Ocena przydatności wskaźników wagowo – wzrostowych i analiza wybranych modeli statystycznych zastosowanych do wyznaczania czynników ryzyka występowania nieprawidłowego ciśnienia tętniczego krwi u dzieci

Zakład Informatyki i Statystyki
Uniwersytetu Medycznego
im. Karola Marcinkowskiego
w Poznaniu

Praca przygotowana
jako rozprawa doktorska
Promotor: prof. dr hab. Jerzy Moczko
Autor: mgr inż. Anna Sowińska

Poznań, 2011

Gorące podziękowania
dla Pani Prof. dr hab. Alicji Krzyżaniak
za wsparcie i ogromną pomoc
w trakcie powstawania pracy doktorskiej

SPIS TREŚCI

Spis oznaczeń i symboli:

A – atrybut

ACC (accuracy) – skuteczność

AUC (area under the curve), W – pole pod krzywą

BMI (body mass index) – wskaźnik masy ciała

BP (blood pressure) – ciśnienie krwi

C – zbiór liści w drzewie klasyfikacyjnym

CB – zbiór funkcji bazowych w metodzie MARSplines

CI (confidence interval) – przedział ufności

CO (cardiac output) – rzut serca

$deg_T(v)$ – stopień wierzchołka v grafu G

E – zbiór krawędzi

GCV (generalized cross validation) – błąd uogólnionego sprawdzianu krzyżowego

$h_m(x)$ – funkcja bazowa

L – funkcja wiarygodności

LR – iloraz wiarygodności

L^T – zbiór liści drzewa T

n – węzeł drzewa T

N_r^T – zbiór węzłów drzewa T

OR (odds ratio) – iloraz szans

P (prevalence) – częstość występowania wyróżnionego wydarzenia

$p(A)$ – prawdopodobieństwo wystąpienia zdarzenia A

$p(k/s)$ – prawdopodobieństwo wystąpienia k pod warunkiem S

$P(Y=1/x_1,x_2,...x_n)$ – model regresji logistycznej

$q(S)$ – miara różnorodności

r – korzeń drzewa

r_p – współczynnik korelacji dla krzywych ROC

s – węzeł sąsiadujący

$S(A)$ – szansa dla przypadku A

SE_W – błąd standrdowy

S_t – zbiór wartości testu t

T – drzewo klasyfikacyjne

t_n – test węzła n

TPR (total peripherial resistance) – całkowity opór obwodu naczyniowego

V – zbiór wierzchołków

WMC (body mass coefficient) – współczynnik masy ciała

WQ – współczynnik Quetelet'a

WR – wskaźnik Rohrer'a

$x_1,x_2,...,x_n$ – zmienne niezależne

Y – zmienna dychotomiczna

Z – test Wald'a

1. Wstęp

Nadciśnienie tętnicze jest chorobą powszechnie występującą w społeczeństwie wysoko uprzemysłowionym. W Polsce jak wynika z badań NATPOL III PLUS z 2002 roku na nadciśnienie tętnicze choruje 29% dorosłych Polaków, u 30% stwierdza się tzw. wysokie prawidłowe ciśnienie tętnicze, a u 20% niskie prawidłowe. Tylko 21% Polaków charakteryzuje się prawidłowymi wartościami ciśnienia tętniczego (wartości optymalne występują u co piątej osoby powyżej 18 roku życia).

Nadciśnienie tętnicze częściej postrzegane jest jako problem zdrowotny osób dorosłych niż dzieci lub młodzieży. Powodem takiego stanu rzeczy może być mniejsze doświadczenie pediatrów w tej problematyce, niż w zakresie częstych chorób dziecięcych. Jednak coraz więcej autorów jak również dane epidemiologiczne zwracają uwagę na możliwość wystąpienia samoistnego nadciśnienia tętniczego już w okresie dzieciństwa. Wskazują na zwiększenie się liczby dzieci i młodzieży z podwyższonymi wartościami ciśnienia, które często współistnieją z otyłością, paleniem papierosów i małą aktywnością fizyczną. Precyzyjne ustalenie częstości występowania nadciśnienia tętniczego w populacji dzieci i młodzieży nastręcza szereg trudności. Wynikają one z trudności określenia uznanych i akceptowanych norm ciśnienia dla tej grupy wiekowej, jak również powodem jest to, że w przedziale wiekowym 0-17 lat występuje szereg podgrup wiekowych, dla których wartości ciśnienia uznane za prawidłowe są różne. Wynika to oczywiście z naturalnego rozwoju dziecka, a także związanej z tym akceleracji związanej z okresem dojrzewania. Z tego względu oprócz uwzględnienia wieku metrykalnego koniecznym staje się odnoszenie wartości ciśnienia tętniczego do parametrów antropometrycznych, a więc głównie masy ciała i wzrostu [2, 5].

Wiadomo, że brak kontroli ciśnienia krwi i prawidłowej terapii u dorosłych stanowi przyczynę wysokiej zachorowalności na zawał, chorobę niedokrwienną serca, czy udar mózgu i te niepokojące skutki zdrowotne nadciśnienia tętniczego powinny mobilizować służbę zdrowia nie tylko do wczesnego rozpoznawania, lecz także do wdrażania działań profilaktycznych [1, 2].

2. Ciśnienie tętnicze

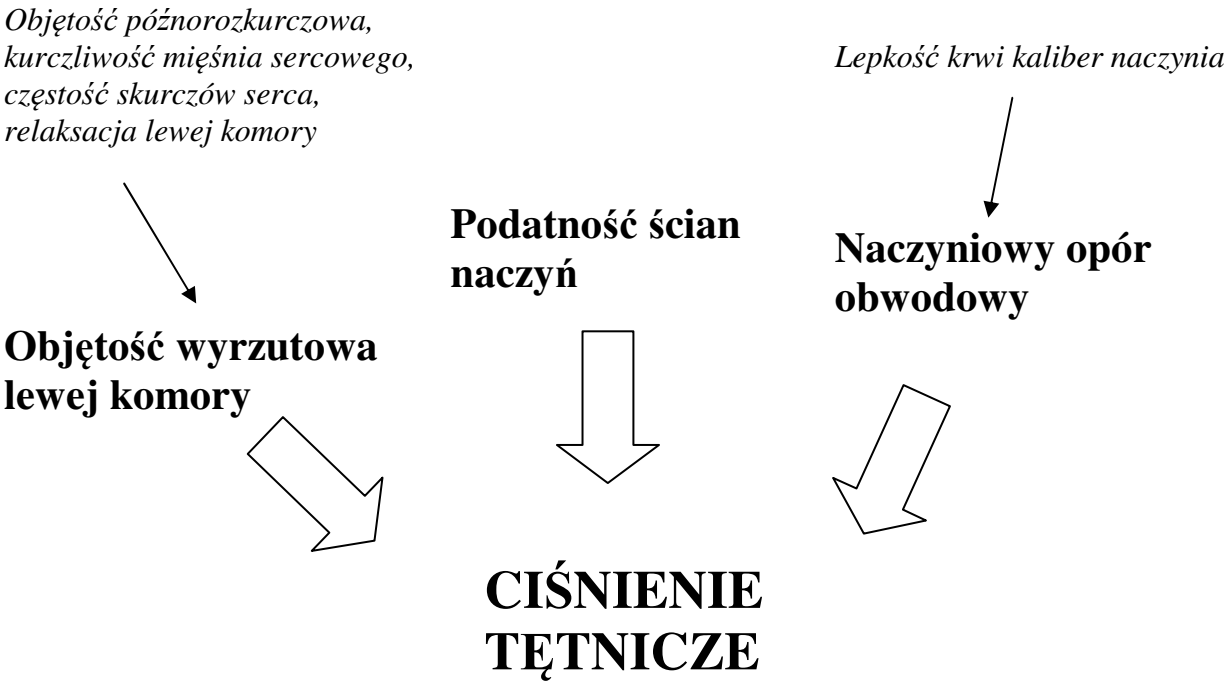
Ciśnienie krwi opisuje się następującym wzorem

$$BP = CO \cdot TPR$$

gdzie, BP (blood pressure) jest to ciśnienie krwi, CO (cardiac output) jest to rzut serca, TPR (total peripherial resistance) jest to całkowity opór obwodu naczyniowego.

Analizując powyższy wzór widzimy, że ciśnienie krwi zależy od rzutu serca i całkowitego obwodowego oporu naczyniowego. Rzut serca jest z kolei wypadkową działania objętości wyrzutowej, czyli objętości krwi wyrzucanej z lewej komory do aorty, w ciągu 1 minuty oraz częstości skurczów serca, która zmienia się u dzieci z wiekiem. Natomiast opór naczyniowy jest zależny od czynnika geometrycznego, czyli promienia naczynia i jego długości oraz lepkości krwi. Receptory wpływają w sposób pośredni na opór naczyniowy poprzez zmianę promienia naczynia. Objętość wyrzutowa lewej komory jest zmienną zależną od kurczliwości mięśnia sercowego i objętości późnorozkurczowej lewej komory. Kurczliwość zależy od zachowania struktury i czynności kardiomiocytów oraz wpływu unerwienia [2, 3, 4].

Rysunek 1. Czynniki determinujące wysokość ciśnienia tętniczego



Wśród mechanizmów regulujących ciśnienie krwi duży udział ma układ nerwowy regulujący napęd autonomiczny i odruchową kontrolę ciśnienia tętniczego oraz nerki – narząd efektorowy regulacji gospodarki wodno – elektrolitowej i narząd endokrynny.

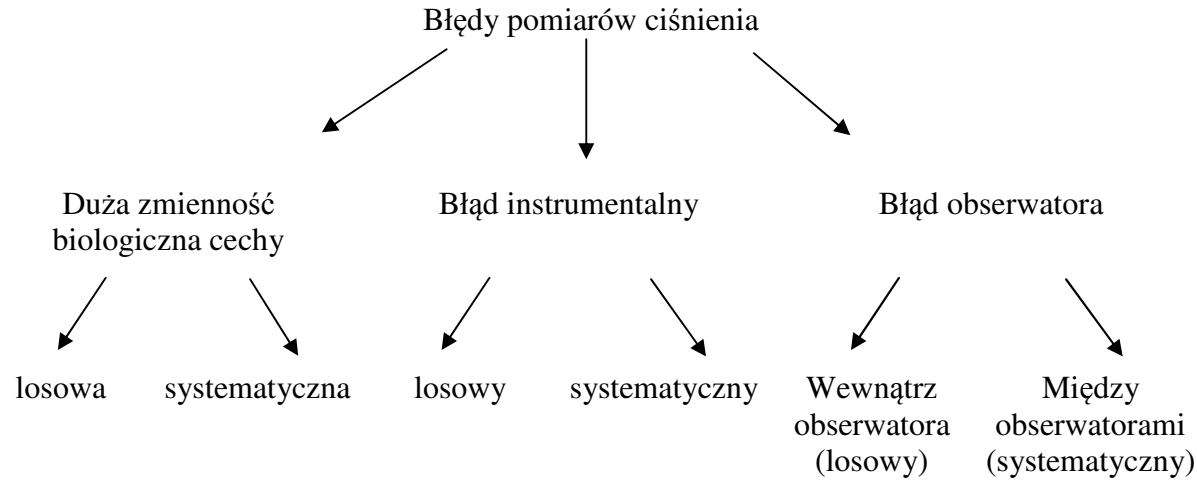
Wartości ciśnienia tętniczego u dzieci są niższe niż u dorosłych i zwiększają się z wiekiem i rozwojem dziecka. Donoszonemu noworodkowi przez pierwszy tydzień życia ciśnienie skurczowe równolegle z rozkurczowym zwiększa się około 1 mm Hg dziennie. W następnych okresach ciśnienie zwiększa się nadal, ale jego tempo jest wolniejsze. U dzieci urodzonych przedwcześnie z niską masą urodzeniową ciśnienie wzrasta szybciej, a wartości ciśnienia wyrównują się z dziećmi urodzonymi o czasie około 4 miesiąca życia i w następnych latach są odwrotnie skorelowane z urodzeniową masą ciała oraz niezależnie z obwodem główki po urodzeniu. To oznacza, że osoby, które urodziły się z małą urodzeniową masą ciała mają na ogół wyższe ciśnienie tętnicze niż osoby urodzone z prawidłową wagą.

Wykazano zależność między wzrostem, szybkością wzrastania, a wysokością ciśnienia tętniczego. Wartości ciśnienia tętniczego u dziewczynek w wieku 7-11 lat mogą być wyższe niż u chłopców, ponieważ u nich wcześniej zaczynają się procesy dojrzewania płciowego. U chłopców natomiast ciśnienie zwiększa się w wieku 12-17 lat, czyli w okresie ich najszybszego wzrostu i dojrzewania płciowego. Już w okresie dzieciństwa dochodzi do zajęcia określonego kanału centylowego wartości ciśnienia tętniczego, który będzie charakterystyczny dla danej osoby. Takie zjawisko nazywane jest utrzymywaniem toru rozwojowego (tracking) [3].

2.1. Pomiar ciśnienia w gabinecie lekarskim

Pomiar ciśnienia tętniczego jest najważniejszym badaniem niezbędnym do rozpoznania nadciśnienia tętniczego. Jeżeli pomiar ma mieć znaczenie kliniczne musi być wykonany bardzo starannie. Zalecenia dotyczące techniki wykonywania pomiarów ciśnienia u dzieci mają na celu wyeliminowanie błędu pomiaru, a stosowanie standardowej techniki pozwala na porównywanie badań. Błędy, które powstają podczas pomiaru ciśnienia mogą wynikać z dużej zmienności biologicznej tej cechy, błędów wynikających z nieprawidłowego sprzętu oraz błędów, które popełnia lekarz lub inna osoba mierząca ciśnienie.

Rysunek 2. Błędy pomiarów



Ciśnienie tętnicze powinno być oznaczone rano po dobrze przespanej nocy, po oddaniu moczu, przed śniadaniem. Wynik dokonanego w tej sytuacji pomiaru odpowiada tzw. ciśnieniu podstawowemu. Warunków tych nie spełnia pomiar dokonany w gabinecie lekarskim, dlatego powinien on zatem być dokonany trzykrotnie podczas wizyty w odstępach kilkuminutowych.

W czasie wizyty lekarskiej pomiaru należy dokonać u pacjenta wypoczętego. Okres wypoczynku w ciepłym pomieszczeniu, w wygodnej pozycji ze wstrzymaniem się od jedzenia i palenia papierosów powinien wynosić 30 min. Pomiaru dokonuje się u pacjenta siedzącego na krześle z wygodnym oparciem. Pacjent powinien usiąść 5-10 min przed dokonaniem pomiaru. Odwiedzione, całkowicie odkryte ramię (nie uciśnięte przez podwinięty rękaw) powinno być dobrze podparte (aby uniknąć napięcia mięsni) i ułożone tak aby mankiet aparatu można było umieścić „na wysokości serca". Pomiaru dokonuje się na prawym ramieniu, ponieważ pomiar na tym ramieniu służył do ustalenia norm odniesienia. Podczas pierwszej wizyty pomiar powinien być dokonany na obydwu ramionach, ewentualną różnicę (>=5 mm Hg) notujemy w karcie pacjenta. W czasie następnych wizyt mierzymy ciśnienie na tym ramieniu, na którym było ono wyższe. Pomiaru najlepiej dokonywać manometrem rtęciowym [3].

2.2. Aparaty do pomiaru ciśnienia

1. Manometry rtęciowe: tego typu aparaty są zaliczane do przyrządów dokładnych i stosowane są do opracowania norm ciśnienia. W czasie pomiaru aparat powinien znajdować się w pozycji pionowej, skala powinna być dokładnie wykalibrowana od 0 do 300 mm Hg z podziałką co 2 mm. Menisk rtęci przed pomiarem powinien znajdować się pozycji zero.
2. Manometry sprężynowe – aneroidy, są one mniej dokładne niż aparaty rtęciowe. Różnice w pomiarach między manometrem rtęciowym, a sprężynowym nie powinny być większe niż 2 mm Hg.
3. Manometry elektroniczne (półautomatyczne) – są mniej precyzyjne, często stosowane w pomiarach domowych. Zaleca się sprawdzanie ich dokładności za pomocą manometru rtęciowego.
4. pomiary ciśnienia metodą ultrasonograficzną (ultradźwiękową) – jest to metoda z wyboru stosowana na oddziałach intensywnej terapii i u małych dzieci oraz niemowląt (tony Korotkowa są u nich słabo słyszalne) [2].

2.2.1. Mankiety

Bardzo ważne znaczenia dla uniknięcia błędu pomiaru ma wybór odpowiedniego rozmiaru mankietu. Prawidłowo dobrany mankiet musi obejmować co najmniej 80% obwodu ramienia i całą stronę dłoniową. Jego szerokość powinna odpowiadać 2/3 długości ramienia licząc od wyrostka barkowego łopatki do wyrostka łokciowego. Połowa szerokości mankietu powinna znajdować się w połowie długości ramienia. Mankiet powinien być ułożony gładko i ściśle przylegać do ramienia, ale go nie uciskać. Zbyt wąski mankiet powoduje znacznie większy błąd pomiaru niż mankiet zbyt szeroki. Mankiet należy w ciągu około 30 sekund napełnić powietrzem do wartości wyżej o 20-30 mm Hg od uzyskanej w chwili zaniku tętna na tętnicy promieniowej, a następnie opróżnić z powietrza z szybkością 2-3 mm Hg/s. Pojawienie się pierwszego stukającego tonu (I faza Korotkowa) odpowiada ciśnieniu skurczowemu, a zanik tonów (V faza Korotkowa) ciśnieniu rozkurczowemu.

Prawidłowo wykonany pomiar u współpracującego pacjenta obarczony jest błędem 2-3 mm Hg. Wyniku pomiaru nie należy zaokrąglać do 10, a podawać z dokładnością do 1 mm Hg.

Tabela 1. Zalecane wielkości mankietów dla dzieci i młodzieży [4] str. 99

Wielkość mankietu	**Szerokość (cm)**	**Długość (cm)**
Noworodki	2,5-4,0	5,0-9,0
Niemowlęta	4,0-6,0	11,5-18,0
Dzieci	7,5-9,0	17,0-19,0
Młodzież	11,5-13,0	22,0-26,0
Szerokie ramię (dorośli z prawidłową masą ciała)	14,0-15,0	30,0-33,0
Grube ramię (dorośli z nadwagą i otyłością)	18,0-19,0	36,0-38,0

Tabela 2. Przyczyny zawyżonej wartości pomiaru ciśnienia tętniczego

- Brak podparcia pleców i/lub ramienia(zwiększone napięcie mięśni)
- Zbyt wąski mankiet aparatu (do 30% błędu)
- Nadmierne wypełnienie mankietu powietrzem (ból – reakcja presyjna)
- zbyt wolne wypuszczanie powietrza z mankietu (zastój żylny – wpływ głównie na ciśnienie rozkurczowe)
- Niepokój, krzyk (niemowlęta)

2.3. Domowy pomiar ciśnienia tętniczego

Przygodny pomiar ciśnienia tętniczego w gabinecie lekarskim bywa obarczony błędem. Jego przyczyną może być zmęczenie pacjenta spowodowane dojazdem do przychodni, czekaniem na wizytę itp. Różnice w wynikach pomiaru mogą być też spowodowane dokonywaniem ich w różnych godzinach przez różne osoby. U znacznego odsetka pacjentów, zwłaszcza tych najmłodszych, wartości ciśnienia tętniczego w pomiarach dokonywanych przez personel medyczny są zawsze wyższe niż w warunkach domowych. Jest to tzw. „nadciśnienie białego fartucha" lub „efekt białego fartucha". Przyczynę tego zjawiska przepisuje się lękowi przed konsekwencjami stwierdzenia wysokich wartości ciśnienia (hospitalizacja, dodatkowe badania, zmiany w trybie życia i diecie) lub lęk przed samym

badaniem. „Efekt białego fartucha" nabiera szczególnego znaczenia u pacjentów z granicznymi wartościami ciśnienia, u których może stać się przyczyną rozpoznania nadciśnienia tętniczego.

Dużą pomocą w rozpoznawaniu lub wyłączeniu efektu białego fartucha są pomiary ciśnienia w warunkach domowych, dokonywane przez rodziców, lub w przypadku starszej młodzieży samych pacjentów [3].

2.4. Przyczyny nadciśnienia tętniczego dzieci i młodzieży

W hipertensjologii wieku dorosłego spotykamy się głównie z nadciśnieniem tętniczym pierwotnym. Postać ta stanowi ok. 90% wszystkich przypadków nadciśnienia tętniczego. Nieco inna sytuacja odnosi się do nadciśnienia w młodej populacji. Według różnych autorów w populacji dzieci i młodzieży do 10 roku życia dominuje nadciśnienie wtórne. Wśród przyczyn je wywołujących są głównie choroby nerek, koarktacja aorty, choroby gruczołów wydzielania wewnętrznego, a także inne rzadsze przyczyny.

Tabela 3. Przyczyny nadciśnienia tętniczego u dzieci

wiek	**przyczyna**
1-6 r. ż.	Choroba miąższu nerek (zmiany strukturalne, zapalne, guzy), koarktacja aorty, zwężenie tętnicy nerkowej, hiperkalcemia, choroby tarczycy, nadmiar mineralokortykoidów
6-10 r. ż.	Zwężenie tętnicy nerkowej, choroby miąższu nerek, nadciśnienie pierwotne, choroby tarczycy, guz chromochłonny, neurofibromatoza i inne (ww.)
powyżej 10 r. ż.	Nadciśnienie pierwotne, choroby miąższu nerek, inne (ww.)

Jest zasadą, że im mniejsze dziecko tym większe prawdopodobieństwo nadciśnienia wtórnego. Nadciśnienie pierwotne, choć może się ujawniać także bardzo wcześnie, to jednak zaczyna dominować po 10 roku życia [5].

3. Regresja logistyczna

Regresja logistyczna jest matematycznym modelem, którego możemy użyć w celu opisania wpływu kilku zmiennych $x_1, x_2,...x_k$ (zarówno ilościowych, jak i jakościowych) na dychotomiczną zmienną y. Regresja ta jest często wykorzystywana w psychologii, w medycynie, w epidemiologii oraz antropologii, ponieważ wiąże w prosty sposób parametry modelu z ilorazem szans (OR), który jest łatwy w interpretacji.

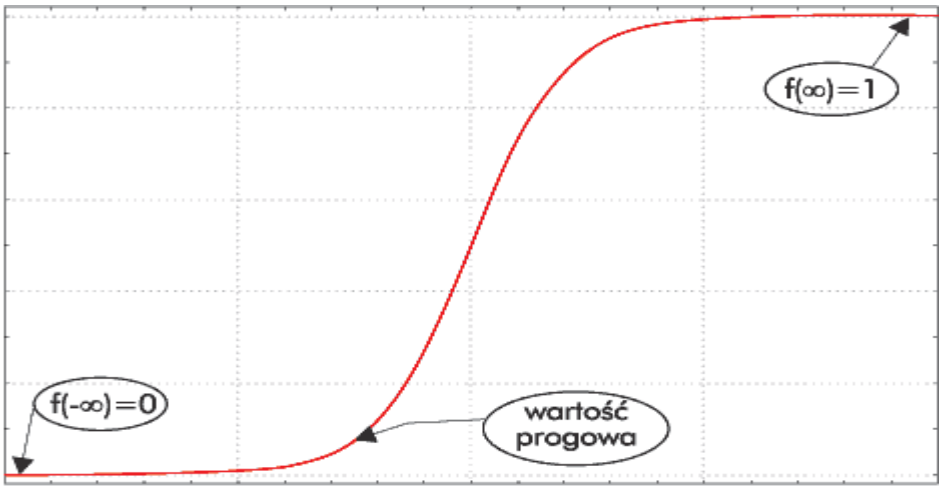
Historia modelu logistycznego ma już wiele lat i sięga XIX wieku. Pionierami byli Verhulst i Pearl, którzy opracowali postać krzywej logistycznej i zastosowali ją w praktyce; jednakże pełny model w postaci stosowanej dzisiaj po raz pierwszy podał i zastosował w 1972 roku Finney.

3.1. Postać funkcji logistycznej

Funkcja logistyczna $f(z)$ jest opisana poniższym wzorem

$$f(z) = \frac{1}{1+e^{-z}}$$

i przyjmuje wartości od 0 do 1. Jej wartość w $-\infty$ jest równa 0, a dla $z = \infty$ jest równa 1.



Kształt funkcji przypomina rozciągniętą literę S. Pokazuje on, że zmiany wartości funkcji są minimalne, jeśli wartości zmiennych są mniejsze od pewnej wartości progowej. Natomiast, gdy próg zostanie przekroczony, wówczas wartość funkcji zaczyna gwałtownie rosnąć osiągają asymptotycznie wartość 1 [20, 21].

3.2. Model logistyczny

Model regresji logistycznej wiążący prawdopodobieństwo jednego z dwóch możliwych wyników zmiennej Y ze zmiennymi wyjaśniającymi jest określony równaniem

$$P(Y=1|x_1,x_2,...,x_n)) = \frac{e^{\left(a_0+\sum_{i=1}^{k} a_i x_i\right)}}{1+e^{\left(a_0+\sum_{i=1}^{k} a_i x_i\right)}}$$

Y – oznacza zmienną dychotomiczną, która przyjmuje *1* najczęściej dla zdarzeń pożądanych np.: przeżycie lub przyjmuje wartość *0* w przeciwnym przypadku np.: zgon.
a_i – są współczynnikami regresji
$x_1, x_2, \ldots, x_k$ – są zmiennymi niezależnymi, które mogą być mierzalne lub jakościowe.

Prawa strona równania to warunkowe prawdopodobieństwo, że zmienna Y przyjmie wartość *1* dla wartości zmiennych niezależnych $x_1, x_2,...,x_k$.
W modelu regresji logistycznej staramy się oszacować współczynniki regresji $a_0, a_1,...,a_k$. Chcemy w ten sposób dopasować jak najlepszy model w oparciu o wartości pewnej grupy danych. Liczebność grupy *n* musi być dostatecznie duża, co oznacza, że $n > 10(k + 1)$, gdzie *k* jest liczbą parametrów. W celu znalezienia estymatorów $a_0, a_1,...,a_k$ stosujemy metodę największej wiarygodności (nie możemy zastosować metody najmniejszych kwadratów, gdyż warunek o stałości wariancji dla zmiennej dychotomicznej *Y* nie jest spełniony). Natomiast do oceny istotności estymatorów będziemy używali testu Wald'a. Obliczamy go dzieląc estymowany współczynnik przez jego błąd standardowy [19]:

$$Z = \frac{a^2}{[SE(a)]^2}$$

Podstawy teoretyczne metody największej wiarygodności opracował twórca analizy wariancji Fisher w 1929 roku. Funkcję wiarygodności definiujemy następującą równością:

$$L = \prod_{i=1}^{n} p\left(y_i | a_1, a_2, ..., a_k\right)$$

gdzie $p(y_i \mid a_1, a_2, \ldots, a_k)$ oznacza prawdopodobieństwo pojawienia się wartości zmiennej zależnej y_i przy danym modelu regresji z parametrami $a_1, \ldots, a_k$. [20, 21].

Metoda największej wiarygodności maksymalizuje funkcję wiarygodności. Oznacza to, że maksymalizuje iloczyn prawdopodobieństw pojawienia się poszczególnych obserwacji z próby przy danych parametrach modelu. Najlepiej jest jako ocenę szacowanych parametrów brać te wartości, dla których wiarygodność jest największa. W długim ciągu doświadczeń im większa wiarygodność (prawdopodobieństwo) zdarzenia, tym większa częstość względna (realizacja). Jeżeli wiarygodność jest mała, to częstość względna wystąpienia zdarzenia jest bliska 0 i wtedy, jeżeli rozważana jest pojedyncza próba możemy w ogóle nie brać pod uwagę możliwości realizacji zdarzenia. Można powiedzieć, że im większa wiarygodność konkretnego modelu, tym większe prawdopodobieństwo, że wartości zmiennej zależnej pojawią się w próbie. Im większa wiarygodność, tym lepsze dopasowanie modelu do danych. Estymatory, które są wyznaczane za pomocą metody największej wiarygodności mają własności, dzięki którym mamy zagwarantowane największe prawdopodobieństwo otrzymania zaobserwowanych wartości zmiennej zależnej. W statystycznych programach komputerowych estymatorów metody największej wiarygodności poszukuje się, maksymalizując funkcję wiarygodności L lub jej logarytm. Z przyczyn obliczeniowych łatwiej jest znaleźć ekstremum funkcji log L niż samej funkcji L [19, 20, 21].

3.3. Iloraz szans

Iloraz szans (*odds ratio*) jest często stosowany w badaniach klinicznych i epidemiologicznych, wraz ze współczynnikami regresji i ich statystyczną istotnością odgrywa ważną rolę w modelu regresji logistycznej.

Aby omówić znaczenie *ilorazu szans* wyjaśnimy na początku pojęcie *szansy.* Szansa jest to stosunek prawdopodobieństwa, że jakieś zdarzenie wystąpi (np. rozwinie się rak płuca), do prawdopodobieństwa, że to zdarzenie nie wystąpi. Dla danego przypadku *A* powyższą definicję możemy zapisać następującym wzorem:

$$Szansa \quad S(A) = \frac{p(A)}{1 - p(A)}$$

Iloraz szans dwóch grup porównywanych *A (grupa narażona na czynnik)* i *B (grupa nienarażona na czynnik)* definiujemy jako stosunek "szansy" wystąpienia *A* do "szansy" wystąpienia *B*, czyli OR_{AxB} (OR od *odds ratio*). Przy tak przyjętym oznaczeniu iloraz szans możemy zapisać w postaci następującego równania:

$$OR_{AxB} = \frac{S(A)}{S(B)} = \frac{p(A)}{1-p(A)} : \frac{p(B)}{1-p(B)}$$

Jeżeli przedstawimy wyniki pomiarów w postaci tabeli:

	Narażenie na czynnik		
	Tak	Nie	Suma
Stan choroby			
Przypadek	a	b	a + b
Kontrola	c	d	c + d
Suma	a + c	b + d	n = a + b + c + d

to iloraz szans definiujemy jako stosunek szansy znalezienia się w grupie narażonej do szansy znalezienia się w grupie nienarażonej.
Szanse znalezienia się w próbach narażonej *(exp)* i nienarażonej *(unexp)* wynoszą:

$$szansa_{\exp} = \frac{\frac{a}{a+c}}{\frac{c}{a+c}} = \frac{a}{c} \qquad szansa_{un\exp} = \frac{\frac{b}{b+d}}{\frac{d}{b+d}} = \frac{b}{d}$$

stąd iloraz szans wynosi:

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \cdot d}{b \cdot c}$$

[19, 20, 21, 22].

