Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Marta Tyce
Nr albumu: 277952

Drzewa decyzyjne z użyciem pakietu R. Zastosowanie w badaniach występowania nawrotu choroby u pacjentek z nowotworem piersi.

Praca licencjacka
na kierunku MATEMATYKA

Praca wykonana pod kierunkiem
dra inż. Przemysława Biecka
Instytut Matematyki Stosowanej i Mechaniki

Sierpień 2011

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

Praca dotyczy drzew decyzyjnych oraz przykładów zastosowania w pakiecie statystycznym R. Praca została podzielona na trzy części. W pierwszym rozdziale omówione zostały zagadnienia takie jak budowa drzewa oraz algorytm lasów losowych. Drugi rozdział poświęcony został funkcjom przydatnym w generowaniu drzew oraz lasów losowych. W trzeciej części pracy przedstawiona została analiza czynników, które mogą mieć wpływ na zwiększone ryzyko pojawienia się nawrotu lub przerzutu u pacjenta w przeciągu 5 lat od rozpoznania choroby.

Słowa kluczowe

drzewo decyzyjne, las losowy, miara różnorodności, indeks Giniego, entropia, predykcja, błąd klasyfikacji

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.2 Statystyka

Klasyfikacja tematyczna

46N30, 62P10

Tytuł pracy w języku angielskim

Classification trees with using statistical package R. Application in breast cancer study.

Spis treści

Wprowadzenie

W mojej pracy przedstawię wstęp do teorii drzew decyzyjnych oraz jej wykorzystywanie za pomocą pakietu R. Drzewa decyzyjne są jedną z najbardziej skutecznych i najpopularniejszych metod klasyfikacji. Polega ona na klasyfikowaniu nowych, przyszłych obserwacji, o których nie mamy informacji o przynależności klasowej. Drzewa znalazły zastosowanie w takich dziedzinach jak botanika i medycyna. Coraz częściej sięga się do nich również w ekonomii, ponieważ są w stanie usprawniać oraz ułatwiać komputerowe procesy podejmowania decyzji.

Temat dotyczący drzew klasyfikacyjnych i rodzin klasyfikatorów został podzielony na dwie części. W swojej pracy zajmę się budową drzew decyzyjnych oraz lasami losowymi, natomiast praca Dominiki Nowickiej dotyczyć będzie reguł przycinania drzew oraz algorytmów bagging i boosting.

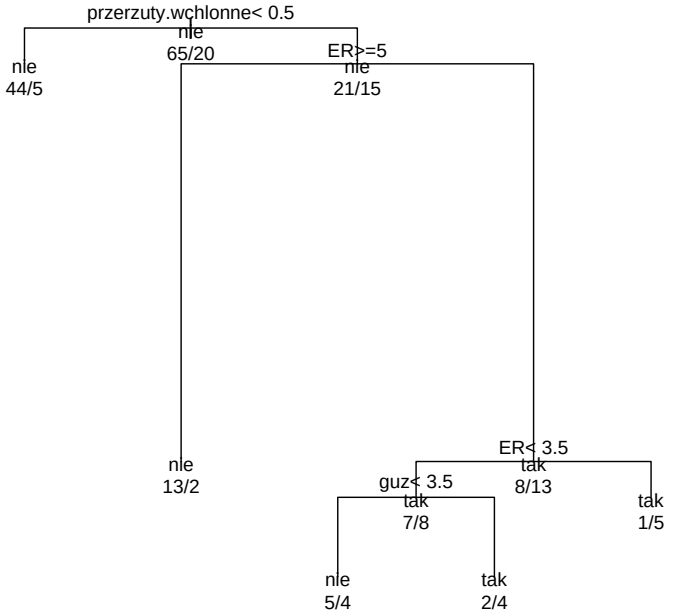
Pierwszy rozdział poświęcę teorii. Omówię proces budowy drzewa decyzyjnego, w jaki sposób wybiera się atrybut oraz najlepszy wskaźnik podziału próby, jakie można wyróżnić miary różnorodności oraz jakie mają one wady. Następnie skupię się na lasach losowych. Rozdział drugi w całości poświęcony zostanie zastosowaniom. Przedstawię różne funkcje, które stosuje się do budowy drzewa decyzyjnego, lasu losowego oraz predykcji danych. Począwszy od wyświetlania drzewa i jego opisu oraz opisu lasu losowego, interpretacji wyników przejdę do predykcji, badania błędów predykcji oraz wyznaczania rankingu zmiennych poprzez wykonanie oceny ich istotności. Rozdział trzeci poświęcony jest analizie danych rzeczywistych dotyczących 85 pacjentek z nowotworem piersi. Dane pochodzą z Dolnośląskiego Centrum Onkologii. Za pomocą drzew klasyfikacyjnych chciałabym sprawdzić, które zmienne mogą mieć wpływ na zwiększone ryzyko pojawienia sie nawrotu lub przerzutu w przeciągu 5 lat od rozpoznania choroby.

Rozdział 1

Teoria

1.1. Wstęp - interpretacja drzewa decyzyjnego

Drzewo jest acyklicznym spójnym grafem skierowanym. Składa się z korzenia oraz gałęzi prowadzących z korzenia do kolejnych węzłów. W korzeniu skupiona jest cała wybrana przez nas próba. Następnie, elementy tej próby są przesuwane w dół drzewa poprzez gałęzie do węzłów. W każdym węźle podejmowana jest decyzja o wyborze gałęzi, w kierunku której przesuną się elementy. W ten sposób próba zostaje podzielona na podgrupy. Pod każdym węzłem wyszczególnione jest kryterium podziału dokonywanego w danym węźle, które jest jednakowe dla wszystkich elementów próby. Ostatecznym etapem budowy są liście drzewa, którym przypisane zostają etykiety (określona zostaje przynależność klasowa).



Rysunek 1.1: Przykładowe drzewo dycyzyjne wyznaczone za pomocą funkcji *rpart()*.

Algorytmy drzew można podzielić na dwie kategorie, dotyczące drzew klasyfikacyjnych (decyzyjnych) oraz drzew regresyjnych. W pierwszym przypadku mamy do czynienia z drzewami, gdzie zmienną dyskryminującą jest zmienna jakościowa. W drugim zmienną dyskryminującą jest zmienna ilościowa. W tym rozdziale głównym źródłem informacji będzie [2].

Posłużę się drzewem z rysunku 1.1 opisującym klasyfikację 85 kobiet chorych na raka piersi pod względem pojawienia się nawrotu w ciągu 5 lat od rozpoznania choroby. Etykieta "tak" oznacza, że nawrót nastąpił. Każda kobieta opisana jest wektorem cech, w tym przypadku są to: występowanie przerzutów w węzłach chłonnych (zmienna przyjmująca wartości 0, 1), wielkość guza (wyrażana w centymetrach) oraz poziom receptora ER (zakres: 0–12). Obok gałęzi podany jest warunek podziału, który musi być spełniony, aby dany element próby losowej trafił do odpowiedniego węzła.

Na przykład, osoba u której pojawiły się przerzuty w węzłach chłonnych, ale wynik ER jest $\geqslant 5$ zostanie zaklasyfikowana do liścia z numerem 6 jako osoba u której nie wystąpił w ciągu 5 lat przerzut. Pod każdym węzłem znajduje się iloraz elementów należących do odpowiednich klas, które znalazły się w węźle. Odwołam się ponownie do węzła z numerem 6. W tym węźle znajduje się 15 elementów wśród których 13 ma przyporządkowaną wartość "nie", pozostałe dwie mają wartości "tak", całemu węzłowi została przypisana etykieta "nie". Można powiedzieć, że te 2 wartości zostały źle zaklasyfikowane, a wartość 2/15 nazwać ułamkiem błędów klasyfikacji popełnionych przez drzewo (2/15 to także oszacowanie na podstawie próby uczącej prawdopodobieństwa błędnej klasyfikacji pod warunkiem wystąpienia przerzutów w węzłach chłonnych oraz wyniku ER $\geqslant 5$).

1.2. Proces budowy drzewa

1.2.1. Wybór atrybutu

Podział w danym węźle odbywa się na podstawie znajdujących się w nim elementów próby uczącej. Polega on na najlepszym podzieleniu próby na części, które następnie przechodzą do węzłów dzieci. Próba znajdująca się w węźle charakteryzuje się pewnym rozkładem klas, czyli tak zwaną różnorodnością klas reprezentowanych przez tę próbę. Rozsądny podział wymaga:

1. podania odpowiedniej miary różnorodności klas w węźle,
2. podania miary różnicy pomiędzy różnorodnością klas w danym węźle oraz klas w węzłach-dzieciach,
3. podania algorytmu maksymalizacji tej różnicy.

