***Model objaśniający ryzyko zachorowania na chorobę serca***

***Przedmiot: Modele parametryczne***

***Marta Denisiuk***

Spis treści

[1. Wstęp 3](#_Toc107100663)

[1.1. Wprowadzenie do tematu 3](#_Toc107100664)

[1.2. Cel badania 3](#_Toc107100665)

[2. Charakterystyka zbioru danych 4](#_Toc107100666)

[2.1. Opis zbioru danych 4](#_Toc107100667)

[3. Przygotowanie danych 5](#_Toc107100668)

[3.1. Braki danych 5](#_Toc107100669)

[3.2. Imputacja danych 5](#_Toc107100670)

[3.3. Podstawowe statystyki 6](#_Toc107100671)

[4. Analiza wstępna 8](#_Toc107100672)

[5. Modelowanie 21](#_Toc107100673)

[5.1. Przygotowanie zbioru danych 21](#_Toc107100674)

[5.2. Budowa modelu logitowego 21](#_Toc107100675)

[5.2.1. Oszacowanie parametrów strukturalnych modeli 21](#_Toc107100676)

[5.2.2. Ocena dopasowania oszacowanych modeli 22](#_Toc107100677)

[5.2.3. Sprawdzenie założeń modelu 22](#_Toc107100678)

[5.2.4. Identyfikacja wartości odbiegających i nietypowych 23](#_Toc107100679)

[5.2.5. Ocena jakości predykcji modelu 23](#_Toc107100680)

[5.3. Budowa modelu probitowego 23](#_Toc107100681)

[5.4. Porównanie modelu logitowego i probitowego 24](#_Toc107100682)

[5.5. Interpretacja modelu 24](#_Toc107100683)

[6. Podsumowanie i wnioski 25](#_Toc107100684)

[Podział pracy 26](#_Toc107100685)

# Wstęp

## Wprowadzenie do tematu

**Choroby serca to liczne schorzenia o charakterze ostrym lub przewlekłym, które obejmują co najmniej jeden z elementów serca. Poza podziałem na ostre i przewlekłe choroby serca, funkcjonuje także podział na choroby wrodzone i nabyte, a także strukturalne i czynnościowe.** Zgodnie ze statystykami choroby serca są jedną z głównych przyczyn zgonów na świecie**, dlatego tak istotne jest jak najszybsze rozpoznanie chorób serca i wdrożenie odpowiedniej terapii.**

Z danych Głównego Urzędu Statystycznego wynika, że co roku z przyczyn kardiologicznych umiera 175 tys. Polaków (dane z 2018 roku). Natomiast w Europie co roku stwierdza się ponad 11 mln nowych przypadków chorób układu sercowo-naczyniowego i 3,9 mln zgonów z ich powodu.

## Cel badania

Celem badania jest określenie jakie zmienne i w jakim stopniu wpływają na wystąpienie choroby serca u człowieka. Utworzenie odpowiedniego modelu pozwoli odpowiedzieć na następujące pytania:

* Jakie czynniki wpływają na ryzyko zachorowania na chorobę serca?
* Jaka grupa osób jest najbardziej narażona na zachorowanie?
* Czy ryzyko zachorowania jest zależne od liczby wypalanych papierosów? Czy liczba nie ma znaczenia?
* Czy osoby z leczonym nadciśnieniem tętniczym są mniej narażone na chorobę serca niż osoby z nieleczonym nadciśnieniem lub osoby bez nadciśnienia tętniczego?
* Czy osoby chorujące na cukrzycę są bardziej narażone na zachorowanie?

# Charakterystyka zbioru danych

## Opis zbioru danych

Zbiór danych pobrano z portalu Kaggle i zawiera on ok. 4200 jednostek oraz 16 zmiennych. Zmienne w bazie danych są ilościowe oraz jakościowe. Zmienna zależną jest zmienna TenYearCHD, która oznacza ryzyko zachorowania na chorobę serca w ciągu 10 lat (1-ryzyko występuje, 0 – ryzyko nie występuje). Zmienne niezależne, które są potencjonalnymi czynnikami ryzyka są podzielone na zmienne demograficzne, behawioralne oraz medyczne – historyczne i bieżące.

* Zmienne demograficzne:
* Sex: Płeć pacjenta (1 – Male, 0 – Female)
* Age: Wiek pacjenta
* Education: Wykształcenie (1 - podstawowe, 2 - średnie, 3 - zawodowe, 4 - wyższe)
* Zmienne behawioralne:
* Current Smoker: Osoba paląca – czy pacjent jest osobą palącą (1 – osoba paląca, 0 – osoba niepaląca)
* Cigs Per Day: Średnia dzienna liczba wypalanych papierosów w sztukach
* Zmienne medyczne historyczne:
* BP Meds: Leczenie ciśnienia krwi – czy pacjent miał leczone ciśnienie krwi (1 – Tak, 0 – Nie)
* Prevalent Stroke: Wystąpienie udaru mózgu – czy pacjent miał udar mózgu w przeszłości (1 – Tak, 0 – Nie)
* Prevalent Hyp: Występowanie nadciśnienia tętniczego – czy pacjent ma nadciśnienie tętnicze (1 – Tak, 0 – Nie)
* Diabetes: Występowanie cukrzycy – czy pacjent ma cukrzycę (1 – Tak, 0 – Nie)
* Zmienne medyczne bieżące:
* Tot Chol: Poziom cholesterolu w mg/dl (miligramy/decylitry)
* Sys BP: Skurczowe ciśnienie krwi w mm Hg
* Dia BP: Rozkurczowe ciśnienie krwi w mm Hg
* BMI: wskaźnik BMI – wskaźnik masy ciała
* HeartRate: Tętno (liczba uderzeń na minutę)
* Glucose: Poziom glukozy w mg/dl (miligramy/decylitry)

# Przygotowanie danych

## Braki danych

Do wykonania dalszych analiz niezbędne jest sprawdzenie, czy występują braki danych. Wykryto łącznie 695 braków danych dla 7 zmiennych (education, cigsPerDay, BPMeds, totChol, BMI, heartRate i glucose). Dla każdej zmiennej z brakami danych obliczono ich procent.

* education --> 2.48%
* cigsPerDay --> 0.68%
* BPMeds --> 1.25%
* totChol --> 1.18%
* BMI --> 0.45%
* heartRate --> 0.02%
* glucose --> 9.16%

Na podstawie uzyskanych wartości procentowych można łatwo zauważyć, że najwięcej braków danych jest dla zmiennych glucose oraz education. Zdecydowano pozostawić wszystkie zmienne i dokonać imputacji danych.

## Imputacja danych

W celu pozbycia się braków danych wykonano imputacje, która polega na sztucznym wstawieniu pewnych wartości do zbioru danych. Technika imputacji danych zależy od typu zmiennej.

Zmienna education jest wyrażona na skali porządkowej i zdecydowano o dokonaniu imputacji poprzez przepisanie wartości z wiersza poprzedniego. Taki sam zabieg został użyty w przypadku zmiennej heartRate, ponieważ brakowało tam tylko jednej obserwacji.

W przypadku zmiennej cigsPerDay wykryto, że braki danych występują tylko wśród osób palących, więc zdecydowano się na imputacje danych z wykorzystaniem średniej liczby dziennie wypalanych papierosów dla osób palących. Taka sama technika imputacji została wykorzystana w przypadku braków dla zmiennych totChol, BMI oraz glucose.

W przypadku braków danych w kolumnie BPMeds zdecydowano, że najlepszą techniką imputacji będzie regresja logistyczna. Zmiennymi objaśniającymi są prevalentHyp, sysBP oraz diaBP.

## Podstawowe statystyki

Po wykonaniu imputacji danych obliczono podstawowe statystyki dla wszystkich zmiennych objaśniających ilościowych. Na podstawie wyników zaprezentowanych w tabeli 3.1. można zauważyć, że średnia wieku badanych osób wynosi prawie 50 lat – najmłodsza badana osoba ma 32 lata a najstarsza 70 lat. Średni poziom cholesterolu wynosi ok. 237 mg/dl i jest to wartość zbyt duża, ponieważ poziom cholesterolu u zdrowego człowieka nie powinien przekraczać 200 mg/dl. Średnia wartość skurczowego ciśnienia krwi, czyli ciśnienia, które powstaje podczas skurczu serca. wynosi ok. 132 mm Hg i jest to wartość mieszcząca się w normie (norma to 120-140 mm Hg). Średnia wartość rozkurczowego ciśnienia krwi, czyli minimalnego ciśnienia podczas fazy spoczynkowej, tj. takiego, które wytwarza się w tętnicach pomiędzy uderzeniami, wynosi ok. 83 mm Hg i jest to wartość prawidłowa (norma do 90 mm Hg). W przypadku BMI, średnia wartość oznacza nadwagę. Średnia wartość tętna wynosi ok. 76 uderzeń na minutę i jest to wartość prawidłowa, ponieważ prawidłowe tętno u osób dorosłych i dzieci powyżej 10 r. ż. ma wartość od 50 do 100 uderzeń/minutę. Średnia wartość poziomu glukozy we krwi wynosi ok. 82 mg/dl i jest to wartość prawidłowa, jeżeli poziom glukozy był mierzony na czczo.

Minimalne wartości dla poszczególnych zmiennych są wartościami, które mogą wystąpić. Maksymalne wartości dla zmiennych totChol, sysBP oraz glucose przyjmują bardzo duże wartości, ale są to wartości obserwowane w przypadku chorób np. hiperglikemii, choroby wieńcowej.

Wszystkie zmienne mają rozkład dodatnio skośny, ponieważ współczynniki skośności są większe od 0. Dla zmiennej glukoza współczynnik przyjmuje bardzo dużą wartość. W dalszym etapie należy przyjrzeć się rozkładowi tej zmiennej.