Iloraz szans równy 1 oznacza równoważność szans porównywanych grup. Iloraz szans większy od 1 wskazuje, że szansa wystąpienia danego zdarzenia (np. zachorowania, zgonu itp.) w grupie *A* jest większa niż w grupie *B*. Natomiast iloraz szans poniżej jedności wskazuje, że w grupie *A* szansa wystąpienia badanego zdarzenia zdrowotnego jest mniejsza niż w grupie *B*.

Jednostkowy iloraz szans pokazuje, jak zmienia się szansa danego zdarzenia przy jednostkowej zmianie zmiennej niezależnej.

4. Drzewa klasyfikacyjne

Drzewa klasyfikacyjne pojawiły się niezależnie w nauczaniu maszynowym, jak i w statystyce. Struktura drzew decyzyjnych pozwala na konstrukcję najogólniejszych reguł, umożliwiając przy tym niezwykle efektywną ich implementację. Stosuje się je do rozwiązywania problemów z dużą ilością danych. Wykorzystuje się do wyznaczania przynależności przypadków lub obiektów do klas jakościowej zmiennej zależnej na podstawie pomiarów jednej lub większej ilości zmiennych objaśniających. Celem analizy opartej na drzewach klasyfikacyjnych jest przewidywanie tzw. predykcja zmiennej wyjściowej na podstawie zmiennych wejściowych (niezależnych) lub wyjaśnianie odpowiedzi zakodowanych w jakościowej zmiennej zależnej.

4.1. Struktura drzewa

W celu sformalizowanego opisu struktury drzewa posłużono się serią definicji przytoczonych z pracy M. Gromady [6] oraz z podręcznika J. Koronackiego i J. Ćwika [7].

Definicja 1. *Drzewem nazywamy dowolny spójny graf acykliczny*
Krawędzie takiego grafu nazywane są *gałęziami*. Wierzchołki, z których wychodzi co najmniej jedna krawędź nazywamy *węzłami*. Wierzchołki nie będące węzłami nazywamy *liśćmi.*

Rozpatrzmy drzewo $T=<V, E>$ o zbiorze wierzchołków V i krawędzi E [6, 7, 9]. W zbiorze V wyróżniamy podzbiór wierzchołków $L^T \subset V$ będących liśćmi drzewa T . Wykorzystując pojęcie stopnia wierzchołka zapisujemy

$L^T:=\{v\in V : deg_T(v) = 1\}$

$deg_T(v)$ – stopień wierzchołka v grafu G

Ustalmy wierzchołek $r\in V$ drzewa T i nazwijmy go korzeniem drzewa T. Oznaczmy przez L_r^T zbiór:

$L_r^T:=L^T \setminus \{r\}$

W szczególnym przypadku korzeń r może być liściem drzewa T. Zbiór L_r^T nie zawiera wtedy korzenia r [6, 7].

Definicja 2 *Zbiór*

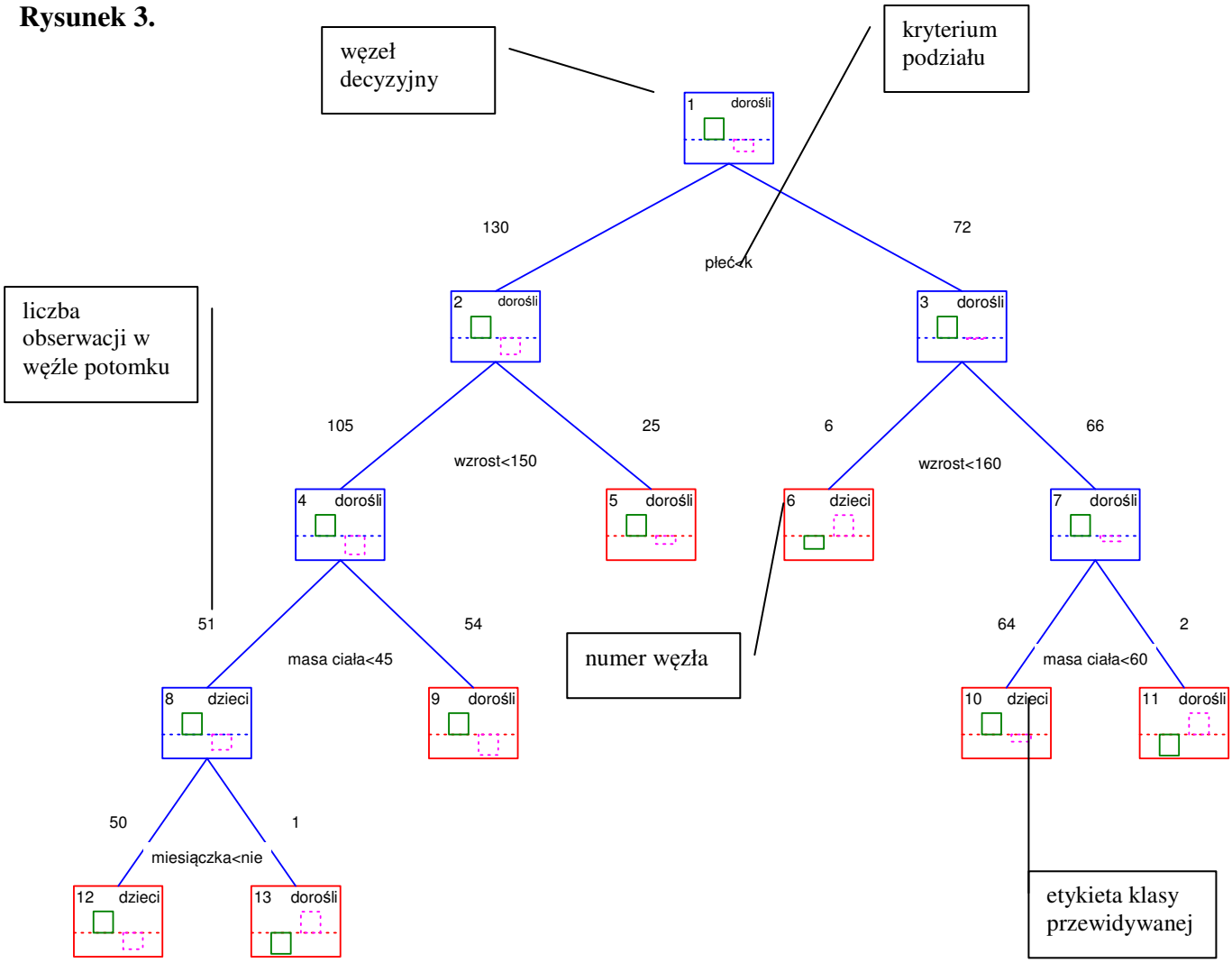
$$N_r^T := V \setminus L_r^T$$

nazywamy zbiorem węzłów drzewa T z ustalonym korzeniem r [6, 7].

Do zbioru węzłów drzewa T zaliczają się wszystkie wierzchołki o stopniu wyższym niż 1 oraz ustalony korzeń $r \in V$.

Dla dowolnych wierzchołków $u, v \in V$ drzewa T istnieje dokładnie jedna u - v droga i jest to droga prosta. W szczególności dla dowolnego liścia $l \in L_r^T$ istnieje dokładnie jedna r - l droga prosta łącząca korzeń r z liściem l. Mówimy, że r – l droga prowadzi od korzenia r, przez węzły, do liścia l.

Rysunek 3.



Każdy węzeł $n \in N_r^T$ posiada dokładnie $deg_T(n)$ wierzchołków sąsiadujących. W przypadku $n \neq r$ istnieje dokładnie jeden węzeł s sąsiadujący z n, który leży na $r - n$ drodze (od korzenia r do węzła n). Pozostałe wierzchołki sąsiadujące z n leżą na drogach od korzenia do liścia, prowadzących przez węzeł n.

Definicja 4 *Poprzednikiem (węzłem macierzystym) wierzchołka $n \in V$ różnego od korzenia r, nazywamy węzeł s sąsiadujący z n leżący na r – n drodze. Piszemy wtedy $s \succ n$. Przyjmujemy, że korzeń r nie posiada poprzedników* [6].

Definicja 5 *Następnikiem (potomkiem) węzła $n \in N_r^T$ nazywamy każdy wierzchołek m (węzeł lub liść) sąsiadujący z n i nie będący jego poprzednikiem. Piszemy wtedy $n \succ m$. Zbiór*

$$n_{\succ} := \{m \in V \ : n \succ m\}$$

nazywamy zbiorem następników węzła n [6].

Definicja 6 *Testem w węźle $n \in N_r^T$ nazywamy każdą funkcję* [6, 7]

$$\mathbf{t}_n : X \to n_{\succ}, \quad gdzie \quad X \ni x \to n_x \in n_{\succ}.$$

Zauważmy, że test jest funkcją przyporządkowaną do węzła, która przeprowadza przykłady na następniki tego węzła.

Definicja 9 *Drzewem klasyfikacyjnym (decyzyjnym) nazywamy każde drzewo T=<V, E> z korzeniem $r \in V$, rodziną testów $\{t_n\}_{n \in N_r^T}$ oraz etykietą liści $c : L_r^T \to C$. Zbiór L_r^T nazywamy zbiorem liści drzewa klasyfikacyjnego T_r. Zbiór N_r^T nazywamy zbiorem węzłów drzewa klasyfikacyjnego T_r.* [6, 7].

Drzewo klasyfikacyjne jest drzewem, które posiada dodatkową interpretację dla węzłów, gałęzi i liści:

- węzły odpowiadają testom przeprowadzonym na wartościach atrybutów przykładów, węzeł drzewa, który nie ma żadnych węzłów macierzystych jest *korzeniem,*
- gałęzie odpowiadają możliwym wynikom tych testów,
- liście odpowiadają etykietom klas danego problemu dyskryminacji (w konwencji drzewo klasyfikacyjne ma więcej niż 1 liść),

- drzewo „rośnie” od góry do dołu (od korzenia do liści).

Zaobserwowane elementy badanej przez nas próby przechodzą wzdłuż gałęzi przez węzły. W węzłach podejmowane są decyzje o wyborze gałęzi, wzdłuż której trwa dalsze przesuwanie. W każdym węźle mamy do czynienia z podziałem elementów docierających na podgrupy (względem zapisanego w nim *kryterium podziału – testu).* Przesuwanie trwa do momentu, gdy napotkamy liść, z etykietą którejś z klas.

Dla każdego liścia istnieje dokładnie jedna droga łącząca go z korzeniem. Zbiór wszystkich dróg może być przekształcony do *zbioru reguł* klasyfikujących przykłady. Klasyfikacja następuje w sposób identyczny jak „robi” to drzewo. Możliwa jest zamiana (konwersja) drzewa decyzyjnego do zbioru reguł. Konwersja wykorzystywana jest przy *przycinaniu* drzewa, które zapobiega *nadmiernemu dopasowaniu*.

4.2. Drzewo jako hipoteza

Poniżej przedstawiona zostanie formalna definicja funkcji klasyfikującej stowarzyszona z drzewem klasyfikacyjnym.

Niech będzie dane drzewo klasyfikacyjne $T_r =< V,E >$ z korzeniem r, rodzina testów $\{t_n\}_{n \in N_r^T}$ i etykietą liści c [6, 7, 9].

Definicja 2.2.1 *Hipoteza h^T reprezentowaną drzewem klasyfikacyjnym T nazywamy przekształcenie zdefiniowane regułą rekurencyjną:*

1. ustalamy $x \in X$, $n_0 = r$

2. $n_{i+1} := t_{ni}(x)$ - wykonuj działanie dopóki wynik nie będzie liściem,

3. jeżeli w k-tym kroku $n_k \in L_r^T$(jest lisciem), to zwróć etykietę liścia $c(n_k) \in C$ [6, 7].

4.3. Metody konstrukcji drzew

Pokazane zostało, że drzewa decyzyjne reprezentują hipotezy. W praktyce bardzo często zachodzi konieczność utworzenia drzewa klasyfikacyjnego dedykowanego do danego problemu dyskryminacyjnego.

Poniżej przedstawione zostaną podstawowe metody konstrukcji drzew. Drzewa te reprezentują hipotezy przybliżające pojęcia docelowe na podstawie dostępnych zbiorów uczących. Rozszerzone zostanie tym samym pojęcie drzewa decyzyjnego do klasyfikatora. Naszym celem będzie zbudowanie drzewa klasyfikacyjnego z możliwie małym błędem rzeczywistym (powstaje podczas testowania obiektami nienależącymi do zbioru przykładów). . W zadaniu budowy drzewa decyzyjnego wyróżnia się cztery podstawowe składowe:

1. Rodzinę $\{t_n^s\}$ testów określających podział w każdym węźle.
2. Zdefiniowane kryterium $\varphi(t_n^s)$ jakości podziału określone dla każdego testu t_s^n w każdym węźle n.
3. Kryterium stopu budowy drzewa.
4. Konstrukcja reguły decyzyjnej (etykiety liści drzewa).

4.3.1. Konstrukcja testów

Dobór odpowiedniego testu jest decyzją ważną, o kluczowym znaczeniu dla późniejszych właściwości drzewa. Test powinien pozwalać na możliwie dokładną klasyfikację dostępnych przykładów. Konstrukcja testów jest wysoce uzależniona od typu testowanego atrybutu. Należy zwrócić uwagę, iż proces doboru testu jest problemem trudnym i kosztownym w realizacji. Niska złożoność obliczeniowa i skalowalność powstającego procesu klasyfikacji jest w tym przypadku priorytetem.

W poniższym tekście testy będą traktowane, jako funkcje zależne jedynie od atrybutu i jego wartości. Zachodzi konieczność wprowadzenia dodatkowych oznaczeń:

$A : X \rightarrow S_A$ - gdzie A atrybut:

$A(x)$ - wartość atrybutu A dla przykładu $x \in X$,

$_A$ - zbiór wartości atrybutu A,

$t : X \rightarrow S_t$ - gdzie t test:

$t(x)$ - wartość testu t dla przykładu $x \in X$,

S_t - zbiór wartości testu t.

Testy dla atrybutów nominalnych:

1. Test tożsamościowy polega na utożsamieniu testu z atrybutem $S_t = S_A$. Taki test jest bardzo wygodny przy drzewach nie będących binarnymi. Pozwala na duży współczynnik rozgałęzienia, co zmniejsza głębokość drzewa i koszt klasyfikacji. Jego mankamentem jest niska stosowalność przy atrybutach o dużej liczbie możliwych wartości.
2. Test różnowartościowy $S_t = \{0, 1\}$. Wybór najlepszego testu równościowego wymaga sprawdzenia co najwyżej wszystkich wartości atrybutu A.
3. Test przynależnościowy jest uogólnieniem testów równościowych. Zauważmy, że dobór najlepszego testu wymaga co najwyżej sprawdzenia wszystkich właściwych podzbiorów zbioru S_A, co przy n możliwych wartościach atrybutu A wymaga $2^{n-1} - 1$ porównań. Jest to zależność wykładnicza (czyli bardzo kosztowna), sugerująca konieczność zaproponowania rozsądnego sposobu wyboru rozpatrywanych zbiorów W jako podzbiorów zbioru S_A. Przy tego rodzaju testach (testy przynależnościowe stosowane są przy konstrukcji klasyfikatora SLIQ i SPRINT) jest to kwestia mająca kluczowy wpływ na dalszą skalowalność procesu klasyfikacji [6, 7].

Testy dla atrybutów ciągłych

1. Test przynależnościowy. W tym przypadku jako podzbiory $W \subset S_A$ bierze się pewne przedziały, gdzie dobór ich „końców" jest istotny. Mankamentem testów przynależnościowych przy ciągłych atrybutach, jest brak uwzględnienia istnienia relacji porządku w zbiorze możliwych wartości analizowanego atrybutu. Konstruuje sie również testy uwzględniające istnienie owej relacji, nazywane testami nierównościowymi (wykorzystywane przy konstrukcji klasyfikatora. SLIQ i SPRINT)

2. Test nierównościowy. Zapisując $S_A = \{w_1, w_2, \ldots, w_n\}$ i przyjmując, ze ciąg $\{w_1, w_2, \ldots, w_n\}$ jest ciągiem uporządkowanym (posortowanym w kolejności rosnącej), możemy stwierdzić, ze dowolna taka wartość w, że $w_i < w < w_{i+1}$ dla ustalonego $i = 1, \ldots, n - 1$, daje jednakowy wynik testu nierównościowego (dzieli zbiór X zawsze w taki sam sposób). Zatem, aby wybrać najbardziej odpowiedni test, wystarczy przeprowadzić tylko $n - 1$ porównań. Zazwyczaj za punkt podziału obiera sie środek przedziału $[w_i, w_{i+1}]$. Przy rozważaniu kwestii skalowalności, należy zwrócić uwagę na koszt sortowania zbioru wartości testowanego atrybutu [6, 7].

4.3.2. Kryteria jakości podziałów

Podpróba docierająca do węzła dzielona jest na części. Proces ten nie powinien być procesem przypadkowym. Zależy nam na podziale, który daje jak najmniejszą różnorodność klas w otrzymanych częściach. Najlepiej byłoby, aby różnica pomiędzy różnorodnością klas w węźle i różnorodnością klas w tych częściach, była możliwie duża.

Definicja 2.3.5 *Każdą funkcję*

$$\phi : G \subset [0,1]^g \to R, \qquad gdzie \qquad (p_1, p_2, ..., p_n) \in G \Leftrightarrow \sum_{k=1}^{g} p_k = 1$$

spełniającą następujące warunki:

1. ϕ *przyjmuje wartość maksymalną w punkcie* $\left(\frac{1}{g}, \frac{1}{g}, \frac{1}{g}, ..., \frac{1}{g}\right) \in G$

2. ϕ *osiąga minimum jedynie w punktach*

$(1, 0, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, 0, \ldots, 1) \in G.$

3. $\phi(p_1, p_2, \ldots, p_g)$ *jest symetryczna ze względu na* $p_1, p_2, \ldots, p_g$

nazywamy funkcją różnorodności klas [6, 7, 11]

$C = \{1,2,...,g\}$; $g \in N$ – zbiór etykiet klas

Główne miary różnorodności klas

W praktyce najczęściej stosuje się niżej wymienione miary różnorodności klas. Indeks Giniego i entropia wykazują większą czułość na zmiany rozkładu klas w próbie.

1. Proporcja błędnych klasyfikacji:

$$q(S) \equiv p(S) := 1 - \max_k p(k|S)$$

2. Indeks Giniego:

$$q(S) \equiv G(S) := 1 - \sum_{k=1}^{g} \left(p(k|S)\right)^2$$

3. Entropia:

$$q(S) \equiv E(S) := 1 - \sum_{k=1}^{g} p(k|S) \ln p\left(k|S\right)$$

$p(k/s)$ – prawdopodobieństwo wystąpienia zdarzenia k pod warunkiem S

[6, 7, 11]

Różnorodność jest tym większa im większa jest wartość miary różnorodności klas $q(S)$. Po dokonaniu podziału w węźle $n \in N_r^T$ zbiór $S_{n\succ m}$ reprezentuje obiekty, które przeszły z węzła n do jego następnika $m \in n_{\succ}$.

Definicja 2.3.8 *Przez miarę zmiany różnorodności klas w węźle* $n \in N_r^T$ *drzewa klasyfikacyjnego* T_r *przy założeniu, że w węźle n znajdują sie wszystkie obiekty z S, rozumie się kryterium oceny podziału w węźle n:*

$$\Delta q\left(S|n\right) := q\left(S\right) - \sum_{m \in n_{\succ},\ P(S_{n\succ m})>0} P\left(S_{n\succ m}|S\right) q\left(S_{n\succ m}\right)$$

[6, 7].

Pisząc $\Delta q\left(S|n\right)$ zakładamy istnienie testu w węźle n. W sytuacji, gdy do węzła przyporządkowany jest zbiór testów, definicja 2.3.8 umożliwia to wybór podziału z największą wartością miary zmiany różnorodności klas. W tym sensie jest to podstawowe kryterium oceny testu w węźle drzewa klasyfikacyjnego.

Dla drzew binarnych Breiman [11] sformułował i udowodnił następujące twierdzenie.

Twierdzenie 2.3.2 (Breiman) *Dla binarnego drzewa T_r i wklęsłej funkcji różnorodności klas zachodzi:*

(i) $\Delta q(S|n) \geq 0$ *dla dowolnego węzła* $n \in N_r^T$ *oraz* $S \in B$, *że* $P(S) > 0$,

(ii) jeżeli $n_{\succ} = \{n_L, n_R\}$, *to równość w* (*i*) *zachodzi wtedy i tylko wtedy, gdy rozkłady klas w S,* $S_{n \succ n_R}$ *są identyczne, tzn.:*

$$\forall k \in \{1,\ldots,g\} \quad p(k|S) = p(k|S_{n \succ n_L}) = p(k|S_{n \succ n_R})$$ [6].

4.3.3. Kryterium stopu i reguła decyzyjna

Budowę drzewa klasyfikacyjnego rozpoczynamy od drzewa złożonego z jednego wierzchołka, do którego przyporządkowujemy zbiór uczący i zbiór dostępnych testów. W dalszych krokach konstruujemy podziały, tworząc węzły i ich następniki. Wraz ze wzrostem drzewa maleje zbiór uczący i zbiór testów docierający na kolejne jego poziomy. Poniżej przedstawione zostaną wytyczne, którymi należy się kierować podczas budowy drzewa. Należy zaniechać konstrukcji podziału w wierzchołku jeżeli:

1. Wystąpienie klasy *k* w podpróbie uczącej dostępnej w wierzchołku jest zdarzeniem z prawdopodobieństwem warunkowym 1.
2. Zastosowanie każdego dostępnego podziału daje zerową lub ujemną miarę zmiany różnorodności klas.
3. Zbiór dostępnych testów jest pusty.

Gdy obiekty w wierzchołku należą do tej samej klasy, to zajdzie przypadek 1. Sytuacja 2 ma miejsce w wierzchołku, w którym zbiór dostępnych testów jest oparty o atrybuty z jednakową wartością dla wszystkich dostępnych przykładów. Warunek 3 bezpośrednio wiąże się z brakiem uzasadnienia dla więcej niż jednokrotnego użycia danego podziału w obrębie jednej ścieżki. Wystąpienie przypadków 2 lub 3 może świadczyć o zajściu jednej z poniższych sytuacji:

- zbiór trenujący nie jest poprawny i zawiera przekłamania,
- zestaw atrybutów nie opisuje obiektów w dostatecznym stopniu i w związku z tym przestrzeń hipotez jest zbyt uboga do reprezentowania pojęcia docelowego,
- przyjęty zbiór dostępnych atrybutów jest niewystarczający.

Definicja 2.4.9 *Jeżeli S jest podpróbą uczącą dostępną w wierzchołku n, a* T *zbiorem dostępnych testów, to kryterium stopu wstrzymujące konstrukcje podziału w n określamy wyrażeniem:*

$$\left(\exists_{k\in C}\, p\left(k|S\right)=1\right)\vee\left(\forall_{t_n\in \mathrm{T}}\Delta q\left(S|n\right)\le 0\right)\vee\left(\mathrm{T}=\varnothing\right)$$

Po wstrzymaniu konstrukcji podziału wierzchołek staje sie liściem, do którego należy przyporządkować etykietę klasy [6].

Definicja 2.3.10 *Jeżeli S jest zbiorem uczącym, to etykietę daną wzorem*

$$c_S := \arg\max_k p\left(k|S\right)$$

nazywamy etykietą większościowej kategorii w zbiorze uczącym S [6, 7].

W sytuacji, gdy zbiór dostępnych obiektów jest zdarzeniem z niezerowym prawdopodobieństwem etykietę wybieramy na podstawie etykiety większościowej kategorii.

Jeżeli nie istnieją wierzchołki, w których kryterium stopu uzna za zasadne utworzenie nowego podziału, to drzewo uznajemy za zbudowane.

4.3.4. Zstępująca konstrukcja drzewa

Istnieje wiele algorytmów uczenia się pojęć wykorzystujących drzewa decyzyjne do reprezentacji hipotez. Każdy z nich dąży do uzyskania struktury o niewielkim rozmiarze z możliwie niewielkim błędem próby, oczywiście zakładając, że błąd rzeczywisty będzie również mały. Poniżej przedstawiona zostanie część wspólna większości tych algorytmów, określana mianem *zstępującej konstrukcji drzewa*. Schemat zakłada rozpoczęcie budowy drzewa od pojedynczego wierzchołka (korzenia), któremu przyporządkowuje się wszystkie

elementy ze zbioru trenującego. Kolejnym krokiem jest ustalenie zasadności utworzenia podziału w węźle. W przypadku decyzji pozytywnej konstruowane są wierzchołki następniki. Nowoutworzone wierzchołki traktowane są jako korzenie „nowych" drzew z przypisanymi częściami próby uczącej. Procedura powtarzana jest do uzyskania liści (wierzchołków bez podziału), dla których określana jest reguła decyzyjna.

4.4. Problem nadmiernego dopasowania

Nadmierne dopasowanie do danych trenujących przejawia się bardzo małym błędem klasyfikacji (często nawet zerowym) na próbie uczącej, lecz zbyt dużym błędem rzeczywistym. Prawdopodobnie tak stworzone drzewo, poprzez bardzo złożoną strukturę, odzwierciedla przypadkowe zależności występujące w zbiorze uczącym. Drzewa klasyfikacyjne są szczególnie narażone na ten problem, gdyż ich struktura umożliwia reprezentację dowolnej hipotezy. Błąd rzeczywisty nadmiernie dopasowanych drzew można zmniejszyć przez ich uproszczenie nazywane *przycięciem*. Drzewo przycięte ma prostszą strukturę, co daje krótszy czas klasyfikacji, ale konsekwencją jest pogorszenie dokładności klasyfikacji zbioru uczącego [8].

4.4.1. Schemat przycinania

W uproszczeniu proces przycinania polega na zastąpieniu drzewa wyjściowego jego poddrzewem. Ujmując to bardziej obrazowo powiemy, że „ucina" się niektóre poddrzewa drzewa wyjściowego, zastępując je liśćmi, którym przypisuje sie etykietę większościowej kategorii wśród obserwacji związanych z tym poddrzewem. Możemy zamiast stosowania schematu przycinania zastosować modyfikację kryterium stopu i w ten sposób zapobiegać nadmiernemu wzrostowi drzewa. Takie postępowanie określa się *przycinaniem w trakcie wzrostu*. Jednak znalezienie odpowiedniego kryterium stopu okazuje się najczęściej trudne i dla ułatwienia przycina się drzewa uprzednio zbudowane.

Przycinanie odbywa sie z pomocą zbioru etykietowanych przykładów, zwanego *zbiorem przycinania*. Pełni on ważną funkcje przy szacowaniu błędu rzeczywistego przyciętego drzewa. Wyróżnia się dwa typy zbiorów przycinania:

1. Zbiór przycinania pochodzi spoza próby uczącej - jeśli mamy dostatecznie duży zbiór uczący.

2. Zbiór przycinania równy jest próbie uczącej - jeżeli nie dysponujemy dużym zbiorem uczącym.

Obecnie znanych jest wiele procedur przycinania drzew. Pierwsza metoda polega na budowaniu drzewa do określonej właściwej wielkości, gdzie właściwą wielkość wyznacza użytkownik na podstawie wiedzy z poprzednich badań lub informacji diagnostycznych uzyskanych w poprzednich analiza. Druga metoda polega na wykorzystaniu dobrze udokumentowanych, ustrukturowanych procedur, które opracowali Breiman i in. (1984) Rozważmy pierwszą strategię - tutaj mamy trzy możliwe opcje:

1. *Sprawdzian krzyżowy na podstawie próby testowej* - Przy tym typie testu drzewo decyzyjne oblicza się na podstawie próby uczącej, a jego trafność przewidywania (zdolność predykcji) testowana na próbie testowej. Innymi słowy sprawdzana jest zdolność przewidywania przynależności do klas na próbie testowej modelu zbudowanego na próbie uczącej. Jeśli *koszty* w próbie testowej są większe niż *koszty* w próbie uczącej (*koszty* równają się proporcji przypadków błędnie zaklasyfikowanych, gdy *prawdopodobieństwa a priori* są oszacowane, a koszty błędnej klasyfikacji są równe), to wskazuje to na słaby wynik i można się spodziewać, że drzewo innej wielkości mogłoby dać lepszy rezultat. Próby uczącą i testową tworzymy gromadząc dwa niezależne zbiory danych.
2. *V-krotny sprawdzian krzyżowy* - tę metodę sprawdzania poprawności drzewa stosujemy, jeżeli nie dysponujemy próbą testową, a próba ucząca jest za mała, aby wyodrębnić z niej taką próbę. Użytkownik określa wartość *V* dla *V-krotnego sprawdzianu krzyżowego.* Wartość *V* wyznacza liczbę podprób losowych wyodrębnianych z próby uczącej. Podpróby powinny być możliwie równe sobie wielkością.. Drzewo klasyfikacyjne określonej wielkości jest obliczane *V* razy. Za każdym razem opuszcza się w obliczeniach jedną z podprób i wykorzystuje się ją jako próbę testową w sprawdzaniu krzyżowym. Zatem widzimy, że każda podpróba jest użyta *V - 1* razy w próbie uczącej i tylko jeden raz w charakterze próby testowej. Następnie dla każdej z *V* prób testowych obliczane są *koszty sprawdzianu krzyżowego,* a te są uśrednione i otrzymujemy *V-krotną* ocenę kosztów *sprawdzianu krzyżowego*, która może być podawana, razem z błędem standardowym
3. *Globalny sprawdzian krzyżowy* – metoda ta polega na powtarzaniu całej analizy określoną liczbę razy eliminując część próby uczącej równą 1 dzielone przez liczbę

powtórzeń. Każda wyeliminowana część próby jest wykorzystana jako próba testowa w *sprawdzianie krzyżowym* wybranego drzewa klasyfikacyjnego.

