Zidentyfikowanie zmiennych mających wpływ na wystąpienie ataku serca

Natalia Machlus, Julia Taborek

Podział pracy:

- Natalia: wykresy i opis zmiennych: target, cp, fbs, resteg, wypełnienie tabel z wynikami modeli logit0, logit1, probit0
- Julia: wykresy i opis zmiennych ca, sex, slope, exang, thal, wypełnienie tabel z wynikami modeli probit1, logit_inter, końcowy logit1

Pozostała część pracy została wykonana podczas wspólnych spotkań na uczelni oraz spotkań online na Teams.

1. Wstępny opis problemu badawczego

Choroby sercowo-naczyniowe, w tym zawał serca, są jednymi z głównych przyczyn zgonów na całym świecie. Zrozumienie czynników ryzyka, które przyczyniają się do wystąpienia zawału serca, jest kluczowe dla poprawy metod profilaktycznych, diagnozy oraz leczenia tych schorzeń. Postęp w technologii medycznej oraz analiza dużych zbiorów danych medycznych umożliwiają identyfikację i ocenę wielu zmiennych, które mogą wpływać na ryzyko wystąpienia zawału serca.

Opis problemu badawczego i cel badania

Zawał serca, czyli nagłe zatrzymanie dopływu krwi do części mięśnia sercowego, może prowadzić do poważnych uszkodzeń serca lub nawet do śmierci. Problem badawczy polega na określeniu, które zmienne mają największy wpływ na zwiększenie prawdopodobieństwa wystąpienia zawału serca. Analiza ta obejmuje zarówno czynniki demograficzne, jak i medyczne.

W tym celu przeanalizowane zostaną następujące zmienne:

a) Zmienne ilościowe:

- `age` - wiek pacjenta [w latach],
- `trestbps` - spoczynkowe ciśnienie krwi podczas przyjęcia do szpitala [mm Hg],
- `chol` - cholesterol w surowicy [mg/dl],
- `thalach` - osiągnięte maksymalne tętno [bpm],
- `oldpeak` - obniżenie odcinka ST wywołane wysiłkiem fizycznym w stosunku do odpoczynku,
- `ca` - liczba głównych naczyń zabarwionych metodą fluoroskopii [wartości od 0 do 3],

b) Zmienne kategorialne:

- `sex`- płeć pacjenta [0: Kobieta, 1: Mężczyzna]
- `cp` - typ bólu w klatce piersiowej [wartości: 0,1,2,3]

    - 0 – typowa dławica piersiowa,
    - 1 – nietypowa dławica piersiowa,
    - 2 – ból niebędący dławicą piersiową,
    - 3 – bezobjawowy,
- `fbs` - poziom cukru we krwi na czczo [1: jeśli > 120 mg/dl, 0: w przeciwnym razie],
- `restecg` - spoczynkowe wyniki elektrokardiogramu [wartości 0,1,2]
    - 0 – normalne,
    - 1 – posiadające nieprawidłowości załamka ST-T (odwrócenia załamka T i/lub uniesienie lub obniżenie odcinka ST o >0,05 mV),
    - 2 – wykazujące prawdopodobne lub pewne przerosty lewej komory według kryteriów Estesa,
- `exang`- dławica wysiłkowa [1: tak, 0: nie]
- `slope` - nachylenie szczytowego odcinka ST podczas ćwiczenia [wartości: 0,1,2]
    - 0 – wznoszące się,
    - 1 – płaskie,
    - 2 – opadające,
- `thal` - typ wady serca [wartości: 0,1,2]
    - 0 – normalny,
    - 1 – usunięta usterka,
    - 2 – wada odwracalna,
- `target` - klasa wyjściowa [wartości 0,1]
    - 0 – mniejsze prawdopodobieństwo zawału serca (< 50% zwężenia średnicy),
    - 1 – większe prawdopodobieństwo zawału serca (> 50% zwężenia średnicy)

Wyniki badania mogą przyczynić się do lepszego zrozumienia mechanizmów prowadzących do zawału serca oraz do opracowania skuteczniejszych strategii profilaktycznych i terapeutycznych.

Dzięki analizie tych zmiennych, badanie dostarczy cennych informacji na temat względnej wagi poszczególnych czynników ryzyka oraz ich wzajemnych interakcji. Ostatecznym celem jest stworzenie modelu predykcyjnego, który pomoże w identyfikacji osób o podwyższonym ryzyku zawału serca, co umożliwi wdrożenie odpowiednich interwencji medycznych w celu zmniejszenia liczby zachorowań i zgonów związanych z tym poważnym schorzeniem.

2. Przygotowanie i prezentacja zbioru danych

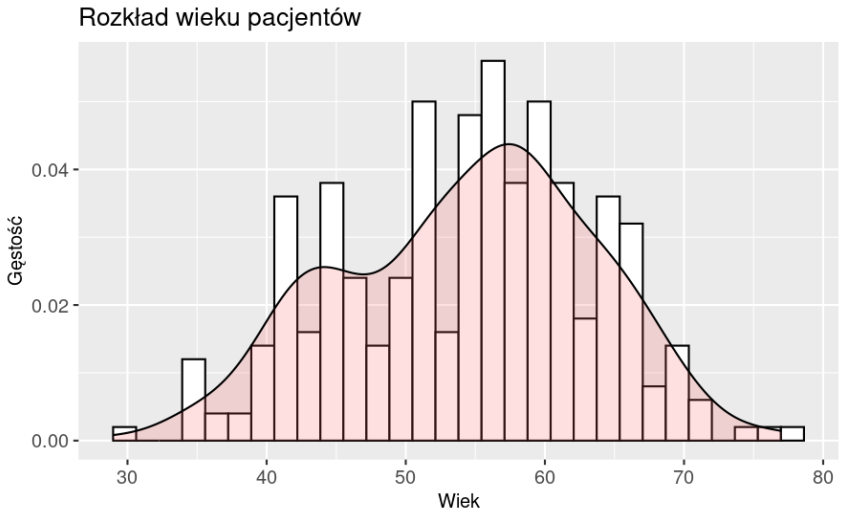
W zbiorze nie wykryto braków danych. Wykryto jeden zduplikowany wiersz. Postanowiono go wyeliminować, ponieważ występował po identycznym wierszu. Można podejrzewać, że był to błąd przy wpisywaniu danych.

2.1. Wstępna analiza zmiennych ilościowych

Zmienne kategorialne zostały zamienione na factory a następnie dla zmiennych ilościowych zostały obliczone podstawowe statystyki.

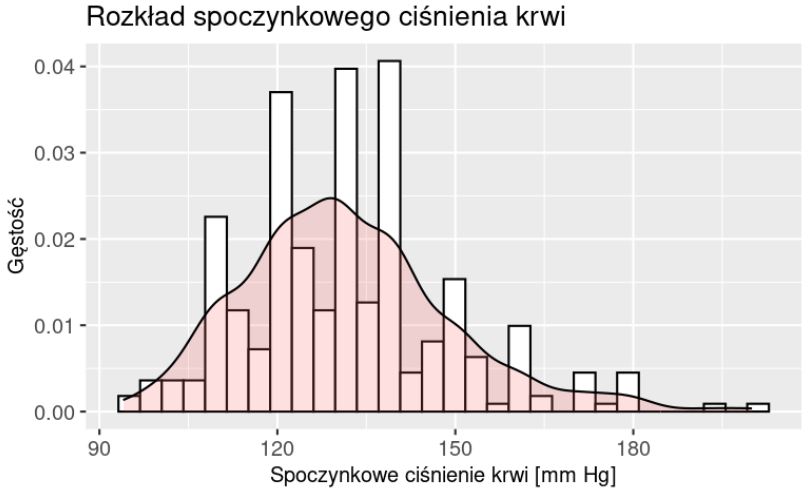
Tabela 1. Podstawowe statystyki zmiennych ilościowych. Źródło: opracowanie własne na podstawie wyników z RStudio.

