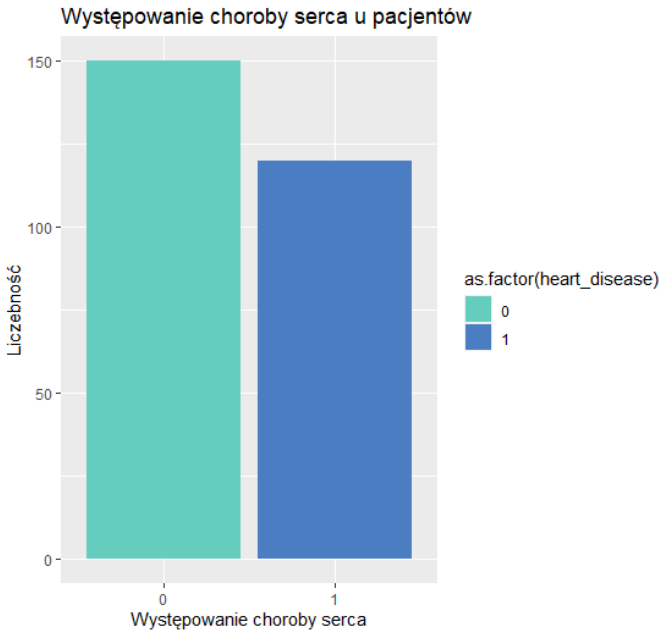
1.1.13 Talasemia



Powyższy wykres pokazuje że:

- mamy trzy unikalne wartości: 3 – normalna, 6 – naprawiony defekt, 7 – odwracalny efekt
- najwięcej mamy pacjentów z wartością 3
- pacjenci z wartością 6 są w mniejszości

1.1.14 Choroba serca



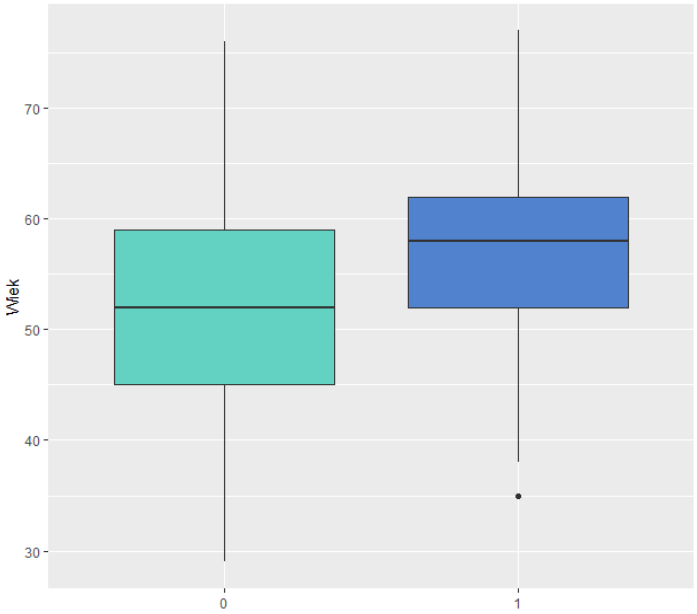
Powyższy wykres pokazuje że:

- ta baza danych jest dobrze zbalansowana pod względem ilości osób zdrowych a osób chorych – 150 osób zdrowych i 120 osób chorych

1.2 Analiza predyktorów ze zmienną celu

W celu tej analizy wykorzystamy wykresy pudełkowe i zestawione wykresy słupkowe. Do analizy czy dana zmienna ma wpływ na występowanie choroby serca użyjemy testu chi-kwadrat oraz test t Welcha. Przyjmiemy istotności jako 0.05.

1.2.1 Wiek



Z powyższego wykresu wynika:

- wartości odstające dla wieku występują tylko w grupie osób posiadających chorobę serca
- grupa osób chorujących ma wyższą medianę wieku równą 58 niż grupa osób zdrowych, której mediana wieku wynosi 52
- rozstęp kwartylny wieku:
    - dla osób zdrowych rozstęp jest pomiędzy 45 a 59 lat
    - dla osób chorych rozstęp jest między 52 a 62 lat (osoby w tym wieku chorują częściej)
    - rozstęp kwartylny wieku dla osób zdrowych ma odrobinę szerszy zakres niż rozstęp dla osób chorych

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

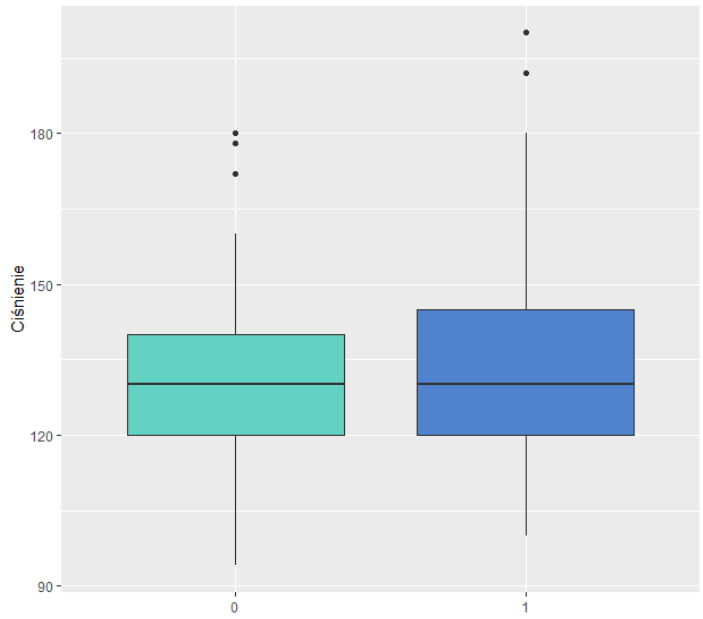
- H0 – nie ma różnicy między średnimi wieku osób chorych i zdrowych

- H1 – jest różnica między średnimi wieku osób chorych i zdrowych

```
Welch Two Sample t-test
data:  d1 and d0
t = 3.6199, df = 266.86, p-value = 0.0003526
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 1.771889 5.998111
sample estimates:
mean of x mean of y
 56.59167  52.70667
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc różnica pomiędzy średnimi wieku osób chorych i zdrowych. Oznacza to, że wiek ma pewien wpływ na występowanie choroby serca.

1.2.2 Ciśnienie



Z powyższego wykresu wynika:

- wartości odstające dla spoczynkowego ciśnienia występują zarówno w grupie osób chorych jak i w grupie osób zdrowych
- więcej wartości odstających jest w grupie osób zdrowych
- mediany obu grup są takie same (zarówno osoby zdrowe jak i chore mają spoczynkowe ciśnienie w okolicy 130)
- rozstęp kwartylny spoczynkowego ciśnienia:

- dla osób zdrowych jest pomiędzy 120 a 140
- dla osób chorych jest pomiędzy 120 a 145
- rozstęp kwartylny ciśnienia dla osób chorych ma odrobinę szerszy zakres niż rozstęp dla osób zdrowych

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

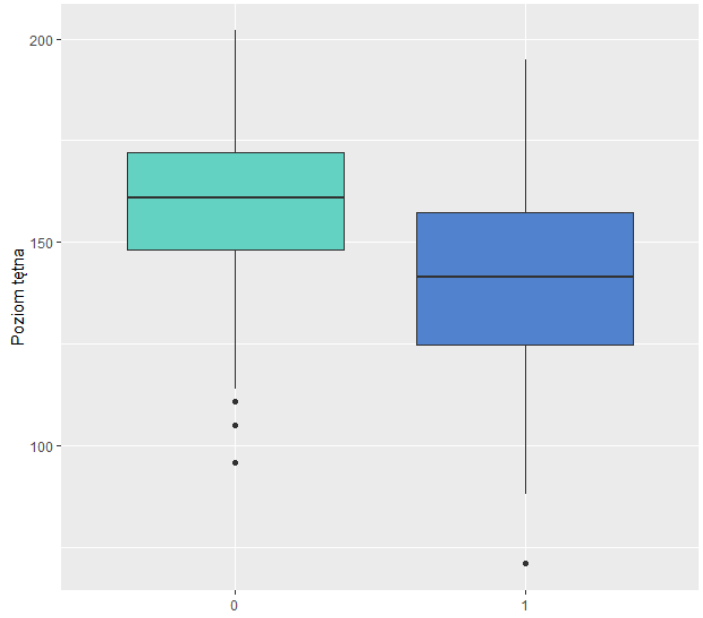
- H0 – nie ma różnicy między średnimi wartościami spoczynkowego ciśnienia osób chorych i zdrowych
- H1 – jest różnica między średnimi wartościami spoczynkowego ciśnienia osób chorych i zdrowych

```
Welch Two Sample t-test

data:  d1 and d0
t = 2.533, df = 235.92, p-value = 0.01196
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 1.238909 9.911091
sample estimates:
mean of x mean of y
 134.4417  128.8667
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc różnica pomiędzy średnimi spoczynkowymi wartościami ciśnienia osób chorych i zdrowych. Oznacza to, że ciśnienie ma pewien wpływ na występowanie choroby serca.