Strategia zaproponowana przez Breimana'a polega na *przycinaniu na podstawie minimalizacji kosztów i złożoności drzewa w sprawdzianie krzyżowym*. Koszty związane z przycinaniem koszt-złożoność są obliczane, gdy drzewo się rozrasta, począwszy od podziału przy węźle źródłowym, aż do momentu, gdy osiągnie maksymalną wielkość, wyznaczoną przez określoną *minimalną liczność (n)*. Gdy do drzewa zostaje dodany każdy następny podział obliczane są koszty dla próby uczącej. Zatem sekwencja ogólnie malejących kosztów (odzwierciedlających lepszą klasyfikację) odpowiada liczbie podziałów drzewa. Koszty dla próby uczącej nazywane są *kosztami resubstytucji.* Oszacowane koszty *sprawdzianu krzyżowego* na podstawie *V-krotnego sprawdzianu krzyżowego* stosowane są do obliczenia kosztów dla węzła źródłowego. Zdefiniowany teraz zostanie parametr zwany parametrem złożoności, którego początkowa wartość wynosi zero. Następnie dla każdego drzewa (łącznie z pierwszym, zawierającym tylko jeden węzeł źródłowy) obliczona zostanie wartość funkcji zdefiniowanej jako koszty dla każdego drzewa plus parametr złożoności razy wielkość drzewa. Potem parametr złożoności będzie nieprzerwanie zwiększany do momentu, w którym wartość funkcji dla największego drzewa przekroczy wartość funkcji dla mniejszego drzewa. Wzięte zostanie mniejsze drzewo jako nowe największe drzewo i ponownie powiększany będzie dalej parametr złożoności, aż wartość funkcji dla największego drzewa przekroczy wartość funkcji dla mniejszego drzewa. Postępując w ten sposób dochodzimy do momentu, aż węzeł źródłowy będzie największym drzewem. Sekwencja największych drzew uzyskanych przy pomocy tego algorytmu ma kilka interesujących własności. Są one zagnieżdżone, ponieważ kolejno przycinane drzewa zawierają wszystkie węzły następnego w kolejności mniejszego drzewa. Początkowo, przy przejściu od jednego do następnego w kolejności, mniejszego drzewa, wiele węzłów często zostaje przyciętych. Dochodząc do węzła źródłowego przycina się mniej węzłów. Przycinana jest również sekwencja największych drzew, gdyż dla każdej wielkości drzewa w tej sekwencji nie ma innego, tej samej wielkości, które miałoby mniejsze koszty [8, 10].

5. Multivariate Adaptive Regression Splines (MARSplines)

Wielozmienna regresja adaptacyjna z użyciem funkcji sklejanych jest uogólnieniem techniki wprowadzonej do szerokiego użytku przez Friedman'a (1991). Służy do rozwiązywania zarówno problemów regresyjnych jak i klasyfikacyjnych. Celem jest znalezienie wartości zmiennych wyjściowych (zależnych) na podstawie zmiennych wejściowych (predykcyjnych).
Multivariate Adaptive Regression Splines (*MARSplines*) jest procedurą nieparametryczną, która nie wymaga założeń dotyczących funkcyjnej zależności między zmiennymi zależnymi a niezależnymi. *MARSplines* modeluje tę zależność za pomocą zbioru współczynników i funkcji bazowych, które są w pełni determinowane przez dane. Przestrzeń wejściowa dzielona jest na obszary, w których określane są osobne funkcje regresyjne lub klasyfikacyjne. Ogólny mechanizm działania *MARSplines* wyobrażać można sobie jako wielokrotną, odcinkową regresję liniową. Takie podejście czyni *MARSplines* szczególnie użytecznym przy większej liczbie wymiarów na wejściu (więcej niż dwie zmienne), kiedy, w przypadku innych technik zagrażać zaczyna problem wymiarowości.

5.1. Algorytm MARSplines

Algorytm *MARSplines* (*Multivariate Adaptive Regression Splines*) to dwuetapowa procedura stosowana sukcesywnie, aż do otrzymania żądanego modelu. W pierwszym etapie zostaje zbudowany model, którego złożoność zwiększamy dodając kolejne funkcje bazowe, aż do osiągnięcia maksymalnego (określonego przez użytkownika) stopnia. Następnie uruchamiana jest procedura wsteczna, usuwania z modelu najmniej znaczących funkcji bazowych, czyli takich, których usunięcie najmniej pogarsza dopasowanie modelu. Implementacja algorytmu przebiega więc w następujących sposób.

Uruchomiany zostaje algorytm z najprostszym modelem, z funkcją bazową o stałej wartości, przeważnie równej 1. Następnie zostaje włączone przeszukiwanie, dla każdej zmiennej i możliwych węzłów, przestrzeni funkcji bazowych. Następuje dodawanie do modelu tych funkcji, które maksymalizują pewną miarę dobroci dopasowania modelu (minimalizują błąd predykcji). Ten krok powtarzany jest aż do osiągnięcia wstępnie

założonego, maksymalnego stopnia złożoności modelu. Na końcu "oczyszcza" się model z funkcji bazowych, które dają za mały wkład do poprawy jakości modelu (w sensie najmniejszych kwadratów).

5.2. Funkcje bazowe

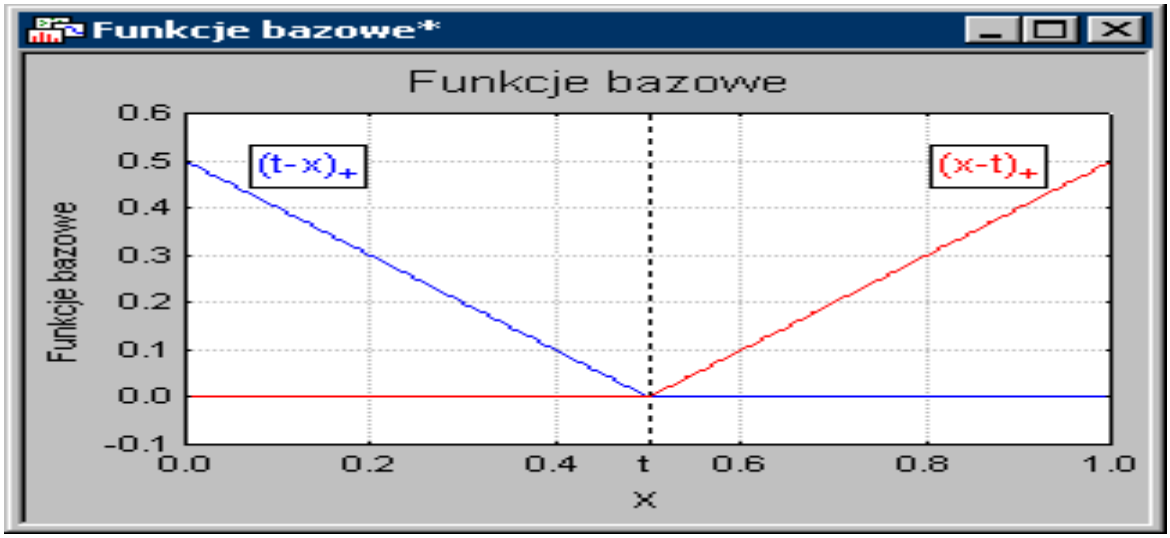
MARSplines używa dwustronnych, obciętych funkcji liniowych (uwidocznionych na **rys. 4**) jako bazowych funkcji dla liniowej lub nieliniowej aproksymacji zależności pomiędzy zmiennymi predykcyjnymi i zmiennymi odpowiedzi. W modelu mamy *2Np* funkcji bazowych, gdzie *N* jest liczbą przypadków, a *p* liczbą predyktorów.

$$(x-t)_+ = \begin{cases} x-t & x>t \\ 0 & pozostale \end{cases} \qquad (t-x)_+ = \begin{cases} t-x & x<t \\ 0 & pozostale \end{cases}$$

t - jest węzłem tworzonym dla każdego punktu x_{ij}

Zbiór funkcji bazowych wygląda następująco

$$CB = \{(x_j - t)_+, (t - x_j)_+\} \qquad t \in \{x_{1j}, x_{2j}, ..., x_{Nj}\}$$
$$j = 1, 2, ..., p$$



Rysunek 4.

Na **rysunku 4.** pokazany jest prosty przykład dwóch funkcji bazowych $(t-x)_+$ *i* $(x-t)_+$ (wg Hastie, i inni 2001). Parametr *t*, to węzeł funkcji bazowej (określający "odcinki" odcinkowej regresji liniowej); położenie węzłów (wartość parametru *t*) wynika z wartości danych. Indeksy "+" za wyrażeniami *(t-x)* i *(x-t)* oznaczają, że bierzemy tylko dodatnią część funkcji liniowej. Zamiast ujemnych wartości przyjmuje się wartość zero, co widać na wykresie [8, 10, 12].

5.3. Model MARSplines.

Algorytm *MARSplines* (*Multivariate Adaptive Regression Splines*) konstruuje modele z dwustronnych, obciętych funkcji bazowych. Funkcje te wraz z parametrami modelu, które znajdowane są za pomocą metody najmniejszych kwadratów pozwalają dokonać predykcji wyjścia na podstawie zmiennych wejściowych. Służą, poprzez liniową lub nieliniową aproksymację, do modelowania rzeczywistej zależności *f(x).*

Model *MARSplines*, dla zmiennych zależnych (wyjściowych) *y*, mający *M* (liczba funkcji bazowych w modelu) wyrażeń, można zapisać następującym równaniem:

$$y = f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

Tak więc, *y* obliczane jest jako funkcja zmiennych predykcyjnych *X* (i ich interakcji). Elementami tej funkcji są, rzędna początkowa (β_0) i ważona (wagami β_m) suma jednej lub wielu funkcji bazowych $h_m(X)$. Sumowanie przebiega przez wszystkie *M* składników modelu. Na model ten możemy patrzeć jak na ważoną sumę funkcji bazowych, wybranych ze zbioru dużej liczby takich funkcji, pokrywających wszystkie wartości, każdego z predyktorów (w zbiorze tym mamy funkcję bazową i parametr *t* dla każdej, poszczególnej wartości, każdego predyktora). Algorytm *MARSplines* przeszukuje przestrzeń wszystkich wartości wejściowych i predykcyjnych (położeń węzłów *t*) jak i interakcji między zmiennymi. Do modelu dodawane są wtedy kolejne funkcje bazowe (wybierane ze zbioru wszystkich dopuszczalnych funkcji), w taki sposób, by maksymalizować ogólny poziom dopasowania (wg minimum sumy kwadratów). Wynikiem tej operacji jest znalezienie najważniejszych zmiennych niezależnych, oraz najważniejszych ich interakcji [8, 10, 12].

Dla rzędu interakcji *K=1* mamy model addytywny, dla rzędu *K=2* model jest parami interakcyjny.

W krokowej procedurze postępującej, do modelu dodawane są kolejne funkcje bazowe. Ilość dodanych funkcji zależy od zadanej, maksymalnej liczby, która powinna być dostatecznie duża (co najmniej dwa razy większa od optymalnej, pod względem minimum kwadratów).

Po zastosowaniu algorytmu postępującego wyboru funkcji bazowych, uruchomiana zostaje procedura wsteczna.. Procedura ta polega na wyrzucaniu kolejno z modelu tych funkcji bazowych, które dają najmniejsze polepszenie dopasowania modelu (w sensie najmniejszych kwadratów). Następnie obliczana jest funkcja błędu najmniejszych kwadratów (odwrotność dopasowania). Miarą dopasowania jest błąd tzw. uogólnionego sprawdzianu krzyżowego (Generalized Cross Validation), który uwzględnia błąd resztowy, oraz złożoność modelu. Jest on wyrażony następującym wzorem:

$$GCV = \frac{\sum_{i=1}^{N} (y_i - f(x_i))^2}{\left(1 - \frac{C}{N}\right)^2}$$

gdzie

$$C = 1 + cd$$

natomiast N jest liczbą przypadków w danych, d jest efektywną liczbą stopni swobody, równą liczbie niezależnych funkcji bazowych. Parametr c służy do sterowania wielkością kary za dodawanie funkcji bazowych. Z doświadczenia wynika, że najlepsze C otrzymuje się przy $2 < d < 3$ [8, 10, 12].

Predyktory jakościowe.

MARSplines jest dostosowany do zadań, w których występują zarówno ilościowe jak i jakościowe zmienne predykcyjne. Jednakże, algorytm podstawowy *MARSplines* zakłada, że predyktory są ilościowe i tak na przykład, obliczonych węzłów program zazwyczaj nie powiąże z kodami klas zmiennych jakościowych.

Wiele zmiennych zależnych (wyjściowych).

Algorytm *MARSplines* może być stosowany, kiedy mamy do czynienia z wieloma zmiennymi zależnymi. W tej sytuacji, przy wielowymiarowym wyjściu, algorytm określa wspólny zbiór funkcji bazowych dla zmiennych niezależnych (predyktorów), lecz osobne zbiory współczynników, dla każdej zmiennej wyjściowej. Takie podejście do wielowymiarowej zmiennej wyjściowej przypomina pewne algorytmy sieci neuronowych, gdzie wielokrotne wyjście obliczane jest na bazie wspólnych neuronów; w *MARSplines*, wielokrotne wyjście obliczane jest ze wspólnych funkcji bazowych, ze specyficznymi (dla każdej zmiennej wyjściowej) współczynnikami.

MARSplines i klasyfikacja.

Ponieważ *MARSplines* może być stosowany do zagadnień z wielowymiarowym wyjściem łatwo jest również zastosować go do zagadnień klasyfikacyjnych. W zagadnieniach klasyfikacyjnych jakościowa zmienna wyjściowa zostaje zakodowana – przetworzona na wielowymiarową zmienną wskaźnikową (1 = przypadek należy do klasy k, 0 = przypadek nie należy do klasy k), następnie procedura *MARSplines* dopasowuje model i oblicza ciągłą predykcję, a w ostatnim etapie przypisuje przypadkom klasy według największych wartości predykcji. Można zauważyć, iż takie zastosowanie daje heurystyczną klasyfikację, która może bardzo dobrze działać w praktyce, jednak należy pamiętać, że prawdopodobieństwa klasyfikacyjne nie są tworzone na podstawie modelu statystycznego.

5.4. Wybór modelu i jego redukowanie

Modele nieparametryczne są bardzo elastyczne, dobrze dostosowują się do danych, co z jednej strony jest zaletą, ale z drugiej strony może prowadzić do niekorzystnego zjawiska nadmiernego dopasowania (przeuczenia, overfitting), o ile się temu nie przeciwdziała. Modele takie, jeżeli dopuścimy dostatecznie dużą liczbę parametrów dość łatwo osiągną zerowy błąd na danych uczących, lecz będą źle działały dla nowych danych (gdyż w modelu nie będzie dobrze zgeneralizowana wiedza pobrana z danych uczących). *MARSplines*, ma tendencję do nadmiernego dopasowywania się do danych. Do zwalczania tego problemu, w *MARSplines* wykorzystana została technika redukcji (pruning), analogiczna do przycinania (w drzewach klasyfikacyjnych), ograniczająca złożoność modelu przez redukowanie liczby funkcji bazowych.

Wybieranie najważniejszych i redukowanie (usuwanie) najmniej ważnych funkcji bazowych, jest operacją, której wynik można wykorzystać do wybrania istotnych predyktorów. Algorytm *MARSplines* wybierze tylko te funkcje bazowe (czyli te zmienne predykcyjne), które dają "mierzalny" wkład do predykcji [8,10,12].

6. Krzywa ROC (Receiver Operating Characteristics)

6.1. Czułość i swoistość, iloraz wiarygodności – dokładność diagnostyczna

Koncepcja dokładności diagnostycznej zakłada przedstawienie jej w kategoriach czułości i swoistości, najczęściej dla jednego, wybranego punktu odcięcia.

Tabela 4 Miary dokładności diagnostycznej testu.

Wynik testu	**Diagnostyka**	
	+	**-**
+	A	B
-	C	D

gdzie

A – liczba wyników prawdziwie dodatnich

B – liczba wyników fałszywie dodatnich

C – liczba wyników fałszywie ujemnych

D – liczba wyników prawdziwie ujemnych

Czułość i swoistość wyliczamy z następujących wzorów

$$czułość = \frac{A}{A+C}, \qquad swoistość = \frac{D}{B+D}$$

Czułość określa zdolność testu do wykrycia choroby w analizowanej grupie chorych. Swoistość odzwierciedla zdolność testu do wykluczenia pacjentów zdrowych. Testy nie mają jednej pary czułości i swoistości, lecz tyle ile jest wartości obserwowanych w wyniku przeprowadzenia testu. Podczas zmiany punktu odcięcia czułość i swoistość również będą się zmieniać i to w przeciwnych kierunkach. Gdy jedna wielkość rośnie, druga będzie maleć. Dla każdego punktu odcięcia mamy więc do czynienia z parami czułości i swoistości. Aby opisać

dokładność diagnostyczną testu musimy mieć obraz par czułości i swoistości dla wszystkich możliwych punktów odcięcia.

Iloraz wiarygodności *LR* przy ustalonym punkcie odcięcia jest to stosunek frakcji prawdziwie dodatniej *(czułość)* do fałszywie dodatniej *(1 – swoistość).*

$$LR = \frac{\frac{A}{A+C}}{\frac{B}{B+D}}$$

Iloraz wiarygodności nie zależy od częstości występowania choroby, natomiast odnosi się do par czułość/swoistość. Iloraz wiarygodności może być zdefiniowany również jako stosunek prawdopodobieństw określonego wyniku testu w obecności i nieobecności choroby [13, 14, 15].

Skuteczność reguły decyzyjnej oblicza się dzieląc przypadki prawidłowo zaklasyfikowane przez wszystkie klasyfikacje reguły

$$ACC = \frac{A+D}{A+B+C+D}$$

[71].

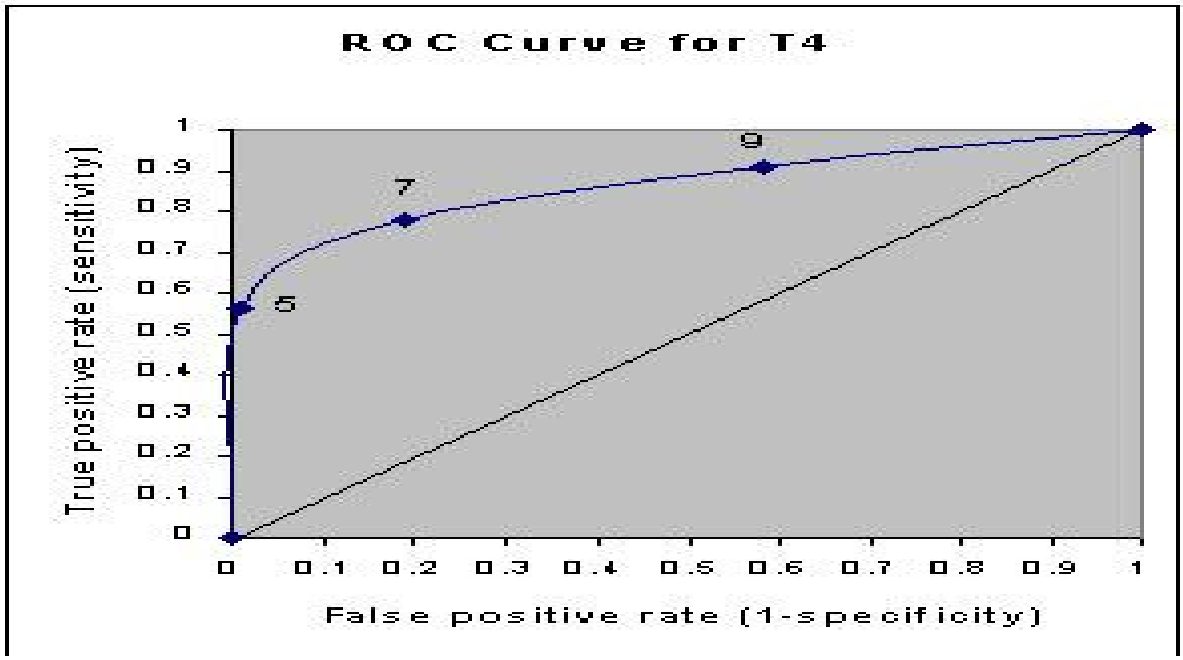
6.2. Prezentacja graficzna krzywych ROC

Krzywe ROC dają pełny obraz całkowitego zakresu par czułości i swoistości, ponieważ wszystkie możliwe pary są odzwierciedlone na wykresie. Krzywe te można skonstruować zarówno dla danych ciągłych jak i dyskretnych. Nanosimy na płaszczyznę punkty o współrzędnych stanowiących parę *(1-swoistość, czułość)* dla wszystkich możliwych punktów odcięcia wartości obserwowanych testu. Na osi rzędnych zaznaczamy frakcję prawdziwie dodatnią (czułość), natomiast na osi odciętych umieszczamy wyniki fałszywie dodatnie (1-swoistość). Połączenie linią łamaną tych punktów jest wykresem krzywej ROC. Gdy punkt odcięcia się zmienia, włączenie do reguły decyzyjnej wyniku prawdziwie dodatniego daje na wykresie krzywej ROC linię pionową, natomiast włączenie fałszywie dodatniego wyniku daje nam linię poziomą.

Na krzywej ROC istnieje kilka punktów o szczególnym znaczeniu interpretacyjnym. Punkt, gdzie *1-swoistość* jest równy *0* oraz czułość również równa *0 - (0, 0),* reprezentuje

sytuację, w której nie mamy wyników pozytywnej klasyfikacji (czułość). Taki klasyfikator nie popełnia błędów w postaci fałszywie dodatnich wyników, ale także powoduje brak przyrostu prawdziwie dodatnich. Odwrotną sytuację reprezentuje punkt *(1, 1*), czyli mamy idealną czułość, ale z drugiej strony istnieje olbrzymi błąd w postaci wyników fałszywie dodatnich. Kolejnym interesującym punktem jest punkt *(0, 1),* który reprezentuje idealną klasyfikację.

Ponieważ wykres krzywej ROC bazuje bezpośrednio na wszystkich wynikach testu, może być nazwany nieparametryczną krzywą ROC Termin nieparametryczny oznacza brak jakichkolwiek parametrów określających przebieg krzywej.



Rysunek 5 Na wykresie przedstawiona została krzywa ROC dla 3 punktów odcięcia [13, 16].

6.3. Pole pod krzywą ROC

Wygodnym sposobem ilościowej oceny dokładności diagnostycznej testu jest wyrażanie jej jedną liczbą. Najpopularniejszą taką miarą jest wielkość pola powierzchni pod krzywą ROC (AUC - area under the curve). Przyjmuje ona wartości od 0,5 (brak zdolności do rozróżniania między dwoma grupami pacjentów) do 1,0 (idealna zdolność dyskryminacyjna). Analityczna metoda obliczania pola jest pokazana przez *Bember* [53] oraz *Halley* i *McNeil* [54]. Ponadto, można je obliczyć pośrednio ze statystyki Wilcoxon'a. W przypadku, kiedy w danych laboratoryjnych nie występują wiązania (pacjent z jednej grupy – chorzy na raka ma

identyczny wynik jak pacjent z drugiej grupy – chorzy z łagodnym rozrostem komórek), koncepcyjnie najprostszą metodą wydaje się wyznaczanie pola jako sumy N prostokątów pod krzywą ROC:

$$W = \sum a_i, \qquad a_i = (x_{i+1} - x_i)y_i, \qquad i = 1,2,,...,N$$

$$SE_W = \sqrt{\frac{W(1-W) + (n_X - 1)(Q_1 - W^2) + (n_N - 1)(Q_2 - W^2)}{n_X n_N}}$$

gdzie W jest wielkością empirycznego pola pod krzywą ROC

n_X – liczebność w grupie badanej

n_N – liczebność w grupie kontrolnej

SE_W – błąd standardowy

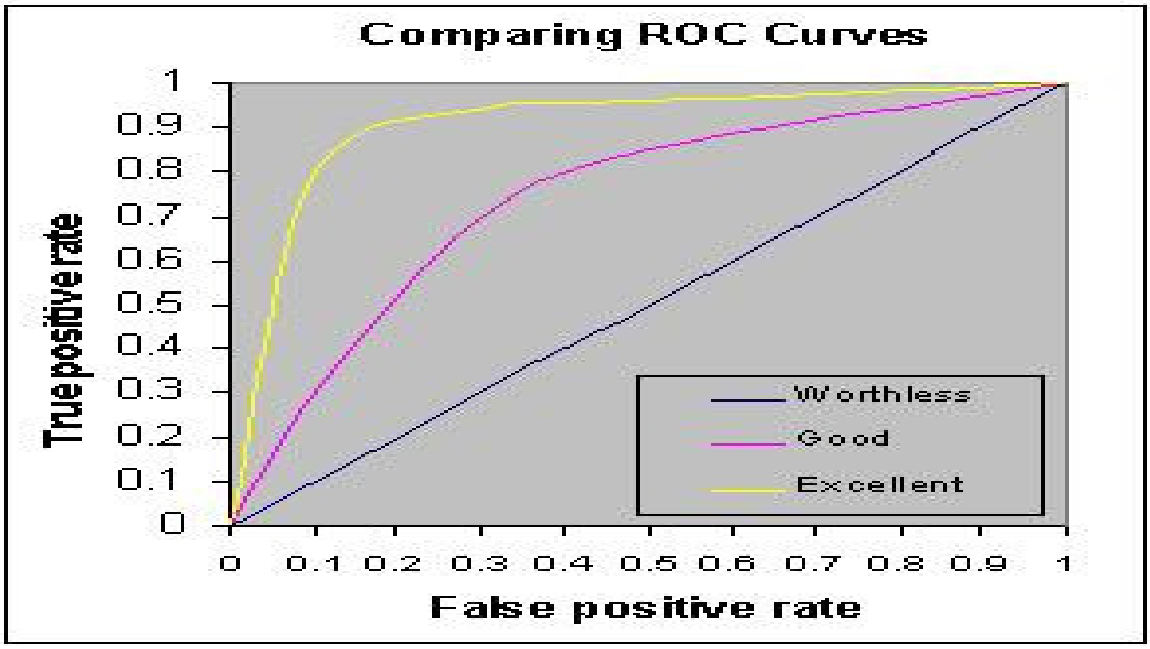
Q_1 i Q_2 wyrażają się następującymi wzorami

$$Q_1 = \frac{W}{2-W}, \qquad Q_2 = \frac{2W^2}{1+W}$$

Ponieważ pole pod krzywą ROC wyrażone jest jedynie liczbą i nie oddaje całości informacji, wskazane jest, aby rozważać je razem z wykresem krzywej ROC. Krzywe ROC mogą się różnić kształtem, lecz mieć podobne pola [13, 16, 26].

6.4. Wykresy krzywych ROC dla wyników kilku testów – statystyczne porównanie testów

Wykresy krzywych ROC dla wyników kilku testów można umieścić wspólnie na jednym wykresie. Pozycja krzywej ROC (bliskość lewego górnego rogu rysunku) daje jakościową informację o dokładności testu. Natomiast położenie dwu lub więcej krzywych dostarcza jakościowego porównania dokładności kilku testów. Krzywa leżąca powyżej od innej, odpowiada testowi o wyższej sile dyskryminacyjnej. Wykres krzywych ROC daje bezpośrednie, natychmiastowe porównania między testami na wspólnej skali, podczas gdy zarówno diagram punktowy jak i histogram częstości wymagają rysunków w różnych skalach. Nie wymaga grupowania pomiarów w klasy jak w przypadku histogramu częstości. Czułość i swoistość można odczytywać wprost z wykresu w przeciwieństwie do histogramu i diagramu punktowego.



Rysunek 6. Na wykresie są przedstawione trzy krzywe ROC, reprezentujące bardzo dobre, dobre i bezwartościowe badania. Żółta krzywa pokazuje doskonały test, krzywa granatowa odzwierciedla bezwartościowe badania, a krzywa różowa badania dobre**.**

Jedną z metod służących do porównywania zdolności różnicowania testów jest porównanie wielkości pól pod krzywymi ROC. Metoda jest zalecana ze względu na ogólny charakter wniosków *Kochańska i wsp* [55].

Kiedy dwa testy opierają się na różnych zbiorach pacjentów (zmienne niepowiązane) wartość sprawdzianu wyraża się następującym wzorem:

$$z = \frac{W_X - W_Y}{SE_{W_X - W_Y}}$$

gdzie :

W_X – pole pod krzywą ROC dla testu X

W_Y – pole pod krzywą ROC dla testu Y

$SE_{W_X - W_Y}$ - błąd standardowy różnicy pól wyrażony wzorem

$$SE_{W_X - W_Y} = \sqrt{SE_{W_X}^2 + SE_{W_Y}^2}$$

SE_{W_X} , SE_{W_Y} są oznaczeniami błędów standardowych odpowiednio dla W_X i W_Y

W przypadku, gdy testy przeprowadzane są na tym samym zbiorze pacjentów (zmienne powiązane) wyniki są skorelowane, a wzór na istotność statystyczną różnic pól pod krzywymi ROC przy danych dwóch testach X, Y wygląda następująco:

$$z = \frac{W_X - W_Y}{\sqrt{SE_{W_X}^2 + SE_{W_Y}^2 - 2r_p SE_{W_X}^2 SE_{W_Y}^2}}$$

gdzie

r_p – jest to współczynnik korelacji między polami W_X i W_Y (*Pearsona* dla danych ciągłych lub τ *Kendalla* dla danych dyskretnych) [13, 16, 17, 18].