Zmienne	Min	1 kwartyl	Mediana	Średnia	3 kwartyl	Max
Age	29,00	48,00	55,50	54,42	61,00	77,00
Trestbps	94,00	120,00	130,00	131,60	140,00	200,00
Chol	126,00	211,00	240,50	246,50	274,80	564,00
Thalach	71,00	133,20	152,50	149,60	166,00	202,00
Oldpeak	0,00	0,00	0,80	1,04	1,60	6,20
Ca	0,00	0,00	0,00	0,72	1,00	4,00



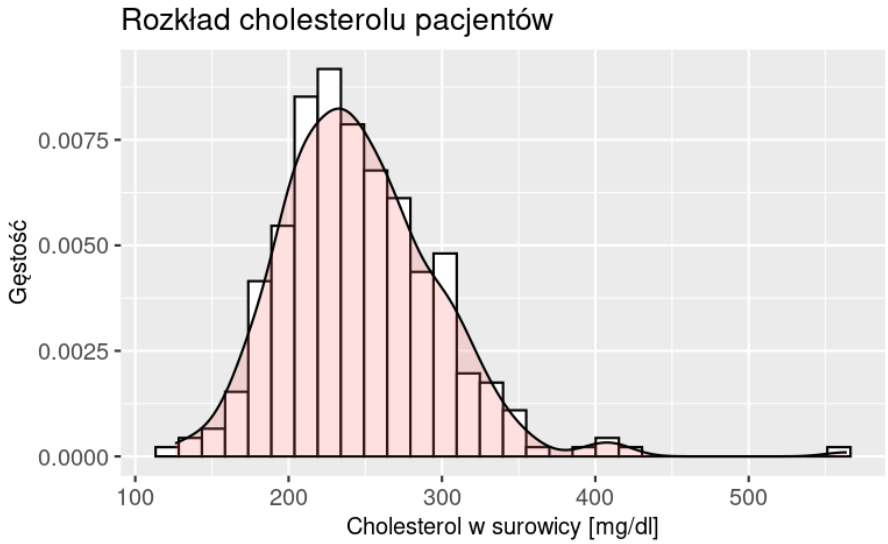
Wykres 1. Rozkład wieku pacjentów. Źródło: opracowanie własne w programie RStudio.

Na podstawie wykresu (*Wykres 1*) oraz statystyk (*Tabela 1*) można zauważyć, że wartość minimalna i maksymalna są odległe od siebie co wskazuje na duży rozrzut wieku w danych. 75% pacjentów jest w wieku powyżej 48 lat. Mediana jest zbliżona do średniej, co wskazuje na w miarę symetryczny rozkład, ale niekoniecznie normalny.



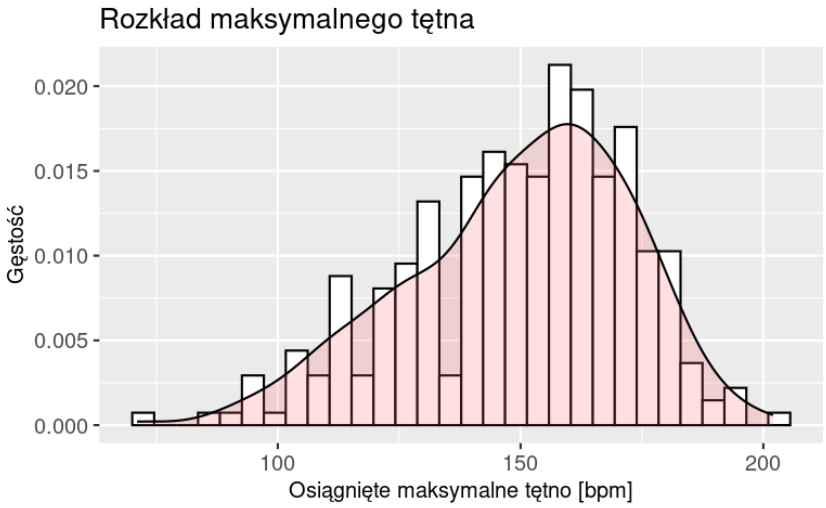
Wykres 2. Rozkład spoczynkowego ciśnienia krwi. Źródło: opracowanie własne w programie RStudio.

Pacjent z najniższym spoczynkowym ciśnieniem krwi w momencie przyjęcia do szpitala posiadał ciśnienie 94 mm Hg. Taka wartość jest na granicy niskiego ciśnienia krwi (hipotensji), co może sugerować problemy zdrowotne, ale w kontekście spoczynkowego ciśnienia krwi, nie jest to ekstremalnie niska wartość. 25% osób ma ciśnienie krwi poniżej 120 mm Hg, co jest uważane za normalne ciśnienie krwi. To sugeruje, że znaczna część populacji ma zdrowe ciśnienie krwi. 75% osób ma ciśnienie krwi poniżej 140 mm Hg. Jest to graniczna wartość dla uznania ciśnienia krwi za wysokie. Wartość powyżej 140 mm Hg jest już uważana za nadciśnienie. Najbardziej nietypowa wartość w zestawie danych to maksymalne ciśnienie krwi wynoszące 200 mm Hg. Jest ona znacznie wyższa niż 3 kwartyl (140 mm Hg) i odbiega od normy. Taka wartość jest uznawana za bardzo wysokie nadciśnienie i wymaga pilnej uwagi medycznej. Interpretacja wykresu (*Wykres 2*) również wskazuje na to, że znaczna część osób ma optymalną wartość ciśnienia. Rozkład ciśnienia jest asymetryczny i prawostronnie skośny – mamy więcej obserwacji o wartościach mniejszych niż średnia.



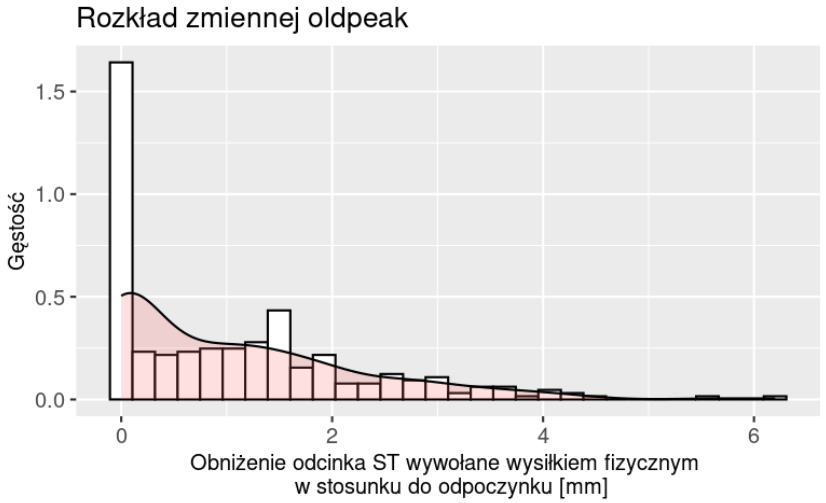
Wykres 3. Rozkład cholesterolu pacjentów. Źródło: opracowanie własne w programie RStudio.

Najniższa wartość cholesterolu w zestawie danych wynosi 126 mg/dl. Jest to wartość w granicach normy (125-200 mg/dl). 75% osób ma poziom cholesterolu powyżej 211 mg/dl – nieco ponad normę. Średnia wartość cholesterolu w surowicy u pacjentów wynosi 246,50 mg/dl. Przekracza ona normę, co może powodować zwiększone ryzyko chorób sercowo-naczyniowych. Średnia jest wyższa od mediany co sugeruje obecność wartości odstających lub prawostronną asymetrię rozkładu, co potwierdza również wykres (*Wykres 3*). Ekstremalne wartości, takie jak 564 mg/dl, są szczególnie niepokojące i wymagają natychmiastowej interwencji. Regularne monitorowanie i odpowiednie zarządzanie poziomem cholesterolu jest kluczowe dla zdrowia sercowo-naczyniowego. Możliwe jednak, że taka wartość jest błędem w danych, dlatego postawiono wykonać test Rosner'a identyfikujący wartości odstające. Test uznał wartość 564 mg/dl za nietypową. W związku z tym postanowiono wykluczyć pacjenta z tą wartością ze zbioru.



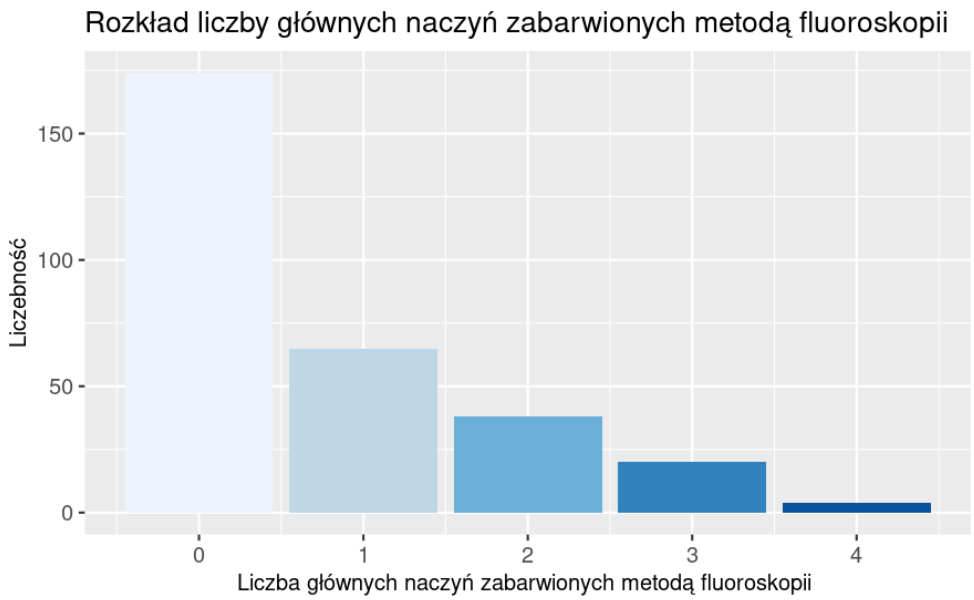
Wykres 4. Rozkład maksymalnego tętna. Źródło: opracowanie własne w programie RStudio.