Rozważam problem dyskryminacji o g klasach, 1,2,...,g, oraz próba ucząca $(\mathbf{x}_i, y_i)$, i=1,...,n. Rozpatruję dany węzeł m. Węzeł ten wyznacza w przestrzeni obserwacji χ pewien obszar R_m. Wektor obserwacji $\mathbf{x}$ znajduje się w węźle m gdy $\mathbf{x} \in R_m \subset \chi$. Niech n_m oznacza liczność próby uczącej, która trafiła do węzła m. Niech

$$\hat{p}_{mk} = \frac{1}{n_m} \sum_{i \in R_m} I(y_i = k) = \frac{n_{mk}}{n_m} \tag{1.1}$$

będzie ułamkiem obserwacji z próby uczącej należących do klasy k w obszarze R_m (gdzie I(A)=1 gdy warunek A zachodzi, I(A)=0 w przeciwnym przypadku). Następnie obserwacje znajdujące się w węźle m klasyfikowane są do klasy

$$k(m) = arg \max_k \hat{p}_{mk}. \tag{1.2}$$

Jest to klasa najliczniejsza w węźle m, czyli klasa najmocniej reprezentowana przez próbę uczącą.

W przypadku gdy

- węzeł m jest liściem – k(m) to ostateczny wynik klasyfikacji każdego wektora $\mathbf{x}$,
- w przeciwnym przypadku k(m) daje wyłącznie informacje, która klasa jest najbardziej reprezentowana w tym węźle. Każdy wektor $\mathbf{x}$, który znalazł się w tym węźle przechodzi do odpowiedniego węzła-dziecka.

1.2.2. Miary różnorodności klas w węźle

Najbardziej rozsądna miara różnorodności klas w węźle to miara, która przyjmuje wartość zerową w momencie gdy wszystkie obserwacje w węźle należą do tej samej klasy, oraz wartość maksymalną, gdy rozkład przynależności do klas jest jednostajny tzn. $p_{m1} = p_{m2} = ... = \frac{1}{g}$.
Popularne miary różnorodności to:
ułamek błędnych klasyfikacji $Q1_m(T)$

$$Q1_m(T) = \frac{1}{n_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk}(m), \tag{1.3}$$

wskaźnik Giniego (indeks Giniego) $Q2_m(T)$

$$Q2_m(T) = \sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{g} \hat{p}_{mk}(1 - \hat{p}_{mk}) = 1 - \sum_{k=1}^{g} \hat{p}_{mk}^2, \tag{1.4}$$

entropia $Q3_m(T)$

$$Q3_m(T) = -\sum_{k=1}^{g} \hat{p}_{mk} \log \hat{p}_{mk}. \tag{1.5}$$

W przypadku klasyfikacji binarnej tzn. gdy są dwie klasy (g=2), podane miary przyjmują następującą postać (gdzie p jest prawdopodobieństwem przynależności do jednej z klas):

$$Q1_m(T) = 1 - \max(p, 1 - p), \tag{1.6}$$

$$Q2_m(T) = 2p(1 - p), \tag{1.7}$$

$$Q3_m(T) = -p \log p - (1 - p) \log(1 - p). \tag{1.8}$$

Niech ustalona zostanie miara różnorodności klas w węzłach budowanego drzewa. Następnie niech:

- m_L i m_P będą kolejno lewym i prawym dzieckiem węzła-rodzica m,
- $\hat{p}_L$ będzie ułamkiem elementów próby uczącej, które przeszły z węzła m do m_L,
- $\hat{p}_R$ będzie ułamkiem elementów próby uczącej, które przeszły z węzła m do m_R,
- n_{m_L} - liczba obserwacji w m_L,
- n_{m_R} - liczba obserwacji w m_R.

$$\hat{p}_L = \frac{n_{m_L}}{n_m}$$
$$\hat{p}_R = 1 - \hat{p}_L$$

Łączna miara różnorodności klas w węzłach-dzieciach węzła m wyraża się wzorem:

$$Q_{m_L,m_R}(T) = \hat{p}_L Q_{m_L}(T) + \hat{p}_R Q_{m_R}(T). \tag{1.9}$$

Miara różnicy między różnorodnością klas w węźle-rodzicu i węzłach dzieciach wyraża się wzorem:

$$\Delta Q_{m,m_L,m_R}(T) = Q_m(T) - Q_{m_L,m_R}(T). \tag{1.10}$$

1.2.3. Wybór najlepszego wskaźnika podziału podpróby

Ten paragraf poświęcony jest poszukiwaniu najlepszego podziału podpróby, opartego na tylko jednym najlepiej wybranym atrybucie. Wśród wszystkich możliwych podziałów podpróby wybrany zostaje ten podział, który maksymalizuje (1.10).
Dla ustalonego atrubutu, który przyjmuje w całej próbie uczącej L różnych wartości, istnieje

$$\frac{1}{2}2^L - 1 = 2^{L-1} - 1$$

różnych podziałów próby znajdującej się w tym atrybucie (istnieje 2^L podzbiorów zbioru L-elementowego, jednakże nie interesuje nas dwukrotnie występujący podział na zbiór pusty oraz całość, ponadto już połowa z nich wyznacza wszystkie możliwe podziały). Należy jednak ograniczyć klasę możliwych podziałów do podziałów monotonicznych, podział monotoniczny to podział typu $x^{(l)} \leqslant$ c lub $x^{(l)} <$ c (oraz $x^{(l)} \geqslant$ c lub $x^{(l)} >$ c). Dzięki takiemu podziałowi liczba sposobów rozdzielenia podpróby wynosi $L - 1$.
Poniższe twierdzenie wyjaśnia w jaki sposób należy postępować w przypadku gdy atrybut jest nominaly, jego wartości nie tworzą żadnego naturalnego porządku (na przykład, kolor oczu lub płeć).

Twierdzenie 1 *Dla entropii i indeksu Giniego mamy:*

(i) $\Delta Q_{m,m_L,m_R}(T) \geqslant 0$ *i równość zachodzi wtedy i tylko wtedy, gdy rozkłady klas w rodzicu i obydwu potomkach są identyczne.*

(ii) Niech g=2 i niech atrybut będzie jakościowy, o L poziomach (wartościach). Załóżmy, że poziomy atrybutu, $x^{(i)}$ *zostały ułożone według rosnących wartości prawdopodobieństw* $p(1|x^{(i)})$*:*

$$p(1|x^{(1)}) \leqslant p(1|x^{(2)}) \leqslant ... \leqslant p(1|x^{(L)}).$$

Wówczas jeden z $L - 1$ *podziałów typu*

$$\{x^{(1)}, ..., x^{(l)}\}, \{x^{(l+1)}, ..., x^{(L)}\}$$

maksymalizuje $\Delta Q_{m,m_L,m_R}(T)$*.*

1.2.4. Wady poszczególnych miar różnorodności

Miara przyrostu informacji- entropia
Powołam się tu na pozycję [6]. W przypadku dwóch atrybutów do wyboru, miara przyrostu informacji wybierze ten o większej liczbie wartości. Jest to problem, w szczególności w sytuacji mocnego zróżnicowania liczności dziedzin atrybutów. Rozważmy skrajny przypadek, w którym

pewien atrybut np. data urodzin, ma tyle różnych wartości, ile jest przykładów uczących. Atrybut ten zostanie wybrany do zbudowania drzewa, ponieważ maksymalizuje on wartość miary entropii. W rezultacie otrzymamy płaskie i szerokie drzewo (każdy liść będzie zawierał pojedynczy przykład – datę), które jest mało czytelne oraz bezużyteczne do predykcji nowych przykładów z innymi wartościami daty niż te, które wystąpiły w zbiorze uczącym.

Ułamek błędnych klasyfikacji
Powołam się tu na przykład z [2], który szczegółowo omówię. Rozpatrzę dwie klasy (g=2), obliczę wartość $\Delta Q_{m,m_L,m_R}(T)$ odpowiednio dla cechy 2 oraz cechy 3, sprawdzając przy tym która z nich bardziej różnicuje cechę 1. Dane przedstawia tabela 1.1 oraz 1.2 (przy czym N oznacza ilość elementów przyporządkowaną do węzła lub klasy).

cecha 1	cecha 2				ogółem	
	TAK		NIE			
	N	%	N	%	N	%
TAK	300	75	100	25	400	50
NIE	100	25	300	75	400	50
ogółem	400	100	400	100	800	100

Tabela 1.1: Opis zmiennej "cecha 2", w kolumnach oznaczonych symbolem % znajdują się wartości określające zawartość elementów danej klasy w stosunku do ilości elementów w węźle wyrażonej w procentach.

cecha 1	cecha 3				ogółem	
	TAK		NIE			
	N	%	N	%	N	%
TAK	200	100	200	33.(3)	400	50
NIE	0	0	400	66.(6)	400	50
ogółem	200	100	600	100	800	100

Tabela 1.2: Opis zmiennej "cecha 3", w kolumnach oznaczonych symbolem % znajdują się wartości określające zawartość elementów danej klasy w stosunku do ilości elementów w węźle wyrażonej w procentach.