Wysoka miara koncentracji dla zmiennej glucose wynika z zastosowanej techniki imputacji danych – dane imputowano przy pomocy średniej. Dla zmiennej totChol kurtoza również jest dość duża i również wynika to z zastosowanej techniki imputacji braków danych średnią.

Na podstawie współczynnika zmienności można stwierdzić, że wszystkie zmienne poza cigsPerDay oraz glucose mają małą zmienność. Zmienna glucose charakteryzuje się przeciętną zmiennością. Natomiast zmienna cigsPerDay ma bardzo silną zmienność.

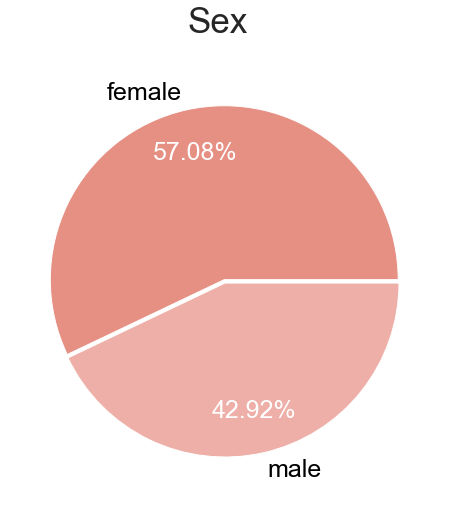
Tabela 3. . Podstawowe statystyki dla zmiennych ilościowych

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Miara | age | cigsPerDay | totChol | sysBP | diaBP | BMI | heartRate | glucose |
| count | 4238 | 4238 | 4238 | 4238 | 4238 | 4238 | 4238 | 4238 |
| mean | 49,58 | 9,06 | 236,72 | 132,35 | 82,89 | 25,8 | 75,88 | 81,97 |
| std | 8,57 | 11,9 | 44,33 | 22,04 | 11,91 | 4,07 | 12,03 | 22,84 |
| min | 32 | 0 | 107 | 83,5 | 48 | 15,54 | 44 | 40 |
| 25% | 42 | 0 | 206 | 117 | 75 | 23,08 | 68 | 72 |
| 50% | 49 | 0 | 234 | 128 | 82 | 25,41 | 75 | 80 |
| 75% | 56 | 20 | 262 | 144 | 89,88 | 28,04 | 83 | 85 |
| max | 70 | 70 | 696 | 295 | 142,5 | 56,8 | 143 | 394 |
| skośność | 0,23 | 1,23 | 0,88 | 1,15 | 0,71 | 0,98 | 0,64 | 6,52 |
| kurtoza | -0,99 | 1,00 | 4,22 | 2,16 | 1,28 | 2,68 | 0,91 | 64,88 |
| Współ. Zmienności | 17,29% | 131,35% | 18,73% | 16,65% | 14,37% | 15,78% | 15,85% | 27,86% |

W tabeli 3.1 przedstawiono podstawowe statystki tylko dla zmiennych ilościowych, ale w zbiorze danych występuje równie dużo zmiennych jakościowych i kolejnym krokiem, będzie wykonanie analizy wstępnej, która pozwoli dokładnie przyjrzeć się obu typom zmiennych.

# Analiza wstępna

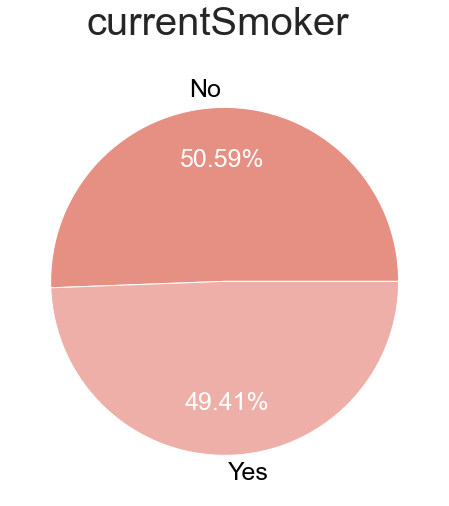
Analiza wstępna ma na celu zaprezentowanie wszystkich zmiennych na wykresach oraz dokładne omówienie przyjmowanych przez nie wartości. Przeprowadzenie jej może pomóc w wykryciu zależności między poszczególnymi zmiennymi i w konsekwencji ułatwić budowanie modelu objaśniającego ryzyko zachorowania na chorobę serca.



Objaśniane zmienne dychotomiczne zaprezentowano na wykresach kołowych.

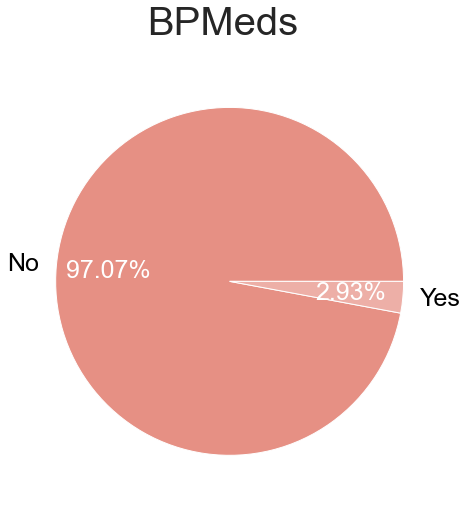
Na wykresie 4.1 przedstawiono rozkład płci w badanym zbiorze. Kobiety stanowią ok. 57% a mężczyźni ok. 43%.

Wykres 4. . Rozkład zmiennej płeć



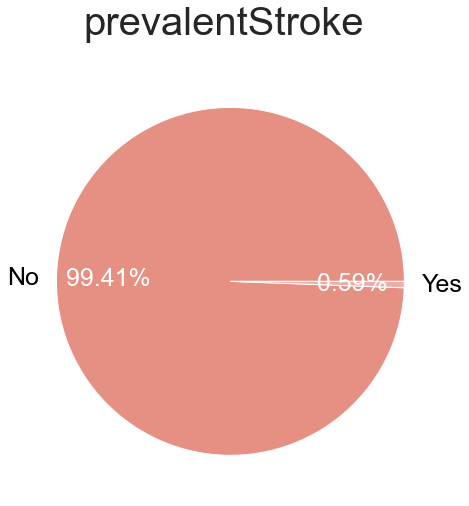
Na wykresie 4.2 zaprezentowano rozkład zmiennej dotyczącej palenia papierosów. Jako aktualni palacze zadeklarowało się ok. 49% badanych. Osoby niepalące stanowią ok. 51%. Omawiana cecha rozkłada się równo po połowie.

Wykres 4. . Rozkład zmiennej currentSmoker



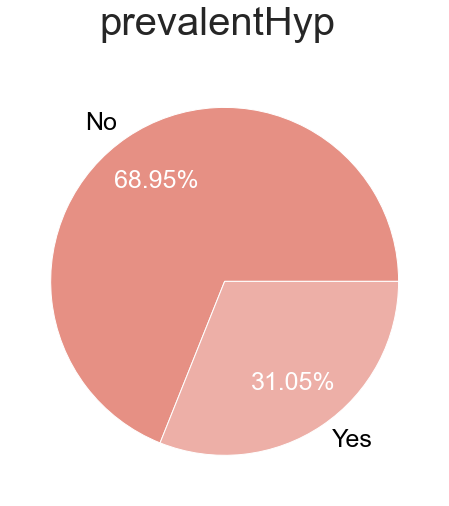
Na wykresie 4.3 pokazano rozkład cechy BPMeds. Zaledwie ok. 3% badanych leczyło ciśnienie krwi. Pozostała część badanych (ok. 97%) nigdy nie poddała się tego typu leczeniu.

Wykres 4. . Rozkład zmiennej BPMeds



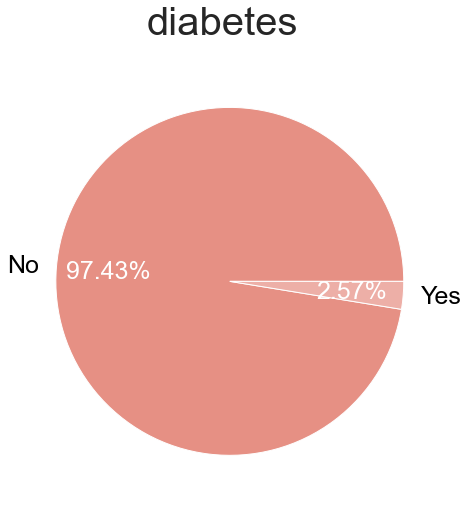
Na wykresie 4.4 przedstawiono rozkład zmiennej prevalentStroke, czyli przeżycie udaru mózgu, który jest stanem zagrażajacym życiu, ponieważ dochodzi do obumarcia częsci mózgu w skutek zatrzymania dopływu krwi do jego tkanek. Zdecydowana większość osób badanych (ok. 99%) nie miała udaru mózgu w przeszłości. Zaledwie 1% badanych osób przeżył ten stan.

Wykres 4. . Rozkład zmiennej prevalentStroke



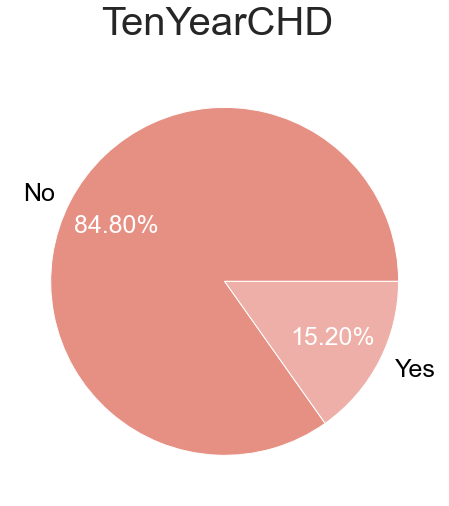
Na wykresie 4.5 można zauważyć, że ok. 31% badanych osób choruje na nadciśnienie tętnicze, czyli chorobę, która charakteryzuje się podwyższonym ciśnieniem krwi. Ok. 69 % badanych nie posiada takiego schorzenia.