Najniższa zarejestrowana wartość maksymalnego tętna wynosi 71 bpm. Jest to wartość niska, co może sugerować niższą zdolność wysiłkową lub inne czynniki wpływające na tętno. 75% pacjentów miało tętno wyższe niż 133,2 bpm. Fakt, że częściej spotykane były wyższe wartości oraz wykres (*Wykres 5*) wskazują na rozkład lewostronnie asymetryczny. Najwyższe zarejestrowane tętno wynosiło 202 uderzeń na minutę.



Wykres 5. Rozkład zmiennej oldpeak. Źródło: opracowanie własne w programie RStudio.

Obniżenie odcinka ST może wskazywać na niedokrwienie mięśnia sercowego, co jest istotne dla diagnozy chorób wieńcowych i planowania leczenia. Na wykresie (*Wykres 5*) ukazana jest wysoka przewaga wartości 0. U 33% pacjentów nie wystąpiło obniżenie odcinka ST w trakcie lub po wysiłku fizycznym, co jest normą lub wskazuje na niewielkie ryzyko choroby niedokrwiennej serca. 75% pacjentów ma wartość oldpeak równą lub mniejszą niż 1,60 mm. To wskazuje, że większość populacji ma umiarkowane obniżenia odcinka ST w odpowiedzi na wysiłek fizyczny. Wartość maksymalna zmiennej oldpeak wynosząca 6,20 mm jest dość wysoka i może sugerować znaczące obniżenie odcinka ST podczas wysiłku fizycznego.



Wykres 6. Rozkład liczby głównych naczyń zabarwionych metodą fluoroskopii. Źródło: opracowanie własne w programie RStudio.

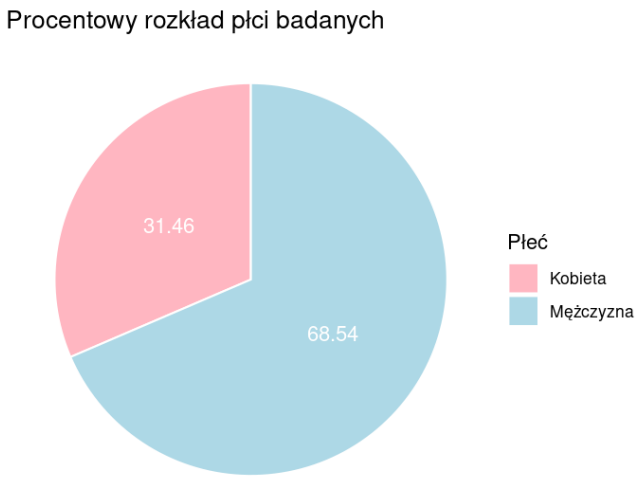
W danych wykryto jedną dodatkową kategorię, która nie była opisana w zbiorze. W ludzkim sercu są trzy główne naczynia wieńcowe, które dostarczają krew do mięśnia sercowego. Odgrywają one kluczową rolę w dostarczaniu natlenowanej krwi do różnych części mięśnia sercowego. Z kolei, liczba głównych naczyń zabarwionych metodą fluoroskopii, jaką podaje się w niektórych zbiorach danych medycznych, czasem uwzględnia dodatkowe odgałęzienia lub istotne gałęzie, dlatego mogą się pojawić wartości od 0 do 4. Dodatkowa wartość może odnosić się do istotnych anatomicznych wariantów lub dodatkowych, istotnych odgałęzień, które są uwzględniane w badaniach obrazowych.

Najwięcej pacjentów (58%) nie miało zabarwionych naczyń metodą fluoroskopii. Wraz ze wzrostem liczby naczyń zabarwionych tą metodą malała liczba pacjentów (*Wykres 6*).

2.2. Wstępna analiza zmiennych kategorialnych

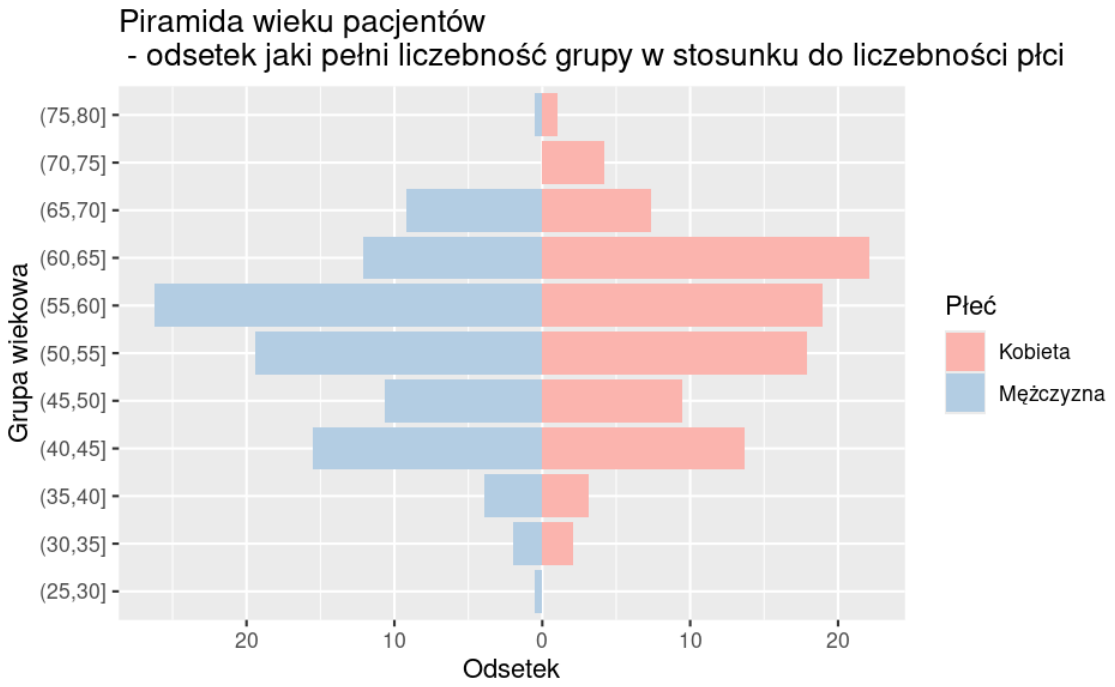
Wykres 7. Rozkład zmiennej objaśnianej. Źródło: opracowanie własne w programie RStudio.

Rozkład zmiennej objaśnianej (*Wykres 6*) wskazuje na to, że w badanej grupie pacjentów istnieje więcej przypadków osób z większym prawdopodobieństwem zawału serca niż tych z mniejszym, choć ta różnica nie jest znacząca. Proporcja przypadków z mniejszym prawdopodobieństwem zawału serca wynosi 45,8%.



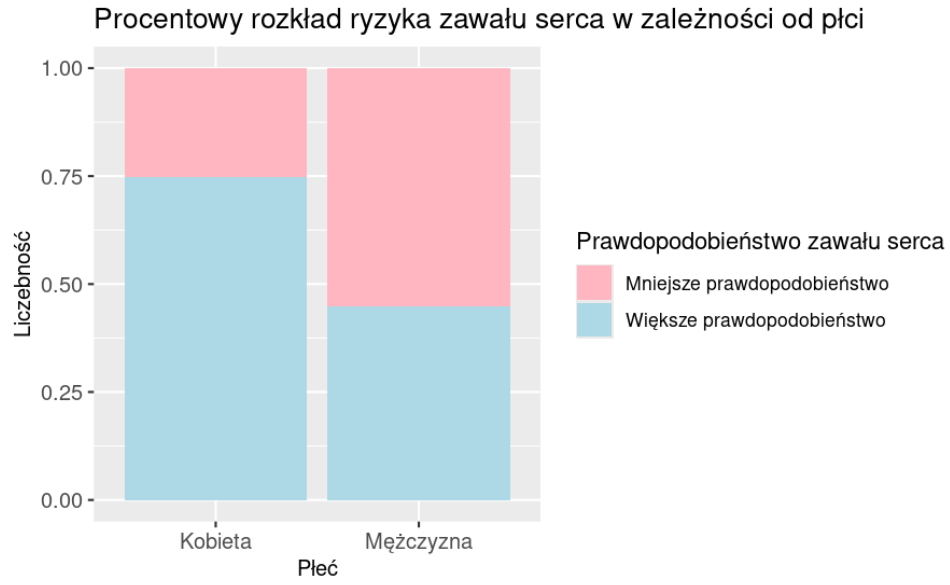
Wykres 8. Rozkład procentowy płci pacjentów. Źródło: opracowanie własne w programie RStudio.