6.5. Optymalny punkt odcięcia

Krzywe ROC zawierają wszystkie możliwe pary czułości i swoistości wyznaczone z danych eksperymentalnych dla badanego testu laboratoryjnego. Jednakże, aby można go było zastosować do dalszego leczenia pacjenta należy wybrać jeden próg decyzyjny odpowiadający optymalnemu punktowi odcięcia. Wymaga to uwzględnienia kosztów niepożądanych błędów, czyli fałszywie dodatniej i fałszywie ujemnej klasyfikacji oraz częstości występowania choroby *P* (prevalence). Ocena względnych kosztów fałszywie dodatniej i fałszywie ujemnej klasyfikacji może być dokonana z punktu widzenia pacjenta, firm ubezpieczeniowych, społeczeństwa itp. Aby połączyć oba czynniki i pokazać fałszywie dodatnich i fałszywie ujemnych w odpowiedniej proporcji musimy uwzględnić drugi czynnik czyli *prevalence* - częstość występowania choroby. Wzór, który łączy oba elementy przyjmuje postać:

$$m = \frac{koszt\ wyników\ falszywie\ dodatnich}{koszt\ wyników\ falszywie\ ujemnych} \cdot \frac{1-P}{P}$$

$$P = (A + C)/(A + B + C + D)$$

gdzie:

m – współczynnik kierunkowy prostej stycznej do krzywej ROC w najlepszym punkcie operacyjnym

W modelu nieparametrycznym dla danych ciągłych bez wiązań wykres krzywej ROC składa się z odcinków linii poziomych lub pionowych (nachylenie = 0 lub nieskończoność). Wiązania w danych laboratoryjnych występują wtedy, gdy pacjent z grupy kontrolnej ma taki sam wynik jak pacjent z grupy badanej. Punkt operacyjny jest wtedy wyznaczony w miejscu, gdzie prosta z nachyleniem równym *m* przesuwając się od lewego górnego rysunku, dotknie wykresu krzywej ROC. Punkt, który daje optymalną kombinację czułości i swoistości przy danym *P* i stosunku fałszywie dodatnich i ujemnych wyników, odpowiada najlepszemu punktowi odcięcia [8].

7. Wskaźniki antropometryczne

W pracy zastosowano następujące wskaźniki proporcji ciała:

1. Wskaźnik Queteleta:

$$WQ = \frac{masa\ ciala\ [g]}{wzrost\ [cm]}$$

2. Wskaźnik masy ciała – Body Mass Index:

$$BMI = \frac{masa\ ciala\ [kg]}{wzrost^2\ [m^2]}$$

3. Wskaźnik Rohrera:

$$WR = \frac{masa\ ciala\ [kg] \cdot 10^5}{wzrost^3\ [cm^3]}$$

4. Współczynnik masy ciała – Body Mass Coefficient

$$WMC = \frac{masa\ ciala^{1,425}\ [kg^{1,425}] \cdot 71{,}84}{wzrost^{1,275}\ [cm^{1,275}]}$$

[4, 23, 24, 25].

8. Cel pracy

Celem pracy jest:

1. Opracowanie i porównanie modeli matematycznych pozwalających na klasyfikację dzieci z nieprawidłowym i prawidłowym ciśnieniem tętniczym krwi.
2. Wyznaczenie tych wielkości, które są najistotniejsze w predykcji nieprawidłowego ciśnienia krwi u dzieci.
3. Wyznaczenie wskaźnika antropometrycznego, który najlepiej opisuje nieprawidłowe ciśnienie krwi u dzieci.

9. Materiał

Badana populacja obejmowała grupę dzieci i młodzieży między 7, a 18 rokiem życia z losowo wybranych wielkopolskich szkół. Badanie obejmowało dzieci i młodzież zdrową bez przewlekłych chorób nerek, serca i endokrynologicznych. Kryteria włączenia i wyłączenia z badań ustalali lekarze pediatrzy w oparciu o badanie ogólno pediatryczne. Badanie zostało przeprowadzone w 2009 roku. Liczebność badanej populacji wg wieku i płci przedstawia poniższa tabela:

Tabela 5. Liczebność badanej populacji

Wiek (w latach)	**Chłopcy**		**Dziewczynki**		**Razem**	
	n	**%**	**n**	**%**	**n**	**%**
7	24	2%	20	1%	44	3%
8	63	5%	74	5%	137	10%
9	46	3%	40	3%	86	6%
10	46	3%	41	3%	87	6%
11	45	3%	60	4%	105	8%
12	45	3%	77	6%	122	9%
13	37	3%	60	4%	97	7%
14	20	1%	25	2%	45	3%
15	15	1%	22	2%	37	3%
16	58	4%	112	8%	170	12%
17	60	4%	114	8%	174	13%
18	98	7%	176	13%	274	20%
Razem	557	40%	821	60%	1378	100%

Badanie zostało przeprowadzone w oparciu o opracowany kwestionariusz KARTA BADANIA PRZESIEWOWEGO NADCIŚNIENIA, który obejmował dane ogólne, uwzględniono tu informacje dotyczące miejsca zamieszkania, rodzaju szkoły, klasy oraz płci i wieku dziecka. Dla ustalenia wieku przyjęto przedział np.: do grupy dzieci 12 letnich zaliczano takie, które miały od 11 lat i 6 miesięcy do 12 lat i 5 miesięcy. W badaniu ujęto również informacje o nałogach dziecka, o warunkach mieszkaniowych, o ocenach z głównych przedmiotów szkolnych.

Badanie obejmowało wywiad rodzinny, który dotyczył występowania chorób układu krążenia w najbliższej rodzinie badanego dziecka. Uwzględniał wiek i wykształcenie rodziców.

Przeprowadzony był również wywiad dotyczący dziecka, czyli przebyte choroby, ilość hospitalizacji, poradnie, zmierzono obwody: ramienia, uda, talii, biodra, zmierzono wagę i wzrost. Trzykrotnie został dokonany pomiar ciśnienia skurczowego, rozkurczowego i tętna w odstępach od dwu do trzech tygodni z odpowiednio dobranym mankietem.

10. Metody

Dla wyznaczenia zmiennej grupującej informującej o tym, czy dziecko ma prawidłowe ciśnienie, czy nieprawidłowe wykorzystano pomiary ciśnienia skurczowego i rozkurczowego. Sprawdzono testem Friedman'a z wielokrotnymi porównaniami Dunn'a, czy istnieją istotne różnice pomiędzy trzykrotnymi pomiarami. Dla ciśnienia skurczowego istotna różnica była między pomiarem pierwszym, a drugim ($p<0.0001$), między pomiarem pierwszym, a trzecim ($p<0.0001$), oraz między pomiarem drugim, a trzecim ($p<0.0001$). Jako ciśnienie skurczowe mające pomóc w pogrupowaniu dzieci wyznaczano medianę z trzech pomiarów. Tym samym testem sprawdzono również istnienie istotnych różnic dla trzykrotnych pomiarów ciśnienia rozkurczowego. Istotna różnica została wykryta między pierwszym pomiarem, a trzecim ($p=0.049$), natomiast nie stwierdzono istotnej różnicy pomiędzy pomiarami pierwszym, a drugim i drugim, a trzecim stąd pomiar drugi został uznany jako grupujący. Ciśnienie skurczowe lub rozkurczowe dzieci uznanych za chore przekracza 90 centyl, pozostałe dzieci uznano za zdrowe. Liczebność w poszczególnych grupach ze względu na płeć i wiek przedstawia poniższa tabela.

Tabela 6

Wiek	**Chłopcy**					**Dziewczynki**				
(w latach)	**0**	**%**	**1**	**%**	**Suma**	**0**	**%**	**1**	**%**	**Suma**
7	24	4%	0	0%	24	20	2%	0	0%	20
8	63	11%	0	0%	63	74	9%	0	0%	74
9	46	8%	0	0%	46	39	5%	1	0%	40
10	41	7%	5	1%	46	41	5%	0	0%	41
11	42	8%	3	1%	45	57	7%	3	0%	60
12	40	7%	5	1%	45	71	9%	6	1%	77
13	32	6%	5	1%	37	53	6%	7	1%	60
14	13	2%	7	1%	20	18	2%	7	1%	25
15	15	3%	0	0%	15	19	2%	3	0%	22
16	46	8%	12	2%	58	88	11%	24	3%	112
17	42	8%	18	3%	60	88	11%	26	3%	114
18	70	13%	28	5%	98	147	18%	29	4%	176
Razem	474	85%	83	15%	557	715	87%	106	13%	821

0 – dzieci z prawidłowym ciśnieniem

1 – dzieci z nieprawidłowym ciśnieniem

W badaniach użyto: testu Grubbs'a, Fisher'a Freeman'a Halton'a oraz Mann'a Whitney'a dla sprawdzenia odpowiednio występowania punktów odstających, ewentualnych zależności oraz różnic pomiędzy zmiennymi.

W celu wyznaczenia modelu pozwalającego na klasyfikacje dzieci zastosowano trzy metody. Pierwsza z nich to regresja logistyczna.

Jako metodę estymacji parametrów w regresji logistycznej wybrano metodę Quasi-Newtona, która w każdym kolejnym kroku w różnych punktach oszacowuje funkcje aby znaleźć pochodne pierwszego i drugiego rzędu, które informują o kierunku zmian i o szybkości zmian nachylenia funkcji. Te informacje wykorzystywane są w celu znalezienia minimum funkcji straty.

Drzewa klasyfikacyjne są drugą techniką użytą w badaniach. Dla znalezienia najlepszego drzewa użyto metodę C&RT, która może być stosowana zarówno dla zmiennych w skali nominalnej jak i porządkowej. Metoda ta sprawdza wszystkie możliwe podziały dla każdej zmiennej predykcyjnej w celu znalezienia podziału, w którym mamy największą redukcję braku dopasowania. Jako miarę dobroci wybrano miarę Gini'ego, która osiąga wartość zero, gdy w danym węźle wystąpi tylko jedna klasa, a osiąga wartość maksymalną, gdy liczności klas w danym węźle są równe. Jako regułę zatrzymania wybrano metodę FACT, która traktuje kompletne drzewo klasyfikacyjne zawierające wszystkie podziały jako drzewo właściwej wielkości. Podziały są robione do czasu gdy węzeł jest czysty – nie zwiera obiektów błędnie zaklasyfikowanych lub jeżeli obejmuje nie więcej niż minimalną liczbę przypadków obliczoną na podstawie frakcji obiektów.

Kolejna metoda, to technika MARSPLINES. Aby znaleźć najlepszy model u chłopców użyto 3296 funkcji bazowych i osiem interakcji. Dla dziewczynek do stworzenia modelu potrzebne były 4824 funkcje bazowe ze stopniem interakcji równym sześć.

W celu wyznaczenia najlepszego wskaźnika antropometrycznego użyto krzywych ROC

Do obliczeń użyto programu Statistica 8. oraz programu StatXact 9.

Pomiary ciśnienia tętniczego krwi zostały przeprowadzone z dokładnością do 1mm Hg, wysokość i obwody ramienia, talii, uda, biodra z dokładnością do 1mm, masa ciała z dokładnością do 10g. Wyniki danych w tabelach są zrzutami ekranowymi zastosowanych pakietów statystycznych dlatego interpretacja każdej wartości powinna być zaokrąglana zgodnie z ogólnymi regułami do odpowiedniej liczby miejsc znaczących.

11. Wyniki

11.1. Analiza wstępna

Sprawdzono testem Fisher'a Freeman'a Halton'a (użyto estymatora asymptotycznego), czy jest zależność między wiekiem, a ciśnieniem określonym w skali nominalnej jako prawidłowe lub nieprawidłowe. Zarówno u chłopców jak i u dziewczynek stwierdzono istotną zależność odpowiednio z p=0.0001 i p=0.0005. Ponieważ liczebności dzieci z nieprawidłowym ciśnieniem były znacząco mniejsze przed piętnastym rokiem życia zarówno u chłopców jak i u dziewczynek, to do dalszej analizy brano pod uwagę tylko te dzieci, które przekroczyły piętnaście lat **tabela 6.**

Za pomocą analizy jednowymiarowej zbadano, czy istnieją różnice i zależności między grupą dzieci z prawidłowym i nieprawidłowym ciśnieniem. Ponieważ zmienne nie mają rozkładu zgodnego z rozkładem normalnym (charakteryzują się silną skośnością rozkładów), wybrano test Mann'a Whitney'a dla stwierdzenia istnienia istotnych różnic powyżej piętnastego roku życia w następujących parametrach.

Tabela 7. Wartość parametrów statystyki opisowej dla badanych wielkości w grupie chłopców z nieprawidłowym ciśnieniem

	N	Średnia	Mediana	Moda	Minimum	Maximum	Odch.Stand.	Skośność	Kurtoza
obwódramienia	58	27,8293	28,00000	29,00000	22,00000	34,0000	2,60896	0,156100	0,112550
obwód talii	58	81,4879	79,10000	78,00000	67,30000	112,0000	9,48099	1,509924	2,512771
obwód bioder	57	100,4772	99,80000	101,0000	86,00000	126,1000	8,06591	0,908276	1,134108
obwód uda	58	52,7569	53,10000	54,00000	35,00000	68,0000	6,45685	0,165624	0,519347
bmi	58	23,7380	23,29790	Multiple	17,80864	34,4591	3,64277	1,079625	1,765077
WQ	58	0,0004	0,00042	Multiple	0,00031	0,0006	0,00007	1,067123	1,582039
WR	58	1,3285	1,29477	Multiple	0,98937	1,8994	0,19812	0,934178	1,258907
WMC	58	46,5814	44,87078	Multiple	29,31148	84,1734	11,56744	1,288675	2,165482
talia/biodra	57	0,8122	0,80952	Multiple	0,70594	0,9464	0,05166	0,543980	0,417118
talia/wysokość	58	0,4558	0,44715	,4166667	0,38023	0,6205	0,04808	1,387799	2,616340
średnia tętno	58	77,3132	77,00000	Multiple	56,66667	108,3333	10,91506	0,653815	0,676898

Tabela 8. Wartość parametrów statystyki opisowej dla badanych wielkości w grupie chłopców z prawidłowym ciśnieniem

	N	Średnia	Mediana	Moda	Minimum	Maximum	Odch.Stand.	Skośność	Kurtoza
obwódramienia	158	26,88418	27,00000	Multiple	20,40000	33,2000	2,916230	0,181846	-0,695999
obwód talii	158	76,82848	76,00000	75,00000	62,70000	104,0000	7,619553	0,818854	0,999007
obwód bioder	158	97,01266	97,00000	100,0000	76,00000	113,0000	6,847369	-0,063845	0,136502
obwód uda	158	51,19747	51,00000	49,00000	40,50000	66,5000	5,097357	0,527474	-0,048729
bmi	158	21,92040	21,31419	Multiple	16,13819	30,0838	3,036797	0,668305	-0,081504
WQ	158	0,00039	0,00039	Multiple	0,00026	0,0005	0,000057	0,536725	-0,026074
WR	158	1,22717	1,20515	Multiple	0,93159	1,7708	0,178569	0,786595	0,365487
WMC	158	41,50477	40,53813	Multiple	22,66899	67,1365	8,777656	0,681235	0,188732
talia/biodra	158	0,79165	0,78351	Multiple	0,68085	0,9753	0,048925	1,046831	1,809933
talia/wysokość	158	0,42956	0,42111	,3977273	0,34783	0,5503	0,040468	0,685668	0,166696
średnia tętno	158	71,19198	70,66667	61,66667	48,00000	99,0000	9,106300	0,569655	0,274168

Tabela 9. Wartość parametrów statystyki opisowej dla badanych wielkości w grupie dziewczynek z nieprawidłowym ciśnieniem

	N	Średnia	Mediana	Moda	Minimum	Maximum	Odch.Stand.	Skośność	Kurtoza
obwódramienia	79	26,26203	26,30000	29,00000	20,00000	32,0000	2,83080	0,011957	-0,578121
obwód talii	79	73,55443	72,00000	Multiple	61,00000	96,0000	8,25031	0,802831	0,022558
obwód bioder	79	99,15823	98,00000	93,00000	67,00000	119,0000	8,37173	-0,452863	1,490316
obwód uda	79	54,46582	55,00000	61,00000	44,60000	66,8000	4,98150	0,155599	-0,774364
bmi	79	23,33381	22,51607	Multiple	17,29631	32,7919	3,79950	0,609728	-0,552816
WQ	79	0,00039	0,00037	Multiple	0,00029	0,0005	0,00006	0,599065	-0,554165
WR	79	1,40538	1,36110	Multiple	1,02345	2,0419	0,24189	0,734308	-0,086337
WMC	79	40,45576	37,89267	Multiple	26,20370	65,5244	9,54661	0,716229	-0,357172
talia/biodra	79	0,74202	0,73786	Multiple	0,65728	0,9701	0,05547	1,579894	3,702430
talia/wysokość	79	0,44248	0,43353	,4335260	0,36477	0,5793	0,05025	0,862841	0,236955
średnia tętno	79	81,87131	80,00000	Multiple	64,00000	120,3333	12,16175	0,739994	0,353096

Tabela 10. Wartość parametrów statystyki opisowej dla badanych wielkości w grupie dziewczynek z prawidłowym ciśnieniem

	N	Średnia	Mediana	Moda	Minimum	Maximum	Odch.Stand.	Skośność	Kurtoza
obwódramienia	323	24,77059	24,60000	24,00000	18,80000	32,0000	2,155904	0,392337	0,499098
obwód talii	323	69,05356	68,50000	69,00000	54,50000	89,9000	6,033808	0,569989	0,385749
obwód bioder	323	94,70186	94,30000	Multiple	77,00000	111,5000	6,016829	0,027237	0,038403
obwód uda	323	50,91796	50,70000	52,00000	40,00000	63,6000	4,262925	0,374317	0,121005
bmi	323	21,10688	20,73514	17,99816	9,96094	30,3737	2,748106	0,528695	1,227326
WQ	323	0,00035	0,00034	,0002970	0,00016	0,0005	0,000047	0,439279	1,357919
WR	323	1,27983	1,25175	1,090798	0,62256	1,8786	0,172264	0,583381	0,977663
WMC	323	34,59965	33,82216	27,39074	11,23138	61,3272	6,699086	0,695466	1,505134
talia/biodra	323	0,72911	0,72289	Multiple	0,64894	0,8739	0,042094	0,765960	0,676035
talia/wysokość	323	0,41843	0,41190	,4285714	0,33851	0,5516	0,035976	0,854662	0,950679
średnia tętno	323	73,55212	72,00000	70,66667	55,00000	101,3333	9,204905	0,422223	-0,171916

Tabela 11. Testowanie istotności różnic statystycznych w grupie chłopców za pomocą testu Mann'a Whitney'a

	Średnia 0	Średnia 1	Min 0	Min 1	Max 0	Max 1	Odch. Stand. 0	Odch. Stand. 1	p-wartość	n 0	n 1
obwód ramienia	26,884	27,829	20,400	22,000	33,200	34,000	2,916	2,609	0,022	158	58
obwód talii	76,828	81,488	62,700	67,300	104,000	112,000	7,620	9,481	<0,0001	158	58
obwód bioder	97,013	100,477	76,000	86,000	113,000	126,100	6,847	8,066	0,010	158	57
BMI	21,920	23,738	16,138	17,809	30,084	34,459	3,037	3,643	<0,0001	158	58
WQ	0,00039	0,0004	0,00026	0,00031	0,0005	0,0006	0,000057	0,00007	0,002	158	58
WR	1,227	1,328	0,932	0,989	1,771	1,899	0,179	0,198	<0,0001	158	58
WMC	41,505	46,581	22,669	29,311	67,137	84,173	8,778	11,567	0,002	158	58
talia/biodra	0,792	0,812	0,681	0,706	0,975	0,946	0,049	0,052	0,003	158	57
talia/wysokość	0,430	0,456	0,348	0,380	0,550	0,620	0,040	0,048	<0,0001	158	58
tętno	71,192	77,313	48,000	56,667	99,000	108,333	9,106	10,915	<0,0001	158	58

0 – prawidłowe ciśnienie
1 – nieprawidłowe ciśnienie

Tabela 12. Testowanie istotności różnic statystycznych w grupie dziewczynek za pomocą testu Mann'a Whitney'a

	Średnia 0	Średnia 1	Min 0	Min 1	Max 0	Max 1	Odch. Stand. 0	Odch. Stand. 1	p-wartość	n 0	n 1
obwód ramienia	24,771	26,262	18,800	20,000	32,000	32,000	2,831	2,916	<0,0001	323	79
obwód talii	69,054	73,554	54,500	61,000	89,900	96,000	8,250	7,620	<0,0001	323	79
obwód bioder	94,702	99,158	77,000	67,000	111,500	119,000	8,372	6,847	<0,0001	323	79
obwód uda	50,918	54,466	40,000	44,600	63,600	66,800	4,982	5,097	<0,0001	323	79
BMI	21,107	23,334	9,961	17,296	30,374	32,792	3,800	3,037	<0,0001	323	79
WQ	0,00035	0,00039	0,00016	0,00029	0,0005	0,0005	0,000047	0,00006	<0,0001	323	79
WR	1,280	1,405	0,623	1,023	1,879	2,042	0,242	0,179	<0,0001	323	79
WMC	34,600	40,456	11,231	26,204	61,327	65,524	9,547	8,778	<0,0001	323	79
talia/biodra	0,729	0,742	0,649	0,657	0,874	0,970	0,055	0,049	<0,0001	323	79
talia/wysokość	0,418	0,442	0,339	0,365	0,552	0,579	0,050	0,040	<0,0001	323	79
tętno	73,552	81,871	55,000	64,000	101,333	120,333	12,162	9,106	<0,0001	323	79

0 – prawidłowe ciśnienie
1 – nieprawidłowe ciśnienie

Nie wykryto istotnych różnic w badanych parametrach zarówno u chłopców jak i u dziewczynek związanych z piciem alkoholu, paleniem papierosów i z wysiłkiem fizycznym.

11.2. Regresja logistyczna

Zaproponowany model logistyczny chłopców w 76% potrafi prawidłowo przewidywać przynależność przypadków do klas, u dziewczynek wynik jest wyższy wynosi 83%. O dobrej klasyfikacji świadczą również ilorazy szans obu modeli (chłopcy 6,54 i dziewczynki 13,35), są one dużo większe od jedności co pozwala nam stwierdzić, że klasyfikacja przypadków za pomocą modelu logistycznego ma sens **tabela 14, 18, 21.**

W obu modelach tętno okazało się istotnym statystycznie parametrem z $p<0,0001$. Drugim parametrem u chłopców, który wszedł do modelu jest wskaźnik wagowo wzrostowy talia/wysokość $p<0,0001$, a u dziewczynek obwód uda $p<0,0001$ **tabela 13, 17.**

Odczytując wyniki z **tabel 13** oraz **17.** widzimy, że jednostkowy iloraz szans u chłopców dla wskaźnika wagowo wzrostowego jest dużo większy od jedności, natomiast dla tętna jest o jedną dziesiątą większy. Ponieważ oba ilorazy są większe od jedności można sformułować wniosek, że im wyższe oba parametry, a w szczególności wskaźnik talia/wysokość, tym większe prawdopodobieństwo wystąpienia nieprawidłowego ciśnienia u chłopców.

U dziewcząt mamy podobną sytuację jednostkowe ilorazy szans dla obu zmiennych są większe od jedności o jedną i o dwie dziesiąte, stąd ten sam wniosek: im wyższe tętno, im wyższy obwód uda tym większa szansa na wystąpienie nieprawidłowego ciśnienia u dziewczynek. Czułość modelu logistycznego dla chłopców wynosi 26%, natomiast swoistość 95%; dla dziewczynek czułość równa jest 25%, a swoistość 95%, w obu przypadkach swoistość jest bardzo dobra.

Możemy stwierdzić, że model zaproponowany przez regresję logistyczną jest odpowiedni dla choroby nadciśnieniowej, gdyż w jej ciężkim i długotrwałym farmakologicznym leczeniu bezbłędna wykrywalność dzieci rzeczywiście zdrowych (swoistość wysoka) jest istotniejsza (nie chcemy szkodzić zdrowym poprzez narażanie ich na ewentualne niepotrzebne badania inwazyjne lub podawanie leków).

Tabele 15, 16, 19 i **20** pokazują jak zmienia się iloraz szans (OR) w zależności od wzrostu tętna, wskaźnika wagowo-wzrostowego oraz obwodu uda.

11.2.1. Chłopcy

Tabela 13. Model regresji logistycznej

Model: Regr. logistyczna (logit)

Strata: Największe prawd. bł. średnkw.skal. Całkowita strata: 109,27959921 Chi2(2)=32,770 p=,00000

	Stała B0	**talia/wysokość**	**tętno**
Ocena	-12,8487	1,511954E+01	0,0697047
Błąd standard.	2,362932	3,904179E+00	0,01731033
Chi-kwadrat Walda	29,56741	1,499746E+01	16,21487
poziom p	5,44E-08	1,078702E-04	5,66789E-05
Iloraz szans z.jedn.	2,63E-06	3,684109	1,072191
-95%CL	2,5E-08	1,675324	1,036224
+95%CL	0,000277	8,101511	1,109408
Iloraz szans zakr.		6,172460E+01	67,05528
-95%CL		7,570475E+00	8,557887
+95%CL		5,032613E+02	525,4113

Tabela 14. Klasyfikacja przypadków przez utworzony model regresji logistycznej

Iloraz szans: 6,5407 % poprawnych: 76,39%

	przewidywane 0	**przewidywane 1**	**Procent**
obserwowane 0	150	8	94,93671
obserwowane 1	43	15	25,86207

Tabela 15. Iloraz szans dla tętna

tętno	**1**	**OR**	**95% CI**
<45-65)	9	wartość referencyjna	
<65-75)	13	1,101694915	0,43 – 2,8
<75-85)	27	2,872340426	1,22 – 6,77
<85-95)	4	6,666666667	1,27 – 35,04
<95-115)	5	6,25	1,4 – 27,93

Tabela 16. Iloraz szans dla wskaźnika wagowo-wzrostowego talia/wysokość

talia/ wysokość	1	OR	95% CI
<0,35-0,4)	5	wartość referencyjna	
<0,4-0,45)	27	2,736	0,98 – 7,67
<0,45-0,5)	18	3,8	1,27 – 11,31
<0,5-0,55)	5	4,75	1,1 – 20,36
<0,55-0,65)	3	22,8	1,97 – 96,6

11.2.2. Dziewczynki

Tabela 17. Model regresji logistycznej

Model: Regr. logistyczna (logit)

Strata: Największe prawd. bł. średnkw.skal. Całkowita strata: 165,18255461 Chi2(2)=68,046 p=,00000

	Stała B0	obwód uda	tętno
Ocena	-15,1158	0,1570187	0,0704257
Błąd standard.	1,965509	0,03069321	0,01330905
Chi-kwadrat Walda	59,14402	26,17086	28,00065
poziom p	1,51E-14	3,14305E-07	1,22072E-07
Iloraz szans z.jedn.	2,72E-07	1,170017	1,072965
-95%CL	5,72E-09	1,101506	1,045255
+95%CL	1,3E-05	1,242791	1,101409
Iloraz szans zakr.		67,22876	99,5984
-95%CL		13,34274	18,02435
+95%CL		338,739	550,3579

Tabela 18. Klasyfikacja przypadków przez utworzony model regresji logistycznej

Iloraz szans: 13,347 Procent poprawnych: 83,33%

	przewidywane 0	przewidywane 1	Procent
obserwowane 0	315	8	97,52322
obserwowane 1	59	20	25,31646

Tabela 19. Iloraz szans dla tętna

tętno	1	OR	95% CI
<55-65)	4	wartość referencyjna	
<65-75)	19	2,356299213	0,77 – 7,22
<75-85)	31	4,931818182	1,66 – 14,64
<85-95)	11	6,663461538	1,94 – 22,85
<95-121>	14	27,5625	7,27 – 104,5

Tabela 20. Iloraz szans dla obwodu uda

obwód uda	1	OR	95% CI
<40-45)	1	wartość referencyjna	
<45-50)	13	2,24137931	0,28 – 18,1
<50-55)	25	3,787878788	0,49 – 29,52
<55-60)	23	10,95238095	1,38 – 86,94
<60-67>	17	26,15384615	3,09 – 98,1

11.2.3. Podsumowanie

Tabela 21 Podsumowanie modeli regresji logistycznej dla chłopców i dziewczynek

	Chłopcy	**Dziewczynki**
zmienne w modelu	talia/wysokość tętno	obwód uda tętno
procent przypadków błędnie zaklasyfikowanych	24%	17%
ilość dzieci z nieprawidłowym ciśnieniem – 1	58	79
ilość dzieci z prawidłowym ciśnieniem – 0	158	323
Procent dzieci z nieprawidłowym ciśnieniem błędnie zaklasyfikowanych	74%	75%
Procent dzieci z prawidłowym ciśnieniem błędnie zaklasyfikowanych	5%	2%
iloraz szans	6,54	13,35

11.3. Drzewa klasyfikacyjne (CRT)

Wygenerowane drzewo klasyfikacyjne chłopców w 73% potrafi prawidłowo przewidywać przynależność przypadków do klas, u dziewczynek wynik jest minimalnie wyższy wynosi 80%. Pierwszym i najistotniejszym kryterium podziału dla chłopców jest BMI (stała podziału równa 23), które dzieli grupę chłopców na zdrowych i na tych z nieprawidłowym ciśnieniem. U dziewczynek najistotniejszą zmienną jest obwód bioder (stała podziału równa 104). Kolejnym istotnym kryterium podziału zarówno dla chłopców jak i dla dziewczynek jest tętno, a następnie dla chłopców wskaźnik wagowo wzrostowy talia/wysokość, a u dziewczynek obwód talii **tabela 22, 25.**

Odczytując wyniki z **tabel 24** oraz **27**. widzimy, że średnie koszty dla prób testowych, zwane globalnymi kosztami sprawdzianu krzyżowego (global CV cost) jak również ich błąd standardowy są zbliżone, a nawet minimalnie mniejsze u chłopców do kosztów sprawdzianu krzyżowego dla wybranego drzewa (CV cost) i odpowiednio jego błędu standardowego. Wskazuje to, że procedura automatycznego wyboru drzewa wybrała odpowiednie drzewo i nie możemy w tym przypadku mówić o przeuczeniu. Zauważyć również można, że koszty w próbie uczącej tzw. koszty resubstytucji (resubstitutive cost) zarówno dla chłopców jak i dziewczynek są zbliżone do kosztów sprawdzianu krzyżowego dla wybranego drzewa (CV cost), co wskazuje, że wybrane drzewo jest drzewem najlepszym. Czułość drzewa klasyfikacyjnego dla chłopców wynosi 54%, natomiast swoistość 79%, dla dziewczynek czułość równa jest 32%, a swoistość 91%. W obu przypadkach swoistość jest bardzo dobra. Model zaproponowany przez drzewa klasyfikacyjne podobnie jak regresja logistyczna jest dobry dla wykrywania choroby nadciśnieniowej, ze względu na wysoką swoistość.