Zacznę od policzenia różnicy między różnorodnością klas na podstawie indeksu Giniego dla cechy 2.
$Q_m(T) = 2p(1-p) = 2 \cdot 0.5 \cdot (1-0.5) = 0.5$
$Q_{m_L,m_R}(T) = \hat{p}_L Q_{m_L}(T) + \hat{p}_R Q_{m_R}(T) = 0.5 \cdot (2 \cdot 0.75 \cdot (1-0.75)) + 0.5 \cdot (2 \cdot 0.25 \cdot (1-0.25)) = 0.375$
$\Delta Q_{m,m_L,m_R}(T) = Q_m(T) - Q_{m_L,m_R}(T) = 0.5 - 0.375 = 0.125$

Następnie dla cechy 3.
$Q_m(T) = 2 \cdot 0.5(1-0.5) = 0.5$
$Q_{m_L,m_R}(T) = 0.75(2 \cdot 0.(3) \cdot (1-0.(3))) + 0.25(2 \cdot 1 \cdot (1-1)) = 0.(3)$
$\Delta Q_{m,m_L,m_R}(T) = Q_m(T) - Q_{m_L,m_R}(T) = 0.5 - 0.(3) \approx 0.167$

Analogicznie dla ułamka błędnych klasyfikacji. Ze względu na cechę 2.
$Q_m(T) = 1 - \max(p, 1-p) = 1 - 0.5 = 0.5$
$Q_{m_L,m_R}(T) = \hat{p}_L Q_{m_L}(T) + \hat{p}_R Q_{m_R}(T) = 0.5(1 - \max(0.75, 1-0.75)) + 0.5(1 - \max(0.25, 1-0.25)) = 0.25$

$\Delta Q_{m,m_L,m_R}(T) = Q_m(T) - Q_{m_L,m_R}(T) = 0.5 - 0.25 = 0.25$

Ze względu na cechę 3.
$Q_m(T) = 1 - 0.5 = 0.5$
$Q_{m_L,m_R}(T) = 0.75(1 - \max(0.(3), 1 - 0.(3))) + 0.25(1 - \max(1, 1 - 1)) = 0.25$
$\Delta Q_{m,m_L,m_R}(T) = Q_m(T) - Q_{m_L,m_R}(T) = 0.5 - 0.25 = 0.25$

Dla miary indeksu Giniego wartość $\Delta Q_{m,m_L,m_R}(T)$ jest większa dla cechy 3. Można powiedzieć, iż cecha 3 lepiej różnicuje cechę 1. Tymczasem wartość $\Delta Q_{m,m_L,m_R}(T)$ jest jednakowa dla obydwu cech, jeżeli obliczamy ją na podstawie ułamka błędnych klasyfikacji. Podsumowując, użycie ułamka błędnych klasyfikacji jest w tym przypadku bezużyteczne. Obliczenie różnicy między różnorodnościami w węźle-rodzicu i węzłach dzieciach nie określa, który atrybut powinniśmy wybrać przy konstrukacji drzewa (podczas gdy wybór miary Giniego jednoznacznie wskazuje cechę 3).

1.3. Lasy losowe

Lasy losowe są uznawane za jedną z najlepszych metod klasyfikacji. Pojedyncze klasyfikatory lasu losowego to drzewa decyzyjne. Algorytym *RandomForest* bardzo dobrze nadaje się do badania próby, gdzie wektor obserwacji jest dużego wymiaru. Ich dodatkową zaletą jest możliwość użycia nauczonego lasu losowego do innych zagadnień niż tylko do klasyfikacji. Przykładowo, na podstawie drzew z lasu można wyznaczyć ranking zmiennych, a tym samym okreslić, które zmienne mają lepsze właściwości predykcyjne (to zagadnienie zostanie omówione w rozdziale 2 mojej pracy).

Algorytm budowy lasu losowego

- Losujemy ze zwracaniem z n-elementowej próby uczącej n wektorów obserwacji. Na podstawie takiej pseudopróby stworzone zostanie drzewo.
- W każdym węźle podział odbywa się poprzez wylosowanie bez zwracania m spośrod p atrybutów, następnie w kolejnym węźle k spośrod m atrybutów itd ($p \gg m \gg k$) (parametr m jest jedynym elementem algorytmu, który trzeba ustalić, wartość dająca dobre wyniki dla modeli decyzyjnych to około $m = \sqrt{p}$, dla modeli regresyjnych $\frac{m}{3}$).
- Proces budowania drzewa bez przycinania trwa, jeżeli to możliwe do momentu uzyskania w liściach elementów z tylko jednej klasy.

Proces klasyfikacji

- Dany wektor obserwacji jest klasyfikowany przez wszystkie drzewa, ostatecznie zaklasyfikowany do klasy w której wystąpił najczęściej.
- W przypadku elementów niewylosowanych z orginalnej podpróby, każdy taki i-ty element zostaje poddany klasyfikacji przez drzewa w których budowie nie brał udziału. Taki element zostaje następnie przyporządkowany klasie, która osiągana była najczęściej (w ten sposób zaklasyfikowane zostały wszystkie elementy z oryginalnej próby).

Rozdział 2

Pakiety i funkcje programu R

2.1. Wprowadzenie

Rozdział 2 jest poświęcony pakietom służącym do budowy modeli drzew klasyfikacyjnych i regresyjnych dostępnych w programie R oraz funkcjom wykorzystywanym w pracy z drzewami. Omówię odpowiednie składnie poleceń. Posłużę się takimi pakietami jak:

- *tree*,
- *rpart*,
- *party*,
- *maptree*,
- *randomForest*.

W celu przedstawienia procesu tworzenia drzew oraz sposobów posługiwania się nimi wykorzystam dane kobiet z nowotworem piersi. Dotyczą one 85 pacjentek, zbadanych pod względem wystąpienia przerzutu w przeciągu 5 lat od rozpoznania choroby. Będę posługiwać się dwoma zestawami danych: próba testowa to 20 obserwacji wybranych losowo spośród 85 nie zawierających braków danych oraz pozostałe 65 obserwacji będą stanowić próbę uczącą. W tabeli 2.1 przedstawiam przegląd wybranych zmiennych, które zostaną dokładniej omówione w rozdziale 3.

`nawrot.5lat`	czy nastąpił nawrót w ciągu 5 lat
`wiek.rozpoznanie`	wiek pacjenta w momencie rozpoznania choroby
`guz`	wielkość guza w cm
`przerzuty.wchlonne`	czy wystąpiły przerzuty w węzłach chłonnych
`rakpiersi.rodzina`	czy pojawił się rak piersi w wywiadzie rodzinnym
`VEGF`	poziom naczyniowo-śródbłonkowego czynnika wzrostu
`HGFR`	poziom receptora czynnika wzrostu hepatocytów

Tabela 2.1: W kolumnie pierwszej wymienione są skróty nazw zmiennych, w drugiej opisy zmiennych.

2.2. Funkcje

2.2.1. Budowa modelu: drzewa decyzyjne

W tym paragrafie podane zostaną funkcje, które uruchamiają procedury budowy modeli drzew klasyfikacyjnych i regresyjnych. Informacje, które tu wykorzystam, pochodzą przede wszystkim z [4]. Ograniczę się do tych funkcji, które wydają mi się bardziej przydatne w dalszej pracy nad danymi oraz ogólnym poznaniu działania drzew. Wymienię tu funkcję `tree` z pakietu *tree*, `rpart` z pakietu *rpart* oraz `ctree` z pakietu *party*. W tabeli 2.2 przedstawiony jest opis parametrów funkcji.

```
tree(formula, data, subset, na.action=na.pass,
     control=tree.control(nobs), method, split=c("gini"))

rpart(formula, data, subset, na.action=na.pass, method,
      split=c("information","gini"), control)

ctree(formula, data, subset, controls)

tree.control(nobs, mincuit=5, minsize=10)

rpart.control(minsplit=20, minbucket=round(minsplit/3),
              cp=0.01, maxdepth=30)

ctree_control(mincriterion = 0.95, minsplit = 20, minbucket = 7, maxdepth = 30)
```

`formula`	symboliczny opis modelu dyskryminacyjnego i regresyjnego
`data`	macierz danych, która uwzględnia zmienne modelu
`subset`	funkcja wskazująca podzbiór obserwacji do klasyfikacji
`na.action`	wyrażenie mające na celu wskazywanie sposobu postępowania w przypadku braków obserwacji
`control`	parametry sterujące procedurą budowy drzewa
`method`	metoda budowy drzewa, jedyna domyślna wartość to podział rekurencyjny
`split`	określenie miary różnorodności
`nobs`	liczba obserwacji w zbiorze uczącym
`mincriterion`	wartości statystyki lub 1 - p-wartość
`mincut`	minimalna liczba obserwacji w węźle, który powstał w wyniku podziału (domyślnie 5)
`minsize`	minimalna liczba obserwacji w węźle, który ulega podziałowi (domyślnie 10)
`minsplit`	minimalna liczba obserwacji w węźle, który ulega podziałowi (domyślnie 20)
`minbucket`	minimalna liczba obserwacji w liściu drzewa (domyślnie minsplit/3)
`cp`	zmiana wartości poprawy modelu (wzór (1.10)), jeżeli podział nie poprawia o wiecej niż cp to nie jest wykonywany
`maxdepth`	maksymalna głębokość drzewa (domyślnie 30)

Tabela 2.2: W kolumnie pierwszej wymienione są skróty nazw zmiennych, w drugiej opisy zmiennych.