Wśród pacjentów przeważali mężczyźni (*Wykres 8*). Stanowili oni 2 razy większy odsetek wszystkich pacjentów niż kobiety.



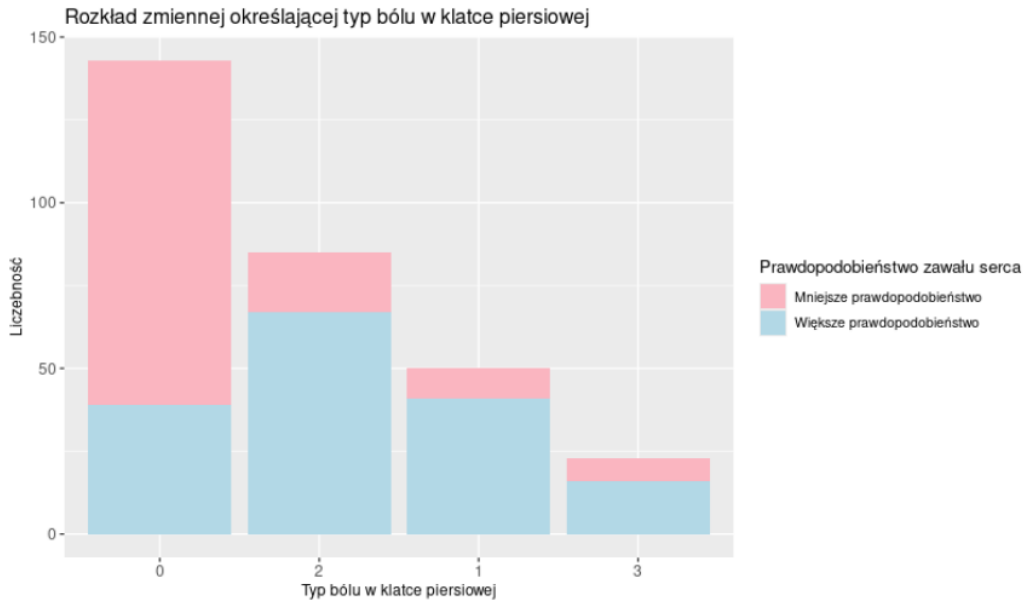
Wykres 9. Piramida wieku pacjentów. Źródło: opracowanie własne w programie RStudio.

Na wykresie (*Wykres 9*) można zauważyć, że nie było żadnej kobiety w wieku od 25 do 30 lat oraz żadnego mężczyzny w wieku 70-75lat. Największy odsetek kobiet (22%) był w wieku od 60 do 65 lat, natomiast ponad 25% wszystkich mężczyzn należało do grupy wiekowej (55,60]. Dodatkowo w przypadku pacjentek 71,6% z nich było w wieku powyżej 50 lat a w przypadku pacjentów odsetek ten wynosił 67%.



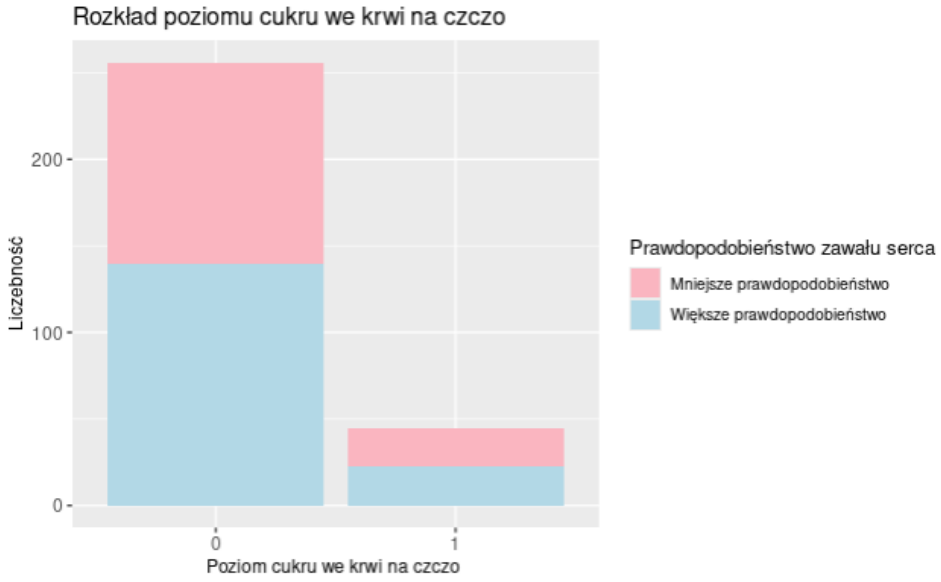
Wykres 10. Procentowy rozkład zawału serca w zależności od płci. Źródło: opracowanie własne w programie RStudio.

Na wykresie (*Wykres 10*) można zauważyć, że większość kobiet (74,74%) znajduje się w zwiększonej grupie ryzyka zawału serca natomiast ponad połowa mężczyzn jest w grupie z mniejszym prawdopodobieństwem.



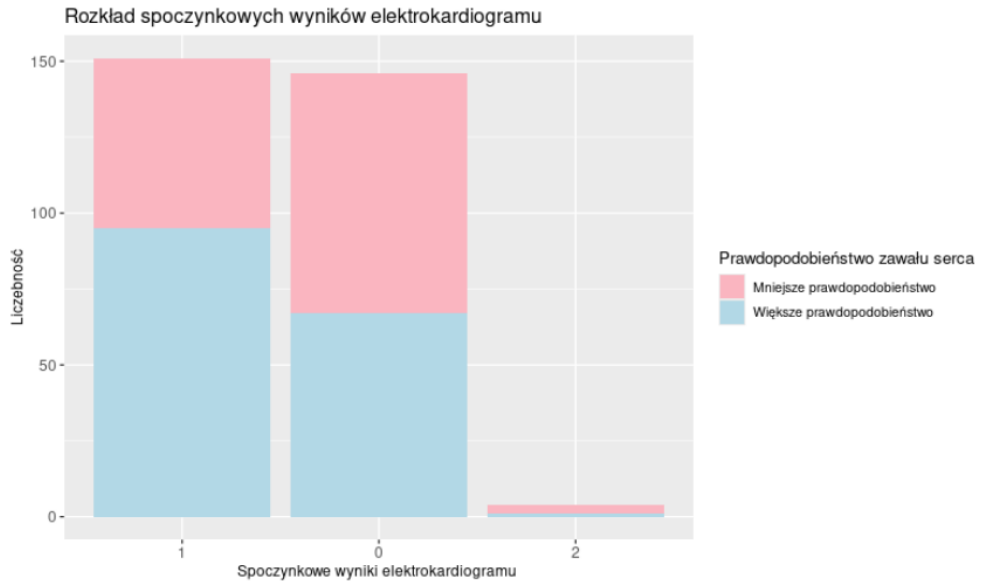
Wykres 11. Rozkład zmiennej określającej typ bólu w klatce piersiowej. Źródło: opracowanie własne w programie RStudio.

Typowa dławica piersiowa (0) jest najczęściej występującym typem bólu w klatce piersiowej o reprezentacji 47,5% w zbiorze (*Wykres 11*). Ból niebędący dławicą piersiową (2) jest również dość powszechny z reprezentacją 28,2%. Nietypowa dławica piersiowa (1) i bezobjawowy typ bólu w klatce piersiowej (3) są mniej powszechne w badanej populacji. Większość pacjentów z typową dławicą piersiową (0) ma niższe prawdopodobieństwo wystąpienia zawału serca. Natomiast pacjenci z bólem klatki piersiowej niebędącym dławicą piersiową (2), pacjenci z nietypową dławica piersiowa (1) oraz pacjenci z bezobjawowym bólem klatki piersiowej (3) mają wyższe prawdopodobieństwo wystąpienia zawału serca.



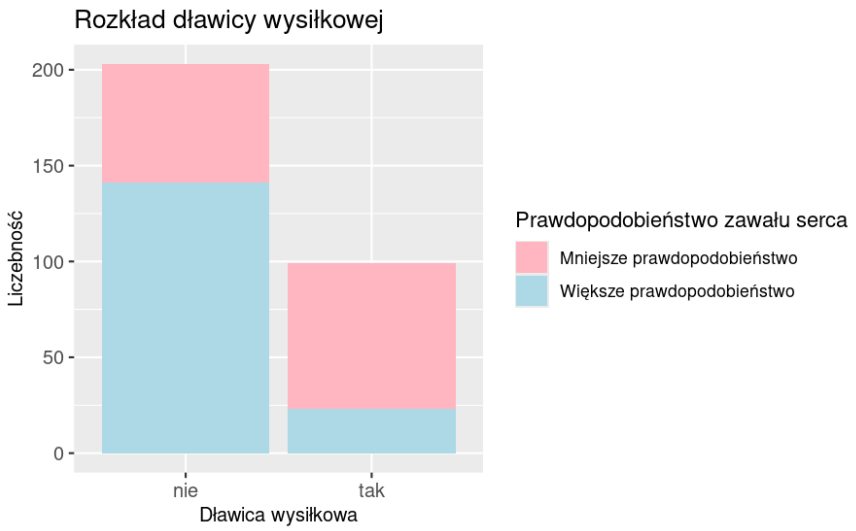
Wykres 12. Rozkład poziomu cukru we krwi na czczo. Źródło: opracowanie własne w programie RStudio.