Wykres 4. . Rozkład zmiennej prevalentHyp



Rozkład zmiennej diabetes został przedstawiony na wykresie 4.6. Zaledwie 3% badanych choruje na cukrzycę, czyli chorobę, która charakteryzuje się zaburzonym wydzielaniem lub działanie insuliny i zbyt wysokim poziomem glukozy we krwi. Wśród badanych ok. 97% nie choruje na tą chorobę.

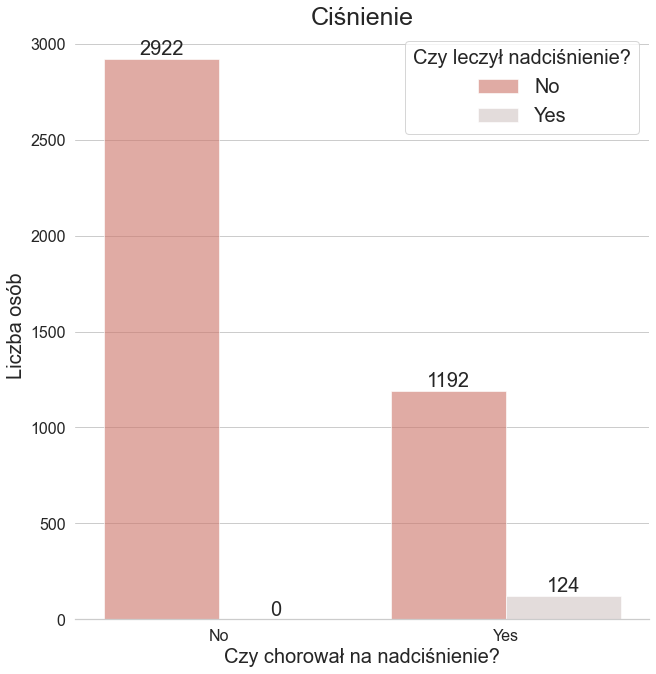
Wykres 4. . Rozkład zmiennej diabetes



Rozkład zmiennej zależnej przedstawiono na wykresie 4.7. Można zauważyć, że ryzyko zachorowania na chorobę serca występuje tylko u ok. 15% badanych. Osoby nienarażone na zachorowanie stanowią ok. 85%.

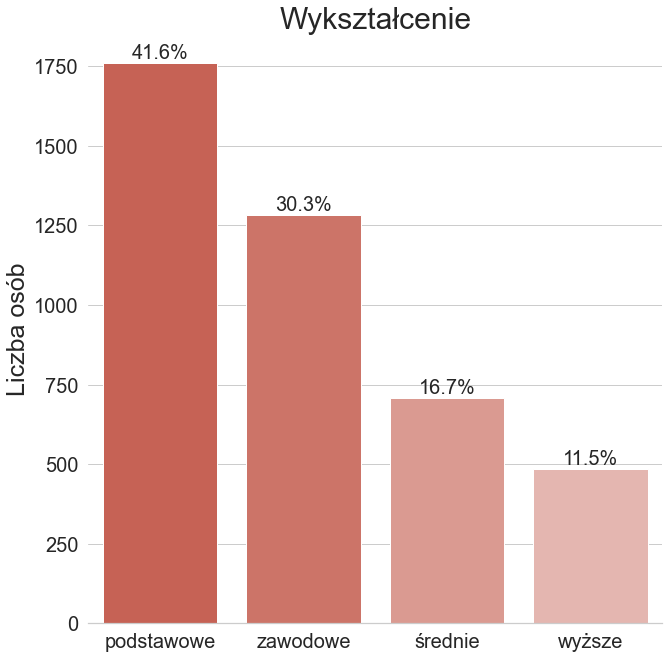
Wykres 4. . Rozkład zmiennej zależnej TenYearCHD

Analizując zmienne przedstawione na wykresach 4.1 – 4.7 zauważono pewną zależność pomiędzy zmiennymi prevalentHyp oraz BPMeds. Na wykresie 4.8 przedstawiono zestawienie osób, które chorują na nadciśnienie tętnicze z podziałem na osoby, które je leczą lub nie. Oczywiście osoby, które nie chorują na nadciśnienie tętnicze nie leczyły ciśnienia krwi. Natomiast wśród osób posiadających te schorzenie, aż 91% osób nie leczyło go.



Wykres 4. . Zależność pomiędzy cechami prevalentHyp oraz BPMeds

Zmienne jakościowe, które zostały wyrażone na skali porządkowej i następnie odpowiednio zakodowane przedstawiono na wykresach słupkowych. Na wykresie 4.9 przedstawiono rozkład zmiennej education. Można zauważyć, że wśród badanych najwięcej osób ma wykształcenie podstawowe (ok. 42%), osoby posiadające wykształcenie zawodowe stanowią ok. 30%, osoby z wykształceniem średnim ok. 17% a osoby z wykształceniem wyższym ok. 12%. Zdecydowana większość osób badanych ma niższe wykształcenie.



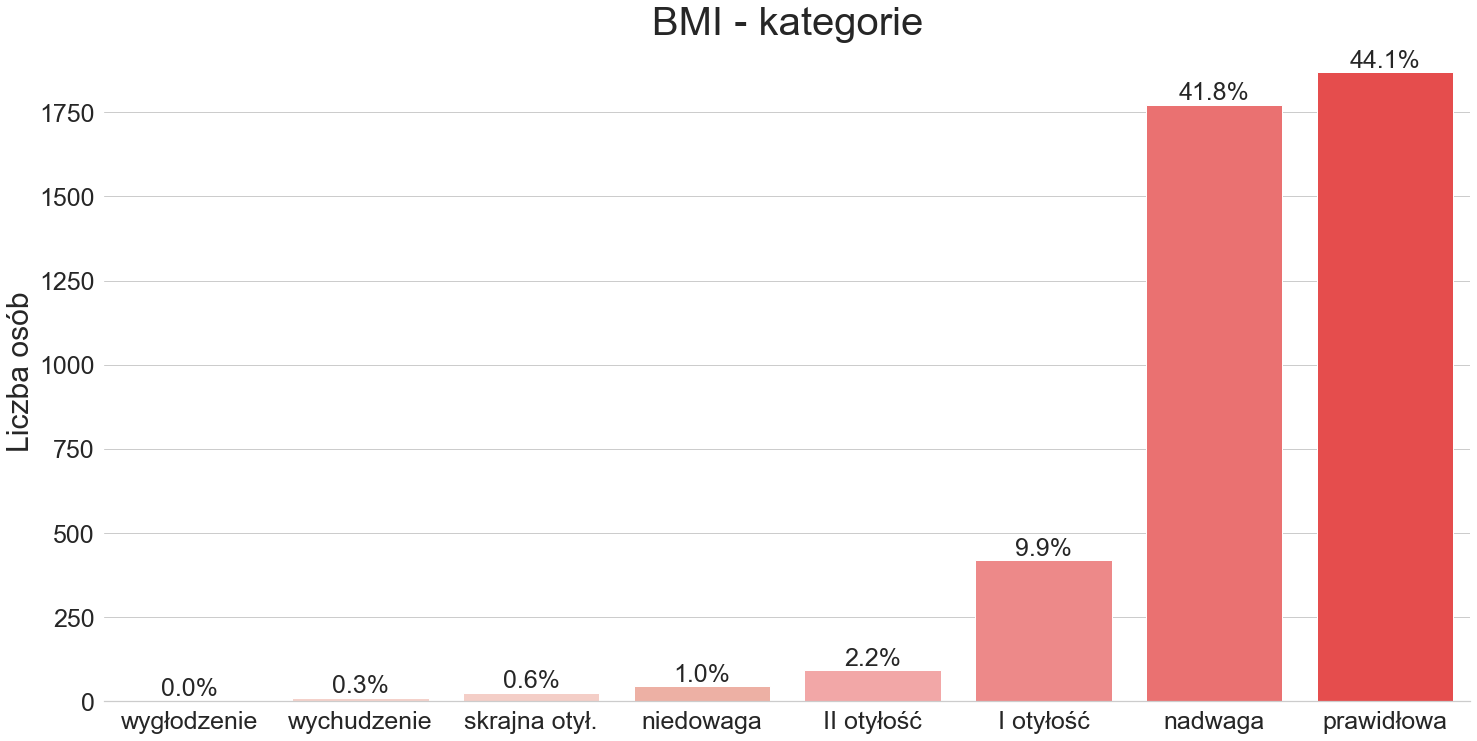
Wykres 4. . Rozkład zmiennej education

Wskaźnik BMI czyli Body Mass Index to nic innego, jak stosunek wagi danej osoby do jej wzrostu. Jest to najpopularniejszy sposób wstępnego diagnozowania nadwagi, otyłości i niedowagi. Podstawowa klasyfikacja BMI dla osób dorosłych wyróżnia osiem oceny wagi:

* mniej niż 16 - wygłodzenie
* 16 - 16.99 - wychudzenie
* 17 - 18.49 - niedowaga
* 18.5 - 24.99 - wartość prawidłowa
* 25 - 29.99 - nadwaga
* 30 - 34.99 - I stopień otyłości
* 35 - 39.99 - II stopień otyłości
* powyżej 40 - otyłość skrajna

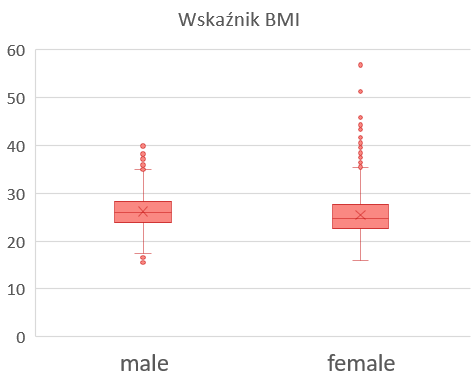
Wskaźnik BMI jest jedynie orientacyjną informacją, ale regularne jego sprawdzanie pomaga kontrolować wagę ciała i uniknąć ryzyka otyłości.