1.2.3 Maksymalny poziom tętna



Z powyższego wykresu wynika:

- wartości odstające dla maksymalnego tętna występują zarówno w grupie osób chorych jak i w grupie osób zdrowych
- więcej wartości odstających jest w grupie osób zdrowych
- grupa osób zdrowych ma wyższą medianę maksymalnego tętna równą 161 niż grupa osób chorych, której mediana maksymalnego tętna jest równa 141.5
- osoby chore mają często niższe tętno maksymalne niż osoby zdrowe
- rozstęp kwartylny tętna:
    - dla osób zdrowych rozstęp jest pomiędzy 148.2 a 172
    - dla osób chorych rozstęp jest pomiędzy 124.8 a 157.2
    - rozstęp kwartylny wieku dla osób chorych ma odrobinę szerszy zakres niż rozstęp dla osób chorych

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

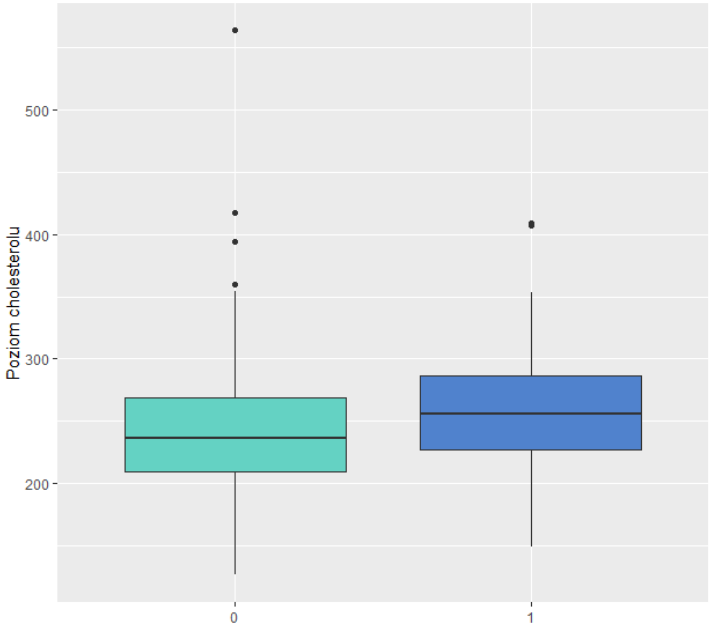
- H0 – nie ma różnicy między średnimi maksymalnymi wartościami tętna osób chorych i zdrowych
- H1 – jest różnica między średnimi maksymalnymi wartościami tętna osób chorych i zdrowych

```
Welch Two Sample t-test

data:  d1 and d0
t = -7.3939, df = 231.07, p-value = 2.604e-12
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -24.66458 -14.28542
sample estimates:
mean of x mean of y
 138.8583  158.3333
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc różnica pomiędzy średnimi maksymalnymi wartościami tętna osób chorych i zdrowych. Oznacza to, że maksymalny poziom tętna ma pewien wpływ na występowanie choroby serca.

1.2.4 Poziom cholesterolu



Z powyższego wykresu wynika:

- wartości odstające dla poziomu cholesterolu występują zarówno w grupie osób posiadających chorobę serca jak i zdrowych
- więcej wartości odstających dla poziomu cholesterolu zawiera grupa zdrowych
- grupa osób chorujących ma wyższą medianę poziomu cholesterolu równą 255.5 niż grupa osób zdrowych, której mediana wieku wynosi 236
- rozstęp kwartylny poziomu cholesterolu:

- dla osób zdrowych rozstęp jest pomiędzy 209 a 268.8
- dla osób chorych rozstęp jest pomiędzy 227.2 a 286.5
- rozstępy kwartylne mają prawie tak samo szerokie zakresy

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

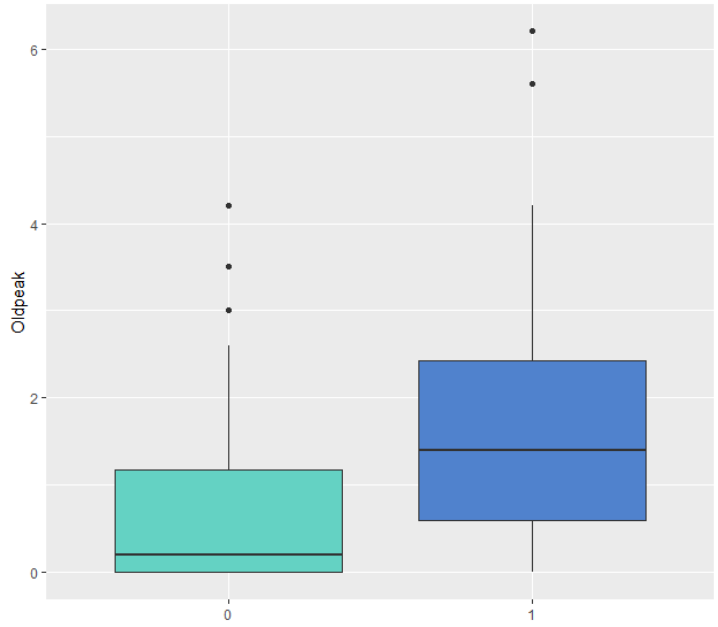
- H0 – nie ma różnicy między średnimi wartościami poziomu cholesterolu osób chorych i zdrowych
- H1 – jest różnica między średnimi wartościami poziomu cholesterolu osób chorych i zdrowych

```
Welch Two Sample t-test

data:  d1 and d0
t = 1.9715, df = 265.06, p-value = 0.04971
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
  0.01582888 24.49083778
sample estimates:
mean of x mean of y
 256.4667  244.2133
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc różnica pomiędzy średnimi wartościami poziomu cholesterolu osób chorych i zdrowych. Oznacza to, że poziom cholesterolu ma pewien wpływ na występowanie choroby serca.

1.2.5 Oldpeak



Z powyższego wykresu wynika:

- wartości odstające dla oldpeak występują zarówno w grupie osób posiadających chorobę serca jak i zdrowych
- więcej wartości odstających dla oldpeak zawiera grupa zdrowych
- grupa osób chorujących ma wyższą medianę oldpeak równą 1.4 niż grupa osób zdrowych, której mediana oldpeak wynosi 0.2
- rozstęp kwartylny oldpeak :
    - dla osób zdrowych rozstęp jest pomiędzy 0 a 1.18
    - dla osób chorych rozstęp jest pomiędzy 0.6 a 2.43
    - rozstęp kwartylny dla osób chorych ma szerszy zakres niż rozstęp dla osób zdrowych