Rozkład zmiennej (*Wykres 12*) wskazuje, że większość badanych pacjentów (85%) ma poziom cukru we krwi na czczo w granicach normy (≤ 120 mg/dl), natomiast 14.9% ma podwyższony poziom cukru we krwi na czczo (> 120 mg/dl). Wysoki poziom cukru we krwi na czczo (> 120 mg/dl) może być objawem cukrzycy lub stanu przedcukrzycowego. Więcej pacjentów z poziomem cukru we krwi na czczo ≤ 120 mg/dl ma wyższe prawdopodobieństwo wystąpienia zawału serca. Natomiast u pacjentów o poziomie cukru we krwi na czczo > 120 mg/dl rozkład zmiennej objaśnianej jest dość równomierny.



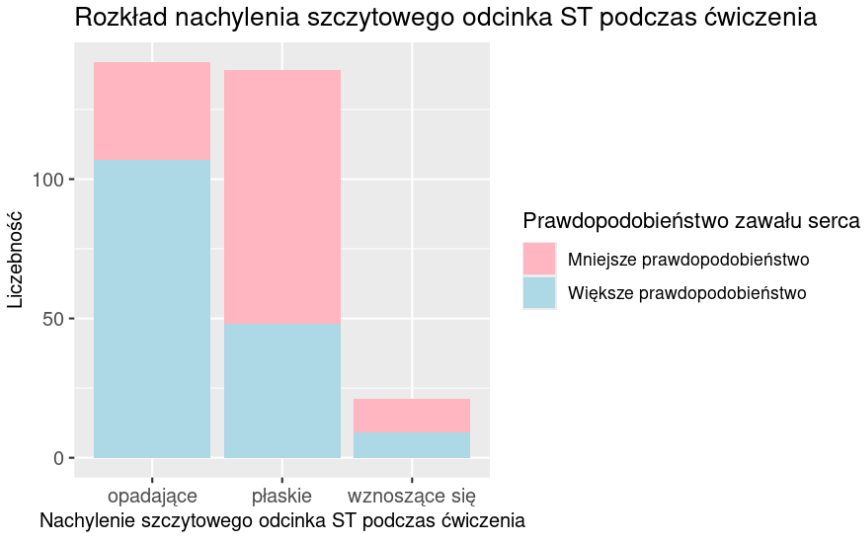
Wykres 13. Rozkład spoczynkowych wyników elektrokardiogramu. Źródło: opracowanie własne w programie RStudio.

Z wykresu (*Wykres 13*) można zauważyć, że ponad połowa badanych (50.2%) ma nieprawidłowości załamka ST-T (1), co może wskazywać na potencjalne problemy z sercem, takie jak niedokrwienie. Znacząca liczba pacjentów (48.5%) ma normalne wyniki elektrokardiogramu (0), co jest wskaźnikiem prawidłowej funkcji serca w spoczynku. Natomiast wyniki elektrokardiogramu wykazujące prawdopodobne lub pewne przerosty lewej komory według kryteriów Estesa (2) są rzadkością. Większość pacjentów z nieprawidłowości załamka ST-T ma wyższe prawdopodobieństwo wystąpienia zawału serca. Pacjenci o normalnych wynikach elektrokardiogramu oraz pacjenci z wynikami wykazującymi prawdopodobne lub pewne przerosty lewej komory w większości mają niższe prawdopodobieństwo wystąpienia zawału serca.



Wykres 14. Rozkład dławicy wysiłkowej. Źródło: opracowanie własne w programie RStudio.

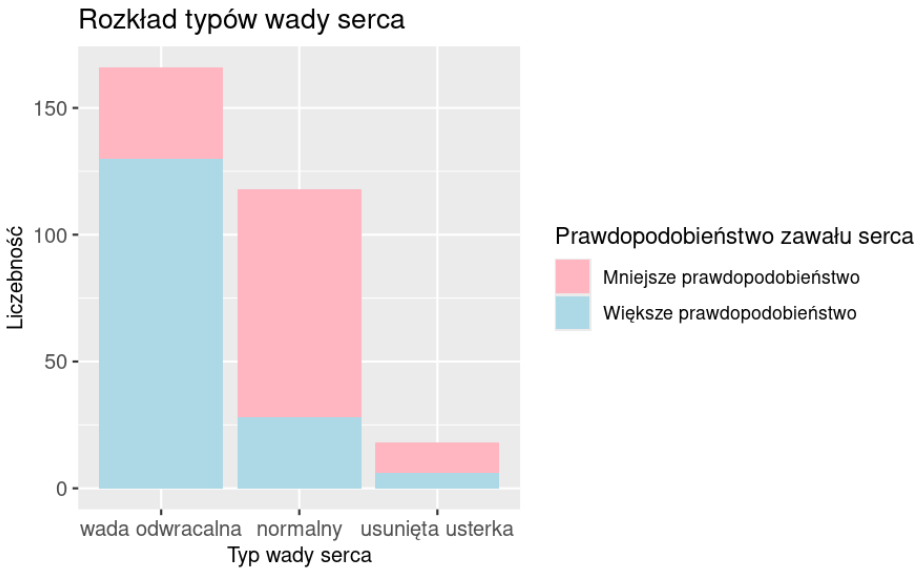
Dławica wysiłkowa wskazuje na to, że tętnice wieńcowe są zwężone lub zablokowane przez miażdżycę, co utrudnia dostarczanie wystarczającej ilości tlenu do serca podczas zwiększonego zapotrzebowania na tlen, jak ma to miejsce podczas wysiłku fizycznego. Z wykresu (*Wykres 14)* można zauważyć, że liczba pacjentów bez dławicy wysiłkowej jest 2 razy większa od liczby pacjentów z dławicą wysiłkową. Wśród grupy reprezentowanej przez większość pacjentów blisko 70% osób ma zwiększone ryzyko zawału serca. Natomiast większość osób z dławicą wysiłkową (76 osób) ma niższe ryzyko zawału serca. Oznacza to, że obecność dławicy wysiłkowej może być związana z niższym ryzykiem zawału serca w tej grupie pacjentów, co może być wynikiem odpowiedniego leczenia i monitorowania stanu zdrowia przedstawicieli tej grupy.



Wykres 15. Rozkład nachylenia szczytowego odcinka ST podczas ćwiczenia. Źródło: opracowanie własne w programie RStudio.

Z wykresu (*Wykres 15*) można zauważyć, że większość pacjentów ma opadające bądź płaskie nachylenie szczytowego odcinka ST podczas ćwiczeń. Nachylenie wznoszące jest rzadkie i posiada je niecałe 25 osób. 'Płaskie' nachylenie jest związane z większą liczbą osób o niższym

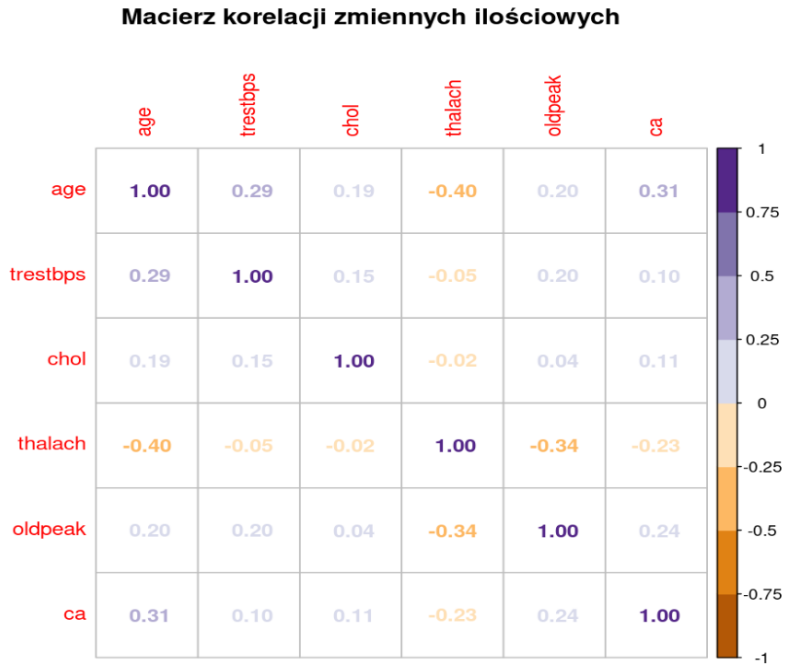
ryzyku zawału serca, co sugeruje, że ta grupa ma relatywnie mniejsze zagrożenie. 'Opadające' nachylenie jest silnie związane z dużą przewagą osób o większym prawdopodobieństwie zawału serca, co może wskazywać na wyższe ryzyko w tej grupie. 'Wznoszące się' nachylenie ma zbliżoną liczbę osób o niższym i wyższym ryzyku zawału serca, co może sugerować brak wyraźnego związku między tym typem nachylenia a ryzykiem zawału serca.