11.3.1. Chłopcy

Tabela 22. Struktura drzewa klasyfikacyjnego

	Lewostr. gałąź	Prawostr. gałąź	n klas 0	n klas 1	Przewid. klasa	Podział stała	Podział zmienna
1	2	3	158	58	0	-22,8725	bmi
2			112	21	0		
3	4	5	46	37	0	-72,1667	tętno
4	6	7	31	12	0	0,4403	talia/wysokość
5			15	25	1		
6			1	4	1		
7	8	9	30	8	0	0,5436	talia/wysokość
8			30	6	0		
9			0	2	1		

Rysunek 7. Interpretacja graficzna drzewa klasyfikacyjnego

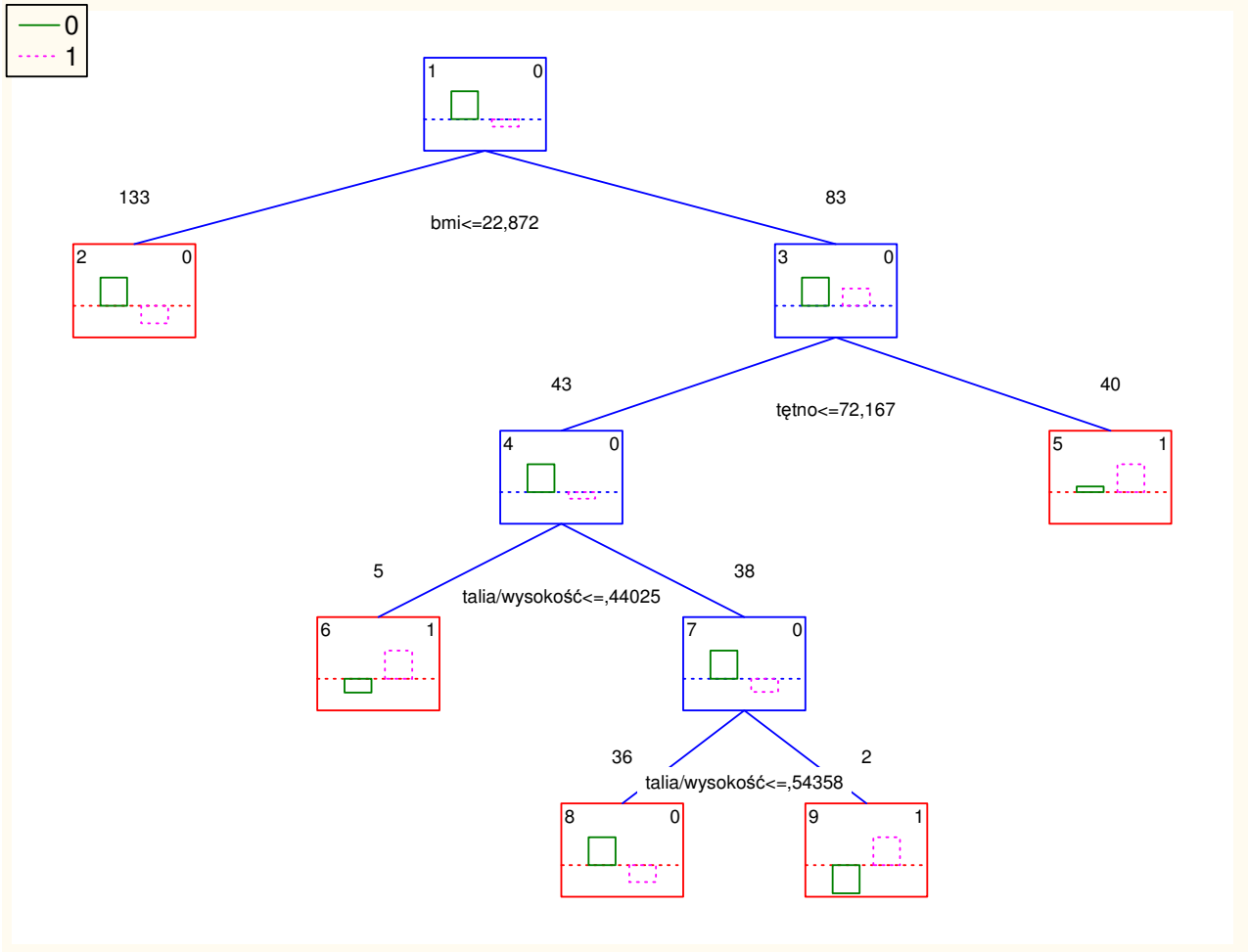


Tabela 23. Udział poszczególnych zmiennych w modelu CRT

0 – udział minimalny

100 – udział maksymalny

	Ranking
obwód ramienia	53
obwód talii	77
bmi	96
talia/wysokość	100
tętno	71

Tabela 24. Błędy i macierze błędnych klasyfikacji w modelu CRT

	końcowe n węzłów	S. krzyż. koszt	Std. błąd	Resub. koszt	Węzeł złożon.
***1**	5	0,296296	0,034172	0,199074	0

Błędne klasyfikacje dla próby uczącej
N próby uczącej = 216

	obserwowane 0	obserwowane 1
przewidywane 0	142	27
przewidywane 1	16	31

Błędne klasyfikacje w globalnym s. krzyż.
Globalne koszty = ,27315; odch. std. = ,03032

	obserwowane 0	obserwowane 1
przewidywane 0	125	26
przewidywane 1	s33	32

11.3.2. Dziewczynki

Tabela 25. Struktura drzewa klasyfikacyjnego

	Lewostr. gałąź	Prawostr. gałąź	n klas 0	n klas 1	Przewid. klasa	Podział stała	Podział zmienna
1	2	3	323	79	0	104,150	obwód bioder
2	4	5	306	53	0	97,667	tętno
3			17	26	1		
4	6	7	304	47	0	72,833	tętno
5			2	6	1		
6			164	11	0		
7	8	9	140	36	0	62,800	obwód talii
8			19	0	0		
9	10	11	121	36	0	0,705	talia/biodra
10			27	14	0		
11	12	13	94	22	0	82,250	obwód talii
12	14	15	93	20	0	103,900	obwód bioder
13			1	2	1		
14			93	19	0		
15			0	1	1		

Rysunek 8. Interpretacja graficzna drzewa klasyfikacyjnego

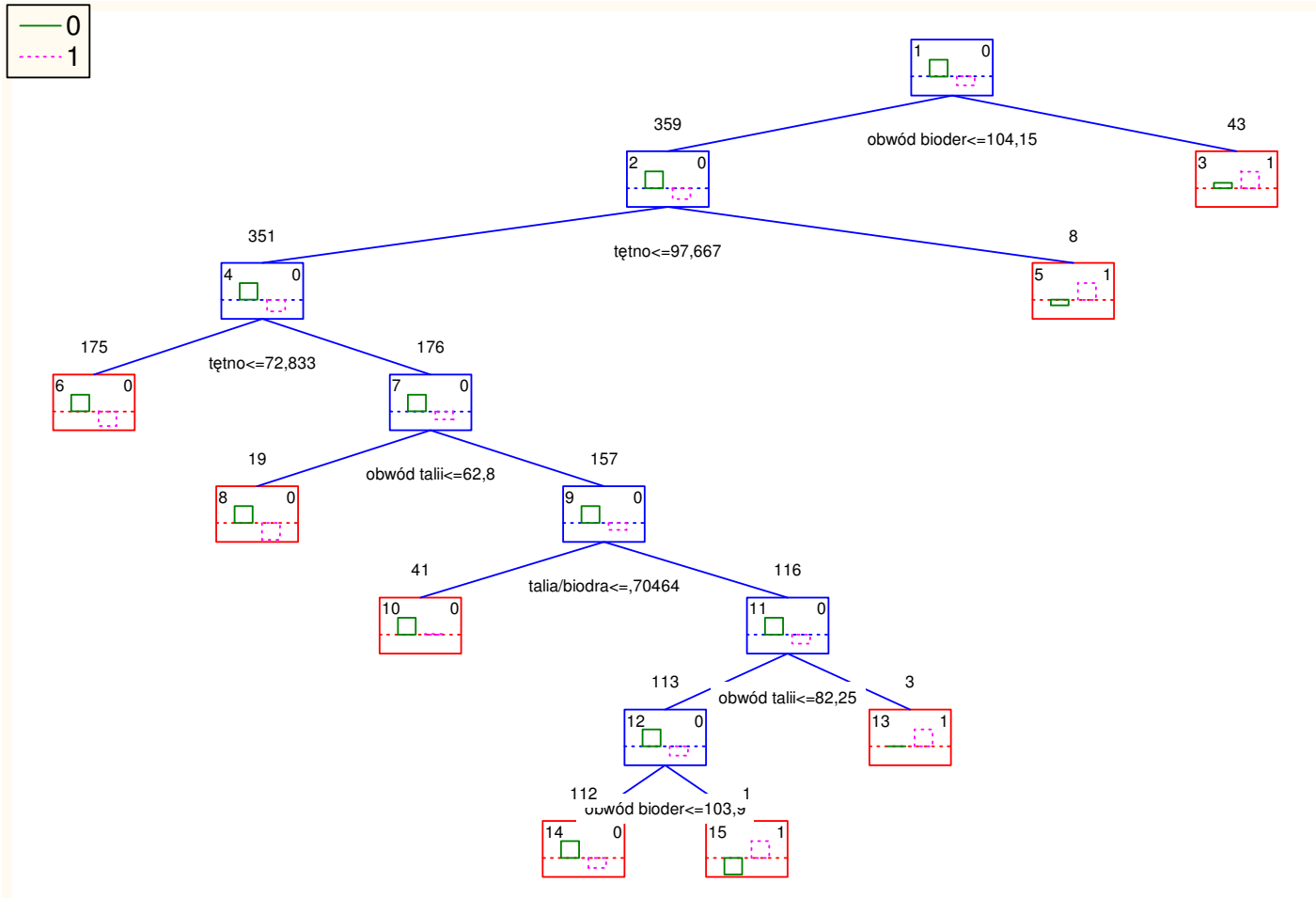


Tabela 26. Udział poszczególnych zmiennych w modelu CRT

0 – udział minimalny

100 – udział maksymalny

	Ranking
obwód talii	73
obwód bioder	85
talia/biodra	71
tętno	100

Tabela 27. Błędy i macierze błędnych klasyfikacji w modelu CRT

	końcowe n węzłów	**S. krzyż. koszt**	**Std. błąd**	**Resub. koszt**	**Węzeł złożon.**
***1**	8	0,181592	0,020165	0,159204	0,00

Błędne klasyfikacje dla próby uczącej
N próby uczącej = 402

	obserwowane 0	**obserwowane 1**
przewidywane 0	303	44
przewidywane 1	20	35

Błędne klasyfikacje w globalnym s. krzyż.
Globalne koszty = ,20398; odch. std. = ,0201

	obserwowane 0	**obserwowane 1**
przewidywane 0	295	54
przewidywane 1	28	25

11.3.3. Podsumowanie

Tabela 28. Podsumowanie modeli drzew klasyfikacyjnych CRT dla chłopców i dziewczynek

	chłopcy	**dziewczynki**
zmienne w modelu	BMI obwód talii obwód ramienia tętno talia/wysokość	obwód bioder obwód talii talia/biodro tętno
ilość węzłów końcowych	5	8
koszt globalnego sprawdzianu krzyżowego	27%	20%
ilość dzieci z nieprawidłowym ciśnieniem – 1	58	79
ilość dzieci z prawidłowym ciśnieniem – 0	158	323
Procent dzieci z nieprawidłowym ciśnieniem błędnie zaklasyfikowanych	46%	68%
Procent dzieci z prawidłowym ciśnieniem błędnie zaklasyfikowanych	21%	9%

11.4. Multivariate Adaptive Regression Splines – MARSplines

Wykonana poniżej analiza MARSplines potrafi w 69% prawidłowo przewidywać przynależność przypadków do klas u chłopców, natomiast u dziewczynek w 76% **tabela 29, 33**. Zarówno u chłopców jak i u dziewczynek zmienna tętno jest predyktorem najczęściej wykorzystywanym w tworzeniu funkcji bazowych. Na drugim miejscu u chłopców jest obwód uda, a u dziewczynek obwód bioder **tabela 31, 35. Tabele 30 i 34** zawierają współczynniki modelu i jego funkcje bazowe. Wartości podświetlone na czerwono wskazują na funkcję bazową typu max(0;zmienna niezależna-węzeł), natomiast wartości niepodświetlone, to funkcje typu max(0;węzeł- zmienna niezależna). Dla prawidłowego ciśnienia u chłopców początek równania modelu MARSplines będzie wyglądał następująco:

*prawidłowe ciśnienie=1,0703+0,0302*max(0;tętno-68,66667)+24,1214*max(0;0,4375-talia/wysokość)-0,3668*max(0;talia/wysokość-0,4375)*max(tętno-80,66667)-...*

Czułość modelu dla chłopców wynosi 61%, natomiast swoistość 95%, dla dziewczynek czułość równa jest 58%, a swoistość 99%. W obu przypadkach swoistość jest bardzo wysoka niemal 100%.

Analizując wykresy można zaobserwować, że największe znaczenie, dla rozpoznawania nieprawidłowego ciśnienia u chłopców jak i dziewczynek ma tętno. Na **rysunku 10** pokazano, że na nadciśnienie największą szansę mają chłopcy z wysokim tętnem i nieco mniejszą szansę z wysokim WMC, a jeszcze mniejszą z podwyższonymi oboma wskaźnikami. Na **rysunku 9** widzimy, że tylko tętno wpływa na wystąpienie nieprawidłowego ciśnienia, natomiast **rysunek 11** wskazuje na wzrost zagrożenia nadciśnieniem u chłopców jeżeli mamy podwyższone tylko tętno lub nieco mniejszą szansę jeżeli jest podwyższona tylko talia/wysokość. Na **rysunku 12** widzimy, że znaczący wpływ na ciśnienie u dziewczynek ma tylko tętno. Natomiast na **rysunkach 13** i **14** widać, że ciśnienie zależy poza tętnem również od obwodu ramienia i obwodu uda.

11.4.1. Chłopcy

Tabela 29. Podsumowanie modelu MARSplines

	Wartość
Niezależnych	8
Zależnych	1
Liczba czynników	15
Liczba funkcji bazowych	35
Rząd interakcji	8
błąd (GCV)	0,309300

Tabela 30. Współczynniki model u MARSplines

	Współcz. 0	Współcz. 1	Węzły obwód talii	Węzły obwód bioder	Węzły obwód uda	Węzły BMI	Węzły WR	Węzły WMC	Węzły talia/wysokość	Węzły tętno
W. wolny	1,0703	-0,0703								
Wsp.1	0,0302	-0,0302								68,66667
Wsp.2	24,1214	-24,1214							0,437500	
Wsp.3	-0,3668	0,3668							0,437500	80,66667
Wsp.4	-2,4370	2,4370						45,37284	0,437500	80,66667
Wsp.5	0,3534	-0,3534		94,2000	52,50000				0,437500	
Wsp.6	-0,0014	0,0014		108,0000				49,55665		68,66667
Wsp.7	-0,0037	0,0037		108,0000				49,55665		68,66667
Wsp.8	-24,5029	24,5029							0,417582	
Wsp.9	0,0293	-0,0293			52,50000					
Wsp.10	-0,0028	0,0028			46,20000					68,66667
Wsp.11	-0,0091	0,0091			46,20000			37,55049		68,66667
Wsp.12	-0,0002	0,0002	62,70000		46,20000			37,55049		68,66667
Wsp.13	-0,0073	0,0073		95,5000	52,50000			39,87382		
Wsp.14	30,6735	-30,6735	84,00000		52,50000	24,05367	1,484428		0,465909	

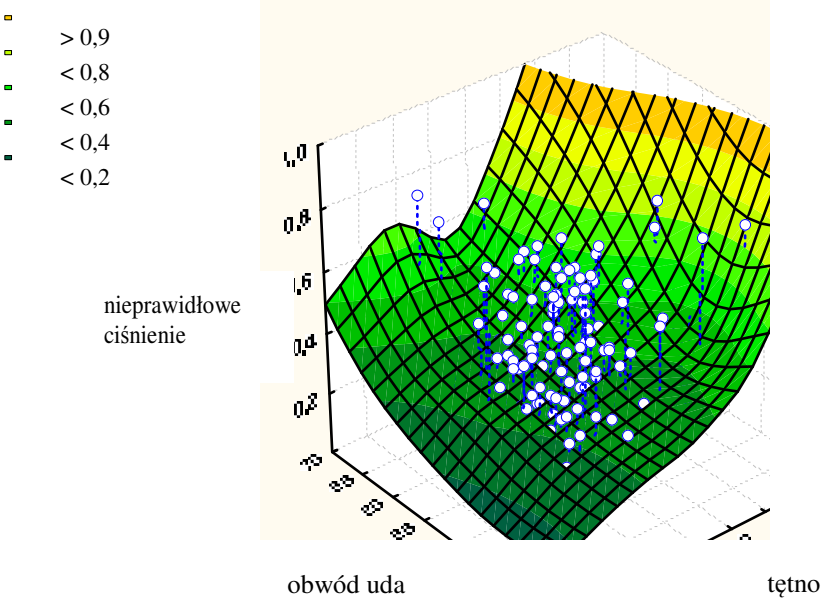
Tabela 31. Udział poszczególnych zmiennych w modelu MARSplines

	Odwołania (do funkcji bazowej)
obtalii	2
obbioder	4
obuda	7
bmi	1
WR	1
WMC	6
talia/wysokość	6
tętno	8

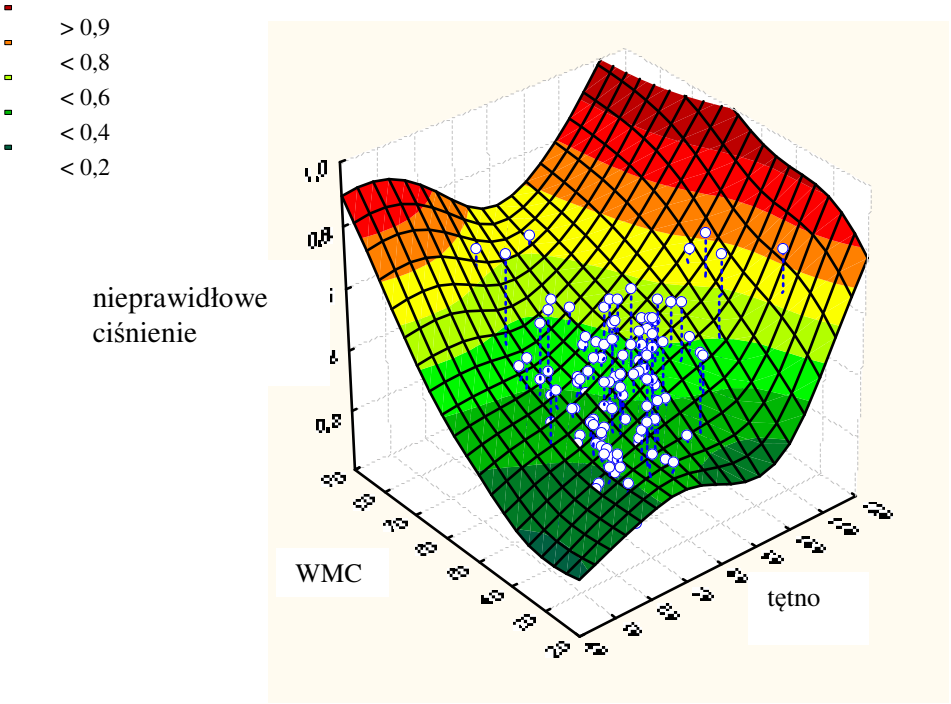
Tabela 32. Macierz błędów modelu MARSplines

	obserwowane 0	obserwowane 1
przewidywane 0	150	22
przewidywane 1	8	35

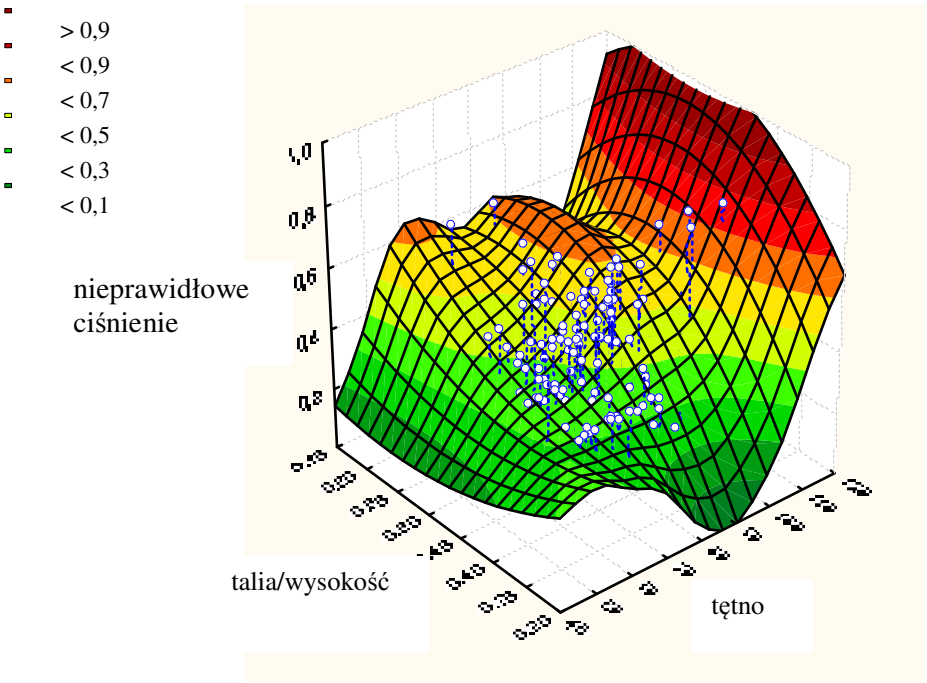
Rysunek 9. Wykres warstwowy ciśnienia, obwodu uda i tętna



Rysunek 10. Wykres warstwowy ciśnienia, WMC i tętna



Rysunek 11. Wykres warstwowy ciśnienia, talii/wysokości i tętna



11.4.2. Dziewczynki

Tabela 33. Podsumowanie modelu MARSplines

	Value
Niezależnych	6
Zależnych	1
Liczba czynników	34
Liczba funkcji bazowych	98
błąd (GCV)	0,241553

Tabela 34. Współczynniki modelu MARSplines

	Współcz. 0	Współcz. 1	Węzły obwód ramienia	Węzły obwód bioder	Węzły obwód uda	Węzły BMI	Węzły WR	Węzły tętno
W. wolny	0,95245	0,0476						
Wsp.1	-0,03914	0,0391		99,30000				
Wsp.2	-0,09736	0,0974		99,30000	55,40000			
Wsp.3	17,64425	-17,6442		99,30000	55,40000		1,398794	
Wsp.4	0,05959	-0,0596	27,00000	99,30000				
Wsp.5	0,43762	-0,4376	27,00000	99,30000			1,694756	
Wsp.6	-0,05178	0,0518		99,30000			1,454377	55,00000
Wsp.7	-0,20026	0,2003					1,301223	95,33333
Wsp.8	0,03605	-0,0360	28,30000					95,33333
Wsp.9	-0,03463	0,0346	27,00000	99,30000			1,694756	92,00000
Wsp.10	0,01205	-0,0120	28,30000		59,40000			95,33333
Wsp.11	0,27130	-0,2713					1,626996	95,33333
Wsp.12	-0,09212	0,0921		99,30000	57,50000		1,454377	
Wsp.13	0,03654	-0,0365	27,00000	99,30000		23,73696		
Wsp.14	0,04329	-0,0433	27,00000	99,30000		23,73696		
Wsp.15	-0,87032	0,8703		99,30000	55,40000	9,96094	1,398794	
Wsp.16	-0,02711	0,0271	26,70000		55,00000	23,77409		75,50000
Wsp.17	-0,02132	0,0213			51,20000	20,58204		75,50000
Wsp.18	1,33727	-1,3373			51,20000		1,491499	75,50000
Wsp.19	-1,03497	1,0350	28,30000					
Wsp.20	0,01137	-0,0114		99,30000	55,40000	22,28752		
Wsp.21	0,00564	-0,0056	26,70000		55,50000			82,33333
Wsp.22	-0,02463	0,0246		99,30000			1,454377	74,00000
Wsp.23	-0,00195	0,0019		99,30000				86,33333
Wsp.24	-0,13054	0,1305	28,30000				1,602646	95,33333
Wsp.25	0,01396	-0,0140	28,30000		57,10000		1,153971	95,33333
Wsp.26	0,00048	-0,0005	24,60000	99,30000	55,40000	19,39058		74,33333
Wsp.27	0,00465	-0,0047		97,40000	51,20000			75,50000
Wsp.28	-0,10247	0,1025	27,00000	99,30000	59,40000	26,48301	1,694756	
Wsp.29	-0,20330	0,2033	27,00000	99,30000	59,40000	25,86923	1,694756	
Wsp.30	-0,08432	0,0843					1,419940	95,33333
Wsp.31	0,01025	-0,0102	26,70000		48,40000			75,50000
Wsp.32	-0,00564	0,0056	26,70000		47,10000			70,33333
Wsp.33	-0,22867	0,2287		99,30000	55,70000			86,33333

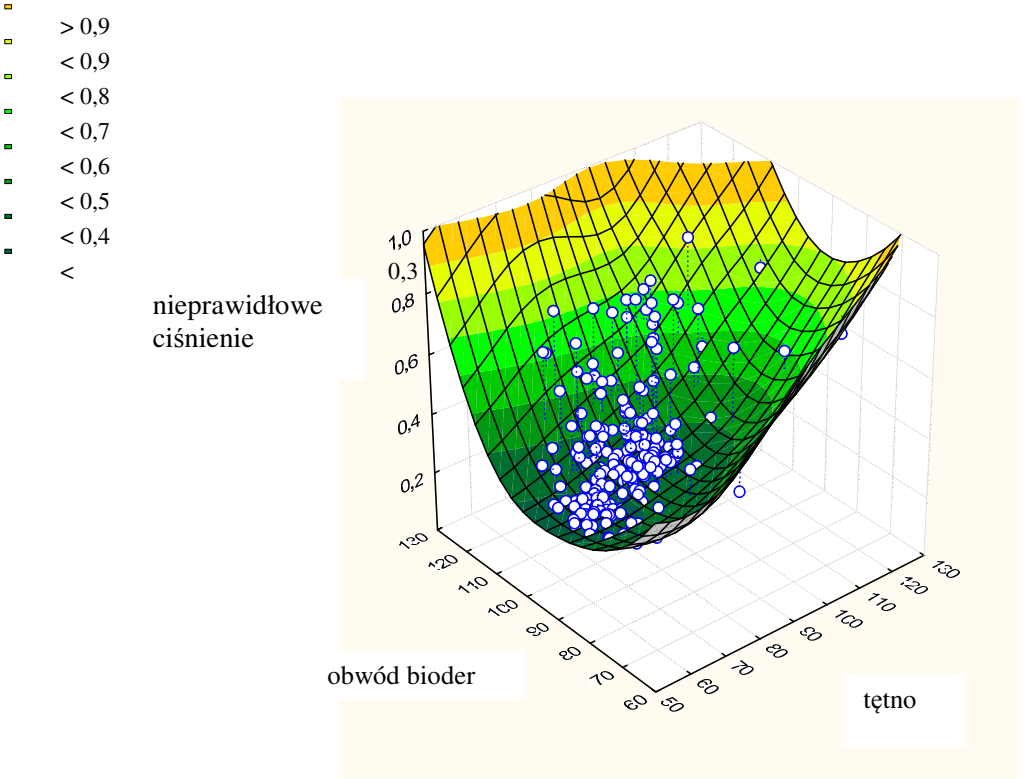
Tabela 35. Udział poszczególnych zmiennych w modelu MARSplines

	Odwołania (do funkcji bazowej)
obramienia	17
obbioder	19
obuda	18
bmi	9
WR	15
tętno	20