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

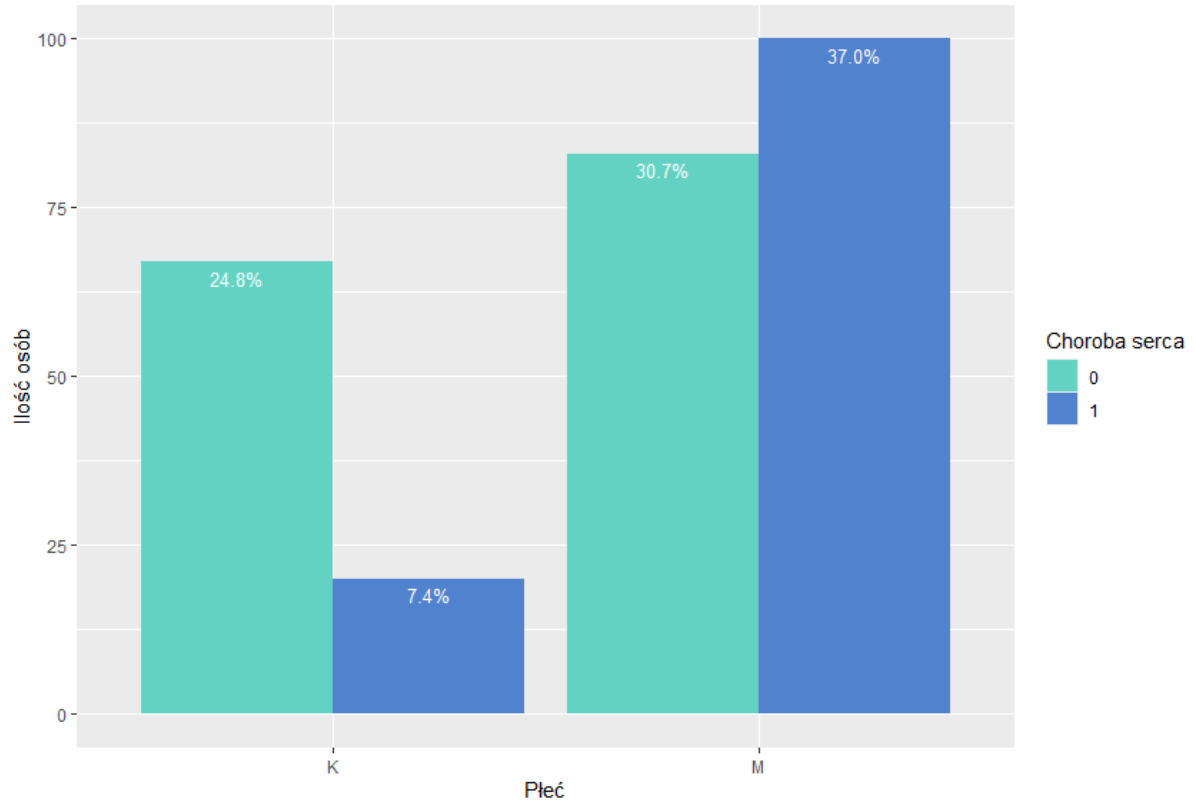
- H0 – nie ma różnicy między średnimi wartościami oldpeak osób chorych i zdrowych
- H1 – jest różnica między średnimi wartościami oldpeak osób chorych i zdrowych

```
Welch Two Sample t-test

data:  d1 and d0
t = 7.1719, df = 190.09, p-value = 1.601e-11
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 0.6970554 1.2259446
sample estimates:
mean of x mean of y
1.5841667 0.6226667
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc różnica pomiędzy średnimi wartościami oldpeak osób chorych i zdrowych. Oznacza to, że oldpeak ma pewien wpływ na występowanie choroby serca.

1.2.6 Płeć



Z powyższego wykresu wynika:

- występowanie choroby serca u kobiet jest rzadsze i wynosi 7.4% badanych pacjentów
- najczęściej choroba serca występuje u mężczyzn, stanowią oni 37% badanych pacjentów

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

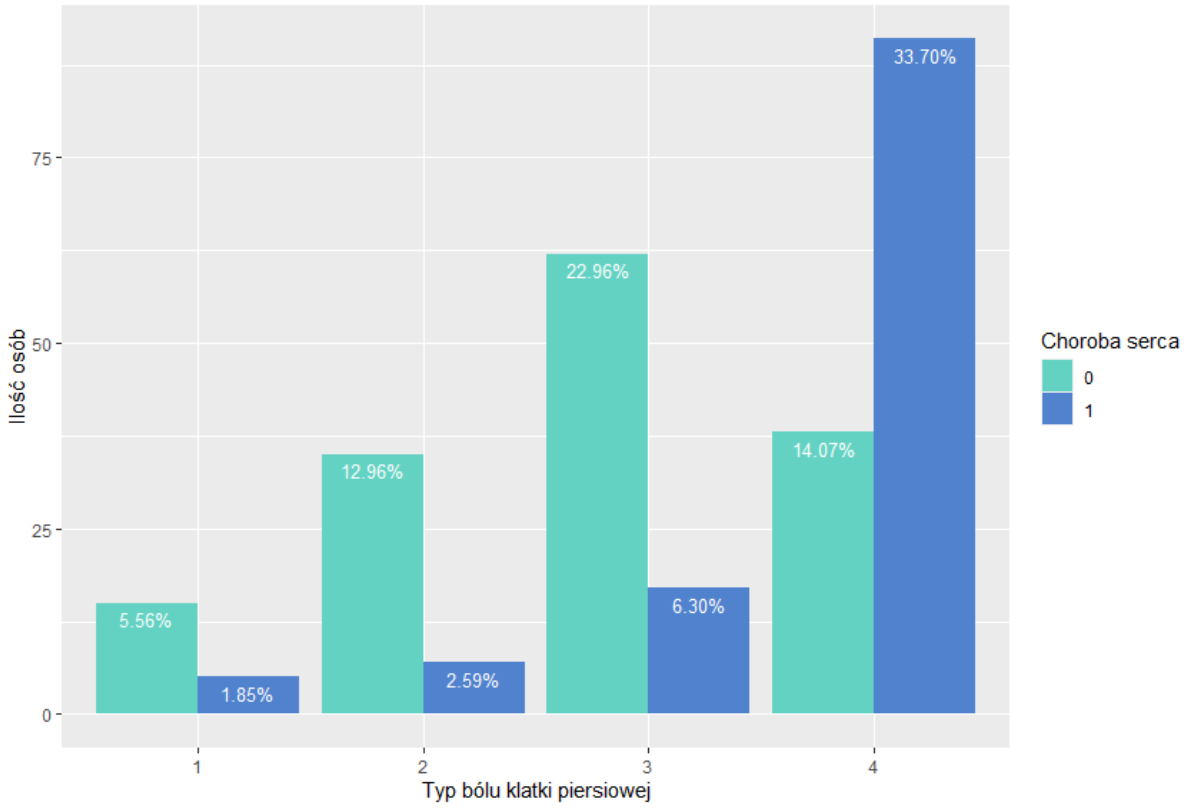
- H0 – nie ma zależności pomiędzy płcią a występowaniem choroby serca
- H1 – jest zależność pomiędzy płcią a występowaniem choroby serca

```
Pearson's Chi-squared test with Yates' continuity correction

data:  c_t
X-squared = 22.667, df = 1, p-value = 1.926e-06
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc zależność pomiędzy płcią a występowaniem choroby serca. Oznacza to, że płeć ma pewien wpływ na występowanie choroby serca.

1.2.7 Typ bólu klatki piersiowej



Z powyższego wykresu wynika:

- ilość chorych osób wzrasta wraz z wzrostem typu bólu (najwięcej chorych dla najmocniejszego bólu: 33.7% badanych, najmniej dla najsłabszego bólu: 1.85% badanych)
- pomiędzy typem 3 i 4 następuje gwałtowny wzrost zachorowań

- liczba zdrowych do pewnego momentu też wzrasta razem ze wzrostem bólu (momentem w którym liczba zdrowych spada jest najmocniejszy ból – 14.07% badanych, co jest całkiem logiczne).

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

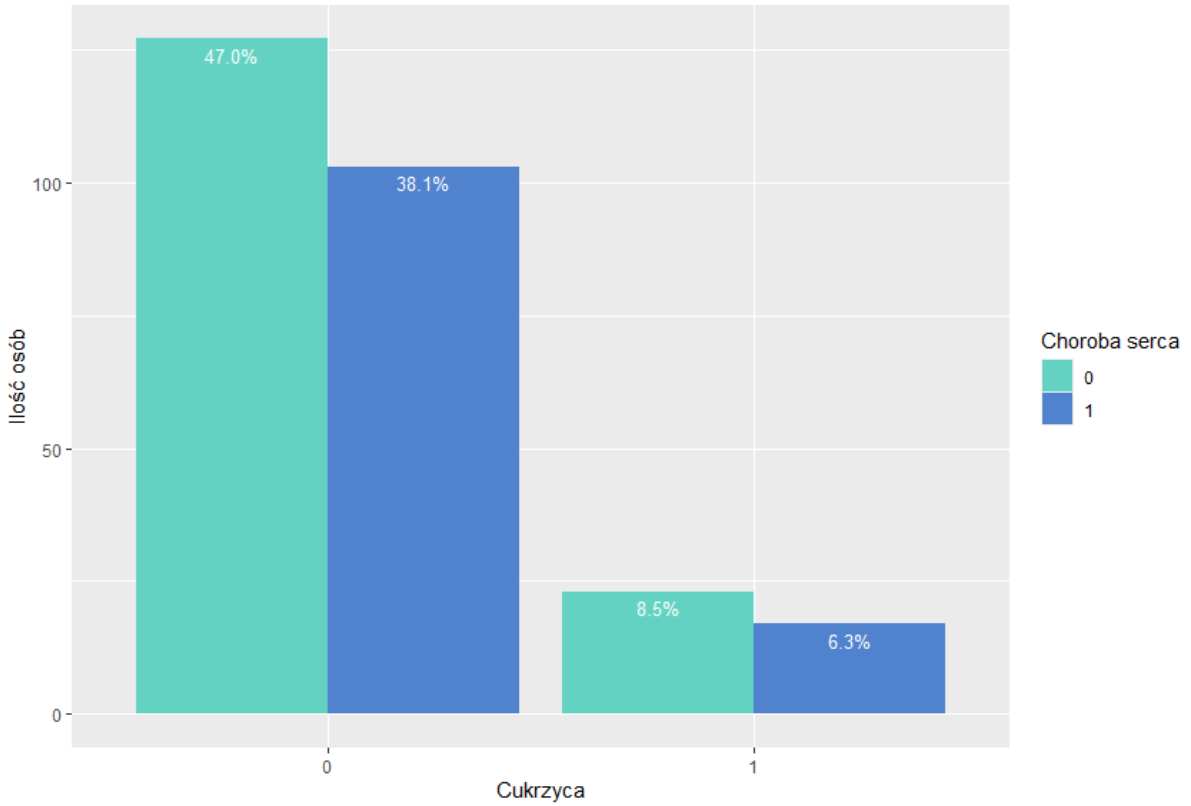
- H0 – nie ma zależności pomiędzy typem bólu a występowaniem choroby serca
- H1 – jest zależność pomiędzy typem bólu a występowaniem choroby serca

```
Pearson's Chi-squared test

data:  c_t
X-squared = 68.588, df = 3, p-value = 8.561e-15
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc zależność pomiędzy typem bólu a występowaniem choroby serca. Oznacza to, że typ bólu ma pewien wpływ na występowanie choroby serca.

1.2.8 Cukrzyca



Z powyższego wykresu wynika:

- występowanie choroby serca jest częstsze u osób, które nie chorują na cukrzycę i wynosi 38.1% badanych pacjentów

- najrzadziej choroba serca występuje u osób które również chorują na cukrzycę, stanowią oni 6.3% badanych pacjentów

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

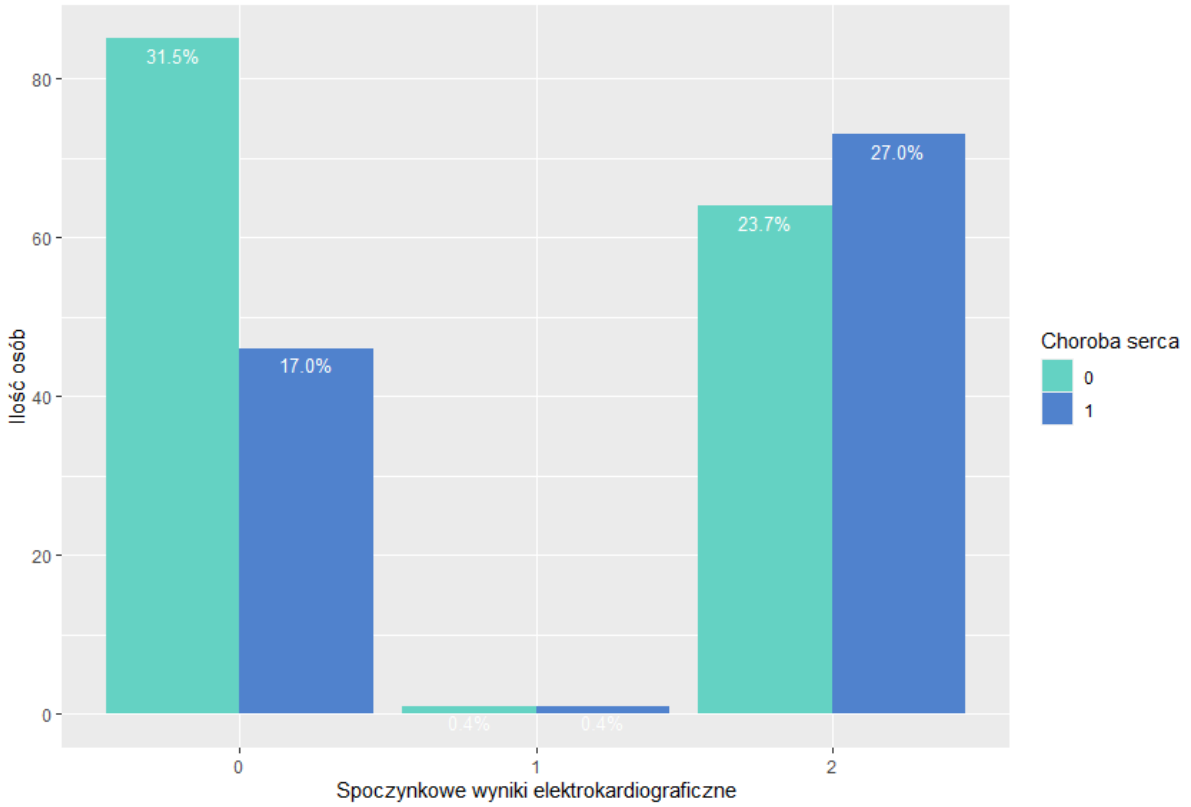
- H0 – nie ma zależności pomiędzy chorowaniem na cukrzycę a występowaniem choroby serca
- H1 – jest zależność pomiędzy chorowaniem na cukrzycę a występowaniem choroby serca