Dane z wyjściowego zbioru (zmienna BMI), które były przedstawione jako zmienna ilościowa, zamieniono na zmienną jakościową porządkową – w ten sposób powstała nowa zmienna catBMI. Zmienną zakodowano z wykorzystaniem przedstawionej klasyfikacji BMI dla osób dorosłych. Na wykresie 4.10 przedstawiono rozkład zmiennej catBMI. Osoby z prawidłową masą ciała stanowią ok. 44% badanych, ale osoby z nadwagą stanowią niewiele mniej, bo ok. 42%. Wśród badanych osób nie występują jednostki wygłodzone, a osoby wychudzone stanowią zaledwie 0,3%. Osoby ze skrajną otyłością stanowią zaledwie 0,6%, osoby z niedowagą 1%. Natomiast osoby z otyłością I lub II stopnia stanowią łącznie ok. 12%.



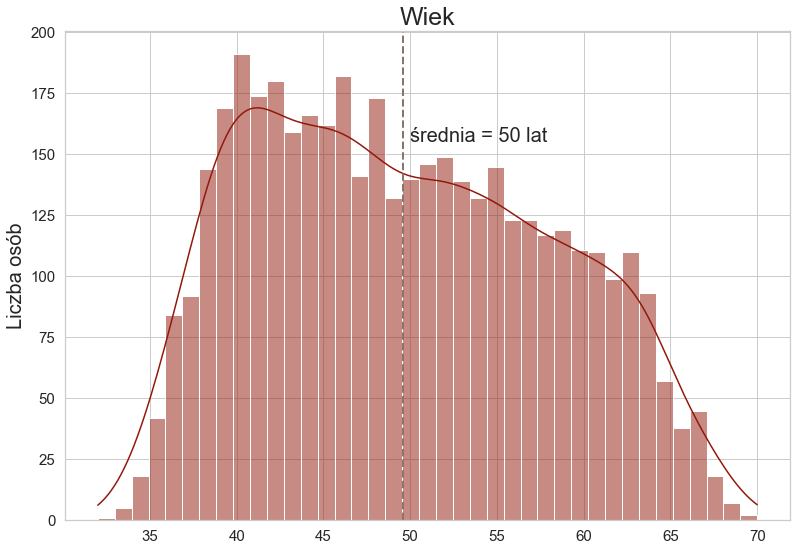
Wykres 4. . Rozkład zmiennej BMI po zakodowaniu

Należy jednak zwrócić uwagę, na fakt, że wskaźnik BMI jest obliczany w tej sam sposób dla kobiet oraz dla mężczyzn, a jak wiadomo, przedstawiciele obu płci różnią się nie tylko sylwetką, ale też proporcją tkanki tłuszczowej do masy mięśniowej. U kobiet występuje fizjologiczne uwarunkowanie do gromadzenia większej ilości tkanki tłuszczowej w organizmie. Jednocześnie mają one proporcjonalnie mniej mięśni. Dlatego należy pamiętać, że jeżeli u kobiety i mężczyzny występuje zbliżony wynik BMI, to istnieje duże prawdopodobieństwo, że u kobiety będzie wyższy poziom tkanki tłuszczowej. Na pudełkowym wykresie 4.11 przedstawiono rozkład zmiennej BMI z podziałem na płeć. Można zauważyć, że średni wskaźnik BMI dla kobiet jest mniejszy niż dla mężczyzn. Mediana również jest mniejsza w przypadku kobiet. Jednak dla kobiet można zaobserwować więcej wartości odbiegających, które są bardzo wysoki i świadczą o skrajnej otyłości.



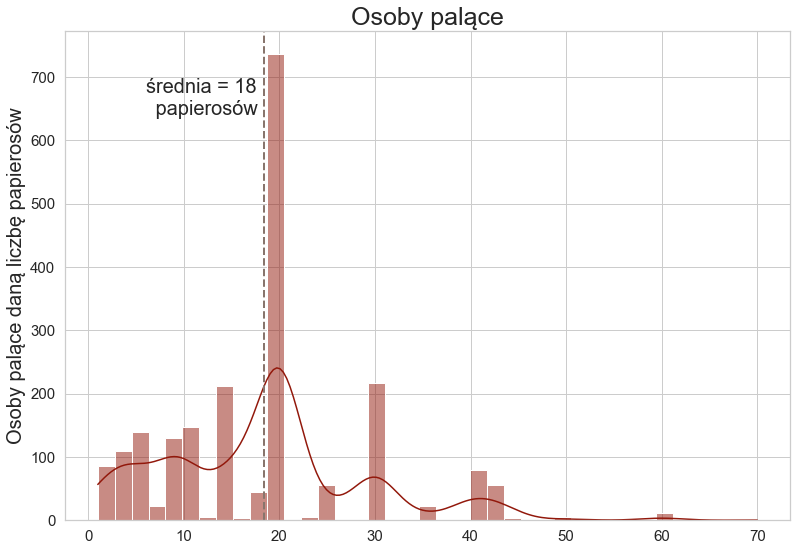
Wykres 4. . Rozkład zmiennej BMI z podziałem na płeć

Pozostałe ilościowe zmiennej objaśniające przedstawiono w formie histogramów. Na wykresie 4.12 przedstawiono rozkład zmiennej age. Średnia wieku osób badanych wynosi 50 lat, ale najwięcej osób ma mniej więcej 40 lat. W rozkładzie można zauważyć delikatną asymetrię prawostronną.



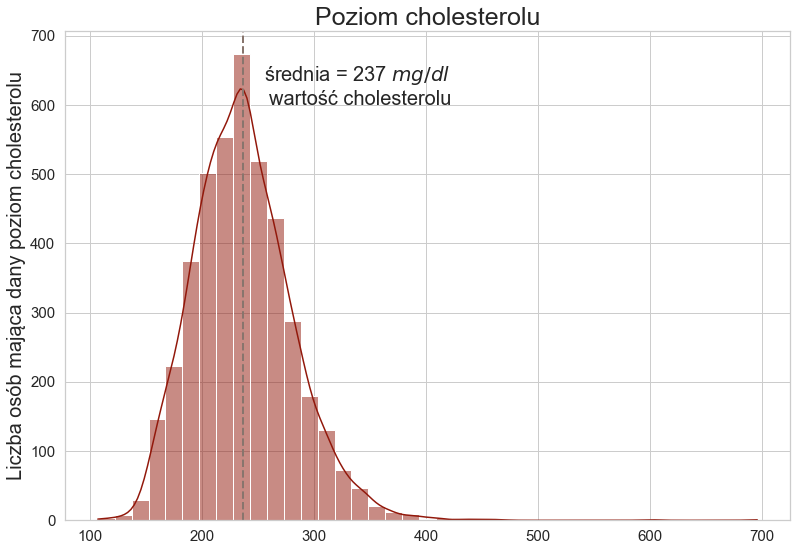
Wykres 4. . Rozkład zmiennej age

Na wykresie 4.13 zaprezentowano rozkład cechy cigsPerDay, ale tylko dla osób palących, która opisuje dzienną liczbę wypalanych papierosów. Można zaobserwować braki pomiędzy poszczególnymi wartościami, ponieważ jest to zmienna skokowa. Najwięcej osób wypala dziennie ok. 20 papierosów. Średnia liczba wypalanych papierosów dla osób palących wynosi natomiast 18 papierosów. Biorąc pod uwagę liczbę papierosów w standardowej paczce, można stwierdzić, że średnio palacze wypalają 1 paczkę papierosów dziennie.



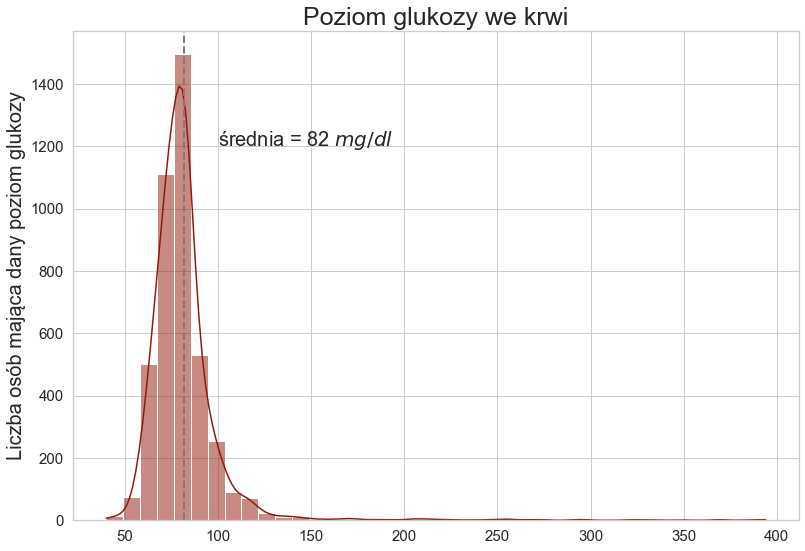
Wykres 4. . Rozkład zmiennej cigsPerDay dla osób palących

Rozkład zmiennej totChol przedstawiono na wykresie 4.14. Można zauważyć, że rozkład cholesterolu w badanej próbie jest zbliżony do rozkładu normalnego. Średni poziom cholesterolu wynosi 237 mg/dl. Cholesterol zalicza się do tłuszczów (lipidów), a więc należy do organicznych związków chemicznych i w związku z tym nie rozpuszcza się w krwi. Cholesterol całkowity u zdrowego człowieka nie powinien przekraczać 200 mg/dl, więc biorąc pod uwagę wartość pierwszego kwartylu (206 mg/dl), można stwierdzić, że ponad 75% badanych osób ma zbyt wysoki poziom cholesterolu.