Przykłady

W celu obejrzenia wyników pracy z pakietem R, najpierw stworzę drzewo (rys. 2.1) wywołując funkcję:

```
> drzewo1 <- rpart(nawrot.5lat~wiek.rozpoznanie+guz+przerzuty.wchlonne+
                   rakpiersi.rodzina+VEGF+HGFR, data=uczaca,
                   method="class", minsplit=5, minbucket=3)
```

Wywołując polecenie `drzewo1` uzyskam opis drzewa:

```
> drzewo1
n= 65

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 65 17 nie (0.73846154 0.26153846)
   2) przerzuty.wchlonne< 0.5 35  5 nie (0.85714286 0.14285714)
     4) guz< 3.25 21  1 nie (0.95238095 0.04761905) *
     5) guz>=3.25 14  4 nie (0.71428571 0.28571429)
      10) wiek.rozpoznanie< 50.5 4  0 nie (1.00000000 0.00000000) *
      11) wiek.rozpoznanie>=50.5 10  4 nie (0.60000000 0.40000000)
        22) wiek.rozpoznanie>=66.5 3  0 nie (1.00000000 0.00000000) *
        23) wiek.rozpoznanie< 66.5 7  3 tak (0.42857143 0.57142857) *
   3) przerzuty.wchlonne>=0.5 30 12 nie (0.60000000 0.40000000)
     6) wiek.rozpoznanie< 52.5 9  1 nie (0.88888889 0.11111111) *
     7) wiek.rozpoznanie>=52.5 21 10 tak (0.47619048 0.52380952)
      14) wiek.rozpoznanie>=59.5 15  6 nie (0.60000000 0.40000000)
        28) wiek.rozpoznanie< 63 4  0 nie (1.00000000 0.00000000) *
        29) wiek.rozpoznanie>=63 11  5 tak (0.45454545 0.54545455)
          58) HGFR< 1.5 3  0 nie (1.00000000 0.00000000) *
          59) HGFR>=1.5 8  2 tak (0.25000000 0.75000000) *
      15) wiek.rozpoznanie< 59.5 6  1 tak (0.16666667 0.83333333) *
```

Kolejno w wierszu odczytujemu: numer węzła, nazwę atrybutu oraz wartość w której wykonany został podział, liczbę obserwacji w węźle, liczbę obserwacji o wartości błędnie zaklasyfikowanych, klasę do jakiej został zaklasyfikowany węzeł oraz prawdopodobieństwa a posteriori dla klas. Symbol * informuje o tym czy dany węzeł jest liściem.

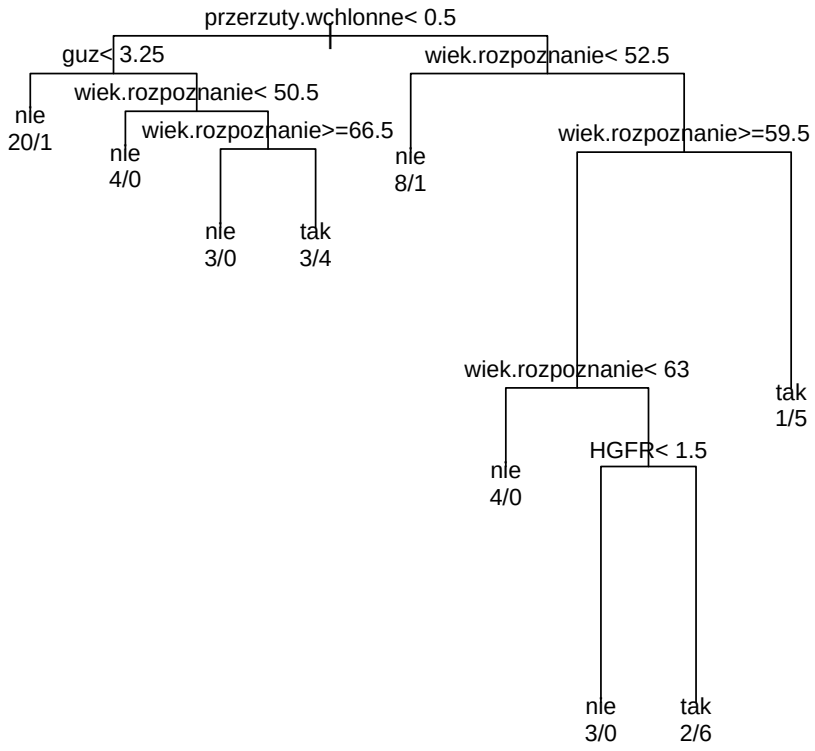
Kolejnym elementem jest wyświetlanie drzewa. Aby to uczynić możemy skorzystać najpierw z funkcji X11(). W ten sposób zostało przygotowane nowe okno do wyświetlenia drzewa. Polecenie `plot` tworzy sam rysunek drzewa (dodatkowo budując drzewo za pomocą funkcji `rpart` otrzymamy drzewo z krawędziami o długości proporcjonalnej do stopnia zróżnicowania obserwacji w węzłach), natomiast polecenie `text` dodaje napisy w węzłach drzewa. W tabeli 2.3 znajduje się opis argumentów.

```
X11(),
plot(drzewo),
text(drzewo,use.n=TRUE,all=TRUE,cex=0.8).
```

```
> X11()
> plot(drzewo1)
> text(drzewo1,use.n=TRUE,all=FALSE,cex=0.7)
```

`use.n=TRUE`	podaje liczebności klas w liściach
`all=TRUE`	podaje klasę w węzłach
`cex`	rozmiar czcionki

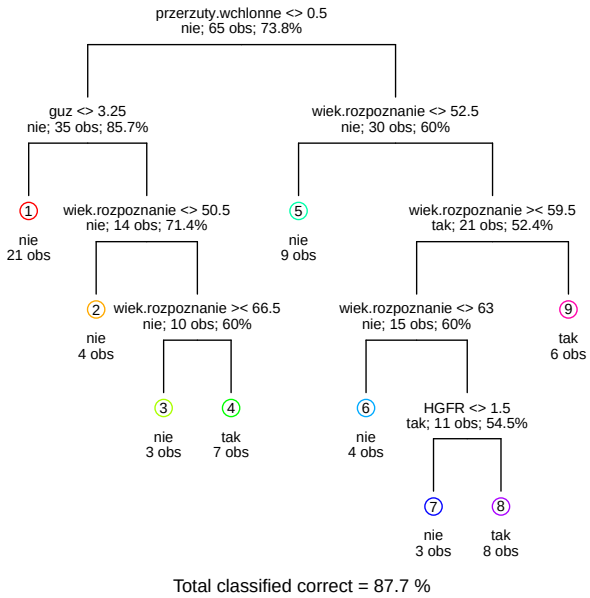
Tabela 2.3: W kolumnie pierwszej wymienione są skróty nazw zmiennych, w drugiej opisy zmiennych.