```
Pearson's Chi-squared test with Yates' continuity correction

data:  c_t
X-squared = 0.0091712, df = 1, p-value = 0.9237
```

P-value > 0.05, nie mamy podstaw aby odrzucić hipotezę zerową. Nie istnieje więc zależność pomiędzy chorowaniem na cukrzycę a występowaniem choroby serca. Oznacza to, że chorowanie na cukrzycę nie ma wpływu na występowanie choroby serca.

1.2.9 Spoczynkowe wyniki elektrokardiograficzne



Z powyższego wykresu wynika:

- najczęściej chorują osoby z wynikiem 2, stanowią oni 27% badanych

- najrzadziej chorują osoby z wynikiem 1, stanowią oni 0.4% (ale może to być spowodowane małą ilością obserwacji z tym wynikiem)

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

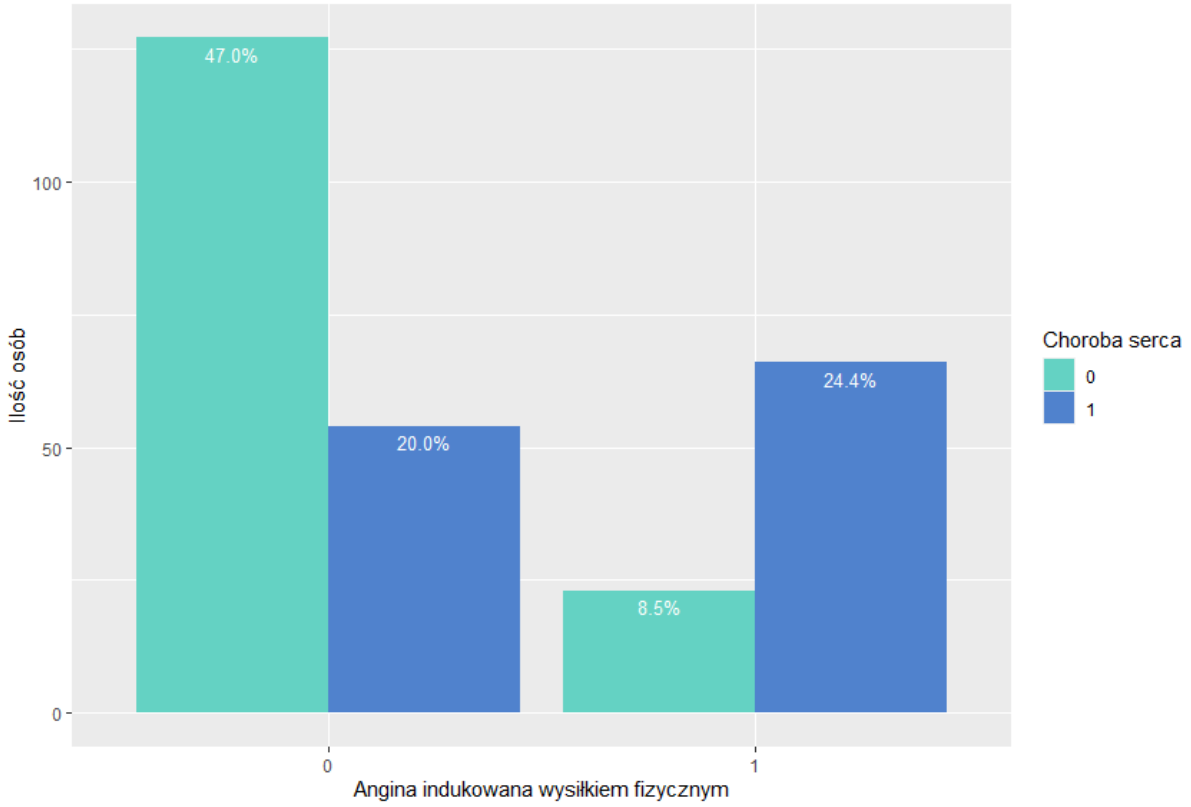
- H0 – nie ma zależności pomiędzy spoczynkowym wynikiem elektrokardiograficznym a występowaniem choroby serca
- H1 – jest zależność pomiędzy spoczynkowym wynikiem elektrokardiograficznym a występowaniem choroby serca

```
Pearson's Chi-squared test

data:  c_t
X-squared = 8.9795, df = 2, p-value = 0.01122
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc zależność pomiędzy spoczynkowym wynikiem elektrokardiograficznym a występowaniem choroby serca. Oznacza to, że spoczynkowy wynik elektrokardiograficzny ma pewien wpływ na występowanie choroby serca.

1.2.10 Angina indukowana wysiłkiem fizycznym



Z powyższego wykresu wynika:

- najczęściej osoby u których wystąpiła angina indukowana wysiłkiem fizycznym posiadają również chorobę serca – 24.4% badanych
- najczęściej gdy u badanego nie wystąpiła angina po ćwiczeniu to również nie posiada choroby serca – 47% badanych