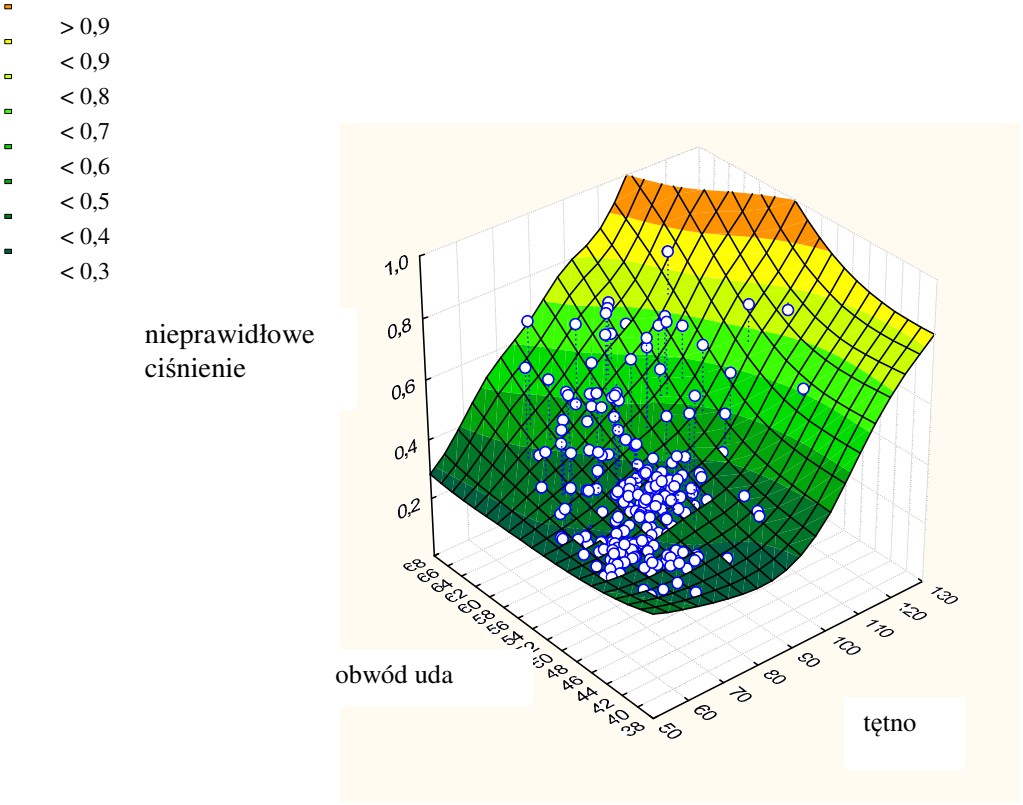
Tabela 36. Macierz błędów w modelu MARSplines

	obserwowane 0	**obserwowane 1**
przewidywane 0	320	33
przewidywane 1	3	46

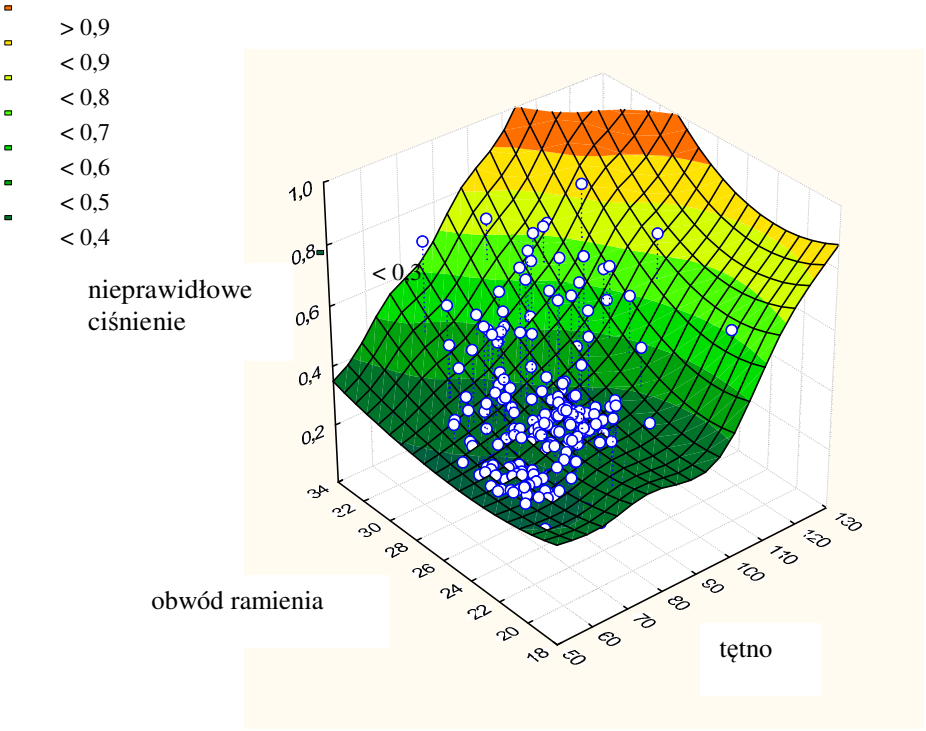
Rysunek 12. Wykres warstwowy ciśnienia, obwodu bioder i tętna



Rysunek 13. Wykres warstwowy ciśnienia, obwodu uda i tętna



Rysunek 14. Wykres warstwowy ciśnienia, obwodu ramienia i tętna



11.4.3. Podsumowanie

Tabela 37. Podsumowanie modeli MARSplines dla chłopców i dziewczynek

	Chłopcy	**Dziewczynki**
zmienne w modelu	talia/wysokość	obwód ramienia
	obwód bioder obwód uda BMI WR tętno obwód talii WMC	obwód bioder obwód uda BMI WR tętno
liczba czynników	15	34
GCV (Generalized Cross Validation) –błąd uogólnionego sprawdzianu krzyżowego	31%	24%
ilość dzieci z nieprawidłowym ciśnieniem – 1	57	79
ilość dzieci z prawidłowym ciśnieniem – 0	158	323
Procent dzieci z nieprawidłowym ciśnieniem błędnie zaklasyfikowanych	38%	42%
Procent dzieci z prawidłowym ciśnieniem błędnie zaklasyfikowanych	5%	1%

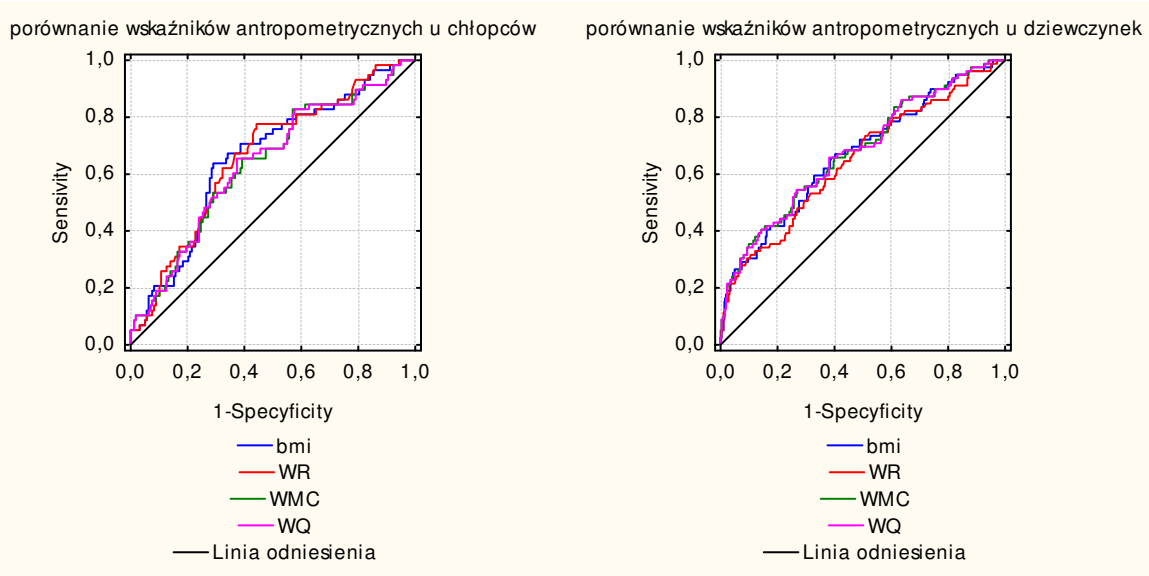
11.5. Krzywe ROC

W powyższym rozdziale zostały porównane i ocenione cztery wskaźniki antropometryczne w celu znalezienia klasyfikatora, który najlepiej rozróżnia dzieci z prawidłowym i nieprawidłowym ciśnieniem.

Analiza wskaźnika BMI metodą krzywych ROC u chłopców wyznacza jako punkt podziału wartość 22,889, co odpowiada 75 centylowi, natomiast u dziewczynek wartość graniczna jest równa 21,355 i mieści się między 50, a 75 centylem **tabela 38, 39**. Odczytując wyniki z **tabel 38** i **39** możemy zauważyć, że zarówno u chłopców jak i u dziewczynek czułość metody jest niska. Swoistość u chłopców jest równa 71%, czyli dobra, natomiast u dziewczynek wynosi tylko 60%. Pola pod krzywymi AUC u chłopców i u dziewczynek są mniejsze od 0,7 czyli prawdopodobieństwo prawidłowego zaklasyfikowania przypadku również jest niskie. Analizując wyniki dotyczące wskaźnika WR u chłopców zauważamy, że wartością progową dla chłopców jest WR równe 1,218 (75 centyl), u dziewczynek 1,249 (50 – 75 centyl). Zarówno u chłopców, jak i dziewczynek czułość metody jest umiarkowanie wysoka i wynosi odpowiednio 78% i 73%. Swoistości są bardzo słabe i wynoszą odpowiednio 56% i 49%. Analizując pola pod krzywymi stwierdzamy, że prawdopodobieństwo prawidłowego zaklasyfikowania przypadku jest niskie. Dla kolejnego wskaźnika WMC krzywe ROC dla chłopców jako wartość progową wyznaczyły punkt 42,519 (75 centyl), u dziewczynek 37,688 (75 centyl). Czułość i swoistość dla chłopców jest słaba wynosi odpowiednio 65% i 60%. U dziewczynek czułość jest bardzo niska 54%, natomiast swoistość jest dobra 73%. Pola pod krzywymi dla obu płci są mniejsze od 0,7, czyli klasyfikacja jest stosunkowo słaba. Ostatni wskaźnik WQ zarówno dla chłopców jak i dla dziewczynek wartość progową wyznaczył w 75 centylu. Czułość i swoistość dla chłopców są niskie 66%, 63%. Dla dziewczynek czułość jest bardzo niska 54%, ale swoistość dobra 73%. Pola pod krzywymi nie przekraczają 0,7.

U chłopców nie można stwierdzić, który ze wskaźników antropometrycznych jest lepszym klasyfikatorem, gdyż nie wykryto istotnych różnic w AUC. U dziewczynek istotną różnicę stwierdzono pomiędzy polami BMI i WR p=0,0261. Pole pod krzywą wskaźnika WR jest istotnie mniejsze od pola pod krzywą dla wskaźnika BMI. Wskaźnik BMI jest lepszym klasyfikatorem u dziewczynek od wskaźnika Rohrer'a.

Rysunek 15. Porównanie pól pod krzywymi ROC dla wskaźników antropometrycznych u chłopców i dziewczynek



Wskaźnikiem, który najlepiej wykrywa nieprawidłowe ciśnienie krwi u chłopców i u dziewczynek jest wskaźnik Rohrer'a. Najwyższą swoistość u chłopców ma wskaźnik BMI, a u dziewczynek wskaźnik masy ciała WMC i wskaźnik Quetelet'a.

11.5.1. Chłopcy

Tabela 38. Wyniki analizy ROC dla chłopców

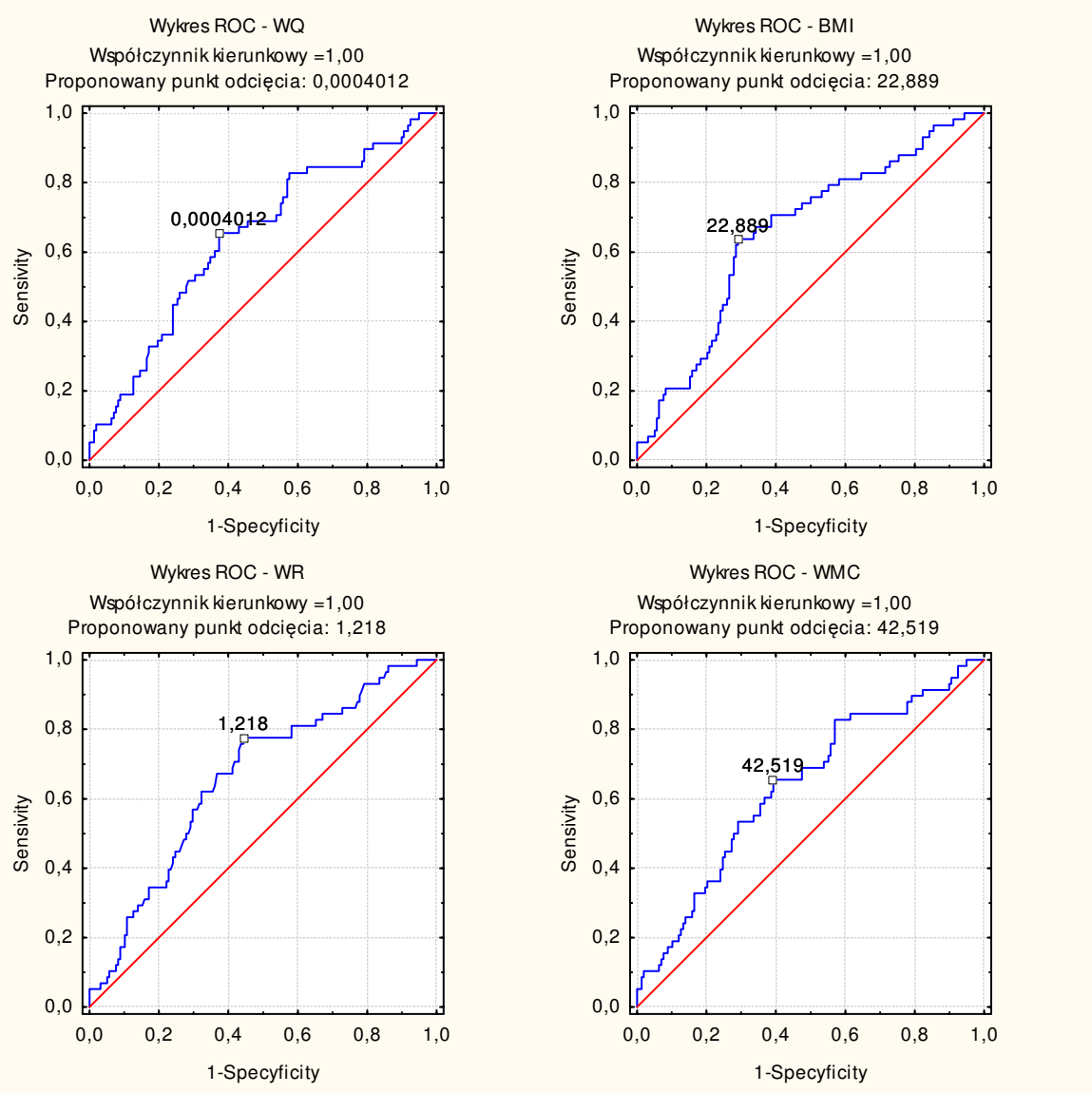
	wartość progowa	True positives	False positives	False negatives	True negatives	Czułość	Swoistość	ACC	AUC
BMI	22,889	37	46	21	112	0,638	0,709	0,690	0,658
WR	1,218	45	70	13	88	0,776	0,557	0,616	0,662
WMC	42,519	38	62	20	96	0,655	0,608	0,620	0,637
WQ	0,000401	38	59	20	99	0,655	0,627	0,634	0,640

AUC (area under the curve) – pole pod krzywą

ACC (accuracy) – skuteczność reguły decyzyjnej

Nie wykryto istotnych różnic w AUC między poszczególnymi wskaźnikami antropometrycznymi.

Rysunek 16. Interpretacja graficzna krzywych ROC u chłopców



11.5.2. Dziewczynki

Tabela 39. Wyniki analizy ROC dla dziewczynek

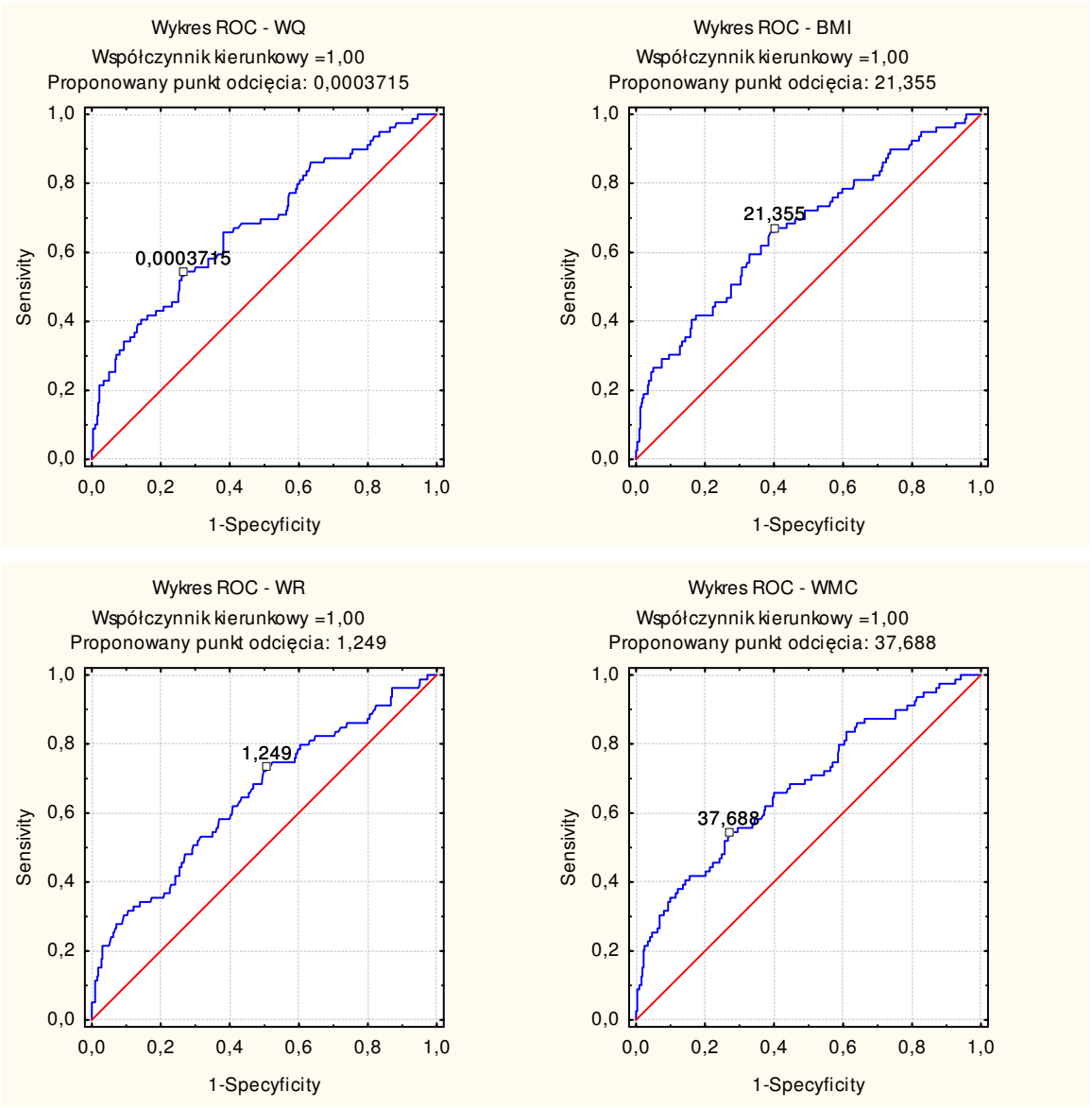
	wartość progowa	True positives	False positives	False negatives	True negatives	Czułość	Swoistość	ACC	AUC
BMI	21,355	53	130	26	193	0,671	0,598	0,612	0,669
WR	1,249	58	164	21	159	0,734	0,492	0,540	0,649
WMC	37,688	43	87	36	236	0,544	0,731	0,694	0,679
WQ	0,0003715	43	86	36	237	0,544	0,734	0,697	0,679

AUC (area under the curve) – pole pod krzywą

ACC (accuracy) – skuteczność reguły decyzyjnej

Wykryto istotną różnicę w AUC między BMI i WR (p=0,0261).

Rysunek 17. Interpretacja graficzna krzywych ROC u dziewczynek



12. Podsumowanie modeli

Tabela 40. Podsumowanie trzech modeli matematycznych przeanalizowanych w pracy

	drzewa klasyfikacyjne		**regresja logistyczna**		**MARSplines**	
	chłopcy	**dziewczynki**	**chłopcy**	**dziewczynki**	**chłopcy**	**dziewczynki**
zmienne	talia/wysokość BMI tętno	tętno obwód bioder obwód talii	talia/wysokość tętno	obwód uda tętno	tętno obwód uda talia/wysokość	tętno obwód bioder obwód uda
czułość	54%	32%	26%	25%	61%	58%
swoistość	79%	91%	95%	95%	95%	99%

Można zauważyć, że zmienna tętno występuje w każdym z zastosowanych modeli zarówno u chłopców jak i u dziewczynek. U chłopców oprócz tętna we wszystkich modelach powtarza się wskaźnik wagowo-wzrostowy talia/wysokość. Można wyciągnąć wniosek, że przyrost tych parametrów u chłopców powoduje największy wzrost prawdopodobieństwa wystąpienia nadciśnienia. U dziewczynek oprócz tętna, które występuje we wszystkich technikach, w dwóch modelach pojawia się obwód bioder i uda. Te parametry mają największy wpływ na wystąpienie nieprawidłowego ciśnienia u dziewczynek.

Najwyższą czułość i swoistość zarówno dla chłopców jak i dla dziewczynek ma model MARSplines, na podobnym poziomie plasuje się swoistość dla regresji logistycznej.

Skuteczność modeli omówionych w pracy można porównać ze sobą stosując metodę Data Miningu – Szybkie wdrażanie modeli predykcyjnych. Jednakże aby móc użyć tę technikę należy w trzech badanych modelach zastosować te same zmienne niezależne. Dlatego też do każdego z modeli porównywanych w pracy (regresja logistyczna, drzewa klasyfikacyjne, MARSplines) zostaną podstawione zmienne niezależne z pozostałych modeli. Powstaną trzy porównania dla każdej grupy zmiennych niezależnych zarówno dla chłopców jak i dla dziewczynek.

Poniżej zostaną porównane trzy omówione techniki klasyfikacyjne u chłopców:

Zmienne niezależne zastosowane w regresji logistycznej dla chłopców to: talia/wysokość, tętno.

Zmienne niezależne zastosowane w drzewach klasyfikacyjnych dla chłopców to: obwód ramienia, talii, BMI, talia/wysokość, tętno.

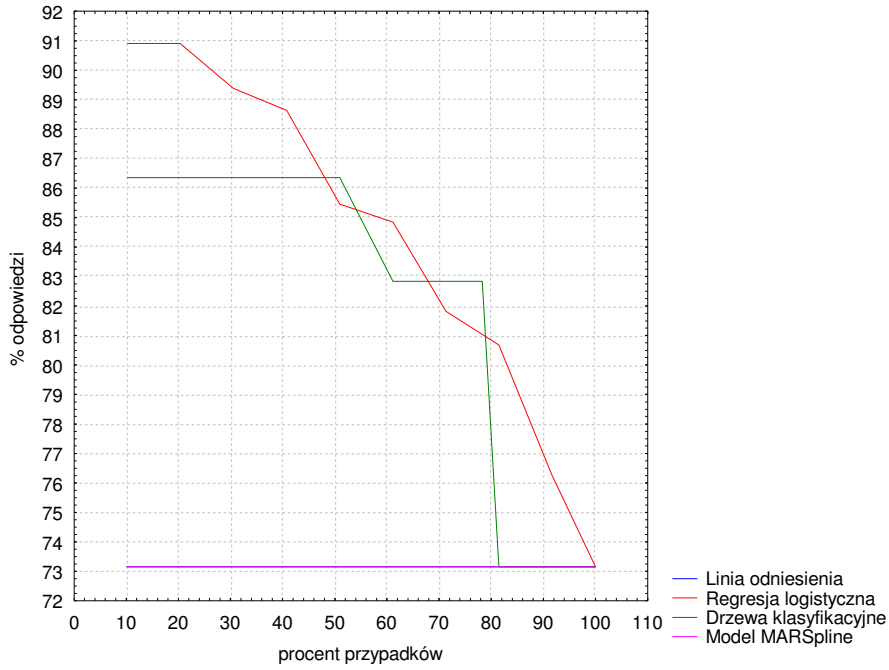
Zmienne niezależne zastosowane w modelu MARSplines dla chłopców to: obwód talii, bioder, uda, BMI, WR, WMC, talia/wysokość, tętno.

Tabela 41. Odsetek błędnych klasyfikacji

	Regresja logistyczna	**Drzewa klasyfikacyjne**	**Model MARSplines**
zmienne niezależne – regresja logistyczna	0,236111	0,217593	0,268519
zmienne niezależne – drzewa klasyfikacyjne	0,245370	0,222222	0,268519
zmienne niezależne – model MARSplines	0,246512	0,218605	0,265116

Odczytując wyniki z **tabeli 41** możemy zauważyć, że najmniejszy błąd klasyfikacyjny pojawia się w modelu drzewa klasyfikacyjne, jest mniejszy od 0,23.

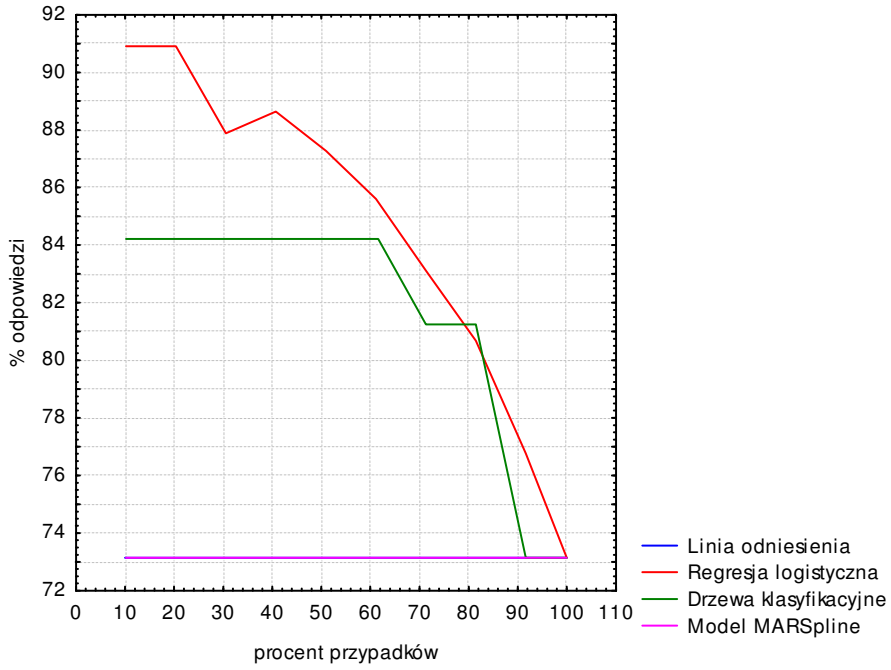
Rysunek 18. Skumulowany wykres przyrostu (% odpowiedzi prawidłowych). Zmienne niezależne – regresja logistyczna



Z **rysunku 18** można odczytać, że biorąc 20 % przypadków najpewniej zaklasyfikowanych za pomocą regresji logistycznej do grupy dzieci z prawidłowym ciśnieniem otrzymamy próbę,

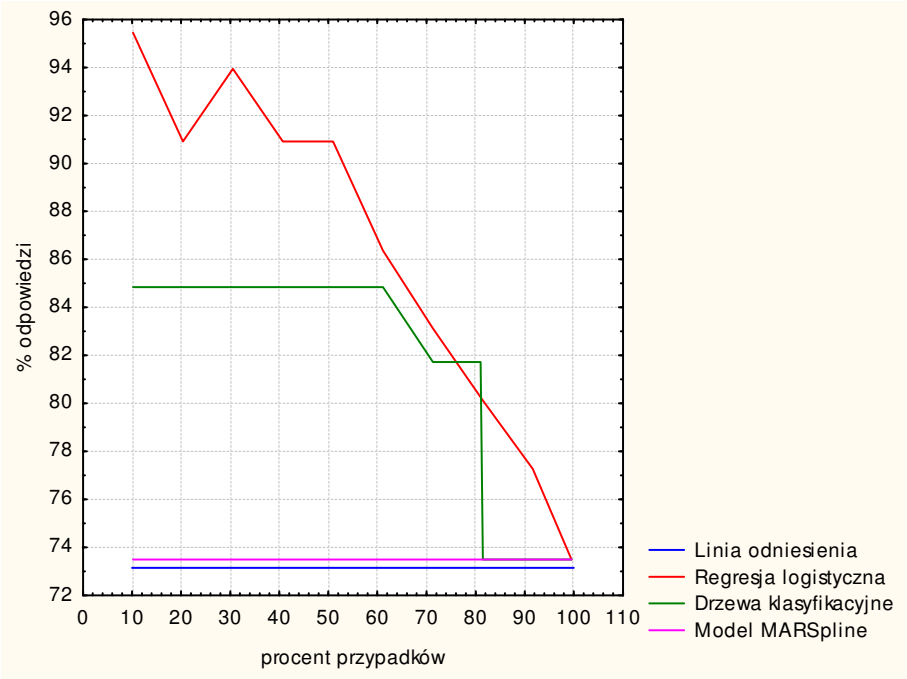
w której co najmniej 90% dzieci należy do wybranej kategorii. Natomiast biorąc 50% przypadków najpewniej zaklasyfikowanych za pomocą drzew klasyfikacyjnych do grupy dzieci z prawidłowym ciśnieniem otrzymamy próbę, w której co najmniej 86% dzieci należy do wybranej kategorii. Odczytując wykres zbudowany dla modelu MARSplines widzimy, że nie ma znaczenia liczba przypadków poprawnie zaklasyfikowanych, gdyż model daje nam stałą próbę w której 74% dzieci należy do grupy z prawidłowym ciśnieniem.

Rysunek 19. Skumulowany wykres przyrostu (% odpowiedzi prawidłowych). Zmienne niezależne – drzewa klasyfikacyjne



Z **rysunku 19** można odczytać, że biorąc 20 % przypadków najpewniej zaklasyfikowanych za pomocą regresji logistycznej do grupy dzieci z prawidłowym ciśnieniem otrzymamy próbę, w której 91% dzieci należy do wybranej kategorii. Natomiast biorąc 60% przypadków najpewniej zaklasyfikowanych za pomocą drzew klasyfikacyjnych do grupy dzieci z prawidłowym ciśnieniem otrzymamy próbę, w której co najmniej 84% dzieci należy do wybranej kategorii. Odczytując wykres zbudowany dla modelu MARSplines widzimy, że nie ma znaczenia liczba przypadków poprawnie zaklasyfikowanych, gdyż model daje nam stałą próbę w której 73% dzieci należy do grupy z prawidłowym ciśnieniem.

Rysunek 20. Skumulowany wykres przyrostu (% odpowiedzi prawidłowych). Zmienne niezależne – model MARSplines



Analizując **rysunek 20** widzimy, że biorąc 10 % przypadków najpewniej zaklasyfikowanych za pomocą regresji logistycznej do grupy dzieci z prawidłowym ciśnieniem otrzymamy próbę, w której 95% dzieci należy do wybranej kategorii. Natomiast biorąc 60% przypadków najpewniej zaklasyfikowanych za pomocą drzew klasyfikacyjnych do grupy dzieci z prawidłowym ciśnieniem otrzymamy próbę, w której co najmniej 85% dzieci należy do wybranej kategorii. Odczytując wykres zbudowany dla modelu MARSplines widzimy, że nie ma znaczenia liczba przypadków poprawnie zaklasyfikowanych, gdyż model daje nam stałą próbę w której 73% dzieci należy do grupy z prawidłowym ciśnieniem.