Wykres 16. Rozkład typów wady serca. Źródło: opracowanie własne w programie RStudio.

W przypadku zmiennej ‘thal’ wykryto kategorię „3”, która nie uwzględniona w opisie. Jednak na stronie, z której pochodzą dane (https://archive.ics.uci.edu/dataset/45/heart+disease) widniała informacja, że oznacza ona normalną wadę serca. W związku z tym postanowiono przekodować ją na kategorię „0”, ponieważ ma ona takie samo znaczenie. Po przekodowaniu najliczniejszą grupą są pacjenci z wadą odwracalną (*Wykres 16*). Wśród nich większość ma zwiększone prawdopodobieństwo zawału serca. Natomiast w przypadku osób z normalną jak i usuniętą usterką większy odsetek stanowią osoby o mniejszym ryzyku zawału serca (kolejno 76% oraz 67%).

2.3. Sprawdzenie korelacji parami zmiennych objaśniających



Wykres 17. Wykres macierzy korelacji zmiennych ilościowych. Źródło: opracowanie własne w programie RStudio.

Zaleca się unikania w modelu zmiennych nadmiernie skorelowanych, tzn. w przypadku współczynnika korelacji Pearsona |r|>=0.7. W analizowanym zbiorze nie ma zmiennych nadmiernie skorelowanych (*Wykres 17*).

W przypadku zmiennej ciągłej i zmiennej dychotomicznej można zastosować test jednorodności rozkładu zmiennej ciągłej w dwóch grupach, np. test Kołmogorowa-Smirnowa.

Tabela 2. Test Kołmogorowa-Smirnowa. Źródło: opracowanie własne na podstawie obliczeń w programie RStudio.

Zmienna ilościowa	Zmienna dychotomiczna	p-value
Thalach	Exang	<0,05
Oldpeak	Exang	<0,05

W testach zostały porównane rozkłady zmiennych ‘thalach’ (osiągnięte maksymalne tętno [bpm]) oraz `oldpeak` (obniżenie odcinka ST wywołane wysiłkiem fizycznym w stosunku do odpoczynku) w 2 grupach w zależności od występowania dławicy wysiłkowej. Rozkłady te statystycznie istotnie różnią się w tych grupach na co wskazuje p-value (*Tabela 2*). Zmienne zostały wybrane do testu ze względu na ich powiązanie z wysiłkiem fizycznym.

Dla par zmiennych jakościowych stosuje się współczynniki kontyngencji, np. V Cramera - silna korelacja, gdy V > 0.5.

Tabela 3. Współczynniki V-Cramera. Źródło: opracowanie własne na podstawie obliczeń w RStudio.

	sex	cp	fbs	restecg	exang	slope	thal	target
sex	1	0,38	0,42	0,38	0,69	0,36	0,51	0,57
cp	0,38	1	0,15	0,37	0,73	0,04	0,11	0,62
fbs	0,42	0,15	1	0,4	0,67	0,12	0,13	0,58
restecg	0,38	0,37	0,4	1	0,69	0,45	0,5	0,58
exang	0,69	0,73	0,67	0,69	1	0,65	0,74	0,71
slope	0,36	0,04	0,12	0,45	0,65	1	0,09	0,54
thal	0,51	0,11	0,13	0,5	0,74	0,09	1	0,62
target	0,57	0,62	0,58	0,58	0,71	0,54	0,62	1

W tabeli powyżej (*Tabela 3*) zostały wyliczone parami współczynniki V-Cramera dla zmiennych. Kolorem czerwonym zostały oznaczone silne korelacje. Zmienna exang (dławica wysiłkowa) jest silnie skorelowana ze wszystkimi pozostałymi zmiennymi, dlatego postanowiono ją wyeliminować ze zbioru zmiennych objaśniających. Dodatkowo zmienna thal (typ wady serca) jest silnie skorelowana ze zmienną sex (płeć), jednak korelacja ta nieznacznie przekracza próg 0,5, więc postanowiono pozostawić obydwie zmienne w zbiorze.

2.4. Podział zbioru danych na uczący i testowy.

Następnie podzielono zbiór danych na zbiór uczący i testowy. Dokonano losowego podziału w proporcji: 70% i 30% odpowiednio.

Tabela 4. Proporcje zmiennej target w zależności od zbioru. Źródło: opracowanie własne na podstawie obliczeń w programie RStudio.

	0	1
Cały zbiór	45,85	54,15
Zbiór uczący	45,02	54,98
Zbiór testowy	47,78	52,22

Proporcje w zbiorach uczącym i testowym są zbliżone do proporcji w całym zbiorze danych (*Tabela 4*).

3. Budowanie modelu

3.1. Model logitowy

Na początku postanowiono zbudować model logitowy ze wszystkimi zmiennymi objaśniającymi (logit0). Wykonano test ilorazu wiarygodności istotności wszystkich zmiennych oraz globalny test Walda. W obydwu przypadkach odrzucono hipotezę zerową na rzecz hipotezy alternatywnej – w modelu istnieją zmienne, które mają istotny wpływ na kształtowanie się prawdopodobieństwa zawału serca. Okazało się jednak, że wiele zmiennych

(*Tabela 5*) nie zostało uznanych za statystycznie istotne. Sprawdzono również założenie o braku współliniowości zmiennych. Jedynie w przypadku zmiennej „slope" wartość miary VIF przekraczała 2,5 – granicę, przy której powinno odrzucać się zmienne w modelu logitowym.

Tabela 5. Model logit0. Źródło: opracowanie własne na podstawie programu RStudio.

Zmienna	Ocena parametru	Błąd standardowy	Statystyka Walda	p-value
wyraz wolny	3,528	3,391	1,040	0,298
age	-0,021	0,030	-0,703	0,482
sex1	-2,105	0,654	-3,220	0,001
cp1	1,739	0,699	2,488	0,013
cp2	2,081	0,592	3,517	0,000
cp3	2,340	0,855	2,738	0,006
trestbps	-0,017	0,015	-1,161	0,246
chol	-0,008	0,005	-1,514	0,130
fbs1	-0,720	0,726	-0,992	0,321
restecg1	0,570	0,486	1,172	0,241
restecg2	-0,406	2,113	-0,192	0,848
thalach	0,031	0,014	2,282	0,023
oldpeak	-0,783	0,306	-2,556	0,011
slope1	-1,700	1,236	-1,375	0,169
slope2	-1,663	1,427	-1,165	0,244
ca	-0,660	0,245	-2,697	0,007
thal1	-0,137	1,102	-0,124	0,901
thal2	1,062	0,520	2,045	0,041

Następnie na podstawie metody krokowej przy minimalizacji kryterium informacyjnego AIC wyestymowano model logit1. Modele uzyskane metodami „both" oraz „backward" były identyczne.

Tabela 6. Model logit1. Źródło: opracowanie własne na podstawie programu RStduio.

Zmienna	Ocena parametru	Błąd standardowy	Statystyka Walda	p-value
wyraz wolny	-0,529	2,042	-0,259	0,796
sex1	-1,865	0,573	-3,255	0,001
cp1	1,754	0,681	2,574	0,010
cp2	1,765	0,532	3,320	0,001
cp3	1,850	0,808	2,291	0,022
chol	-0,011	0,005	-2,247	0,025
thalach	0,030	0,012	2,479	0,013
oldpeak	-0,684	0,244	-2,804	0,005
ca	-0,704	0,229	-3,074	0,002
thal1	-0,061	0,997	-0,061	0,951
thal2	1,214	0,472	2,574	0,010

Ponownie test ilorazu wiarygodności wykazał, że w modelu istnieją zmienne istotne statystycznie. Testy lokalne Walda (*Tabela 6*) wykazały, że wszystkie parametry z wyjątkiem wariantu thal1 są istotne. Wartości miary VIF współliniowości zmiennych nie przekraczały już granicy 2,5 zatem zmienne nie są współliniowe.

3.2. Model probitowy

Następnie wyestymowano model dwumianowy probitowy ze wszystkimi zmiennymi objaśniającymi (probit0).

Tabela 7. Model probit0. Źródło: opracowanie własne na podstawie programu RStudio.