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

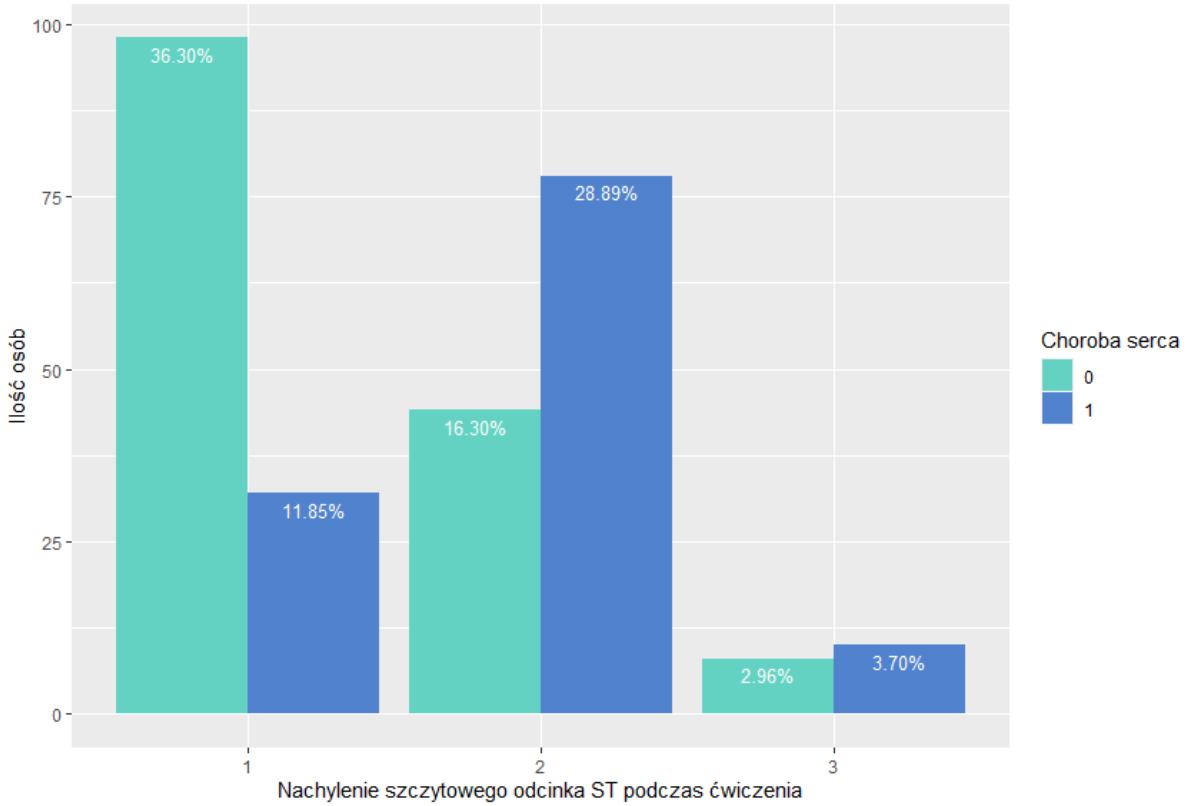
- H0 – nie ma zależności pomiędzy anginą indukowaną wysiłkiem fizycznym a występowaniem choroby serca
- H1 – jest zależność pomiędzy anginą indukowaną wysiłkiem fizycznym a występowaniem choroby serca

```
Pearson's Chi-squared test with Yates' continuity correction

data:  c_t
X-squared = 45.692, df = 1, p-value = 1.384e-11
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc zależność pomiędzy anginą indukowaną wysiłkiem fizycznym a występowaniem choroby serca. Oznacza to, że angina indukowana wysiłkiem fizycznym ma pewien wpływ na występowanie choroby serca.

1.2.11 Nachylenie szczytowego odcinka ST podczas ćwiczenia



Z powyższego wykresu wynika:

- najczęściej choroba serca wystąpiła u badanych u których podczas ćwiczenia nachylenie szczytowego odcinka ST wyniosło 2 – 28.89% obserwacji, a najrzadziej gdy wyniosło 3 – 3.7% obserwacji
- najczęściej osoby zdrowe miały nachylenie równe 1 – 36.3%

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

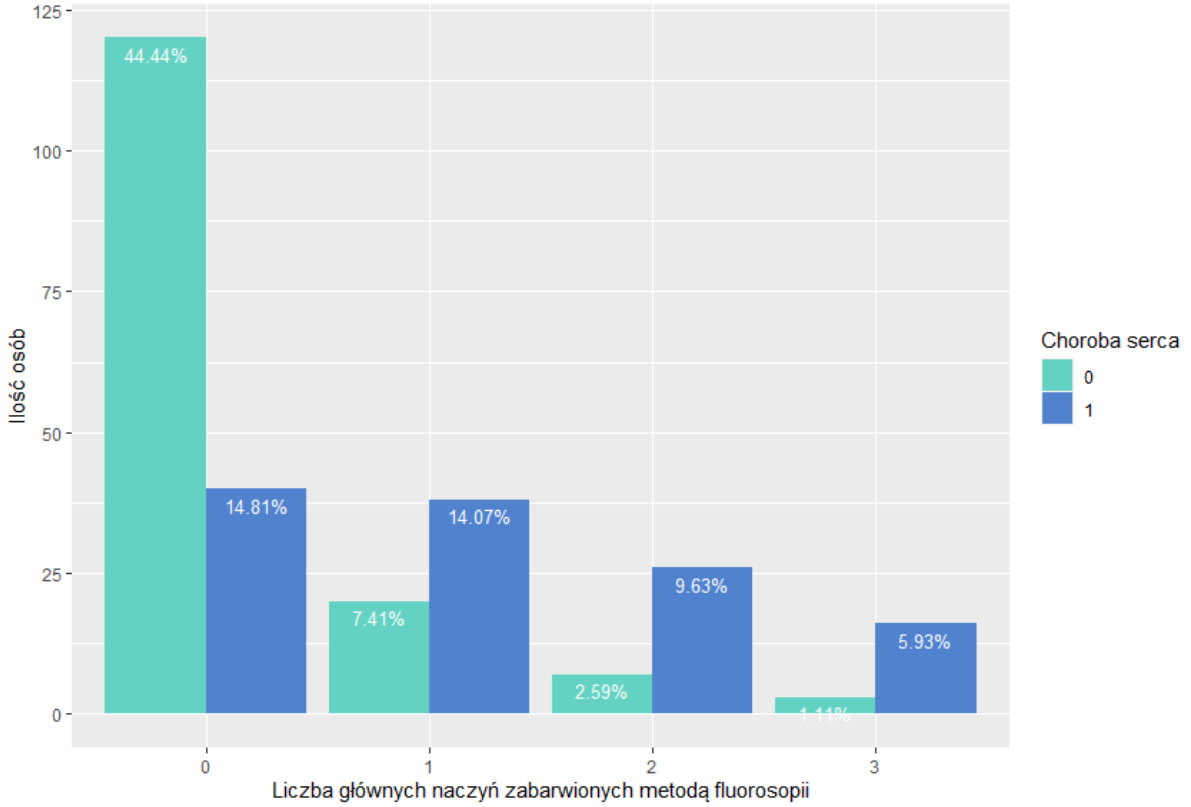
- H0 – nie ma zależności między nachyleniem szczytowego odcinka ST podczas ćwiczenia a występowaniem choroby serca
- H1 – jest zależność między nachyleniem szczytowego odcinka ST podczas ćwiczenia a występowaniem choroby serca

```
Pearson's Chi-squared test

data:  c_t
X-squared = 40.37, df = 2, p-value = 1.713e-09
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc zależność między nachyleniem szczytowego odcinka ST podczas ćwiczenia a występowaniem choroby serca. Oznacza to, że nachyleniem szczytowego odcinka ST podczas ćwiczenia ma pewien wpływ na występowanie choroby serca.

1.2.12 Liczba głównych naczyń zabarwionych metodą fluoroskopii



Z powyższego wykresu wynika:

- najczęściej choroba serca wystąpiła gdy liczba głównych naczyń zabarwionych metodą fluoroskopii wyniosła 0 i 1 – odpowiednio 14.81% i 14.07% obserwacji
- gdy liczba zabarwionych naczyń wynosi 0 jest więcej osób zdrowych niż chorych (zdrowych 44.44% obserwacji). W reszcie przypadków jest na odwrót.
- można zauważyć że coraz to większa ilość zabarwionych naczyń jest rzadziej spotykana wśród obserwacji