Podsumowując wyniki otrzymane za pomocą metody szybkiego wdrażania modeli predykcyjnych dla chłopców możemy zauważyć, że dla przypadków, którym modele dawały najwyższe prawdopodobieństwo klasyfikacyjne regresja logistyczna daje największą poprawność klasyfikacji przypadków z prawidłowym ciśnieniem bez względu na to jakie zmienne niezależne zastosowano w modelu. Natomiast najmniejszy błąd klasyfikacyjny mają drzewa klasyfikacyjne **tabela 41.**

Poniżej zostaną przedstawione wyniki otrzymane po zastosowaniu metody szybkiego wdrażania modeli predykcyjnych dla dziewczynek:

Zmienne niezależne zastosowane w regresji logistycznej dla dziewczynek to: obwód uda, tętno.

Zmienne niezależne zastosowane w drzewach klasyfikacyjnych dla dziewczynek to: obwód talii, bioder, talia/bioder, tętno.

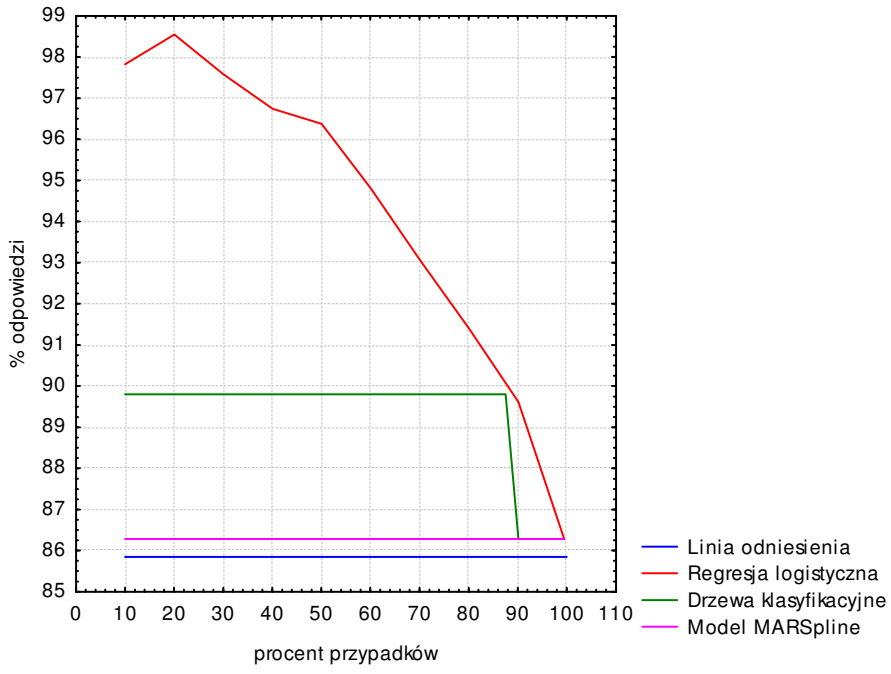
Zmienne niezależne zastosowane w modelu MARSplines dla dziewczynek to: obwód ramienia, bioder, uda, BMI, WR, tętno.

Tabela 42. Odsetek błędnych klasyfikacji dla dziewczynek

	Regresja logistyczna	**Drzewa klasyfikacyjne**	**Model MARSplines**
zmienne niezależne – regresja logistyczna	0,166667	0,196517	0,196517
zmienne niezależne – drzewa klasyfikacyjne	0,169154	0,174129	0,196517
zmienne niezależne – model MARSplines	0,171642	0,196517	0,196517

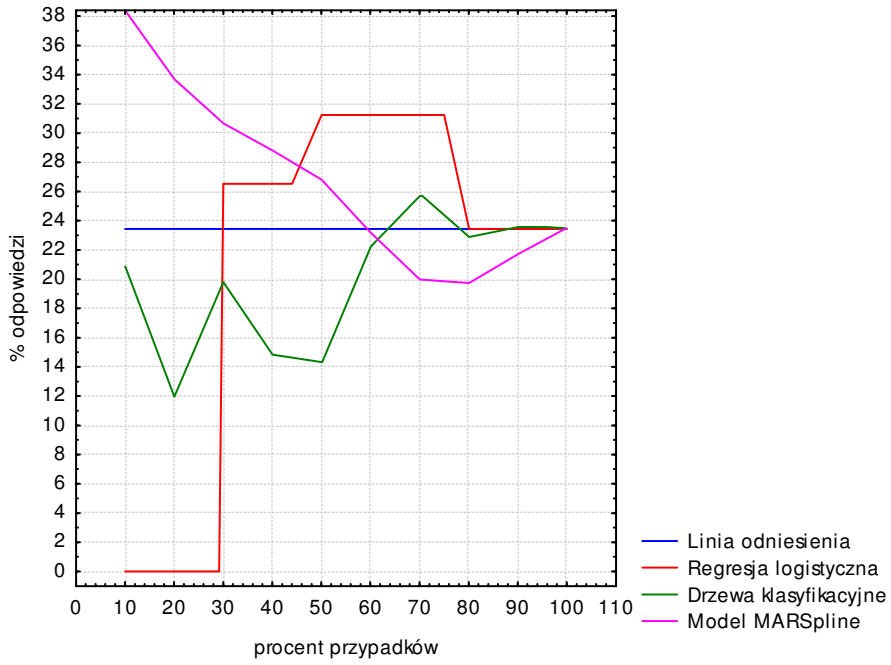
Odczytując wyniki z tabeli możemy zauważyć, że najmniejszy błąd klasyfikacyjny pojawia się w modelu regresji logistycznej, jest równy 0,166667.

Rysunek 21. Skumulowany wykres przyrostu (% odpowiedzi prawidłowych). Zmienne niezależne – regresja logistyczna



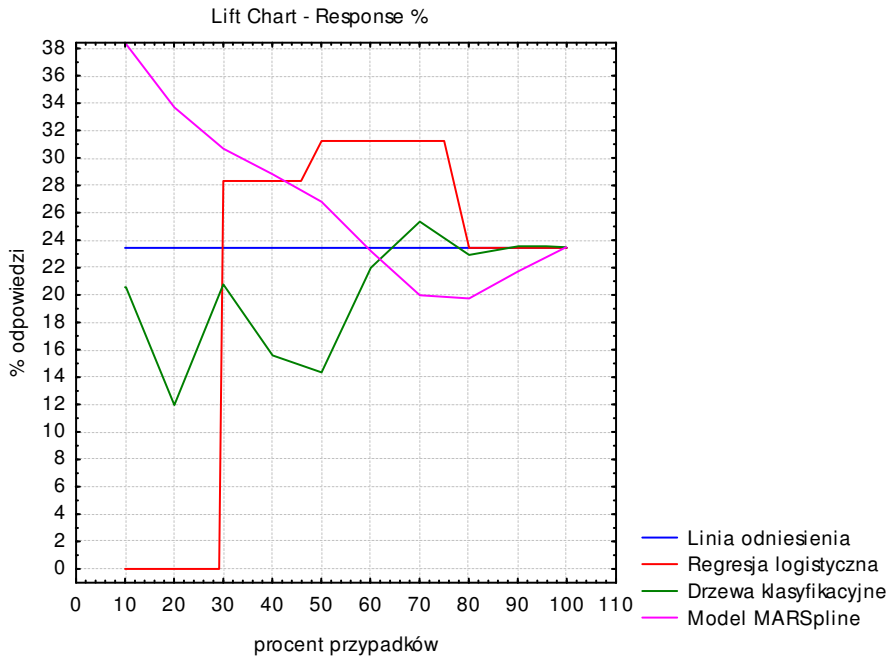
Analizując **rysunek 21** widzimy, że biorąc 20 % przypadków najpewniej zaklasyfikowanych za pomocą regresji logistycznej do grupy dzieci z prawidłowym ciśnieniem otrzymamy próbę, w której 98% dzieci należy do wybranej kategorii (prawidłowe ciśnienie krwi). Natomiast biorąc 90% przypadków najpewniej zaklasyfikowanych za pomocą drzew klasyfikacyjnych do grupy dzieci z prawidłowym ciśnieniem otrzymamy próbę, w której co najmniej 90% dzieci należy do wybranej kategorii. Odczytując wykres zbudowany dla modelu MARSplines widzimy, że nie ma znaczenia liczba przypadków poprawnie zaklasyfikowanych, gdyż model daje nam stałą próbę w której 86% dzieci należy do grupy z prawidłowym ciśnieniem.

Rysunek 22. Skumulowany wykres przyrostu (% odpowiedzi prawidłowych). Zmienne niezależne – drzewa klasyfikacyjne



Z **rysunku 22** można odczytać, że porównywane trzy modele statystyczne (w których jako zmienne niezależne użyto obwodu talii, bioder oraz wskaźnik talia/biodro i tętno) słabo klasyfikują przypadki. Wykresy trzech modeli znajdują się przeważnie poniżej linii odniesienia, która odpowiada klasyfikacji losowej.

Rysunek 23. Skumulowany wykres przyrostu (% odpowiedzi prawidłowych). Zmienne niezależne – model MARSplines



Z **rysunku 23** można odczytać, że porównywane trzy modele matematyczne (w których jako zmienne niezależne użyto obwodu ramienia, bioder, uda oraz wskaźnika BMI, WR i tętna) słabo klasyfikują przypadki. Powyższy wykres wygląda podobnie jak **rysunek 22**. Wykresy trzech modeli znajdują się przeważnie poniżej linii odniesienia, która odpowiada klasyfikacji losowej.

Podsumowując wyniki otrzymane za pomocą metody szybkiego wdrażania modeli predykcyjnych dla dziewczynek możemy zauważyć, że dla przypadków z najwyższym prawdopodobieństwem klasyfikacyjnym największą poprawność klasyfikacji przypadków z prawidłowym ciśnieniem daje regresja logistyczna, ale tylko jeżeli jako zmienne niezależne przyjmiemy obwód uda oraz tętno. Natomiast jeżeli użyjemy innych zmiennych niezależnych wartość klasyfikacyjna modeli jest na pograniczu losowej klasyfikacji. Najmniejszy błąd klasyfikacyjny ma regresja logistyczna, bez względu na to jakie zmienne niezależne zastosowano w modelu **tabela 42.**

Biorąc pod uwagą wyniki powyższej analizy oraz badania przeprowadzone w pracy dochodzimy do wniosku, że szybkie wdrażanie modeli predykcyjnych jako metoda badawcza i porównująca modele nie spełnia oczekiwań. W założeniach techniki porównywane modele muszą mieć te same zmienne niezależne. Przeprowadzone badania w niniejszej pracy wykazały, że aby mieć najlepszą klasyfikację przypadków każdy z trzech użytych modeli musi mieć inne zmienne niezależne („własne"). Potwierdzenie tego wniosku możemy znaleźć odczytując wyniki z **tabel 41** i **42**. Widzimy, że najmniejszy błąd modele uzyskują mając „własne" zmienne niezależne.

13. Dyskusja

Rozpatrywane w badaniu nadciśnienie tętnicze często występuje wśród dorosłych Polaków. W niektórych województwach choroba obejmuje nawet połowę populacji dorosłych mężczyzn (woj. śląskie: 49%, woj. wielkopolskie: 50%) oraz prawie 40% dorosłych kobiet (woj. śląskie: 38%, woj. wielkopolskie: 37%) [61]. U dzieci i młodzieży liczba chorujących jest znacznie mniejsza. Kształtuje się na poziomie 2–15% [59, 60] w zależności od regionu, w którym prowadzone były badania. W niniejszym badaniu procent dzieci chorujących na nadciśnienie mieści się w podanym powyżej przedziale i wynosi on 7,12%. Natomiast dzieci z nieprawidłowym ciśnieniem, czyli mieszczącym się powyżej 90 centyla jest niemal dwukrotnie więcej 13,7%. W Polsce badania epidemiologiczne dotyczące ciśnienia krwi u dzieci były prowadzone już w 60 – tych latach [4]. Jednym z pierwszych naukowców w Polsce, który przeprowadził przekrojowe badania ciśnienia tętniczego krwi u dzieci był H. Wojdon [62] (przebadał 2382 dziewczynki z Poznania w wieku od 13 do 17 lat). W latach 60 –tych i 70 – tych takie badania prowadzili J. Kopczyński [63] w Warszawie, T. Gerkowicz [64] w Lublinie, J. Chodakowska [65] w Warszawie. Kolejni naukowcy zajmujący się zagadnieniem ciśnienia tętniczego krwi u dzieci byli: I. Kowalik [66] w roku 1980 w Wielkopolsce i Kaszubach, J. Lipiec [67] w roku 1981 w Łodzi, T. Wyszyńska [68] w roku 1981 w Warszawie, J. Baszczyński [69] w 1982 roku w Łodzi i wielu innych. We wszystkich badaniach do pomiaru wykorzystano aparat rtęciowy i w większości mankiet dostosowany był do obwodu ramienia. Najnowsze przekrojowe badania były prowadzone przez A. Krzyżaniak [4] w 1999, a w 2004 roku ze współpracą z Zakładem Rozwoju Dzieci i Młodzieży Instytutu Matki i Dziecka w Warszawie. Zmierzono ciśnienie tętnicze krwi u dzieci z Poznania. W 1999 roku w badaniu wzięło udział 2955 chłopców i 2934 dziewczynek w wieku od 10 do 18 lat. W 2004 roku zmierzono ciśnienie 899 chłopcom i 906 dziewczynkom między 7, a 18 rokiem życia [2].

Badania przekrojowe są ważnym elementem służącym do oceny stanu zdrowia dzieci i młodzieży. Dzięki takim badaniom jesteśmy w stanie wyliczyć mierniki stanu zdrowia. Najczęściej stosowanymi miernikami są wskaźniki morfologiczne. Niestety nadal rzadko mierzone jest ciśnienie tętnicze krwi, które nadal przez wielu uważane jest za mało ważny miernik fizjologiczny. Na podstawie tych wskaźników jesteśmy w stanie zaobserwować, czy wykształciła się z biegiem lat jakaś pozytywna lub negatywna tendencja w rozwoju i czy należy ją wspierać, czy jej przeciwdziałać.

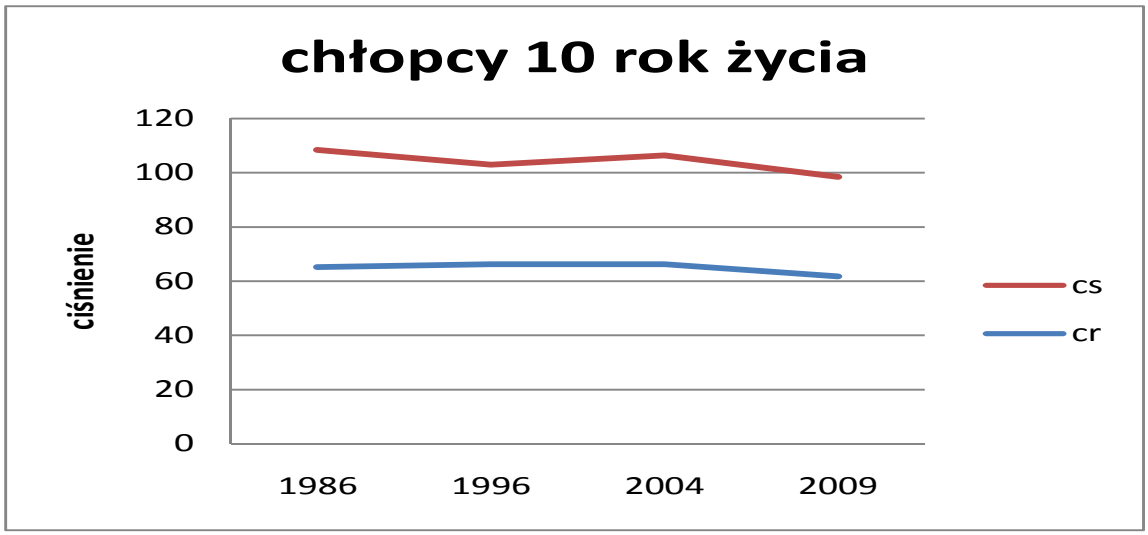
W **tabeli 43** zebrane są średnie i odchylenia standardowe ciśnień skurczowych i rozkurczowych w latach 1986, 1996, 2004 i 2009 [2, 4, 70] u chłopców i u dziewczynek. Nie można było stwierdzić, czy średnie ciśnień w poszczególnych latach różnią się istotnie od siebie ze względu na brak danych dotyczących liczebności. Do porównań użyto tylko wartości ciśnień dla 10 i 18 roku życia ze względu na to, że w okresie dojrzewania występują różnego rodzaju wahania ciśnień związane z zachodzącymi w tym okresie zmianami w organizmie dziecka.

Na wykresach można zaobserwować jak kształtuje się ciśnienie w poszczególnych latach. U chłopców jak i u dziewczynek w dziesiątym roku życia z biegiem lat zauważamy nieznaczny spadek obu ciśnień w szczególności skurczowego. Analizując wykres dotyczący osiemnastego roku życia u chłopców stwierdzamy wzrost ciśnienia skurczowego, natomiast rozkurczowe pozostaje cały czas na tym samym poziomie. U dziewczynek w osiemnastym roku życia ciśnienia przez cały okres obserwacji nie przejawiają wyraźnej tendencji wzrostowej ani malejącej.

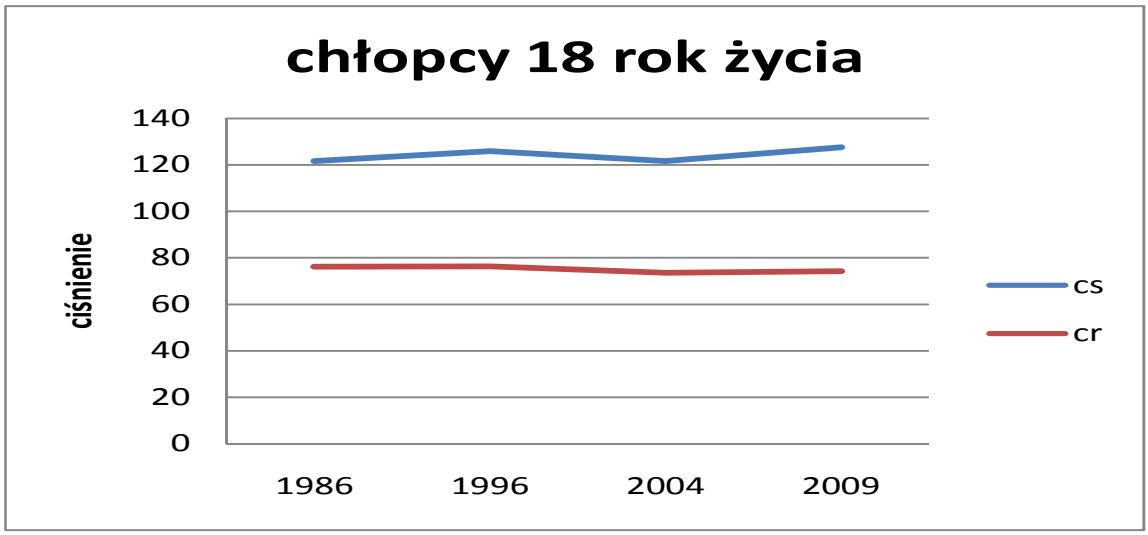
Tabela 43. Przegląd ciśnienia skurczowego i rozkurczowego w latach od 1986 do 2009 u chłopców i dziewczynek

chłopcy		**ciśnienie skurczowe**		**ciśnienie rozkurczowe**		**dziewczynki**		**ciśnienie skurczowe**		**ciśnienie rozkurczowe**	
wiek	**rok**	$\bar{x}$	**sd**	$\bar{x}$	**sd**	**wiek**	**rok**	$\bar{x}$	**sd**	$\bar{x}$	**sd**
10	1986	108,44	9,31	65,18	7,37	10	1986	110,77	11,75	66,89	7,95
	1996	102,97	12,84	66,23	8,68		1996	102,89	13,59	65,84	8,79
	2004	106,4	7,73	66,2	6,53		2004	105,3	8,38	65,7	6,9
	2009	98,41	6,83	61,74	5,57		2009	96,75	5,89	60,04	4,65
18	1986	121,56	13,96	76,19	7,74	18	1986	117,22	9,55	75,31	6,12
	1996	125,99	11,85	76,34	8,45		1996	114,32	10,68	69,83	8,22
	2004	121,6	12,07	73,6	8,1		2004	113,2	10,11	71,4	7,91
	2009	127,61	12,71	74,32	7,05		2009	115,45	12,4	69,57	8,03

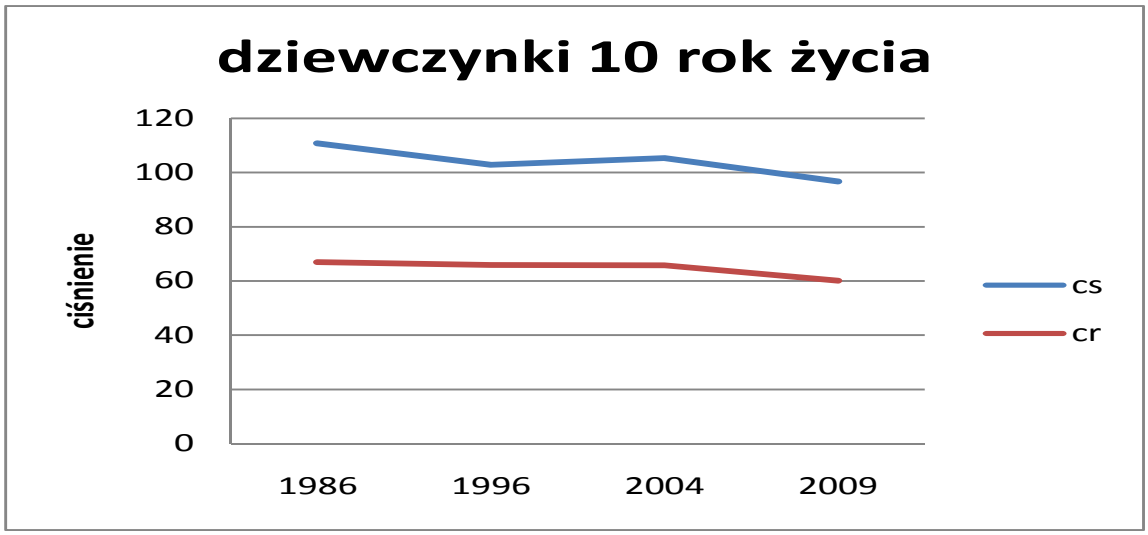
Rysunek 24. Wykres ciśnienia skurczowego i rozkurczowego w latach 1986, 1995, 2004, 2009 u 10 – letnich chłopców



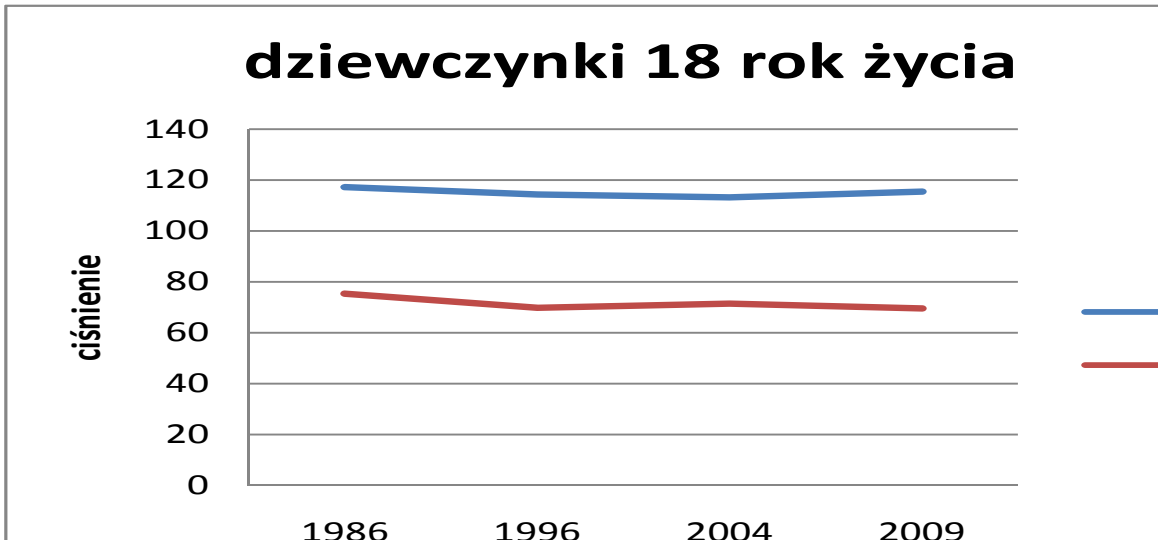
Rysunek 25. Wykres ciśnienia skurczowego i rozkurczowego w latach 1986, 1995, 2004, 2009 u 18 – letnich chłopców



Rysunek 26. Wykres ciśnienia skurczowego i rozkurczowego w latach 1986, 1995, 2004, 2009 u 10 – letnich dziewczynek



Rysunek 27. Wykres ciśnienia skurczowego i rozkurczowego w latach 1986, 1995, 2004, 2009 u 18 – letnich dziewczynek



Do roku 1977 nadciśnienie nie było osobno zdefiniowane dla dorosłych, dzieci i młodzieży. Stosowano te same kryteria oceny ciśnienia dla każdej z grup. Dopiero pod koniec lat 70-tych tzw. amerykańska Grupa Robocza ds. Kontroli Ciśnienia u Dzieci przeprowadziła badania, dzięki którym odkryto, że ciśnienie zależy nie tylko od wieku, ale również od płci i od wzrostu. Dzięki temu odkryciu po raz pierwszy ustalono osobne kryteria definiujące nieprawidłowe ciśnienie i nadciśnienie dla dzieci i młodzieży. Przyjęto wówczas, że o nieprawidłowym ciśnieniu u dzieci i młodzieży należy mówić wówczas, gdy ciśnienie skurczowe lub rozkurczowe przekracza 90 centyl, natomiast jeżeli któreś z ciśnień przekraczało 95 centyl wtedy uznawano, że młody człowiek ma nadciśnienie [27, 28, 43].

W latach 70-tych próbowano również ustalić przyczyny występowania nadciśnienia u dzieci [27]. Wówczas nie występowała jeszcze zgodność poglądów na genetyczne uwarunkowanie tętniczego ciśnienia krwi, czy występowania nadciśnienia. Dominowało wówczas kilka poglądów. W pierwszym wysokie ciśnienie krwi uważane było za skutek wpływu genu dominującego o niepełnej penetracji. W drugim dopatrywano się przyczyn w czynnikach genetycznych i częściowo środowiskowych, jeśli pominąć przyczyny patologiczne. Przez jeszcze innych, którzy nie odrzucając znaczenia czynnika genetycznego w regulacji tętniczego ciśnienia krwi, powstawanie nadciśnienia widziało głównie w czynnikach środowiskowych [30], a przede wszystkim bytowych [31, 32, 35]. W obecnie przeprowadzonych badaniach stwierdza się, że czynnik genetyczny ma ok. 20% wpływ na ciśnienie tętnicze krwi i w wielu pracach potwierdza się jego ważny udział [27, 33, 34]. Rozpatrując znaczenie czynnika genetycznego w kształtowaniu ciśnienia krwi trzeba przypomnieć, że skurczowe ciśnienie jest w pewnej stałej proporcji do ciężaru ciała obu płci [4]. Stwierdzono, że masa ciała ma udział w kształtowaniu poziomu ciśnienia krwi właściwego dla danego osobnika. Im większa masa ciała tym większe ciśnienie krwi musi powstać w naczyniach aby krew dotarła do każdej komórki organizmu [29]. Znajduje to potwierdzenie w badaniach. Wskazują one, że wskaźnik wagowo - wzrostowy ciała jest w pewnym związku z ciśnieniem tętniczym. Uznaje się, że dzieci otyłe i z nadwagą mają większą szansę na wystąpienie nadciśnienia [36, 37, 34, 38, 39, 40, 41, 42]. Zaobserwowano także, że stosunkowo niskie ciśnienie wykazują osobnicy o ektomorficznym typie budowy, podczas gdy wysokie ciśnienie jest właściwe raczej osobnikom mezo – i endomorficznym.

Interesującym aspektem badań zależności ciśnienia od masy ciała i wzrostu jest analiza ich zależności w oparciu o wskaźniki wagowo – wzrostowe. W większości badań rozpatruje się głównie zależność najbardziej znanego wskaźnika wagowo - wzrostowego BMI od ciśnienia tętniczego krwi. Wykazano, że u dzieci i młodzieży z nieprawidłowym ciśnieniem

lub z nadciśnieniem BMI istotnie różni się od dzieci zdrowych – jest wyższe [44, 45, 47, 48, 49, 50, 51], co potwierdza analiza przeprowadzona w niniejszej pracy. Rzadziej prowadzone są badania na pozostałych wskaźnikach wagowo – wzrostowych takich jak Rohrer'a, Quetelet'a lub WMC.

Stale dyskutowany jest wybór najlepszego ze wskaźników wagowo – wzrostowych, który nie tylko opisze otyłość, ale również wykaże zależność ciśnienia od masy ciała. Wskaźnikiem, który najlepiej opisuje otyłość u dorosłych jest BMI [25, 52]. Wśród pediatrów do najbardziej znanych zaliczany był wskaźnik Quetelet'a i Rohrer'a [23]. Wskaźnik Rohrer'a znany jest jako wskaźnik smukłości. W publikacjach coraz częściej spotykanym wskaźnikiem wagowo – wzrostowym jest współczynnik masy ciała. Wskaźnik nie tylko opisuje otyłość, ale również może być stosowany w ocenie ubytku masy ciała [4, 24].

W powyższych badaniach wykazano, że u chłopców nie istnieje wskaźnik, który istotnie różniłby się od pozostałych. Stwierdzono natomiast, że najlepszą czułość wykazuje wskaźnik Rohrer'a, a swoistość wskaźnik BMI. U dziewczynek natomiast stwierdzono istotną różnicę między wskaźnikami BMI, a Rohrer'a. Zaobserwowano, że wskaźnik BMI lepiej od wskaźnika Rohrer'a klasyfikuje ciśnienie tętnicze krwi u dziewczynek. Z przeprowadzonych badań wynika również, że wskaźniki WMC i Quetelet's u dziewczynek mają najlepszą swoistość. Wykazany wpływ wskaźnika Quetelet'a na ciśnienie tętnicze krwi znajduje potwierdzenie w pracy A. Krzyżaniak [4], która w swoich epidemiologicznych badaniach zaobserwowała jego wpływ na ciśnienie skurczowe i rozkurczowe.