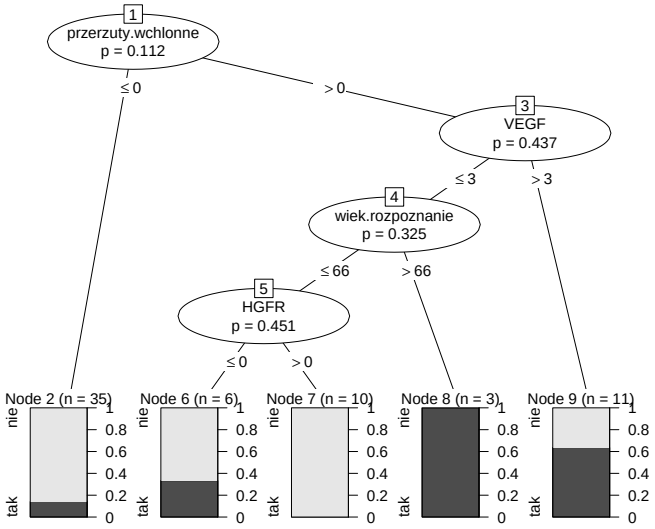
Rysunek 2.1: Przykładowe drzewo klasyfikacyjne wyznaczone funkcją *rpart()*

Inny wykres uzyskam korzystając z funkcji `draw.tree()` z pakietu *maptree* (rysunek 2.2). Parametr `nodeinfo` odpowiada za opis węzła. Dzięki niemu można uzyskać dodatkowe informacje na temat ilości elementów poprawnie zaklasyfikowanych, podanej w procentach, czyli także prawdopodobieństwa a posteriori dla klas. Można również użyć funkcji `ctree()` w celu lepszego graficznego przedstawienia wyników (rysunek 2.3). Poniżej przedstawiam sposoby wywołania obydwu funkcji.

```
> drzewo2 <- draw.tree(drzewo1,cex=0.7,nodeinfo=TRUE)
> drzewo3 <- ctree(nawrot.5lat~wiek.rozpoznanie+guz+przerzuty.wchlonne+
                   rakpiersi.rodzina+VEGF+HGFR, data=uczaca,
                   controls=ctree_control(minsplit=2,minbucket = 1,
                   maxdepth = 10, mincriterion = 0.50))
> plot(drzewo3)
```



Rysunek 2.2: Przykładowe drzewo klasyfikacyjne wyznaczone funkcją *rpart()*, wywołane funkcją *draw.tree()*



Rysunek 2.3: Przykładowe drzewo klasyfikacyjne wyznaczone funkcją *ctree()*

2.2.2. Budowa modelu: lasy losowe

```
randomForest(formula, data=NULL, na.action=na.fail, x, y=NULL, xtest=NULL,
             ytest=NULL, ntree=500, mtry)

rpart.control(minsplit=20, minbucket=round(minsplit/3),
              cp=0.01, maxdepth=30)
```

formula	symboliczny opis modelu dyskryminacyjnego i regresyjnego
data	macierz danych, która uwzględnia zmienne modelu
x	macierz danych albo macierz zmiennych odwzorowująca zbiór uczący
y	wektor wartości zmiennej objaśnianej dla zbioru uczącego
xtest	macierz danych albo macierz zmiennych odwzorowująca zbiór uczący, zawierająca dane dla zbioru testowego
ytest	wektor wartości zmiennej objaśnianej dla zbioru testowego
ntree	liczba drzew (domyślnie 500)
mtry	liczba zmiennych wybieranych losowo do stworzenia drzewa (domyślnie dla modelu klasyfikacyjnego – pierwiastek z liczby zmiennych, dla modelu regresyjnego – (liczba zmiennych)/3)

Tabela 2.4: W kolumnie pierwszej wymienione są skróty nazw zmiennych, w drugiej opisy zmiennych.

Przykłady

Lasy losowe dostępne są za pomocą funkcji `randomForest`, która znajduje się w pakiecie *randomForest*. Poniżej przedstawiam przykład jej użycia. Przy konstrukcji lasów losowych uzyskiwane są losowe wersje zbioru danych. Aby móc odtworzyć uzyskane wyniki należy skorzystać z funkcji

```
> set.seed()
```

Poniżej przedstawiam funkcje, które umożliwiają otrzymanie lasu losowego.

```
> las <- randomForest(nawrot.5lat~wiek.rozpoznanie+guz+przerzuty.wchlonne+
                      rakpiersi.rodzina+VEGF+HGFR,data=uczaca, ntree=500)
> las
```

W wyniku otrzymuję informacje na temat wygenerowanego lasu.

```
Call:
 randomForest(formula = nawrot.5lat ~ wiek.rozpoznanie + guz +
 przerzuty.wchlonne +  rakpiersi.rodzina + VEGF + HGFR,
 data = uczaca, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 35.38%
Confusion matrix:
    nie tak class.error
nie  42   6       0.125
tak  17   0       1.000
```

Wartość 0.125 to błąd klasyfikacji dla pierwszej klasy, Wartość 1.000 – błąd dla drugiej klasy. Ogólnie błąd klasyfikacji dla wszystkich danych jest na poziomie 35.38 %.
Ustawienie parametru `do.trace` pozwala na uzyskanie wglądu w proces redukcji błędu klasyfikacyjnego, obliczenego dla zbioru OOB po dodaniu kolejnych drzew. Wywołanie oraz fragment wyniku przedstawiam poniżej.

```
> las <- randomForest(nawrot.5lat~wiek.rozpoznanie+guz+przerzuty.wchlonne+
                      rakpiersi.rodzina+VEGF+HGFR,data=uczaca, ntree=500,
                      do.trace=25)

ntree      OOB      1      2
   25:  35.38% 16.67% 88.24%
   50:  33.85% 18.75% 76.47%
   75:  35.38% 16.67% 88.24%
  100:  36.92% 18.75% 88.24%
  125:  40.00% 18.75%100.00%
  (...)
  375:  35.38% 12.50%100.00%
  400:  35.38% 12.50%100.00%
  425:  35.38% 12.50%100.00%
  450:  35.38% 12.50%100.00%
  475:  38.46% 16.67%100.00%
  500:  35.38% 12.50%100.00%
```

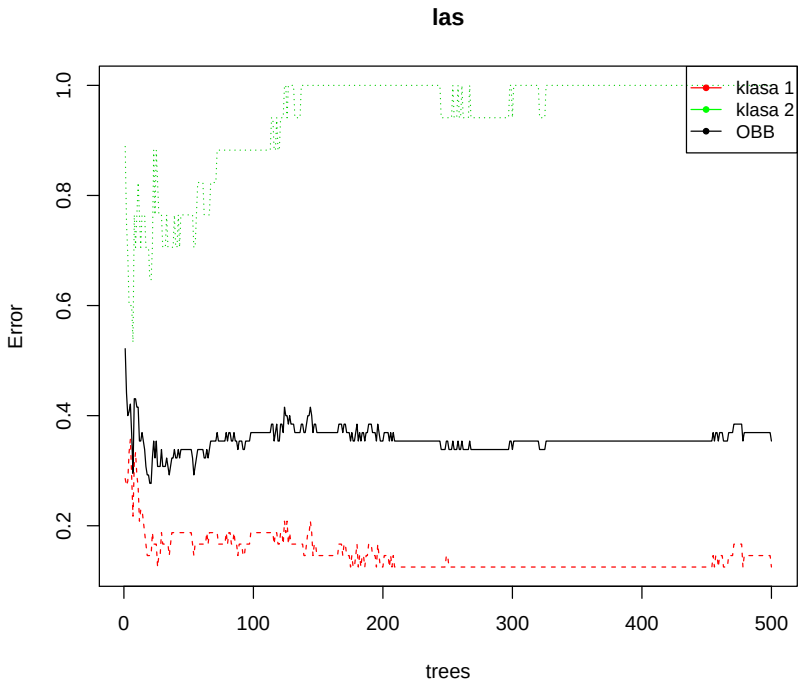
Można skorzystać z procedury, która pozwala nam obserwować wpływ liczby poszczególnych modeli na wielkość błędu klasyfikacji (rysunek 2.4).

```
> plot(las)
> leg <- c("klasa 1","klasa 2","OBB" )
> legend("topright",legend=leg, col=c("red","green","black"), lty=1,
         pch=20, cex=0.9)
```