Zmienna	Ocena parametru	Błąd standardowy	Statystyka Walda	p-value
wyraz wolny	2,095	1,884	1,112	0,266
age	-0,012	0,017	-0,713	0,476
sex1	-1,239	0,360	-3,446	0,001
cp1	1,046	0,388	2,693	0,007
cp2	1,169	0,325	3,601	0,000
cp3	1,345	0,486	2,766	0,006
trestbps	-0,009	0,008	-1,099	0,272
chol	-0,005	0,003	-1,604	0,109
fbs1	-0,429	0,403	-1,063	0,288
restecg1	0,304	0,270	1,125	0,261
restecg2	-0,311	1,140	-0,273	0,785
thalach	0,017	0,007	2,270	0,023
oldpeak	-0,445	0,168	-2,649	0,008
slope1	-0,927	0,653	-1,421	0,155
slope2	-0,914	0,750	-1,219	0,223
ca	-0,368	0,135	-2,729	0,006
thal1	-0,123	0,620	-0,199	0,842
thal2	0,596	0,294	2,024	0,043

Model (*Tabela 7*) posiadał wiele zmiennych niemających istotnego wpływu na zmienną objaśnianą, dlatego zastosowano metodę krokową. Ponownie modele uzyskane metodami „both" oraz „backward" były identyczne.

Tabela 8. Model probit1. Źródło: opracowanie własne na podstawie programu RStudio.

Zmienna	Ocena parametru	Błąd standardowy	Statystyka Walda	p-value
wyraz wolny	-0,193	1,143	-0,169	0,866
sex1	-1,090	0,317	-3,437	0,001
cp1	1,035	0,372	2,786	0,005
cp2	1,039	0,297	3,503	0,000
cp3	1,112	0,462	2,406	0,016
chol	-0,006	0,003	-2,309	0,021
thalach	0,016	0,007	2,505	0,012

oldpeak	-0,392	0,135	-2,896	0,004
ca	-0,397	0,127	-3,131	0,002
thal1	-0,108	0,564	-0,191	0,848
thal2	0,676	0,270	2,502	0,012

Otrzymany model probit1 (*Tabela 8*) jest odpowiednikiem modelu logit1 – posiada te same zmienne objaśniające i jedynie wariant thal1 nie jest istotny statystycznie przy modelowaniu prawdopodobieństwa zawału serca.

3.3. Model dwumianowy logitowy z interakcją

Jako kolejny wyestymowano model logitowy uwzględniający interakcję (logit_iter0) dla tych samych zmiennych objaśniających co w przypadku modelu logit1. Model ten miał wszystkie parametry istotne jednak bardzo wysoką wartość kryterium informacyjnego AIC. W kolejnym modelu (logit_inter1) postanowiono wyeliminować zmienną „chol", ponieważ miała ona najwyższą wartość p-value (z wyjątkiem wariantu thal1) w modelu logit1. Na podstawie wartości p-value wybrano istotne efekty interakcji i uwzględnione je w ostatecznym modelu z interakcjami (logit_inter2).

Tabela 9. Model logit_inter2. Źródło: opracowanie własne na podstawie programu RStudio.

Zmienna	Ocena parametru	Błąd standardowy	Statystyka Walda	p-value
wyraz wolny	-2,011	1,886	-1,067	0,286
sex1	-1,605	0,587	-2,735	0,006
cp1	0,729	0,771	0,945	0,344
cp2	0,565	0,624	0,906	0,365
cp3	3,303	1,668	1,981	0,048
thalach	0,027	0,013	2,131	0,033
oldpeak	-1,395	0,508	-2,748	0,006
ca	-1,968	0,605	-3,255	0,001
thal1	0,736	1,809	0,407	0,684
thal2	0,441	0,690	0,639	0,523
cp1:ca	1,470	0,801	1,835	0,067
cp2:ca	2,440	0,792	3,081	0,002
cp3:ca	-0,898	1,281	-0,701	0,483
oldpeak:thal1	-2,672	4,302	-0,621	0,535
oldpeak:thal2	1,231	0,603	2,041	0,041

Wpływ zmiennej „thal" okazał się nieistotny statystycznie (*Tabela 9*), jednak interakcja wariantu thal2 ze zmienną oldpeak była istotna statystycznie, dlatego postanowiono pozostawić tą zmienną w modelu.

4. Porównanie i ocena modeli

Tabela 10. Miary dobroci dopasowania modeli. Źródło: opracowanie własne na podstawie obliczeń w programie RStudio.

Model	kryterium_AIC	McFadden	Cragg_Uhler
Logit1	161,92	0,52	0,68
Probit1	161,02	0,52	0,68
Logit_inter2	155,26	0,57	0,73

W tabeli *(Tabela 10)* zostały porównane modele pod względem dobroci dopasowania. Miary te bazują na największej wiarygodności. Pożądane są jak najmniejsze wartości kryterium AIC i wartości miar pseudo-R2 bliskie 1. Najmniejszą wartość kryterium AIC jak i pseudo-R2 posiada model z interakcją. Wyniki dla modelu logit1 oraz probit1 są zbliżone.

Następnie zostały stworzone tablice trafności predykcji modeli oraz na ich podstawie obliczone miary jakości predykcji (*Tabela 11*).

Tabela 11. Miary jakości predykcji modeli dla zbioru uczącego. Źródło: opracowanie własne na podstawie obliczeń w programie RStudio.

Model	ACC	ER	SENS	SPEC	PPV	NPV
Logit1	0,86	0,14	0,84	0,87	0,89	0,82
Probit1	0,82	0,18	0,78	0,88	0,89	0,76
Logit_inter2	0,87	0,13	0,88	0,86	0,89	0,85

Miary oparte na tablicy trafności:

- Zliczeniowy R2 (Accurancy) = udział liczby trafnie sklasyfikowanych jednostek w ogólnej liczbie jednostek
- Wskaźnik błędu (ER – Error Rate) = udział liczby źle sklasyfikowanych jednostek w ogólnej liczbie jednostek
- Czułość (Sensitivity) = udział liczby trafnie oszacowanych 1 w liczbie wszystkich obserwowanych 1
- Swoistość (Specificity) = udział liczby trafnie oszacowanych 0 w liczbie wszystkich obserwowanych 0
- Dodatnia zdolność predykcyjna (PPV – Positive Predictive Value) = udział liczby trafnie oszacowanych 1 w liczbie wszystkich prognozowanych 1
- Ujemna zdolność predykcyjna (NPV – Negative Predictive Value) = udział liczby trafnie oszacowanych 0 w liczbie wszystkich prognozowanych 0

Kolorem zielonym zostały oznaczone najlepsze wartości każdej miary. Ponownie najlepiej wypadł model z interakcjami. Modele nie różniły się jedynie ze względu na miarę PPV.

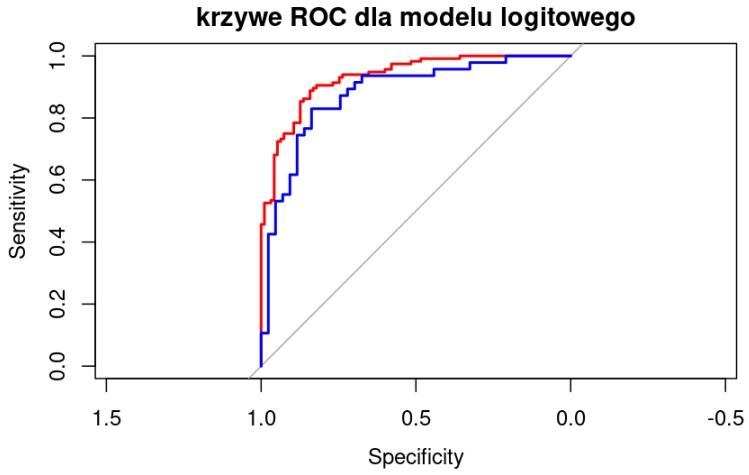
Następnie zostały porównane miary jakości predykcji na zbiorze testowym (*Tabela 12*).

Tabela 12. Miary jakości predykcji modeli dla zbioru testowego. Źródło: opracowanie własne na podstawie obliczeń w programie RStudio.

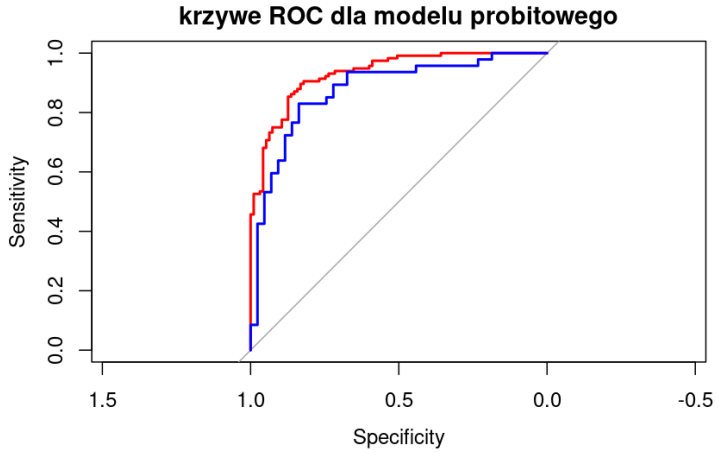
Model	ACC	ER	SENS	SPEC	PPV	NPV
Logit1	0,83	0,17	0,83	0,84	0,85	0,82
Probit1	0,80	0,20	0,74	0,86	0,85	0,76
Logit_inter2	0,77	0,23	0,83	0,70	0,75	0,79

Na zbiorze testowym najlepiej wypada model logitowy.