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

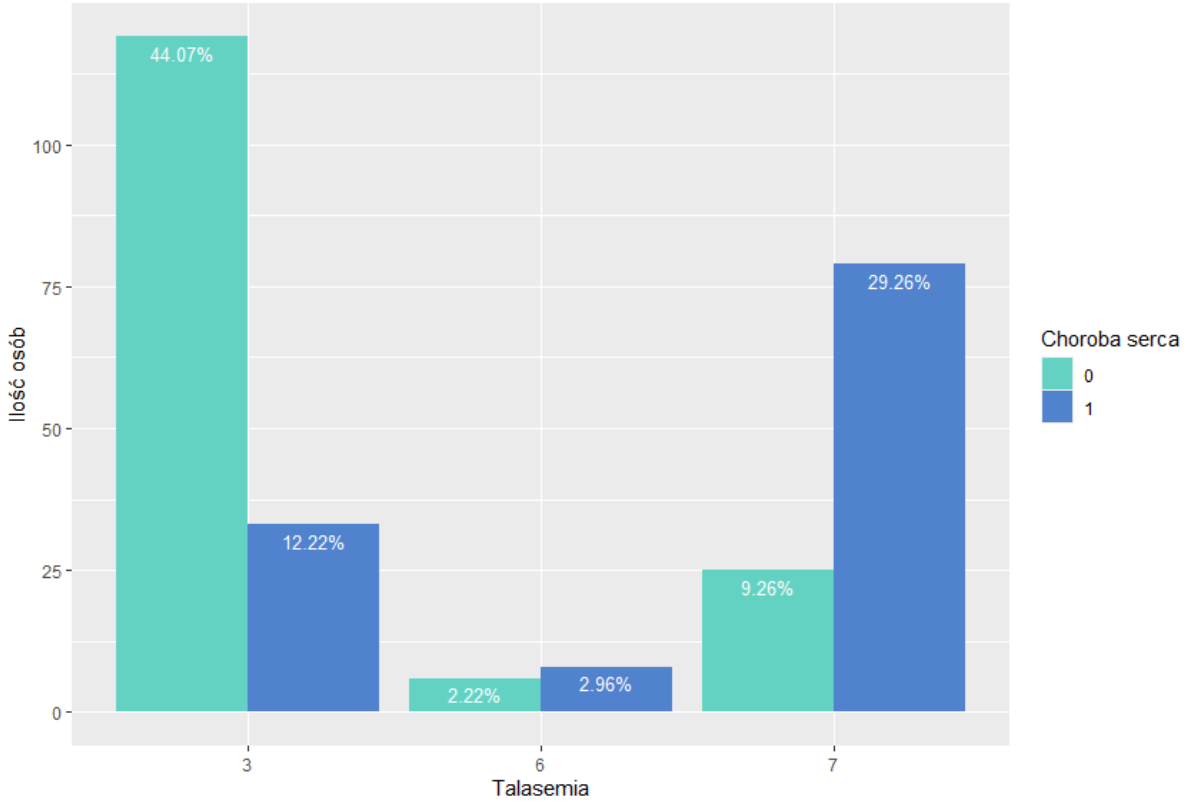
- H0 – nie ma zależności pomiędzy liczbą głównych naczyń zabarwionych metodą fluoroskopii a występowaniem choroby serca
- H1 – jest zależność pomiędzy liczbą głównych naczyń zabarwionych metodą fluoroskopii a występowaniem choroby serca

```
Pearson's Chi-squared test

data:  c_t
X-squared = 62.863, df = 3, p-value = 1.437e-13
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc zależność między liczbą głównych naczyń zabarwionych metodą fluoroskopii a występowaniem choroby serca. Oznacza to, że liczbą głównych naczyń zabarwionych metodą fluoroskopii ma pewien wpływ na występowanie choroby serca.

1.2.13 Talasemia



Z powyższego wykresu wynika:

- najczęściej osoby chore mają wartość talasemii równą 7, czyli odwracalny defekt – 29.26% obserwacji
- najrzadziej osoby chore mają talasemię równą 6, czyli naprawiony defekt – 2.96%
- przy talasemii równej 3, czyli prawidłowej, ilość osób zdrowych przewyższa ilość osób chorych – zdrowi stanowią 44.07% obserwacji

Określamy hipotezę zerową H0 oraz hipotezę alternatywną H1:

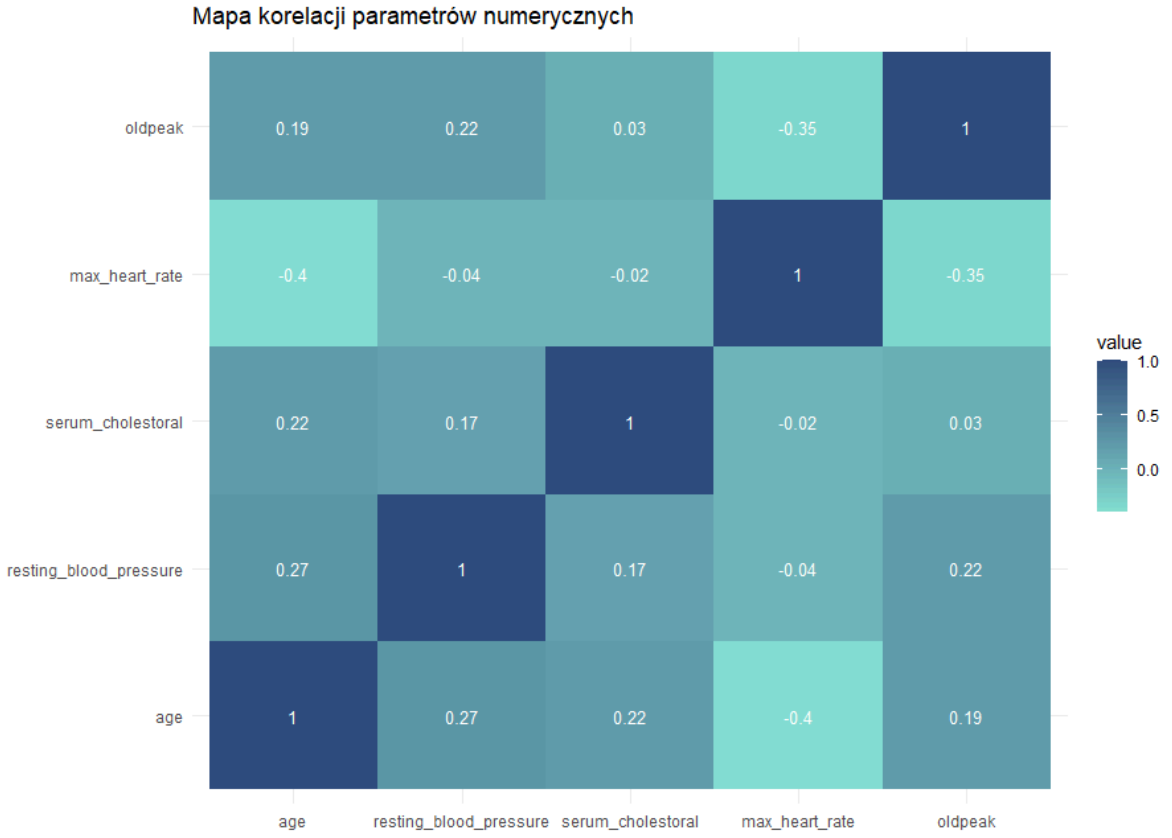
- H0 – nie ma zależności pomiędzy talasemią a występowaniem choroby serca
- H1 – jest zależność pomiędzy talasemią a występowaniem choroby serca

```
Pearson's Chi-squared test

data:  c_t
X-squared = 74.569, df = 2, p-value < 2.2e-16
```

P-value < 0.05, odrzucamy hipotezę zerową. Istnieje więc zależność między talasemią a występowaniem choroby serca. Oznacza to, że talasemia ma pewien wpływ na występowanie choroby serca.

1.3 Mapa ciepła dla parametrów numerycznych



Najsilniejsza pozytywna korelacja występuje pomiędzy wiekiem a ciśnieniem (nie jest jednak ona jakoś szczególnie duża, wynosi 0.27). Po za tym warto też zwrócić uwagę na korelację cholesterolu z wiekiem i oldpeak z ciśnieniem (są one na drugim miejscu jeżeli chodzi o najsilniejszą pozytywną korelację – wartość 0.22)
Najsilniejsza negatywna korelacja występuje pomiędzy wiekiem a tętnem oraz tętnem a oldpeak (wartości te to kolejno -0.4 i -0.35, jest to już wartość warta odnotowania, widać że te zmienne mają widoczną zależność)

1.3 Podsumowanie

Predyktory mające wpływ na występowanie choroby serca:

- age
- pressure
- serum_cholestoral
- max_heart_rate
- oldpeak
- gender
- chest_pain_type
- resting_elect
- angina
- slope
- vessel
- thalassemia

2 Budowa modeli klasyfikujących

Zbudujemy dwa modele, jeden oparty o MLP, a drugi o lasy losowe.

2.1 Metoda MLP

Do budowy modelu MLP skorzystamy z pakietu neuralnet

```
library(neuralnet)
library(plyr)
library(caret)
```

Wyodrębniamy ze zbioru zmienną celu i usuwamy ją oraz zbędne predyktory ze zbioru danych

```
y <- dane$heart_disease
dane = subset(dane, select = -c(fasting_blood_sugar))
dane = subset(dane, select = -c(heart_disease))
```

Badamy działanie modelu dla klasyfikacji binarnej (przewidujemy czy dany pacjent choruje na serce czy nie). Aby zmienne pasowały do klasyfikacji binarnej wykonujemy polecenie:

```
y_klas = revalue(as.factor(y), replace = c('1'='tak','0'='nie'))
```

Dzielimy dane na część uczącą i testową

```
podzial <- factor(sample(0:1, replace = TRUE, size = nrow(dane),
prob = c(0.7,0.3)), levels = c(0, 1))
levels(podzial) <- c("train", "test")

dane_train <- dane[podzial == "train",]
dane_test <- dane[podzial == "test",]
y_reg_train <- y_reg[podzial == "train"]
y_reg_test <- y_reg[podzial == "test"]
y_klas_train <- y_klas[podzial == "train"]
y_klas_test <- y_klas[podzial == "test"]
```

Należy teraz znormalizować zmienne numeryczne, a zmienne kategoryczne zamienić na zmienne 0-1. Napiszemy funkcję wybierające i przekształcające kolumny odpowiednich typów. Zmienne ilościowe normalizujemy do przedziału $[-1, 1]$

```
norm <- function(x){
    return(2*(x-min(x))/(max(x)-min(x))-1)
}
trans_num <- function(X){
    num_cols <- unlist(lapply(X, is.numeric))
    X <- apply(X[,num_cols],2,norm)
    return(X)
}
```

Zmienne jakościowe zamieniamy na zmienne zerojedynkowe