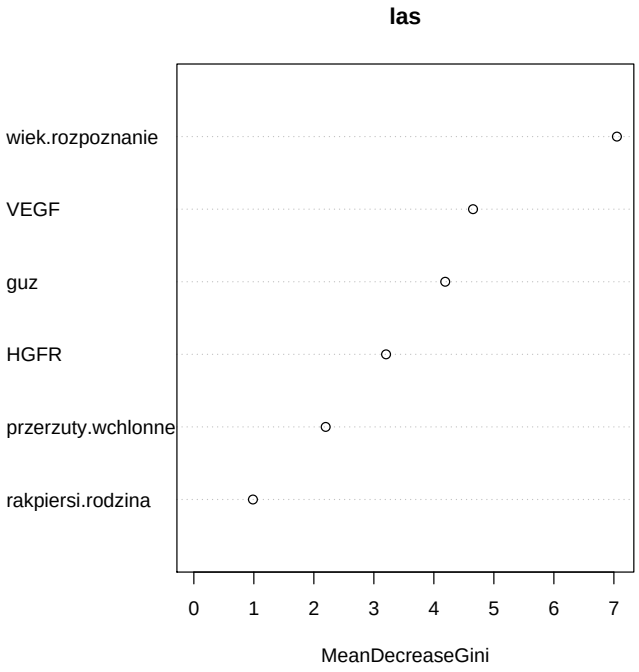
Dodatkową zaletą jest możliwość użycia lasu losowego do innych zagadnień niż tylko do klasyfikacji. Przykładowo, na podstawie drzew z lasu można wyznaczyć ranking zmiennych tzn. wykonać ocenę istotności zmiennych poprzez obliczenie średniej zmiany indeksu Giniego dla każdej zmiennej, a tym samym okreslić, które zmienne mają lepsze właściwosci predykcyjne. Proces ten polega na obliczeniu różnicy między różnorodnością klas w węźle-rodzicu i węzłach dzieciach (wzór (1.10)) dla pewnej zmiennej, dla każdego drzewa z lasu. Następnie wszystkie te wartości są sumowane. Wyznaczając taką średnią zmianę indeksu dla każdej ze zmiennych otrzymamy ranking zmiennych według ich własności predykcyjnych. Taką informacje uzyskuję poprzez funkcję `importance` z pakietu *randomForest*. Można tę informację również przedstawić graficznie (przykład jest przedstawiony na rysunku 2.5) za pomocą funkcji `varImpPlot`. Informacje o funkcjach `importance` oraz `varImpPlot` uzyskałam z [1].

```
> importance(las)
                    MeanDecreaseGini
wiek.rozpoznanie           7.0504493
guz                        4.1890669
przerzuty.wchlonne         2.1970843
rakpiersi.rodzina          0.9845307
VEGF                       4.6525242
HGFR                       3.2034030
```

```
> varImpPlot(las)
```



Rysunek 2.4: Wykres błędu klasyfikacji w zależności od liczby drzew.



Rysunek 2.5: Wykres istotności zmiennych wykonany za pomocą funkcji varImpPlot.

2.2.3. Predykcja: drzewa decyzyjne

Uzyskane modele klasyfikacyjne oraz regresyjne można zastosować do predykcji danych tzn. określania wartości zmiennej zależnej dla obserwacji należących do zbioru rozpoznawalnego.

```
predict(tree, newdata, type=c("vector","prob","class","matrix"), na.action=na.pass)
```

`tree`	obiekt w postaci drzewa
`newdata`	zbiór rozpoznawany tzn. zbiór, który będzie podlegać klasyfikacji
`type`	forma w jakiej mają być podane wyniki, do wyboru mamy: `"vector"`– do każdej z obserwacji przyporządkowany jest numer klasy `"prob"`– prawdopodobieństwo a posteriori dla każdej z klas `"class"`– nazwa klasy w której znalazła się dana obserwacja `"matrix"`– wynikiem jest macierz w której kolejnych kolemnach uzyskujemy numer klasy, liczebność klasy oraz prawdopodobieństwo a posteriori
`na.action`	wyrażenie mające na celu wskazywanie sposobu postępowanie w przypadku braków wartości zmiennych w zbiorze `newdata`

Tabela 2.5: W kolumnie pierwszej wymienione są skróty nazw zmiennych, w drugiej opisy zmiennych.

Przykłady

Spróbuję przewidzieć teraz wynik testu dla próby "testowa". Obecnie na podstawie 65 danych mam zbudowane drzewo. Według wyników tego drzewa chcę przyporządkować obserwacje, w tym przypadku stwierdzić jakie jest prawdopodobieństwo, że dana osoba będzie miała nawrót raka w ciągu 5 lat. Korzystam z funkcji (domyślnie ustawiony parametr type="prob").

```
> predict(drzewo1,testowa)
```

Poniżej przedstawiony jest fragment wyniku.

```
         nie        tak
81 0.9523810 0.04761905
6  0.4285714 0.57142857
74 0.1666667 0.83333333
83 0.9523810 0.04761905
9  0.8888889 0.11111111
48 0.1666667 0.83333333
46 0.4285714 0.57142857
37 1.0000000 0.00000000
35 0.8888889 0.11111111
58 0.4285714 0.57142857
```

Za pomocą funkcji

```
> predict(drzewo1,testowa,type="vector")
```

otrzymuję wyniki postaci:

```
81  6 74 83  9 48 46 37 35 58 51 22 28 64 72 12  4 19 15 76
 1  2  2  1  1  2  2  1  1  2  1  2  1  1  1  1  1  2  1  1  .
```

Gdzie w górnym wierszu występuje numer obserwacji, natomiast w dolnym numer klasy, do której dana obserwacja została przydzielona. Korzystając z parametru `class` zamiast `vector` w dolnym wierszu otrzymałabym nazwę klasy – w tym przypadku "tak" lub "nie".

Przydatna jest także umiejętność przedstawienia wyników w postaci tabelki oraz zliczenie błędów predykcji.

```
> tab <- table(predict(drzewo1,testowa,type="class"),testowa$nawrot.5lat)
> tab
```

Wynikiem jest tabelka. Wiersz u góry oznacza, w której klasie była obserwacja, natomiast kolumna po lewej oznacza, do której klasy przydzielił ją klasyfikator.

```
     nie tak
nie   11   2
tak    6   1
```

Można zliczyć ilość błedów predykcji za pomocą

```
> sum(tab)-sum(diag(tab))
```

W wyniku otrzymuję 8. Oznacza to, że na 20 elementy, które należało przyporządkować, źle przyporządkowanych zostało 8.

2.2.4. Predykcja: lasy losowe

```
predict(model, newdata, type="response")
```

`model`	obiekt klasy `randomForest`
`newdata`	zbiór rozpoznawany tzn. zbiór, który będzie podlega klasyfikacji
`type`	forma w jakiej mają być podane wyniki, do wyboru mamy: `"response"`– do każdej z obserwacji przyporządkowana jest nazwa klasy `"prob"`– prawdopodobieństwo a posteriori dla każdej z klas `"vote"`– liczba głosów dla każdej klasy

Tabela 2.6: W kolumnie pierwszej wymienione są skróty nazw zmiennych, w drugiej opisy zmiennych.

Przykłady

Wywoływanie funkcji jest analogiczne jak w przypadku drzew (domyślnie ustawiony paramentr type="response").

```
> predict(las,testowa)
```

Poniżej przedstawiam fragment wyniku oraz tabelkę, którą wykorzystam do zliczenia błędów predykcji.

```
 81   6  74  83   9  48  46  37  35  58  51  22  28  64  72  12   4  19  15  76
nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie nie
```

```
> tab <- table(predict(las,testowa),testowa$nawrot.5lat)
> tab

      nie tak
  nie  17   3
  tak   0   0

> sum(tab)-sum(diag(tab))
```

W wyniku otrzymuję 3. W porównaniu do przykładu z poprzedniego podrozdziału las losowy poprawił jakość predykcji.

Rozdział 3

Analiza danych

3.1. Wprowadzenie

Rozdział 3 poświęcony jest analizie danych rzeczywistych dotyczących 85 pacjentek z nowotworem piersi. Rak gruczołu sutkowego powszechnie znany jako rak piersi, jest najczęstszym nowotworem złośliwym występującym u kobiet. Wiedza o przyczynach raka piersi oraz świadome postępowanie może znacznie zwiększyć szansę uniknięcia choroby i ewentualnego jej pokonania, jeżeli się pojawi. Diagnoza raka piersi nie oznacza konieczności całkowitej amputacji piersi. Jeżeli tylko jest to możliwe, stosuje się mniej radykalny zabieg operacyjny i inne, uzupełniające sposoby leczenia takie jak chemioterapia, hormonoterapia oraz radioterapia. Za pomocą drzew klasyfikacyjnych chciałabym sprawdzić jakie cechy mogą mieć wpływ na zwiększone ryzyko pojawienia się nawrotu lub przerzutu w przeciągu 5 lat od rozpoznania choroby.

3.2. Opis danych

Dane dotyczące raka piersi zawierają znaczną ilość zmiennych. Po przeanalizowaniu kontekstu medycznego ograniczyłam się do 23 zmiennych, które zostały przedstawione w tabeli 3.1 oraz 3.2.

skrót	pełna nazwa
wiek.rozpoznanie	wiek w momencie rozpoznania
guz	wielkość guza w cm
przerzuty.wchlonne	przerzuty w węzłach chłonnych
rakpiersi.rodzina	rak piersi w wywiadzie rodzinnym
rakinne.rodzina	inne nowotwory w wywiadzie rodzinnym
akt.horm	długość okresu aktywności hormonalnej
menopauza	menopauza
porody	liczba porodów
poronienia	liczba poronień
bmi	wartość BMI
chemioterapia	zastosowano chemioterapię
hormonoterapia	zastosowano hormonoterapię
radioterapia	zastosowano radioterapię
nawrot.5lat	nawrót nastąpił w ciągu 5 lat

Tabela 3.1: Opis zmiennych modelu.