Ciekawym aspektem badań prowadzonych nad ciśnieniem u dzieci powinno być zaprojektowanie modelu matematycznego, który pomagałby w znalezieniu czynników wpływających na wystąpienie nieprawidłowego ciśnienia krwi lub czynników wpływających na wystąpienie sercowo naczyniowych powikłań.

Jednym z badaczy zajmującym się tym problemem jest J. Nawarycz [56], który w swoich badaniach użył rozmytego systemu stratyfikacji ryzyka. System opiera się o bazę wiedzy ekspertów oraz reguły wnioskowania rozmytego dotyczącego w jego pracy trzech niezależnych czynników (atrybutów): BMI (body mass index), WC – otyłość brzuszna i podwyższonego ciśnienia tętniczego krwi. Na podstawie tego modelu starał się zaklasyfikować badane dzieci do czterech grup: brak ryzyka, średnie ryzyko, wysokie ryzyko i bardzo wysokie ryzyko wystąpienia sercowo naczyniowych powikłań.

Tym problemem zajęła się również A. Krzyżaniak [4]. Dla znalezienia czynników ryzyka zastosowała model regresji wielokrotnej. Wykazała i przy tym potwierdziła wyniki z wcześniejszych badań, mówiące że na ciśnienie krwi wpływa wiek, masa ciała oraz choroby

nadciśnieniowej mogą być również podwyższone wartości wskaźnika wagowo – wzrostowego Quetelet'a oraz WMC – wskaźnika masy ciała.

W niniejszej pracy spośród wielu możliwych modeli matematycznych do szczegółowej analizy zostały wybrane trzy. Za ich pomocą starano się znaleźć czynniki, które wykazywałyby największy wpływ na występowanie nieprawidłowego ciśnienia u dzieci. Podział na grupy, czyli na dzieci z prawidłowym i nieprawidłowym ciśnieniem został przeprowadzony na podstawie pracy A. Krzyżaniak „Ciśnienie tętnicze u dzieci i młodzieży" Poznań 2004 [2]. Tabela 44 i 45 przedstawia różnice jakie powstały w grupach po zaklasyfikowaniu dzieci wg nowych niedostępnych przy powstawaniu tej pracy ogólnopolskich norm opublikowanych w 2009 roku.

Tabela 44. Porównanie liczebności w grupie chłopców z prawidłowym i nieprawidłowym ciśnieniem między normą z 2004 roku, a normą z 2009 roku

	norma ciśnienia wg badań z 2004 roku		**norma ciśnienia wg badań z 2009 roku**		**różnica w liczebnościach**
wiek	**0**	**1**	**0**	**1**	
16	46	12	44	14	**2**
17	42	18	36	24	**6**
18	70	28	63	35	**7**
Razem	**158**	**58**	**143**	**73**	**15**

Tabela 45. Porównanie liczebności w grupie dziewczynek z prawidłowym i nieprawidłowym ciśnieniem między normą z 2004 roku, a normą z 2009 roku

	norma ciśnienia wg badań z 2004 roku		**norma ciśnienia wg badań z 2009 roku**		**różnica w liczebnościach**
wiek	**0**	**1**	**0**	**1**	
16	88	24	97	15	**9**
17	88	26	88	26	**0**
18	147	29	131	45	**16**
Razem	**323**	**79**	**316**	**86**	**7**

W niniejszej pracy błąd pierwszego rodzaju α oznacza uznanie zdrowego dziecka za dziecko z nieprawidłowym ciśnieniem krwi, natomiast błąd drugiego rodzaju β jest równoznaczny z uznaniem dziecka z nieprawidłowym ciśnieniem za dziecko zdrowe.

H_0: brak odchyleń od normy

H_1: istnieją istotne odchylenia od normy

		rzeczywistość	
		H_0: prawdziwa	*H_1: fałszywa*
model	*H_0: prawdziwa*	nie ma błędu	błąd II rodzaju β
	H_1: fałszywa	błąd I rodzaju α	nie ma błędu

Po konsultacji z lekarzami zdecydowano, że popełnienie błędu pierwszego rodzaju α jest groźniejsze w konsekwencjach niż popełnienie błędu drugiego rodzaju. Dlatego można stwierdzić, że modele użyte w tej pracy są odpowiednimi metodami dla wykrywania nieprawidłowego ciśnienia krwi u dzieci ze względu na wysoką swoistość. Najlepszą techniką okazały się MARSplines'y, które mają najwyższą zarówno czułość jak i swoistość. Za pomocą odpowiednio dobranych parametrów dla tych modeli udało się wyłonić u chłopców dwa najistotniejsze predyktory mogące wpływać na nieprawidłowe ciśnienie krwi: podwyższone tętno oraz podwyższony wskaźnik wagowo – wzrostowy talia/wysokość. U dziewczynek z kolei najistotniejszymi czynnikami wybranymi przez te modele są: podwyższone tętno oraz podwyższone obwody bioder i uda. W literaturze przedmiotu i w dostępnych bazach Uniwersytetu nie znaleziono analiz, które mogłyby potwierdzić lub zanegować powyższe wyniki, nie znaleziono również badań dotyczących wpływu pomiarów antropometrycznych na ciśnienie tętnicze krwi. Badania epidemiologiczne dotyczą głównie wykrywania zaburzeń w rozwoju fizycznym dzieci i młodzieży na podstawie tychże wskaźników [57, 58].

Badania przeprowadzone w niniejszej pracy pozwoliły wygenerować trzy modele matematyczne, które klasyfikują dzieci na te z prawidłowym i z nieprawidłowym ciśnieniem. W przypadku drzew klasyfikacyjnych stałe podziału zaproponowane przez model okazały się trudne w interpretacji, gdyż nie są liczbami całkowitymi **tabela 22 i 25**. Aby zmienić te mało intuicyjne i problematyczne dla lekarzy kryteria podziału program Statistica 8.0 udostępnia interakcyjne drzewa klasyfikacyjne. Za pomocą tej aplikacji badacz jest w stanie sam zadecydować o wartości stałych podziału.

Na **rysunku 43 i 44** przedstawione zostały drzewa klasyfikacyjne dla chłopców i dziewczynek z całkowitymi stałymi podziału, a **tabele 46, 47** przedstawiają ilości błędnych klasyfikacji dla poszczególnych grup.

Rysunek 43. Wykres drzewa klasyfikacyjnego dla chłopców

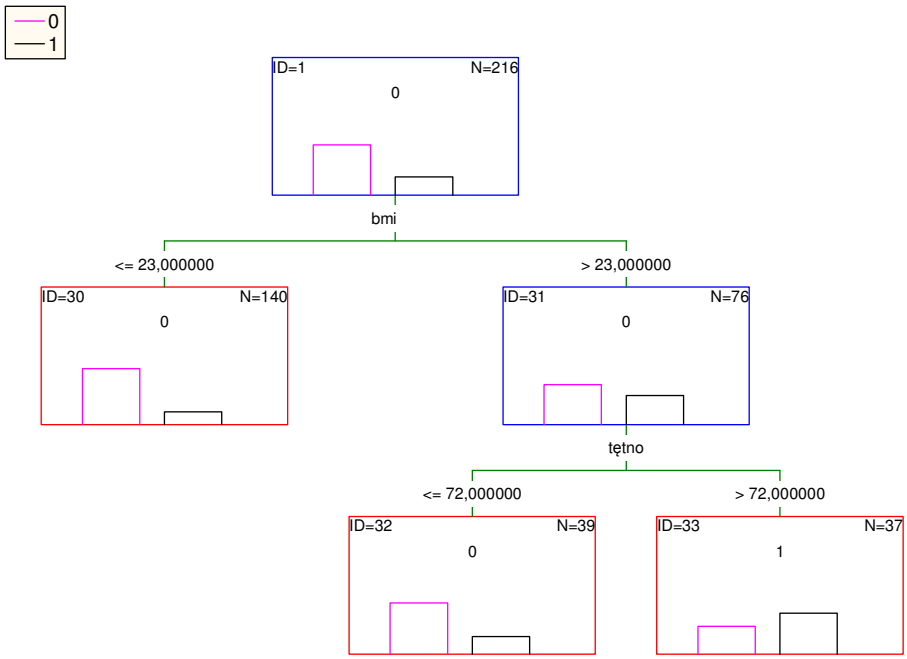


Tabela 46. Macierz błędnych klasyfikacji dla uzyskanego metodą interakcyjną drzewa w grupie uczącej u chłopców

Błędne klasyfikacje dla próby uczącej
N próby uczącej = 216

	obserwowane 0	**obserwowane 1**
przewidywane 0	143	36
przewidywane 1	15	22

Rysunek 44. Wykres drzewa klasyfikacyjnego dla dziewczynek

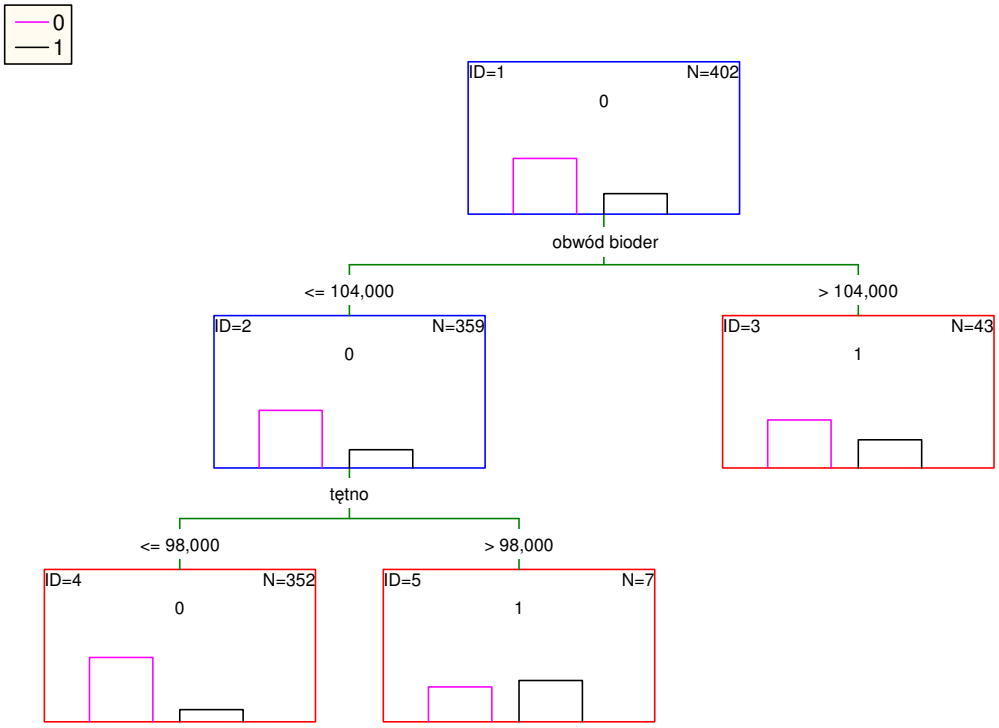


Tabela 47. Macierz błędnych klasyfikacji dla uzyskanego metodą interakcyjną drzewa w grupie uczącej u dziewczynek

Błędne klasyfikacje dla próby uczącej
N próby uczącej = 402

	obserwowane 0	**obserwowane 1**
przewidywane 0	306	46
przewidywane 1	17	33

Porównując wyniki otrzymane w rozdziale *11.3.* **tabela 24, 27** z otrzymanymi powyżej **tabela 43 i 44** możemy zaobserwować nieznaczne pogorszenie klasyfikacji u chłopców dla nowo utworzonych drzew. U dziewczynek klasyfikacja oryginalnych drzew jest podobna, do klasyfikacji drzew interakcyjnych. Interpretacja stałych podziału w drzewach interakcyjnych jest prostsza i łatwiejsza do zapamiętania, a ponieważ nastąpiło tylko nieznaczne pogorszenie klasyfikacji możemy wykorzystywać nowe stałe w praktyce.

Tabela 48. Porównanie poprawności klasyfikacji drzew dla grupy uczącej

	drzewo oryginalne		**drzewo interakcyjne**	
	% poprawnych 0	**% poprawnych 1**	**% poprawnych 0**	**% poprawnych 1**
chłopcy	84	66	80	59
dziewczynki	87	64	87	66

W niniejszej pracy udało się wyłonić model matematyczny który najlepiej klasyfikuje dzieci na dzieci zdrowe i z nieprawidłowym ciśnieniem krwi. Udało się ustalić zmienne, które są najistotniejsze, które znacząco wpływają na ciśnienie krwi u dzieci i mogą być wykorzystane w algorytmie postępowania testu przesiewowego nadciśnienia. Otrzymane wyniki mogą ułatwić znalezienie dzieci narażonych na wystąpienie nieprawidłowego ciśnienia krwi i tym samym zapobiec konsekwencjom wystąpienia sercowo - naczyniowych powikłań.

14. Wnioski

1. Porównując trzy modele klasyfikacyjne stwierdzono, że spośród analizowanych modeli model MARSplines jest optymalną techniką z najlepszą czułością i swoistością.
2. W predykcji nieprawidłowego ciśnienia krwi największe znaczenie u chłopców ma podwyższone tętno oraz podwyższony wskaźnik talia/wysokość. U dziewczynek zmiennymi wskazującymi na możliwość wystąpienia nieprawidłowego ciśnienia krwi są podwyższone tętno oraz podwyższone obwody bioder i uda.
3. Wskaźnikiem proporcji wagowo – wzrostowych, który najlepiej opisuje u chłopców nieprawidłowe ciśnienie krwi jest wskaźnik Rohrer'a, który ma najlepszą czułość i wskaźnik BMI, który ma najlepszą swoistość. U dziewczynek najlepszym wskaźnikiem jest wskaźnik Rohrer'a, który ma najlepszą czułość i wskaźnik Quetelet'a, który ma najlepszą swoistość.

15. Bibliografia

1. Pediatria PRAKTYCZNA Tom 6 nr 4 Poznań 1997 s. 4,5
2. Krzyżaniak A., Ciśnienie tętnicze u dzieci i młodzieży, praca zbiorowa, Poznań 2004
3. Wyszyńska T., Litwin M., Nadciśnienie tętnicze u dzieci i młodzieży, Warszawa 2002 s. 9-60
4. Krzyżaniak A., Ciśnienie tętnicze krwi dzieci i młodzieży miasta Poznania w latach 1986 i 1996 uwarunkowania, kierunek zmian, normy, Poznań 1999
5. Bryl W., Wywiad rodzinny, wskaźniki antropometryczne, wybrane parametry metaboliczne i skuteczność leczenia przeciwnadciśnieniowego u młodzieży z pierwotnym nadciśnieniem tętniczym, Poznań 2006, s. 6-20
6. Gromada M., Zalety i ograniczenia drzew klasyfikacyjnych, , 2003
7. Koronacki J., Ćwik J., Statystyczne systemy uczące się, WNT Warszawa 2005
8. Podręcznik STATISTICA
9. Teres Kenneth Sorensen, Gerrit K. Janssens, Data mining with genetic algorithms on binary, European Journal of Operational Research, 151, (2003), s. 253-264
10. Hastie T., Tibshirani R., Fiedman J., The Elements of Statistical Learning, 2001, s. 266-272, 283-290, 276-278
11. Breiman L., Friedman J.H., Charles J., Classification and Regression Trees, Stone Chapman & Hall, New York 1984
12. Friedman J. H. Multivariate adaptive regression splines, Department of Statistics Tech. Report 102 Rev, August 1990
13. Kochańska-Dziurowicz A., Klimek K., Mielniczuk M., Przydatność krzywych ROC w ocenie stosowanych testów laboratoryjnych, DIAGN.LAB., 1999, s. 83-104
14. Fawcett T., ROC Graphs: Notes and Practical Considerations for Researches, Kluwer, Academic Publishers Netherland, 2004
15. Fawcett T., An introduction to ROC analysis, Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306, USA, 19 December 2005
16. Klimek K., Przydatność krzywych ROC w diagnostyce różnicowej łagodnego rozrostu i raka gruczołu krokowego. Meta – analiza, Kraków 2000 Statsoft Polska
17. Bamber C. D., Math J.,The area above the ordinal dominance graph and area below the receiver operating characteristic, Psychol., 1975, s. 12, 387

18. Hanley J.A., McNeil B. J., The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology, 1982, s. 143, 29
19. Kleinbaum D. G., Klein M., Logistic Regression, Departament of Epidemiology Emory University Atlanta, USA 1994
20. Stanisz A., Przystępny kurs statystyki z wykorzystaniem programu STATISICA PL na przykładach z medycyny tom II, Kraków 2000
21. Stanisz A., Regresja logistyczna, Medycyna Praktyczna 2001/10
22. Petrie A., Sabin C., Statystyka medyczna w zarysie, tłumaczenie Jerzy Moczko, Wydawnictwo lekarskie PZWL, Warszawa 2006
23. Wolański N., Metody kontroli i normy rozwoju dzieci i młodzieży, PZWL Warszawa 1975
24. Książyk J., Współczynnik masy ciała – propozycja nowej metody oceny stanu odżywiania. Pediatria Polska 1995, s. 70, 4, 347-351
25. Garrow J. S., Webster J., Quetelet's index as a measure of fatness. Int. J. Obes. 1985, s. 9, 2 147-153
26. Włodarski M., Porównanie parametrycznej i nieparametrycznej metody obliczania krzywej ROC na przykłądzie zbioru sygnałów elektroretinograficznych, Metoda Informatyki Stosowanej, nr 2/2009 (19), Polska Akademia Nauk, s. 177-192
27. Falkner B., Sadowski Robert H., Hypertension in Children and Adolescent, The American Journal of Hypertension, 1995 December, vol. 8, no. 12, part 2, s. 106S-110S
28. Cromwell Polly F., Maunn N., Zolkowski-Wynne J., Evaluation and Management of Hypertension in Children and Adoles-cents (Part two): Evaluation and Management, Journal of Pediatric Health Care, September/October 2005, s. 309-313
29. Wolański N., Podobieństwo tętniczego ciśnienia krwi między rodzicami i ich dziećmi w różnej fazie rozwoju osobniczego, Przegląd antropologiczny, Tom 37 z. 1, Poznań 1971
30. Miall W.E., Oldham P.D. Brit med. J., t. 1, 1963 s. 471-283
31. Napoleon W., Pyżuk M., Przegląd Antropologiczny, t. 35, s. 437, 1969
32. Napoleon W., Pyżuk M Mat, i Prace Antrop., t. 78, 1969
33. Coody Deborah K., Yetman Robert J., Portman Ronald J., Hypertension in Children, Journal of Pediatric Health Care, 1995, January-February, s. 3-11
34. Simsolo Rosa B., Romo Miriam M., Rabinovich Laura, Bonanno Mariela, Grunfeld Beatriz, Family History of Essential Hypertension Versus Obesity as Risk Factor for

Hypertension in Adolescents, American Journal of Hupertension, Ltd. 1999, s. 260-263

35. Cromwell Polly F., Maunn Nancy, Zolkowski-Wynne Joanna, Evaluation and Management of Hypertension in Children and Adoles-cents (Part two): Evaluation and Management, Journal of Pediatric Health Care, September/October 2005, s. 309-313
36. Majewski Marek, Szajner – Milart Irena, Ciśnienie tętnicze a otyłość u dzieci i młodzieży szkolnej – badania epidemiologiczne, Pediatria Polska 1991 LXVI 3-4
37. Maggio Albane B. R., Aggoun Ycine, Marchand M., Martin E., Herrmann Francois, Beghetti Maurice, Farpour-Lambert Nathalie J., Associations among Obesity, Blood Pressure, and Left Verticular Mass, The Journal of Pediatrics, April 2008, s. 489-493
38. Neutel Joel M., Smith David H. G., Weber Michael A., Is High Blood Pressure a Late Manifestation of the Hypertension Syndrome?, American Journal of Hypertension, Ltd. 1999, s. 215-223
39. Sorof Jonathan M., Poffenbarger Tim, Franco Kathy, Portman Ronald, Evaluation of White Coat Hypertension in Children: Importance of the Definitions of Normal Ambulatory Blood Pressure and the Severity of Casual Hypertension, The American Journal of Hypertension, Ltd. 2001, s.855-860
40. How common is hypertension in adolescents?, The Journal of Pediatrics, June 2007, s. 640
41. Diller Philip M., Huster Gertrude A., Leach Alan D., Laskarzewski Peter M., Sprecher Dennis L., Definition and applicaction of the discretionary screening indicators according to the National Cholesterol Education Program for Children and Adolescents, The Journal of Pediatrics, vol. 126, no. 3, s. 345-351
42. Matsuoka Seiji, Awazu Midori, Masked hypertension in children and young adults, Pediatr NEphrol 2004 vol. 19, s. 651-654
43. Pediatrics Report of the Second Task Force on Blood Pressure Control in Children - 1987 Valume 79 JAnuary 1987 nr 1
44. Beall C. M., Gebremedhin A., Brittenham G. M., Decker M. J., Shambeo M., Blood Pressure among Ethiopians on the Simien Plateau, Annals of human biology, 1997, vol. 24, no. 4, s. 333-342
45. Portman Ronald J., McNiece Karen L., Swiford Rita D., Braun Michael C., Samuels Joshua A., Pediatric Hypertension: Diagnosis, Evaluation, Management, and Treatment for the Primary Care Physician, Curr Probl Pediatr Adolesc Health Care, August 2005, s. 262-289

46. Rashid Asrar, Ivy D. Dunbar, Pulmonary hypertension in children, Current Pediatrics 2006, 16, s. 237-247
47. Landsberg Lewis, Hypertension in the obese patient : pathogenic mechanism, cardiovascular risk and treatment, AJH-May 2003-vol. 16, no. 5, part 2
48. Oparil Suzanne,Treating Hypertension in the obese patient, AJH-May 2003-vol. 16, no. 5, part 2
49. Litwin Mieczysław, Śladowska Joanna, Syczewska Małgorzata, Niemirska Anna, Daszkowska Jadwiga, Antoniewicz Jolanta, Wierzbicka Aldona, Wawer Zbigniew T., Different BMI cardiovascular risk thresholds as markers of organ damage and metabolic syndrome in primary hypertension, Pediatr Nephrol 2008, vol. 23, s. 787-796
50. Robinson Renee F., Batisky Donald L., Hayes John R., Nahata Milap C., Mahan John D., Body mass index in primary and secondary pediatric hypertension, Pediatr. Nephrol 2004, vol. 19, s. 1379-1384
51. Sorof Jonathan M., Turner Jennifer, Poffenbarger Tim, Portman Ronald J., Influance of BMI on blood pressure varies by ethnicity in school-aged children, AJH May 2003, vol. 16, s. 494
52. Grajek S., Paradowski S., Cieślicka T., Ocena zależności ciśnienia tętniczego od masy ciała przy użyciu różnych wskaźników wagowo – wzrostowych, Prz. Epid. 1981, s. 35, 4, 489-495
53. J. Math, The area above the ordinal dominance graph and area below the receiver operating characteristic curve, Psychol., 1975, s. 12, 387
54. The meaning and use of the area under a recciver operating characteristic (ROC) curve, Radiology 1982, s. 143, 29
55. Kochańska – Dziurowicz A., Mielniczuk M. R., Stojko A., Kaletka J.,The clinical utility of measuring free-to-total prostate-specific antigen (PSA) ratio and PSA density in differentrating between benign prostactic hyperplsia and prostate cancer, Br. J. Urology 1998, s. 81, 834
56. Tadeusz Nawarycz, Krzysztof Pytel, Lidia Ostrowska – Nawarycz, System rozmyty do stratyfikacji ryzyka sercowo – naczyniowego, Proceedings of the 2nd Polish and International PD- Forum Conference on Computer Science. Smardzewice – Łódź, Poland, 2006
57. Jodkowska Maria, Woynarowska Barbara, Oblacińska Anna, Test przesiewowy do wykrywania zaburzeń w rozwoju fizycznym u dzieci i młodzieży w wieku szkolnym,

Warszawa 2007, Instytut Matki i Dziecka Zakład Ochrony i Promocji Zdrowia Dzieci i Młodzieży, Samodzielna Pracownia Medycyny Szkolnej, konsultacje - Prof. dr hab. Jadwiga Charzewska, Komitet Antropologii Polskiej Akademii Nauk

58. Charzewska J., Bergman P., Kaczanowski K., Piechaczek H., Otyłość - epidemią XXI wieku. Dziewiąte Warsztaty Antropologiczne Im. Profesora Janusza Charzewskiego, Warszawa 2006, AWF
59. Kowalska Małgorzata, Krzych Łukasz J., Siwik Paulina, Zawiasa Agnieszka, Uwarunkowania występowania nadciśnienia tętniczego u chłopców i dziewcząt w wieku szkolnym w województwie śląskim, nadciśnienie tętnicze, 2008, tom 12, nr 4, s. 269-276
60. Januszewicz Andrzej, Nadciśnienie tętnicze u dzieci i młodzieży, Medycyna Praktyczna, 2002, www.mp.pl
61. Tykarski A., Posadzy-Małaczyńska A., Wyrzykowski B. i wsp., Rozpowszechnienie nadciśnienia tętniczego oraz skuteczność jego leczenia u dorosłych mieszkańców naszego kraju, wyniki programu WOBASZ Pol. 2005; 63 (supl. IV): s. 614–661.
62. Wojdon – Machała H., Próba ustalenia norm ciśnienia tętniczego dla dziewcząt w okresie pokwitania, Pediat. Pol. 1970, s. 45, 9, 1071-1080
63. Kopczyński J., Ciśnienie tętnicze w zbiorowości młodzieży, Prz. Epid. 1968, s. 22, 3, 311-320
64. Gerkowicz T., Szajner – MilartI., Jabłońska K.,Olajossy M., Zachowska B., Nowak L., Bocheńska E., Badania epidemiologiczne ciśnienia tętniczego u dzieci z wybranych szkół i przedszkoli miasta Lublina, Pol. Tyg. Lek. 1974, 29, 17/18, s. 695-698
65. Chodakowska J., Czarnecki W., Januszkiewicz P., Deka A., Nadciśnienie tętnicze u licealistów warszawskich, Pol. Tyg. Lek. 1977, s. 32, 31, 1191-1194
66. Kowalik I., Próba ilościowej oceny wpływu czynników genetycznych i środowiskowych na zróżnicowanie ciśnienia tętniczego krwi, Prz. Antrop. 1984, 99, 5, s. 315-324
67. Lipiec J., Częstość występowania podwyższonego ciśnienia tętniczego krwi i jego rozkład dzieci i młodzieży, Prz. Pediat. 1981, 11, 4, s. 375-379
68. Waszyńska T., Nadciśnienie tętnicze u dzieci i młodzieży szkolnej – ocena częstości występowania i przyczyn, Pediat. Pol. 1985, 2, s. 169-176

69. Baszczyński J., Sordyl E., Karpiński E., Sobuś W., Szydłowski A., Żytkiewicz B., Nadciśnienie tętnicze krwi u chłopców w wieku 9-19 lat w regionie uprzemysławianym, Zdr. Pub. 1982, s. 93, 5-6, 231-233
70. Krzyżaniak A., Paluszak W., Bortkiewicz E., Wojdon – Machała H., Mroziński B., Maciejewski J., Nadciśnienie tętnicze u dzieci i młodzieży, Przekrojowe badania kształtowania się wartości ciśnienia wśród młodzieży miasta Poznania, Pediat. Prakt., 1997, 5, 4, s. 5-25
71. Harańczyk G., Krzywe ROC, czyli ocena jakości klasyfikatora i poszukiwanie optymalnego punktu odcięcia, StatSoft Polska 2010

Streszczenie

Przedmiotem rozprawy jest zastosowanie, ocena i porównanie trzech modeli matematycznych w celu znalezienia tego, który najlepiej klasyfikuje dzieci na dzieci z prawidłowym i nieprawidłowym ciśnieniem tętniczym krwi. Celem jest wyznaczenie zmiennych, które są najistotniejsze w predykcji nieprawidłowego ciśnienia krwi oraz wyznaczenie wskaźnika wagowo – wzrostowego spośród czterech porównywanych: wskaźnik Quetelet'a, wskaźnik masy ciała BMI, wskaźnik Rohrer'a, współczynnik masy ciała WMC, który najlepiej opisywałby ciśnienie tętnicze krwi u dzieci.

W badaniu wzięło udział 1378 dzieci między 7, a 18 rokiem życia z Wielkopolski. Do analizy wzięto tylko dzieci powyżej piętnastego roku życia ze względu na małe liczebności w grupie badanej w pozostałych latach.

W pracy porównywano regresję logistyczną z drzewami klasyfikacyjnymi i z modelem MARSplines. Model MARSplines został wybrany jako optymalny z najwyższą czułością i swoistością.

Najistotniejszymi zmiennymi wybranymi spośród analizowanych, które mogą pomóc w predykcji nieprawidłowego ciśnienia krwi okazały się dla chłopców tętno i wskaźnik wagowo – wzrostowy talia/wysokość, a dla dziewczynek tętno, obwód uda i obwód bioder.

Za pomocą krzywych ROC ustalono, że w grupie chłopców najlepszą czułość ma wskaźnik Rohrer'a, a swoistość wskaźnik BMI. W grupie dziewczynek najwyższą czułość wykazał wskaźnik Rohrer'a, a swoistość wskaźnik Ouetelet'a.

Summary

The subject area of presented doctoral thesis is application, evaluation and comparison of three mathematical models. The main goal is to find such a model which is the best in disrimination children into children with normal blood pressure and children with abnormal blood pressure. The aim is to find variable which is the most significant to predict abnormal blood pressure and to find such a index of four which were compared: Quetelet's, BMI, Rohrer's, WMC which would be the best in description of blood pressure in children.

The research group consists of 1378 children between 7 and 18 years old. Data analysis group consists of children who are older then 15 years old because of small amount of children in younger case groups.

In doctoral thesis the results of the most frequently used logistic regression method was compared to results of classification trees and to MARSplines techniques. The MARSplines turned out to be the best technique with the highest sensitivity and specificity.

The most significant variables which were chosen of those analyzed appeared to be appropriately: for boys- heart rate and waistline/height index, for girls: heart rate and hip size and thigh size.

On the basis of ROC curves we found that in boy's group the best sensitivity occurs for Rohrer's index and the best specificity has BMI index. In girl's group the best sensitivity was proved for Rohrer's index, while the best specificity has Quetelet's index.