```
library(varhandle)
trans_kat <- function(X){
    kat_cols <- unlist(lapply(X, is.factor))
    Y <- list()
    for(i in colnames(X[,kat_cols])){
        Y <- cbind(Y, to.dummy(X[,i], prefix = i))
    }
    return(Y)
}
```

Złączamy oba przekształcone zbiory i przekształcamy w ramkę danych

```
predyktory <- function(X){
    X <- cbind(trans_num(X),trans_kat(X))
    X <- apply(X,2,as.numeric)
    return(data.frame(X))
}
predyktory(dane_train)
```

Jako hiper parametr act.fct - różniczkowalna funkcja aktywacji, użyjemy funkcji softplus

```
beta <- 10
softplus <- function(x) log(1+exp(beta*x))/beta
softplus(1)
```

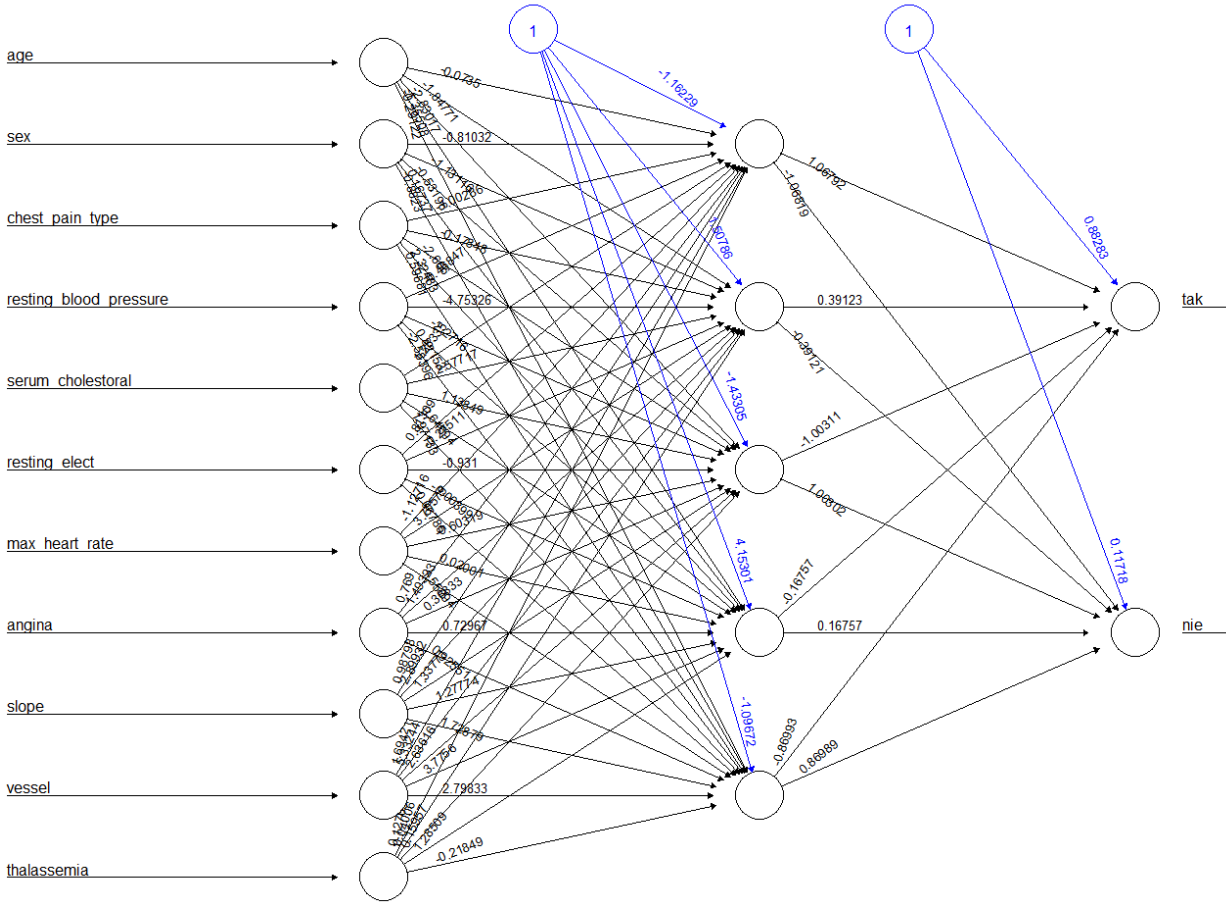
Klasyfikacja binarna – przygotujmy zbiór uczący dodając binarną zmienną celu

```
dane_klas_train <- predyktory(dane_train)
dane_klas_train$heart_disease <- y_klas_train
```

Zbudujemy sieć z pięcioma neuronami w warstwie ukrytej, z softplus jako funkcję aktywacji, maksymalną liczbą kroków uczenia sieci jako pięć do potęgi piątej

```
siec_klas <- neuralnet(heart_disease~., dane_klas_train , hidden
= 5, act.fct = softplus, stepmax = 5e5)
```

Zbudowany model wygląda następująco:



Zbadajmy trafność na zbiorze testowym. Wynikiem działania sieci (net.result) jest dwukolumnowa macierz. Napiszemy funkcję, która przypisze takiej macierzy wektor odpowiednich klas przewidywanych. Dla uproszczenia przekształćmy macierz prawdopodobieństw w ramkę danych, której nagłówkiem będą nazwy klas.

```
przewidywania <- function(prob){
    klasy <- colnames(prob)
    maksima <- apply(prob, 1, max)
    prog <- c()
    for(i in 1:nrow(prob)){
        prog <- c(prog, klasy[prob[i,] == maksima[i]])
    }
    return(factor(prog, levels = klasy))
}

dane_klas_test <- predyktory(dane_test)
prob <- predict(siec_klas, dane_klas_test)
prob <- data.frame(prob)
colnames(prob) <- levels(y_klas_train)
dane_klas_test$mlp_PV <- przewidywania(prob)
confusionMatrix(y_klas_train, dane_klas_train$mlp_PV, positive =
"tak")
confusionMatrix(y_klas_test, dane_klas_test$mlp_PV, positive =
"tak")
```

Dla zbioru testowego wynik mamy następujący:

```
Confusion Matrix and Statistics
          Reference
Prediction nie tak
       nie  32   3
       tak  16  15

               Accuracy : 0.7121
                 95% CI : (0.5875, 0.817)
    No Information Rate : 0.7273
    P-Value [Acc > NIR] : 0.666965

                  Kappa : 0.4079

 Mcnemar's Test P-Value : 0.005905

            Sensitivity : 0.8333
            Specificity : 0.6667
         Pos Pred Value : 0.4839
         Neg Pred Value : 0.9143
             Prevalence : 0.2727
         Detection Rate : 0.2273
   Detection Prevalence : 0.4697
      Balanced Accuracy : 0.7500

       'Positive' Class : tak
```

Dla zbioru uczącego wynik mamy następujący:

```
Confusion Matrix and Statistics
          Reference
Prediction nie tak
       nie 111   4
       tak  16  73

               Accuracy : 0.902
                 95% CI : (0.8527,
0.9391)
    No Information Rate : 0.6225
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7976

 Mcnemar's Test P-Value : 0.01391

            Sensitivity : 0.9481
            Specificity : 0.8740
         Pos Pred Value : 0.8202
         Neg Pred Value : 0.9652
             Prevalence : 0.3775
         Detection Rate : 0.3578
   Detection Prevalence : 0.4363
      Balanced Accuracy : 0.9110

       'Positive' Class : tak
```

Narysujmy krzywe ROC:

```
train_roc = roc(y_klas_train ~ y_klas_train_score, plot = TRUE,
print.auc = TRUE, legacy.axes = TRUE)
test_roc = roc(y_klas_test ~ y_klas_test_score, plot = TRUE,
print.auc = TRUE, legacy.axes = TRUE)
```

2.2 Lasy losowe

W analogiczny sposób przygotowujemy dane:

```
daneL <- read.csv("heart_disease.csv", header = TRUE, sep = ";")
daneL <- subset(daneL, select=-c(fasting_blood_sugar))
daneL_train <- daneL[podzial == "train",]
daneL_test <- daneL[podzial == "test",]
```

Budowa modelu

```
daneL_train$heart_disease <- factor(daneL_train$heart_disease)
daneL_test$heart_disease <- factor(daneL_test$heart_disease)
fc <- heart_disease~age + sex + chest_pain_type +
resting_blood_pressure + serum_cholestoral + resting_elect +
max_heart_rate + angina + oldpeak + slope + vessel + thalassemia
model_klas <- randomForest(fc, data = daneL_train)
model_klas
```

Nasz model wygląda następująco:

```
randomForest(formula = fc, data = daneL_train)
               Type of random forest:
classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 18.14%
Confusion matrix:
    0  1 class.error
0 100 15   0.1304348
1  22 67   0.2471910
```

Zawiera on 500 drzew i 3 wartości w każdym rozgałęzieniu. Wskaźnik błędu wynosi 18.14%

Sprawdźmy ważność zmiennych

```
model_klas$importance
```

```
                        MeanDecreaseGini
age
8.002136
sex
2.793333
chest_pain_type
14.194164
resting_blood_pressure
7.599824
serum_cholestoral
8.444385
resting_elect
1.692584
```