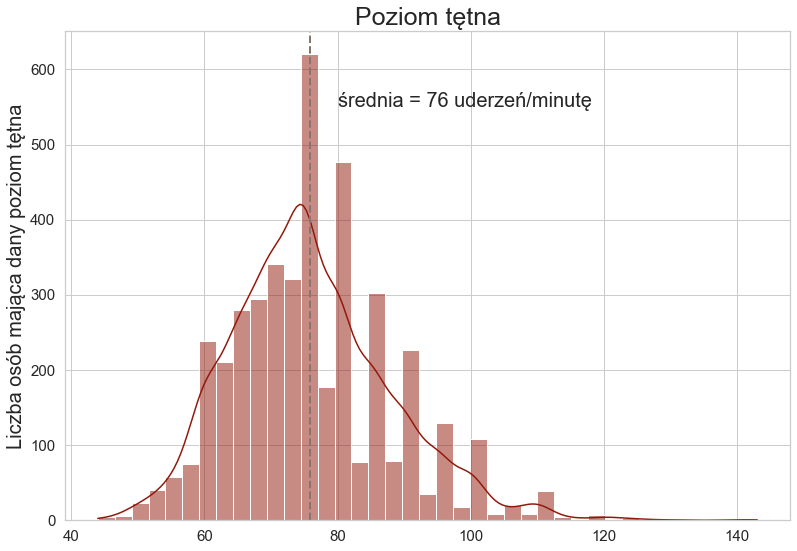
Wykres 4. . Rozkład zmiennej totChol

Glukoza to cukier prosty i dla większości żywych organizmów stanowi główne źródło energii. Innymi słowy, glukoza stanowi „paliwo” dla ludzi i zwierząt, dlatego też jej przyswajanie we właściwych ilościach jest kluczowe dla naszego funkcjonowania.  Nieprawidłowy poziom glukozy we krwi rozpoczyna się od 100-125 mg/dl - mówimy wówczas o stanie przedcukrzycowym.  Diagnoza cukrzycy pojawia się w sytuacji, gdy w wynikach badań krwi glukoza przekroczy poziom 126 mg/dl. Na wykresie 4.15 przedstawiono rozkład zmiennej glucose. Wykres jest bardzo skupiony wokół średniej, ale wynika to z zastosowanej techniki imputacji danych – imputacji średnią. Średnia wartość glukozy we krwi wynosi 82 mg/dl i mieści się w normie. Na podstawie wykresu 4.15 można łatwo zauważyć, że w badanej próbie jest mało osób chorujących na cukrzycę.



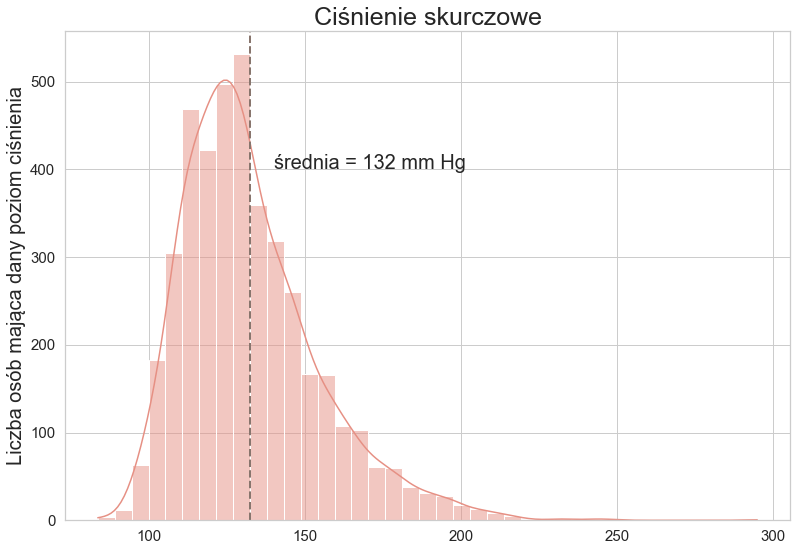
Wykres 4. . Rozkład zmiennej glucose

Innym słowem określającym tętno jest puls – najprościej mówiąc jest to liczba uderzeń serca na minutę. Wartości tętna spoczynkowego są cechą indywidualną i różnią się między poszczególnymi osobami. Są też niższe podczas spoczynku, a wyższe np. w trakcie aktywności fizycznej. Prawidłowe tętno różni się w zależności od wieku, ale w badanej próbie znajdują się tylko osoby dorosłe. Normą dla osób dorosłych jest 60-90 uderzeń serca w ciągu minuty (zazwyczaj ok. 70 uderzeń na minutę). Na wykresie 4.16 przedstawiono rozkład zmiennej heartRate. Średnia wartość tętna to 76 uderzeń na minutę. Można zauważyć, że zdecydowana większość badanych osób mieści się w normie.



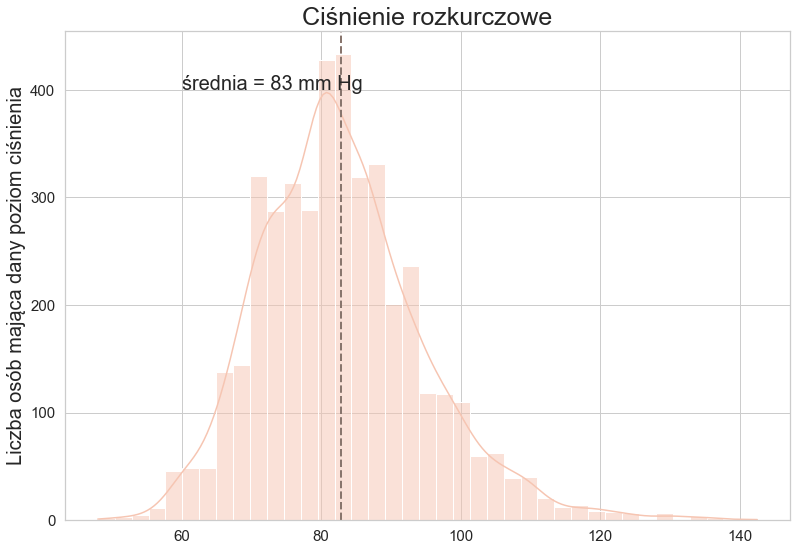
Wykres 4. . Rozkład zmiennej heartRate

Ciśnienie tętnicze krwi to siła, z jaką ciecz - czyli krew - naciska na ściany naczynia. Wyróżnia się ciśnienie skurczowe i ciśnienie rozkurczowe. Ciśnienie skurczowe to największe ciśnienie, jakie panuje w układzie krwionośnym. Mierzone jest ono w czasie skurczu lewej komory serca – maksymalnego wyrzutu krwi. Za prawidłową wartość skurczowego ciśnienia krwi uważa się wartość pomiędzy 120–129 mm Hg. Wysokie ciśnienie skurczowe to takie powyżej 140 mm Hg. Na wykresie 4. 17 przedstawiono rozkład zmiennej sysBP. Średnia wartość ciśnienia skurczowego wśród badanych osób wynosi 132 mm Hg i jest to wartość wykraczająca poza normę. Biorąc pod uwagę medianę dla omawianej zmiennej (128 mm Hg), można stwierdzić, że mniej więcej połowa badanych osób ma zbyt wysokie ciśnienie skurczowe. Natomiast na podstawie zmiennej prevalentHyp wiadomo, że tylko ok. 31% osób choruje na nadciśnienie tętnicze.