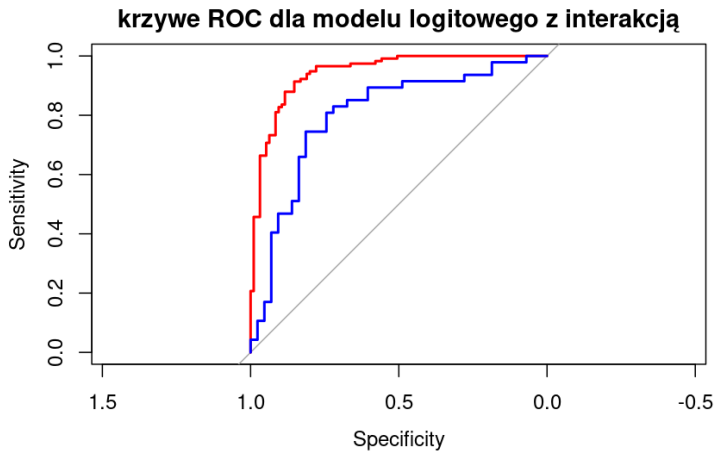
Wykresy krzywej ROC dla poszczególnych modeli



Wykres 18. Krzywa ROC dla modelu logitowego. Źródło: opracowanie własne w programie RStudio.



Wykres 19. Krzywa ROC dla modelu probitowego. Źródło: opracowanie własne w programie RStudio.



Wykres 20. Krzywa ROC dla modelu logitowego z interakcją. Źródło: opracowanie własne w programie RStudio.

Wykresy krzywej ROC dla modeli logitowego i probitowego (*Wykres 18,Wykres 19*) są zbliżone natomiast krzywe ROC dla modelu z interakcją (*Wykres 20*) różni się od pozostałych. Następnie zostały obliczone miary trafności oparte na krzywej – AUC (pole pod krzywą).

Tabela 13. Miary AUC. Źródło: opracowanie własne na podstawie obliczeń w programie RStudio.

	AUC dla zbioru uczącego	AUC dla zbioru testowego
Logit1	0,93	0,88
Probit1	0,93	0,88
Logit_inter2	0,94	0,80

Pola pod krzywą dla zbioru uczącego (*Tabela 13*) są zbliżone jednak w przypadku zbioru testowego model z interakcją wypada najgorzej.

5. Interpretacja modelu

Zdecydowano się interpretować model logitowy (logit1), ponieważ ma najlepszą zdolność predykcyjną na zbiorze testowym. Model probitowy ma zbliżone wartości kryterium informacyjnego, miar pseudo-R2 oraz AUC, jednak ma on gorszą interpretację. Model z interakcją najlepiej ze wszystkich radzi sobie na zbiorze uczącym, lecz najgorzej na zbiorze testowym co może świadczyć o przeuczeniu.

Przed przystąpieniem do interpretacji modelu logitowego zmieniono grupę referencyjną na grupę najmniejszego ryzyka zawału serca (mężczyźni, typowa dławica piersiowa, usunięta usterka serca oraz zerowy poziom zmiennych ilościowych).

Tabela 14. Ostateczny model. Źródło: opracowanie własne na podstawie programu RStudio.

zmienna	ocena parametru	błąd standardowy	statystyka walda	p-value	iloraz szans
wyraz wolny	-2,455	2,046	-1,200	0,230	0,086
sex0	1,865	0,573	3,255	0,001	6,459
cp1	1,754	0,681	2,574	0,010	5,775
cp2	1,765	0,532	3,320	0,001	5,844
cp3	1,850	0,808	2,291	0,022	6,363
chol	-0,011	0,005	-2,247	0,025	0,989
thalach	0,030	0,012	2,479	0,013	1,030
oldpeak	-0,684	0,244	-2,804	0,005	0,505
ca	-0,704	0,229	-3,074	0,002	0,494
thal0	0,061	0,997	0,061	0,951	1,063
thal2	1,275	1,007	1,266	0,206	3,577

Pomimo, że zmienna thal jest nieistotna (*Tabela 14*) postanowiono pozostawić ją w modelu, ponieważ jej poziom równy 1 był istotny statystycznie.

Interpretacja współczynników modelu:

- Wyraz wolny

  Szanse zwiększonego prawdopodobieństwa zawału serca w grupie referencyjnej wynosiłyby 0,086 jednak wyraz wolny nie ma interpretacji, ponieważ niemożliwy jest zerowy poziom cholesterolu w surowicy [mg/dl] oraz aby maksymalne tętno wynosiło 0 bpm.
- Sex0

  Szanse zwiększonego ryzyka zawału serca w przypadku kobiet są średnio prawie 6,5-krotnie większe niż dla mężczyzn przy stałości pozostałych zmiennych (ceteris paribus).
- Cp1

  Szanse zwiększonego ryzyka zawału serca są średnio 5,7-krotnie wyższe dla osób z nietypową dławicą piersiową niż dla osób z typową dławicą piersiową ceteris paribus.
- Cp2

  Szanse zwiększonego ryzyka zawału serca są średnio prawie 6-krotnie większe dla osób z bólem niebędącym dławicą piersiową niż dla osób z typową dławicą piersiową ceteris paribus.
- Cp3

  Szanse zwiększonego ryzyka zawału serca są średnio ponad 6-krotnie większe dla osób z bezobjawowym bólem klatki piersiowej niż dla osób z typową dławicą piersiową ceteris paribus.
- Chol

  Wraz ze wzrostem cholesterolu w surowicy o 1 mg/dl szanse zwiększonego ryzyka zawału serca maleją średnio o 1%.
- Thalach

  Wraz ze wzrostem maksymalnego osiągniętego tętna o 1 bpm szanse zwiększonego ryzyka zawału serca wzrosną średnio o 3%.

- Oldpeak

  Wraz ze wzrostem obniżenia odcinka ST o 1 jednostkę szanse prawdopodobieństwa większego ryzyka zawału serca zmaleją średnio o 49,5%.
- Ca

  Wraz ze wzrostem liczba głównych naczyń zabarwionych metodą fluoroskopii o 1 naczynie szanse zwiększonego ryzyka zawału serca zmaleją średnio o 50,5%.
- Thal0

  Szanse zwiększonego ryzyka zawału serca są średnio o 6% większe dla osób z normalnym typem wady serca niż dla osób z usuniętą usterką serca. Jednak nie jest to istotny statystycznie efekt.
- Thal2

  Szanse zwiększonego ryzyka zawału serca są średnio o 3,5-krotnie większe dla osób z odwracalną wadą serca niż dla osób z usuniętą usterką serca. Jednak nie jest to efekt istotny statystycznie.

6. Zakończenie

Celem było stworzenie modelu predykcyjnego, który pomoże w identyfikacji osób o podwyższonym ryzyku zawału serca, co umożliwi wdrożenie odpowiednich interwencji medycznych w celu zmniejszenia liczby zachorowań i zgonów związanych z zawałem serca. W trakcie analizy wykryto zmienne, które miały istotny statystycznie wpływ na kształtowanie prawdopodobieństwa zawału serca. Były to następujące zmienne:

- `chol` - cholesterol w surowicy [mg/dl],
- `thalach` - osiągnięte maksymalne tętno [bpm],
- `oldpeak` - obniżenie odcinka ST wywołane wysiłkiem fizycznym w stosunku do odpoczynku,
- `ca` - liczba głównych naczyń zabarwionych metodą fluoroskopii [wartości od 0 do 4],
- `sex`- płeć pacjenta [0: Kobieta, 1: Mężczyzna],
- `cp` - typ bólu w klatce piersiowej [wartości: 0,1,2,3]
  - 0 – typowa dławica piersiowa,
  - 1 – nietypowa dławica piersiowa,
  - 2 – ból niebędący dławicą piersiową,
  - 3 – bezobjawowy,
- `thal` - typ wady serca [wartości: 0,1,2]
  - 0 – normalny,
  - 1 – usunięta usterka,
  - 2 – wada odwracalna,

Porównane zostały 3 modele – dwumianowy logitowy, dwumianowy probitowy oraz dwumianowy logitowy z interakcją. Najlepszym modelem na zbiorze testowym okazał się model logitowy. Osiągał on dokładność na poziomie 83%. Na jego podstawie w grupie największego ryzyka zawału serca są kobiety z bezobjawowym bólem klatki piersiowej.

Spis tabel

Spis wykresów