Kolejne kolumny przedstawiają skróty nazwy, którymi posługuję się w pakiecie R, pełną nazwę oraz krótki opis, jeżeli wprowadzenie go jest konieczne dla dalszego rozumienia tematu. Cechy `porody` , `poronienia` oraz `bmi` zawierają niewielkie braki danych.

skrót	pełna nazwa	opis
VEGF	VEGF	naczyniowo-śródbłonkowy czynnik wzrostu
HGFR	HGFR	receptor czynnika wzrostu hepatocytów
grading	grading	stopień złośliwości histologicznej nowotworu
rozprezajacy	rozprężajacy typ wzrostu	guz nowotworowy rozpycha się w otoczeniu, w którym wzrasta wywierając na nie ucisk, doprowadzając do zaniku otaczających tkanek
naciekajacy	naciekający typ wzrostu	wzrost guza niszczący prawidłowe tkanki, związany z wnikaniem komórek nowotworowych do naczyń limfatycznych, krwionośnych oraz przestrzeni około pni nerwowych
cialka.apoptyczne	ciałka apoptotyczne	pęcherzyki powstałe w wyniku zmian w strukturze błony komórkowej
komorki.jasne	obecność komórek jasnych	jeden z typów komórek
MRP2	MRP2 IRS	rodzaj białka
ER	ER IRS	typ receptora
PR	PR IRS	typ receptora
HER	HER IRS	typ receptora
CK	CK 5/6 IRS	enzym
podscielisko	podścielisko	błona podśluzowa
wimentyna	wimentyna I - podścielisko	rodzaj białka

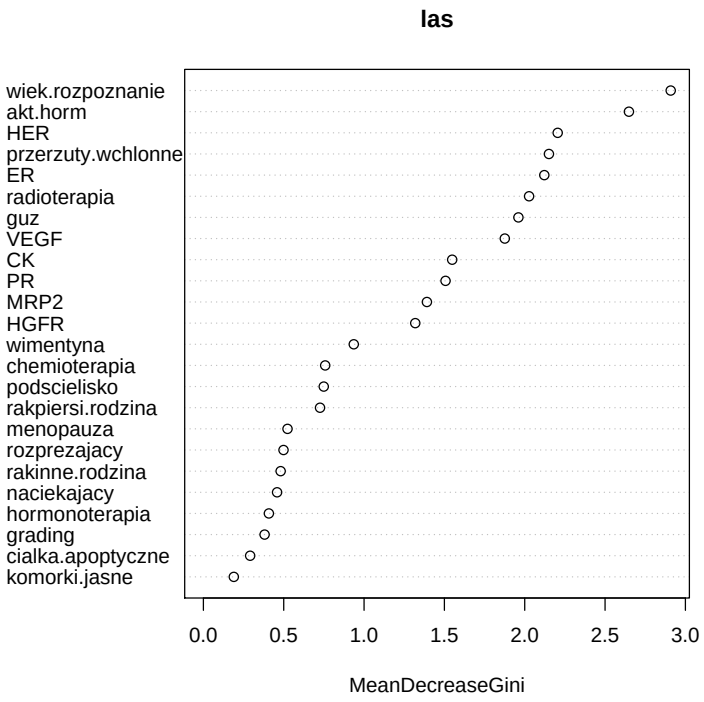
Tabela 3.2: Opis zmiennych modelu.

3.3. Badanie istotności zmiennych

Na wstępie sprawdzam istotność atrybutów według algorytmu lasów losowych. Użyję funkcji `importance()` oraz `varImpPlot()` z pakietu *randomForest*. Następnie wybrane przeze mnie cechy posłużą mi do analizy danych. Funkcja `randomForest` nie akceptuje danych z brakami, w tym przypadku nie uwzględniłam liczby porodów i liczby poronień.

```
> las <- randomForest(nawrot.5lat~VEGF+HGFR+wiek.rozpoznanie+guz+
przerzuty.wchlonne+rakpiersi.rodzina+rakinne.rodzina+menopauza+
akt.horm+radioterapia+chemioterapia+hormonoterapia+grading+
rozprezajacy+naciekajacy+cialka.apoptyczne+komorki.jasne+MRP2+
grading+ER+PR+HER+CK+podscielisko+wimentyna, data=dane)

importance(las)
                    MeanDecreaseGini
VEGF                       1.8760010
HGFR                       1.3186978
wiek.rozpoznanie           2.9085252
guz                        1.9606247
przerzuty.wchlonne         2.1505696
```

```
rakpiersi.rodzina          0.7260320
rakinne.rodzina            0.4808059
menopauza                  0.5235987
akt.horm                   2.6484525
radioterapia               2.0270352
chemioterapia              0.7581435
hormonoterapia             0.4075479
grading                    0.3809706
rozprezajacy               0.4986220
naciekajacy                0.4586434
cialka.apoptyczne          0.2910322
komorki.jasne              0.1894015
MRP2                       1.3910846
ER                         2.1217886
PR                         1.5071128
HER                        2.2051667
CK                         1.5484864
podscielisko               0.7491864
wimentyna                  0.9359392

>varImpPlot(las)
```



Rysunek 3.1: Wykres istotności zmiennych wykonany za pomocą funkcji `varImpPlot`.

Po dokonaniu oceny istotności zmiennych poprzez obliczenie średniej zmiany indeksu Giniego dla każdej zmiennej, wybieram spośród nich nowe zmienne do dalszej analizy, które przedstawia tabela 3.3.

skrót	informacje o zmiennej	skrót	informacje o zmiennej
wiek.rozpoznanie	Minimum 29.00 1.Kwantyl 47.00 Mediana 55.00 Średnia 55.61 3.Kwantyl 64.00 Maksimum 86.00	HER	Minimum 0.000 1.Kwantyl 0.000 Mediana 2.000 Średnia 4.165 3.Kwantyl 8.000 Maksimum 12.000
guz	Minimum 0.000 1.Kwantyl 2.500 Mediana 3.000 Średnia 3.234 3.Kwantyl 4.000 Maksimum 7.500	VEGF	Minimum 0.0 1.Kwantyl 0.0 Mediana 2.0 Średnia 2.4 3.Kwantyl 4.0 Maksimum 9.0
przerzuty.wchlonne	0 – 49 osób 1 – 36 osób	radioterapia	0 – 48 osób 1 – 37 osób
ER	Minimum 0.000 1.Kwantyl 0.000 Mediana 3.000 Średnia 3.365 3.Kwantyl 4.000 Maksimum 12.000	PR	Minimum 0.000 1.Kwantyl 0.000 Mediana 4.000 Średnia 4.353 3.Kwantyl 8.000 Maksimum 12.000
CK	Minimum 0.000 1.Kwantyl 0.000 Mediana 0.000 Średnia 0.918 3.Kwantyl 2.000 Maksimum 9.000	MRP2	Minimum 1.000 1.Kwantyl 4.000 Mediana 6.000 Średnia 6.518 3.Kwantyl 8.000 Maksimum 12.000
HGFR	Minimum 0.000 1.Kwantyl 0.000 Mediana 2.000 Średnia 1.847 3.Kwantyl 4.000 Maksimum 4.000	akt.horm	Minimum 13.00 1.Kwantyl 30.00 Mediana 33.00 Średnia 33.15 3.Kwantyl 38.00 Maksimum 45.00
porody	Minimum 0.000 1.Kwantyl 1.000 Mediana 2.000 Średnia 1.917 3.Kwantyl 2.000 Maksimum 5.000 brak danych 1	poronienia	Minimum 13.00 1.Kwantyl 0.0000 Mediana 0.0000 Średnia 0.5952 3.Kwantyl 1.0000 Maksimum 4.0000 brak danych 1

Tabela 3.3: Szczegółowy opis wybranych zmiennych.