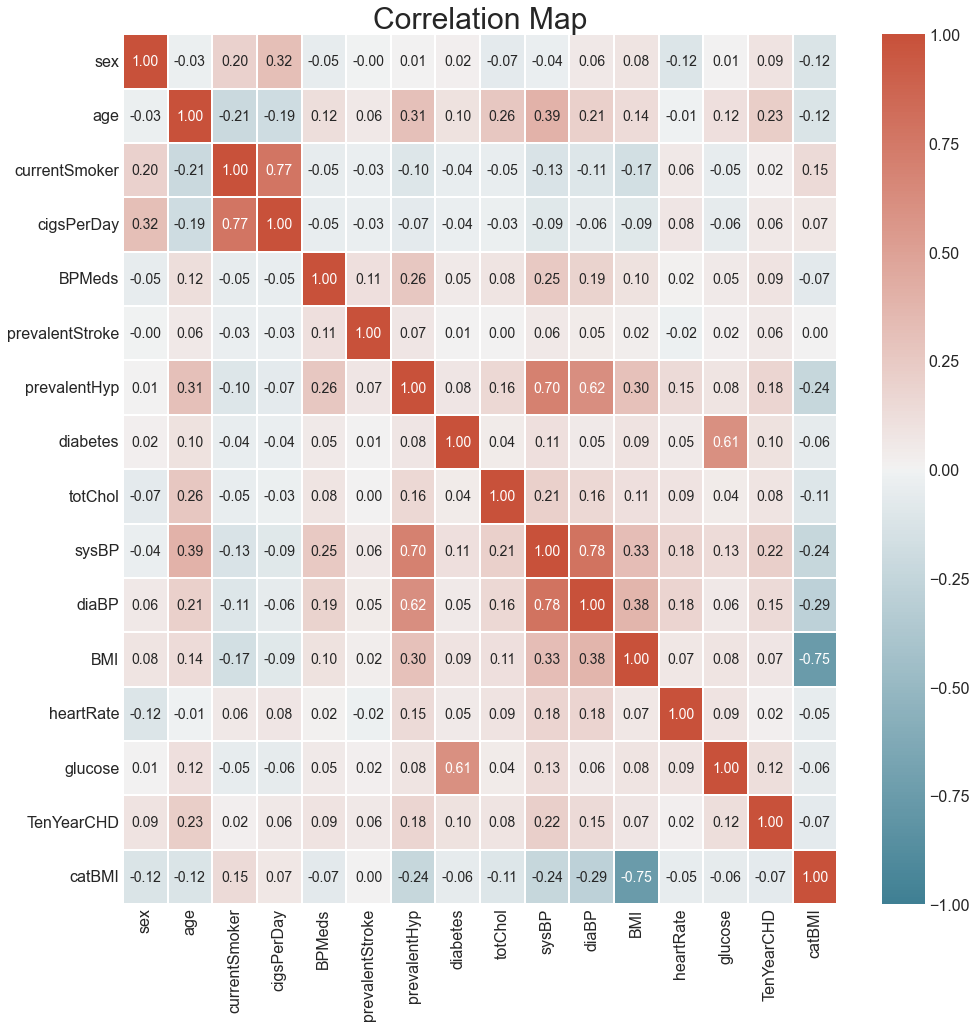
Wykres 4. . Rozkład zmiennej sysBP

Drugim parametrem określającym ciśnienie krwi jest ciśnienie rozkurczowe notowane jest tuż przed wyrzutem krwi z serca i jest ciśnieniem najniższym, jakie może być zanotowane w tętnicach. Powodem, dla którego nie spada ono do zera w czasie rozkurczu, jest fakt, że ściany tętnic oddają zmagazynowaną energię i wywierają nacisk na znajdującą się w środku krew. Za prawidłowe rozkurczowe ciśnienie krwi uważa się wartość pomiędzy 80-84 mm Hg, natomiast wysoka wartość to wartość powyżej 90 mm Hg. U osób powyżej 40 roku życia to ciśnienie skurczowe krwi, czyli wartość wyższa, ma większe znaczenie niż ciśnienie rozkurczowe. To dlatego, że jest to lepszy wskaźnik ryzyka udaru mózgu i zawału serca. Na wykresie 4.18 przedstawiono rozkład zmiennej diaBP. Średnia wartość rozkurczowego ciśnienia krwi wynosi 83 mm Hg i mieści się w określonych normach.



Wykres 4. . Rozkład zmiennej diaBP

Ostatnim etapem analizy wstępnej jest obliczenie korelacji i przedstawienie jej w postaci heatmap. Ten krok pomoże w identyfikacji zmiennych nadmiernie skorelowanych – korelacja równa lub powyżej 0,7, ponieważ nadmiernie skorelowane zmienne nie mogą być wykorzystane podczas budowania modelu objaśniającego ryzyko zachorowania na chorobę serca. Na wykresie 4.19 przedstawiono mapę korelacji wszystkich zmiennych. Można zauważyć, że zdecydowana większość zmiennych jest ze sobą słabo skorelowana. Najwyższa korelacja wynosząca 0,78 występuje pomiędzy zmiennymi sysBP oraz diaBP – jedną z tych zmiennych należy wyeliminować. Zmienne nadmiernie skorelowane (korelacja równa 0,77) to również zmienne cigsPerDay oraz currenSmoker – korelacja w tym przypadku jest oczywista, ponieważ osoby niepalące przyjmują zerowe wartości dla zmiennej cigsPerDay. Zmienne prevalentHyp oraz sysBP również mają wysoką wartość korelacji (równa 0,7) – zmienna sysBP jest nadmiernie skorelowana z dwiema zmiennymi, więc ją eliminujemy. Ostatnia wysoka korelacja (równa 0,75) występuje pomiędzy zmiennymi BMI oraz catBMI – w tym przypadku nadmierne skorelowanie również jest oczywiste, ponieważ zmienna catBMI została utworzona na podstawie zmiennej BMI. Na tym etapie zauważono, że zmienna catBMI niesie więcej informacji i podjęto decyzję o eliminacji zmiennej BMI.



Wykres 4. . Mapa korelacji wszystkich zmiennych

Analiza wstępna pozwoliła na zapoznanie się ze zmiennymi z analizowanego zbioru, wykrycie zależności pomiędzy zmiennymi i podjęcie pewnych decyzji w związku z budowaniem modelu.

# Modelowanie

## Przygotowanie zbioru danych

Do zbudowania modelu logitowego niezbędne jest odpowiednie przygotowanie zbioru danych. Pierwszym krokiem jest sprawdzenie rozkładów zmiennych ilościowych. Wykonano to korzystając ze testu Kołomogorowa-Smirnowa, który jest używany dla dużych zbiorów danych. Uzyskane wyniki zaprezentowano w tabeli 5.1. Odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej dla wszystkich zmiennych, ponieważ p-value wynosi 0. Algorytmy uczenia maszynowego wymagają skalowania danych, ponieważ pomaga to w szybszym zbieganiu się gradientu w kierunku minimum. Na podstawie tych informacji wykonano normalizację danych, ponieważ rozkład zmiennych nie jest znany, wiemy tylko, że nie mają rozkładu normalnego.

Tabela 5. . Wyniki testu Kołomogorwa-Smirnowa

|  |  |  |
| --- | --- | --- |
| Zmienna | Wartość statystyki | p-value |
| age | 1,0 | 0,0 |
| cigsPerDay | 0,5 | 0,0 |
| totChol | 1,0 | 0,0 |
| diaBP | 1,0 | 0,0 |
| hearRate | 1,0 | 0,0 |
| glucose | 1,0 | 0,0 |

Kolejnym krokiem przygotowania zbioru danych jest odpowiednie kodowanie zmiennych – zmienne jakościowe zakodowano przy użyciu funkcji as.factor w programie RStudio. Łącznie jako zmienne kategoryczne zapisano 7 zmiennych objaśniających: Sex, Education, currentSmoker, BPMeds, prevalentHyp, prevalentStroke oraz diabetes, a także zmienną objaśnianą TenYearCHD.

Następnie podzielono zbiór danych na zbiór uczący (70%) oraz zbiór testowy (30%). Należy sprawdzić czy proporcje zmiennej objaśnianej są takie same we wszystkich trzech zbiorach. W tabeli 5.2 zaprezentowano uzyskane wartości, które pozwalają stwierdzić, że proporcje się zgadzają i zbiór danych został prawidłowo podzielony.

Tabela 5. . Proporcje zmiennej objaśnianej w poszczególnych zbiorach danych

|  |  |  |  |
| --- | --- | --- | --- |
| Wartość | Zbiór danych | Zbiór uczący | Zbiór testowy |
| No | 0,848 | 0,848 | 0,849 |
| Yes | 0,152 | 0,152 | 0,151 |

## Budowa modelu logitowego

### Oszacowanie parametrów strukturalnych modeli

Podstawą modelu logitowego jest logit, czyli logarytm ilorazu prawdopodobieństwa przyjęcia oraz nieprzyjęcia wartości 1 przez zmienną endogeniczną, to znaczy logarytmu szansy przyjęcia wartości 1 przez zmienną objaśnianą. Nasz model to model z parametryzacją z poziomem odniesienia polegającą na zakodowaniu zmiennych jakościowych zerojedynkowych – 1, gdy zdarzenie wystąpiło lub 0, gdy zdarzenie nie wystąpiło.

Zdecydowano się na budowę modelu zaczynając od modelu ze wszystkimi zmiennymi, ale z wykluczeniem zmiennych nadmiernie skorelowanych, które zostały zidentyfikowane podczas analizy wstępnej. Następnie, na podstawie wartości statystyki Walda oraz p-value, eliminowano zmienne nieistotne statystycznie. Jako poziom istotności przyjęto α = 0,1. Hipotezy dla testu z wykorzystaniem statystyki Walda prezentują się następująco:

Parametry modelu 1 estymowano bez zmiennych sysBP, BMI oraz currentSmoker – zmienne wyeliminowane z powodu nadmiernej korelacji. Uzyskane wyniki zaprezentowano na zdjęciu 5.1. Zmiennych istotnych statystycznie, przy przyjętym poziomie istotności równym 0,1, jest 7 – są to zmienne: sex, age, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, glucose. Jedna kategoria zmiennej catBMI jest istotna statystycznie (catBMIwcyhudzenie), ale pozostałe kategorie tej zmiennej nie są istotne statystycznie, więc całościowo zmienna ta nie jest istotna. Kryterium AIC dla modelu 1 wynosi 2293,3.

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Oszacowane parametry dla modelu 1

Dla modelu 1 obliczono także współczynnik VIF, który informuje o współliniowości predyktorów – brak współliniowości jest jednym z założeń modelu GLM. Na zdjęciu 5.2 zaprezentowano obliczone współczynniki VIF dla wszystkich zmiennych z modelu 1. Wszystkie wartości wynoszą ok. 1-2 i oznacza to, że predyktory nie są współliniowe (jako wartość graniczną dla VIF przyjmujemy 4).