```
max_heart_rate
13.881928
angina
5.301689
oldpeak
8.670463
slope
5.189902
vessel
9.845913
thalassemia
13.436511
```

Widzimy że nasz model za predyktory najważniejsze w przewidywaniu uznał:

- chest_pain_type
- max_heart_rate
- thalassemia

Najmniej ważne natomiast są:

- resting_elect
- sex

Korekta poziomu pomiarów dla zmiennych jakościowych

```
daneL_test <- rbind(daneL_train[1,], daneL_test)
daneL_test <- dane_test[-1,]
```

Zastosowanie modelu

```
daneL_train$rf_PV <- predict(model_klas,  type = "class")
daneL_train$rf_prob1 <- predict(model_klas,  type = "prob")[,2]
daneL_test$rf_PV <- predict(model_klas, daneL_test, type =
"class")
daneL_test$rf_prob1 <- predict(model_klas, daneL_test, type =
"prob")[,2]
```

Ocena jakości

```
confusionMatrix(table(daneL_train$heart_disease,daneL_train$rf_PV
), positive = "1")
```

```
Confusion Matrix and Statistics

      0   1
  0 100  15
  1  22  67

               Accuracy : 0.8186
                 95% CI : (0.7588,
0.869)
    No Information Rate : 0.598
    P-Value [Acc > NIR] : 1.108e-11

                  Kappa : 0.628

 Mcnemar's Test P-Value : 0.3239

            Sensitivity : 0.8171
            Specificity : 0.8197
         Pos Pred Value : 0.7528
         Neg Pred Value : 0.8696
             Prevalence : 0.4020
         Detection Rate : 0.3284
   Detection Prevalence : 0.4363
      Balanced Accuracy : 0.8184

       'Positive' Class : 1
```

```
confusionMatrix(table(daneL_test$heart_disease,daneL_test$rf_PV),
positive = "1")
```

```
Confusion Matrix and Statistics

     0  1
  0 29  6
  1  6 25

               Accuracy : 0.8182
                 95% CI : (0.7039,
0.9024)
    No Information Rate : 0.5303
    P-Value [Acc > NIR] : 9.93e-07

                  Kappa : 0.635

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8065
            Specificity : 0.8286
         Pos Pred Value : 0.8065
         Neg Pred Value : 0.8286
             Prevalence : 0.4697
         Detection Rate : 0.3788
   Detection Prevalence : 0.4697
      Balanced Accuracy : 0.8175

       'Positive' Class : 1
```

Krzywa ROC

Dla zbioru uczącego:

```
tarin_roc = roc(dane_train$heart_disease ~ dane_train$rf_prob1,
plot = TRUE, print.auc = TRUE, legacy.axes = TRUE)
```

Dla zbioru testowego:

```
test_roc = roc(dane_test$heart_disease ~ dane_test$rf_prob1, plot
= TRUE, print.auc = TRUE, legacy.axes = TRUE)
```

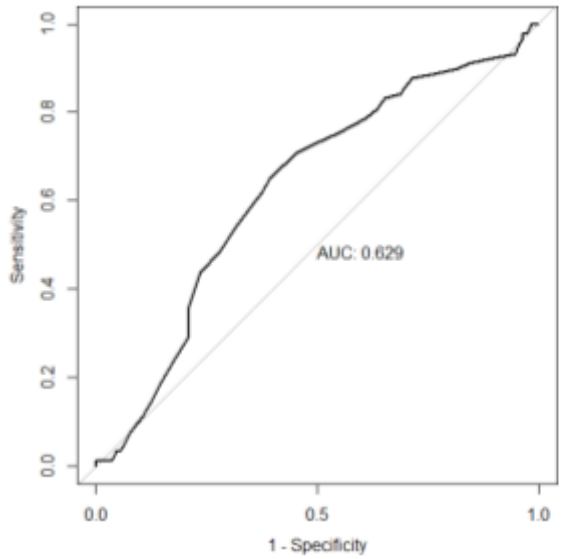
3 Porównanie modeli

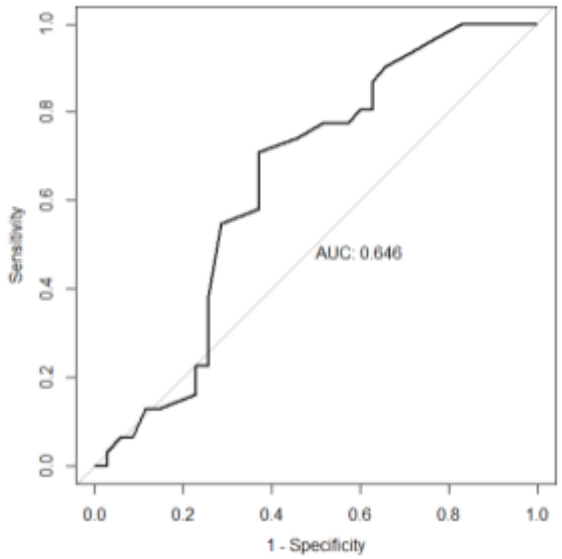
Tak wyglądają współczynniki dla modeli:

- MLP
    - Zbiór uczący
        - Trafność: 90.2%
        - Czułość: 94.81%
        - Swoistość: 87.40%
        - Precyzja: 82.02%
        - Wskaźnik F1: 87.95%
    - Zbiór testowy
        - Trafność: 71.21%
        - Czułość: 83.33%
        - Swoistość: 66.67%
        - Precyzja: 48.39%
        - Wskaźnik F1: 61.22%
- Las losowy
    - Zbiór uczący
        - Trafność: 81.86%
        - Czułość: 81.71%
        - Swoistość: 81.97%
        - Precyzja: 75.28%
        - Wskaźnik F1: 78.36%
    - Zbiór testowy
        - Trafność: 81.82%
        - Czułość: 80.65%
        - Swoistość: 82.86%
        - Precyzja: 80.65%
        - Wskaźnik F1: 80.65%

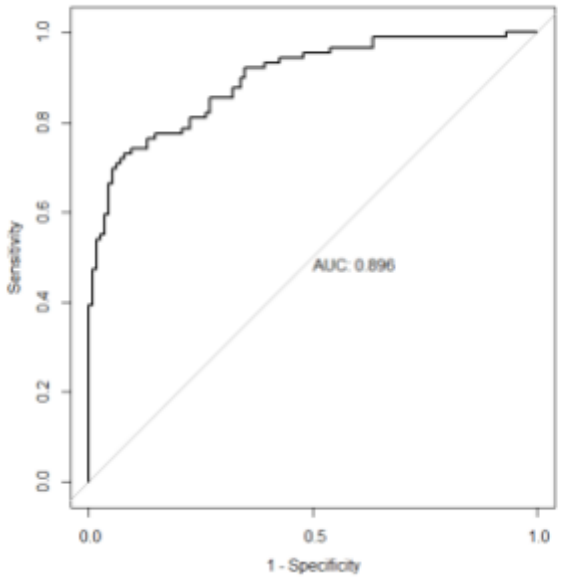
Tak wyglądają krzywe ROC:

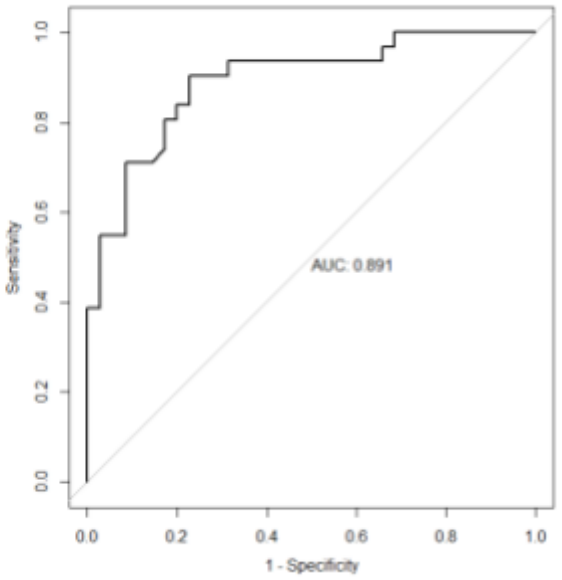
- MLP (od lewej uczący i testowy)





- Las losowy (od lewej uczący i testowy)





Model MLP radzi sobie gorzej jeśli chodzi o zbiór testowy w porównaniu do zbioru treningowego. Klasyfikuje poprawnie mniej w porównaniu do klasyfikacji danych uczących. Dla przykładu trafność dla zbioru uczącego wynosi 90.2% gdzie dla testowego jest to 71.21%. Jest to różnica około 19%. Wartości AUC dla obu zbiorów są zbliżone, testowy – 0.64, uczący – 0.62.

Model lasu losowego radzi sobie podobnie zarówno dla zbioru uczącego jak i testowego, różnice we wskaźnikach są minimalne. Wartości AUC są zbliżone: 0.896 i 0.891.

Widać że model stworzony z lasów losowych radzi sobie zdecydowanie lepiej (wykresy ROC dla lasu bardziej zbliżają się do lewej i górnej krawędzi i wartości AUC są większe co jest bliższe idealnemu klasyfikatorowi). Model MLP natomiast z wartościami AUC 0.64 i 0.62 zbliża się bardziej ku klasyfikatorowi losowemu. Nie jest to zbyt dobry model klasyfikujący.