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Współczynnik VIF dla predyktorów modelu 1

Model 2 zbudowano bez zmiennych sysBP, BMI oraz cigsPerDay – jest to drugi model ze wszystkimi zmiennymi, poza zmiennymi nadmiernie skorelowanymi, ale w tym przypadku wyeliminowano zmienną cigsPerDay zamiast zmiennej currentSmoker. Taki zabieg ma na celu porównanie, który model jest lepszy. Oszacowane parametry dla modelu 2 zaprezentowano na zdjęciu 5.3. Przy przyjętym poziomie istotności α, 6 zmiennych jest istotnych statystycznie – są to: sex, age, currentSmoker, BPMeds, prevalentStroke, prevalentHyp. Kryterium AIC dla modelu 2 wynosi 2308,6 i jest większe niż dla modelu 1, więc zdecydowano się wybrać model 1 jako model „bazowy” bez zmiennych nadmiernie skorelowanych.

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Oszacowane parametry dla modelu 2

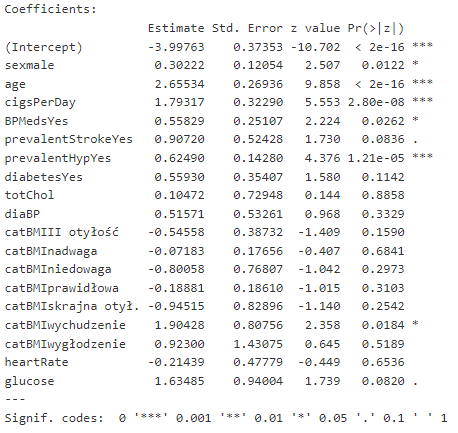
Dla modelu również obliczono współczynniki VIF i wyniki zaprezentowano na zdjęciu 5.4. W przypadku modelu 2 predyktory również nie są współliniowe, ponieważ wartości nie przekraczają nawet 2.

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Współczynnik VIF dla predyktorów modelu 2

Na podstawie wartości p-value z modelu 1 zauważono, że zmienne education, diabetes, totChol, diaBP, heartRate oraz catBMI są nieistotne statystycznie. W związku z tym kolejnym krokiem jest estymacja parametrów modelu 1 bez zmiennej education i sprawdzenie istotności parametrów. Na zdjęciu 5.5 zaprezentowano oszacowane parametry dla modelu 3. Widać, że zmienne sex, age, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp oraz glucose są istotne statystycznie, ponieważ ich wartości p-value są mniejsze od założonego poziomu istotności 0,1.



Zdjęcie 5. . Oszacowane parametry dla modelu 3

W dalszym ciągu zbyt wiele parametrów jest nieistotnych statystycznie i w kolejnym kroku należy odrzucić następną zmienną. Kryterium AIC dla modelu 3 ma wartość 2289,4, więc jest to mniejsza wartość niż dla modelu 2 – oszacowany model bez zmiennej education jest lepszy. Zdecydowano się na oszacowanie modelu 4 dodatkowo bez zmiennej catBMI. Oszacowane parametry dla modelu 4 zaprezentowano na zdjęciu 5.6. Zmienne istotne statystycznie to: sex, age, cigsPerDay, BPMeds, prevalentHyp oraz glucose. W dalszym ciągu aż 5 zmiennych jest nieistotnych statystycznie. Wartość kryterium AIC dla modelu 4 wynosi 2286,6, więc po raz kolejny oszacowano lepszy model niż poprzedni.

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Oszacowane parametry dla modelu 4

Model 5 oszacowano dodatkowo bez zmiennej totChol. Oszacowane parametry przedstawiono na zdjęciu 5.7. Zmienne istotne statystycznie to: sex, age, cigsPerDay, BPMeds, prevalentHyp oraz glucose. Wartość kryterium AIC dla modelu 5 wynosi 2284,6. W dalszym ciągu w modelu znajdują się 4 zmienne nieistotne statystycznie.

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Oszacowane parametry dla modelu 5

Model 6 został oszacowany bez zmiennej heartRate. Oszacowane parametry przedstawiono na zdjęciu 5.8. Zmienne istotne statystycznie to: sex, age, cigsPerDay, BPMeds, prevalentHyp oraz glucose. W modelu 3 znajdują się 3 zmienne nie istotne statystycznie. Wartość kryterium AIC dla modelu 6 wynosi 2282,8.

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Oszacowane parametry dla modelu 6

Model 7 oszacowano bez zmiennej diaBP. Na zdjęciu 5.9 zaprezentowano parametry strukturalne dla tego modelu. Zmienne istotne statystycznie to sex, age, BPMeds, prevalentHyp oraz glucose. W modelu 7 tylko 2 zmienne są nieistotne statystycznie a wartość kryterium AIC wynosi 2281,6.

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Oszacowane parametry dla modelu 7

Model 8 został oszacowany bez zmiennej diabetes. Na zdjęciu 5.10 przedstawiono oszacowane parametry strukturalne modelu. Zmienne sex, age, cigsPerDay, BPMeds, prevalentHyp oraz glucose są istotne statystycznie. W modelu pozostała tylko jedna zmienna nieistotna statystycznie i wyeliminowano ją w następnym modelu. Kryterium AIC dla modelu 8 wynosi 2281,8 – jest to wartość minimalnie większa niż dla modelu 7.

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Oszacowane parametry dla modelu 8

Model 9 estymowano bez zmiennej prevalentStroke. Oszacowane parametry strukturalne zaprezentowano na zdjęciu 5.11. Istotne statystycznie są wszystkie parametry poza. Kryterium AIC ma wartość 2282,1 i jest to wartość większa niż dla modelu 8, więc ten model jest gorzej dopasowany.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Zdjęcie 5. . Oszacowane parametry dla modelu 9

W programie RStudio oszacowano więcej modeli, ponieważ do modelu 9 krokowo dodawano wcześniej wyeliminowane zmienne, ale w innej kolejności. Wszystkie te modele były gorzej dopasowane niż już oszacowano modele.

Dodatkowo próbowano oszacować modele z interakcjami, ale wyniki nie były zadowalające, więc zrezygnowano z ich prezentacji w projekcie.

### Ocena istotności parametrów strukturalnych modeli

Do oceny istotności parametrów strukturalnych modeli stosuje się test ilorazu wiarygodności oraz test Walda. W programie RStudio wykonano testy dla oszacowanych modeli. W celu zobrazowania toku myślenia zdecydowano się na zaprezentowanie wyników testu ilorazu wiarygodności dla modelu 7 na zdjęciu 5.12 oraz wyników testu Walda dla tego samego modelu na zdjęciu 5.13. Na podstawie wartości p-value dla testu ilorazu wiarygodności, która wynosi 0, odrzucono hipotezę zerową na rzecz hipotezy alternatywnej, która mówi, że wszystkie zmienne są istotne statystycznie.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Zdjęcie 5. . Wyniki testu ilorazu wiarygodności dla modelu 7

Wynik testu Walda pokrywa się z wynikami testu ilorazu wiarygodności i również informuje o istotności wszystkich parametrów.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Zdjęcie 5. . Wyniki testu Walda dla modelu 7

### Ocena dopasowania oszacowanych modeli

Ocena dopasowania modelu jest jednym z etapów statystycznej oceny modelu i wykorzystuje się do tego miary takie jak:

* Kryteria informacyjne – w naszym modelu będziemy posługiwać się kryterium informacyjnym Akaike (AIC). Im mniejszą wartość przyjmuje kryterium, tym model jest lepiej dopasowany do danych empirycznych,
* Miary pseudo-R2 – miar tego typu istnieje bardzo dużo, ale na potrzeby projektu wykorzystano miarę McFaddena oraz miarę Cragga-Uhlera. Im większa wartość miar pseudo-R2, tym model jest lepiej dopasowany,
* Statystyka odchylenia – im większa wartość statystyki, tym model gorzej dopasowany.

Dla wszystkich oszacowanych modeli obliczono wymienione powyżej miary i wyniki zaprezentowano na zdjęciu 5.14. Najlepiej dopasowanym modelem jest model 7, ponieważ ma najmniejsza wartość kryterium AIC. Jeżeli chodzi o miary pseudo-R2 to najwyższe wartości przyjmują dla modelu 1. W pracy jednak kierowano się głownie kryterium informacyjnym AIC.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Zdjęcie 5. . Miary dopasowania modeli logitowych

### Ocena jakości predykcji modelu

Kolejnym etapem jest ocena jakości predykcji modelu. Wykorzystano następujące miary oparte na tablicy trafności, która została zaprezentowana na zdjęciu 5.15, a jako punkt odcięcia p\* przyjęto proporcje z próby uczącej:

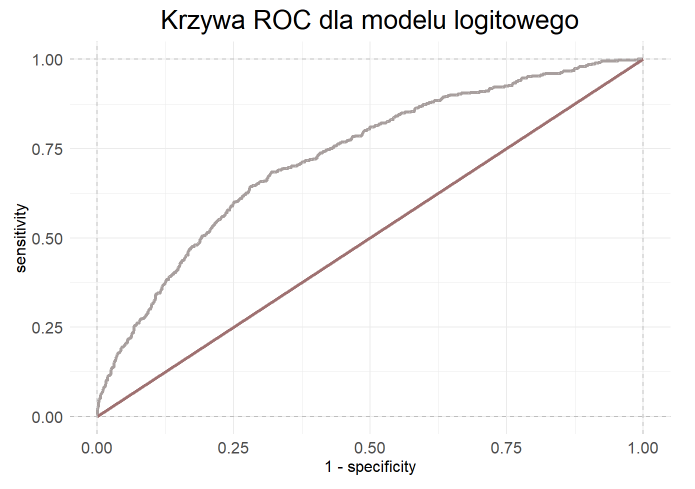
* ACC - zliczeniowy R2 (Accurancy)
* ER – wskaźnik błędu
* SENS (czułość)
* SPEC (swoistość)
* PPV (dodatnia zdolność predykcyjna)
* NPV (ujemna zdolność predykcyjna)



Zdjęcie 5. . Miary jakości predykcji dla modelu 7

Model ten można użyć w praktyce, ponieważ odznacza się stosunkowo wysokim wskaźnikiem ACC (zliczeniowy R2=67,3%) przy niskim ERR (wskaźnik błędu=32,7%). Ze względu na wyższy wskaźnik swoistości od czułości oraz NPV od PPV, model ten lepiej przewiduje 0 (niewystępowanie zjawiska).

Drugim sposobem oceny jakości predykcji jest krzywa ROC oraz pole powierzchni pod krzywą ROC (AUC). Krzywa ROC zawiera więcej informacji niż tablica trafności, ponieważ przedstawia siłę predykcji dla wszystkich możliwych punktów odcięcia *p*\*. Wykres krzywej ROC zaprezentowano na zdjęciu 5.16.



Zdjęcie 5. . Wykres krzywej ROC dla modelu 7

Pole pod krzywą ROC wynosi 0,729 i oznacza to, że model w sposób akceptowalny wyjaśnia badane zjawisko.

## Budowa modelu probitowego

Model probitowy opiera się na odwrotności funkcji dystrybuanty rozkładu normalnego, czyli na probicie. Oszacowano model probitowy dla takich samych zmiennych niezależnych, jakich użyto dla najlepszego modelu logitowego, czyli dla modelu 7.

Oszacowane parametry strukturalne dla modelu probitowego przedstawiono na zdjęciu 5.17. Zmienne istotne statystycznie to: sex, age, cigsPerDay, BPMeds, prevalentHyp oraz diabetes. Zmienna diabetes nie była istotna w modelu 7 logitowym, ale była istotna zmienna glucose, która nie jest istotna statystycznie w modelu probitowym. Interpretacja modelu logitowego polega na określeniu, które zmienne są stymulantami, a które destymulantami. W oszacowanym modelu probitowym wszystkie zmienne mają dodatnie parametry strukturalne, więc wszystkie są stymulantami.

Obraz zawierający stół

Opis wygenerowany automatycznie

Zdjęcie 5. . Oszacowane parametry dla modelu probitowego

## Porównanie modelu logitowego i probitowego

Uzyskane modele, czyli najlepszy model logitowy (model 7) i model probitowy, porównano biorąc pod uwagę miary oceny dopasowania (kryterium AIC oraz miary pseudo-R2). Wyniki użyte do porównania zamieszczono na zdjęciu 5.18. Model probitowy ma mniejszą wartość kryterium AIC oraz większe miary pseudo-R2, więc jest lepiej dopasowany do danych empirycznych. Model logitowy ma jednak ciekawszą interpretację a różnice w dopasowaniu są niewielkie, więc ostatecznie do interpretacji wybrano model logitowy.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Zdjęcie 5. . Miary dopasowania modelu logitowego i probitowego

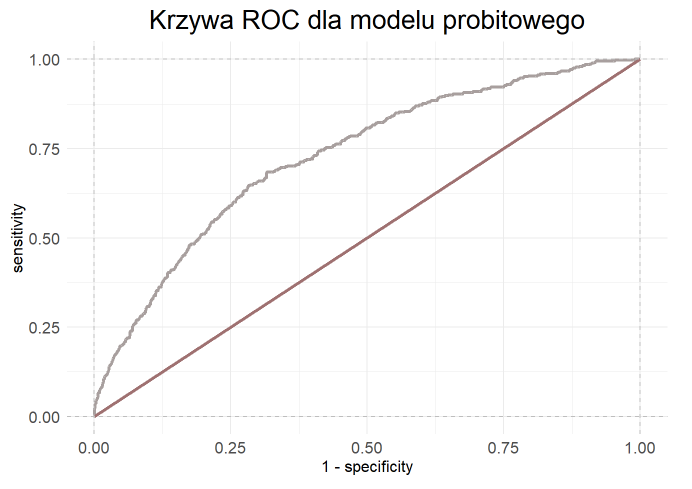
Model logitowy i probitowy porównano również pod względem jakości predykcji. Miary oparte na tablicy trafności zaprezentowano na zdjęciu 5.19. Model logitowy ma nieco większe wartości ACC, czyli oznacza to, że minimalnie lepiej przewiduje wartości zmiennej objaśnianej. Warto jednak podkreślić, że te różnice są minimalne.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Zdjęcie 5. . Porównanie modeli pod względem jakości predykcji na podstawie miar opartych na tablicy trafności

Krzywą ROC dla modelu logitowego zaprezentowano powyżej na zdjęciu 5.16 a krzywą dla modelu probitowego zaprezentowano na zdjęciu 5.20. Pole pod krzywą ROC dla modelu probitowego wynosi 0,7293, więc model w sposób akceptowalny wyjaśnia badane zjawisko. Pola pod krzywymi ROC dla modelu probitowego i logitowego sa takie same.



Zdjęcie 5. . Krzywa ROC dla modelu probitowego

## Interpretacja modelu logitowego

Model logitowy interpretuje się korzystając z pojęcia ilorazu szans. Z uwagi na postać modelu logitowego, do interpretacji oszacowanych parametrów niezbędne jest obliczenie wartości eβi.

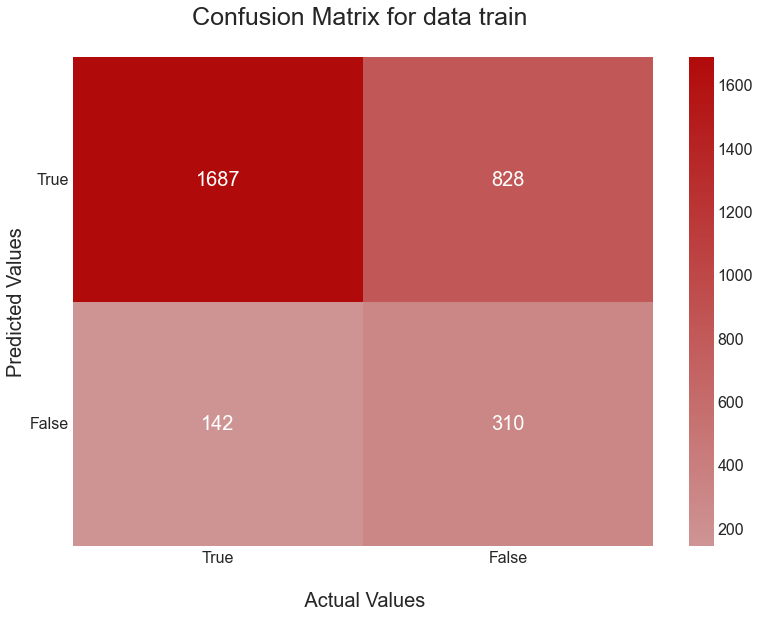
Interpretacja parametrów dla najlepszego modelu:

* exp (-4,006) = 0,0183 -> jest to szansa na zachorowanie na chorobę serca w grupie referencyjnej, więc w przypadku szacowanego modelu ten parametr nie ma interpretacji, ponieważ nie badaliśmy osób w wieku 0 z zerowym poziomem glukozy,
* exp (0,3579) = 1,43 -> szansa na zachorowanie na chorobę serca jest o ok. 43% wśród mężczyzn niż wśród kobiet, ceteris paribus,
* exp (2,6516) = 14,177 -> wraz ze wzrostem wieku o 1 rok, szansa zachorowanie na chorobę serca zwiększa się 14-krotnie, ceteris paribus,
* exp (1,7161) = 1,563 -> wraz ze wzrostem dziennej liczby wypalanych papierosów o jedną sztukę, szansa zachorowania na chorobę serca wzrasta o ok. 56%, ceteris paribus,
* exp (0,5686) = 1,766 -> osoby leczące się na nadciśnienie mają mniejszą szansę na zachorowanie na chorobę serca niż osoby nieleczące się o ok. 43%, ceteris paribus,
* exp (0,7002) = 2,028 -> osoby mające nadciśnienie mają o ok. 100% większe szansa na zachorowanie na chorobę serca niż osoby, które nie mają tego schorzenia, ceteris paribus,
* exp (1,5542) = 1,732 -> wraz ze wzrostem poziomu glukozy we krwi o 1 jednostkę, szansa na zachorowanie na chorobę serca wzrasta o ok. 73%, ceteris paribus.

# Podsumowanie i wnioski

Na ryzyko zachorowania na chorobę serca wpływa wiek, płeć, występowanie chorób takich jak cukrzyca, nadciśnienie. Model objaśniający ryzyko zachorowania na chorobę serca może być pomocny w określaniu grupy ryzyka oraz zapobieganiu chorobie. Osoby znajdujące się w grupie ryzyka powinny być pod stałym nadzorem lekarza.

Oszacowany model logitowy w 67% poprawnie przewidział wartości zmiennej objaśnianej, czyli 10-letnie ryzyko zachorowania na chorobę serca. Poprawnym modelem jest również model probitowy, ale jego interpretacja nie jest tak precyzyjna jak modelu logitowego. Na zdjęciu 6.1 zaprezentowano rozkład predykcji dla modelu 7 logitowego dla danych uczących. Liczba poprawnie oszacowanych 1 wynosi 310, a liczba poprawnie oszacowanych 0 wynosi 1687.



Zdjęcie 6. . Rozkład predykcji dla modelu 7

Zmienne objaśniające, które są istotne statystyczne zgodnie z oszacowanym modelem to zmienne: płeć, wiek, leczenie nadciśnienia, występowanie nadciśnienia, liczba wypalanych dziennie papierosów i poziom glukozy we krwi.