3.4. Analiza danych

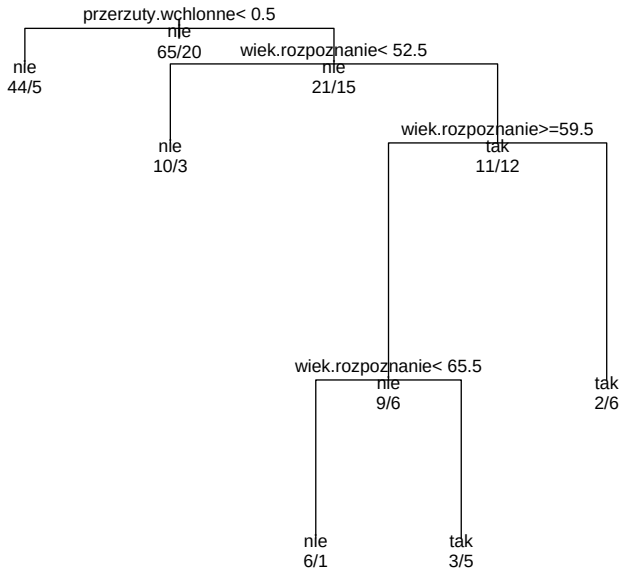
Punkt 1 W pierwszym paragrafie skupię się przede wszystkim na zmiennych takich jak przerzuty w węzłach chłonnych, wielkość guza oraz wiek rozpoznania nowotworu w zależności od zmiennej jakościowej – nawrót w przeciągu 5 lat.

```
> drzewo11 <- rpart(nawrot.5lat~przerzuty.wchlonne+wiek.rozpoznanie+guz,
                    data=dane, minsplit=5,minbucket = 5)
> plot(drzewo11)
> text(drzewo11,use.n=TRUE,all=TRUE,cex=0.8)
> drzewo11
n= 85

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 85 20 nie (0.7647059 0.2352941)
   2) przerzuty.wchlonne< 0.5 49  5 nie (0.8979592 0.1020408) *
   3) przerzuty.wchlonne>=0.5 36 15 nie (0.5833333 0.4166667)
     6) wiek.rozpoznanie< 52.5 13  3 nie (0.7692308 0.2307692) *
     7) wiek.rozpoznanie>=52.5 23 11 tak (0.4782609 0.5217391)
      14) wiek.rozpoznanie>=59.5 15  6 nie (0.6000000 0.4000000)
        28) wiek.rozpoznanie< 65.5 7  1 nie (0.8571429 0.1428571) *
        29) wiek.rozpoznanie>=65.5 8  3 tak (0.3750000 0.6250000) *
      15) wiek.rozpoznanie< 59.5 8  2 tak (0.2500000 0.7500000) *

> drzewo12a <- ctree(nawrot.5lat~guz+przerzuty.wchlonne, data=dane,
                     controls=ctree_control(minsplit=2, minbucket = 1,
                     maxdepth = 10,mincriterion = 0.50))
> drzewo12b <- ctree(nawrot.5lat~guz+przerzuty.wchlonne, data=dane,
                     controls=ctree_control(minsplit=2, minbucket = 1,
                     maxdepth = 10, mincriterion = 0.90))
> plot(drzewo12a)
> plot(drzewo12b)
```

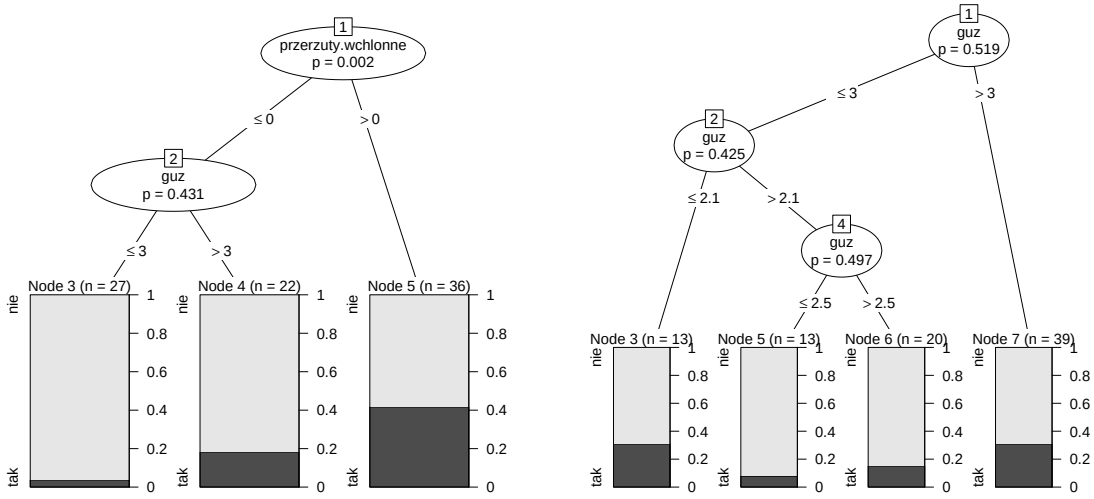
Wnioski

Skupię się na rysunku 3.2 oraz jego opisie, W liściu 2 znalazło się 49 obserwacji, które zostały przyporządkowane do klasy "nie". Wśród nich jest 5 obserwacji źle zaklasyfikowanych. Prawdopodobieństwo znalezienia się w liściu 2 i nie wystąpienia przerzutu w ciągu 5 lat wynosi 0.8979592. Prawdopodobieństwo to w przypadku wystąpienia przerzutów w węzłach chłonnych jest niższe i wynosi 0.5833333. Korzystając z testu proporcji (`prop.test()`) dla parametrów z liścia 2 oraz 3 uzyskałam p-value równe 0.001807. Można więc stwierdzić, że przerzuty w węzłach chłonnych są cechą, która ma znaczny wpływ na wystąpienie przerzutów w ciągu 5 lat od rozpoznania.

Na podstawie węzłów numer 6, 7, 14, 15 można dostrzec zwiększenie zagrożenia dla osób pomiędzy 52.5 a 59.5 wiekiem rozpoznania choroby. Biorąc pod uwagę węzły 6,7 widać jak w miarę wzrostu wieku rozpoznania choroby zwiększa się prawdopodobieństwo a posteriori zaklasyfikowania do klasy "tak" (kolejno 0.2307692,0.5217391 w przypadku przerzutów w węzłach chłonnych). Jednakże, (węzły 14,15) dla wieku $\leqslant 59.5$ prawdopodobieństwo nawrotu zmniejszyło się do 0.4000000. Na podstawie węzłów 28 oraz 29 widać także wzrost prawdopodobieństwa nawrotu dla osób u których rozpoznano raka w późnym wieku (wiek $\leqslant 65.5$). Po przeprowadzeniu testów proporcji dla węzłów 6, 7 otrzymałam p-value 0.2682, dla węzłów 6, 14 p-value = 0.5819, dla węzłów 6, 29 p-value = 1, dla węzłów 14, 15 p-value = 0.795. Reasumując, nie ma zależności pomiędzy wiekiem rozpoznania choroby a nawrotem w ciągu 5 lat, mimo że można zaobserwować na podstawie prawdopodobieństw a posteriori wyższe ryzyko dla wieku $\in [52.5,59.5)$.



Rysunek 3.2: Wykres `drzewo11` zależności nawrotu choroby od przerzutów w węzłach chłonnych oraz wieku rozpoznania choroby wykonany za pomocą `rpart()`



(a) Wykres `drzewo12a` zależności nawrotu choroby od przerzutów w węzłach chłonnych oraz wielkości guza wykonany za pomocą `ctree()`.

(b) Wykres `drzewo12b` zależności nawrotu choroby od wielkości guza.

Rysunek 3.3: Wykresy wykonane za pomocą `ctree()`.

Dla mojego zbioru danych zmienna dotycząca wielkości guza nie daje wiarygodnych informacji na temat nawrotu. Na podstawie rysunku 3.3a, przy dużej p-wartości (na poziomie 0.431) widać nieznaczny wzrost ryzyka nawrotu nowotworu wraz ze wzrostem wielkości guza. Jednakże wyniki z rysunku 3.3b nie dają jednoznacznych informacji. Trudno jest więc stwierdzić, czy dla mniejszego guza szansa na nawrót rzeczywiście się zmniejsza.

Punkt 2 W tym paragrafie skupię się na parametrach takich jak ER, PR, HER, CK, VEGF, HGFR oraz MRP2.

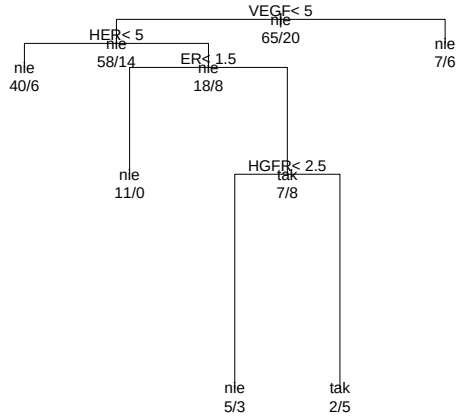
```
> drzewo21a <- rpart(nawrot.5lat~VEGF+HGFR+HER+PR+CK+ER+MRP2, data=dane,
                     minsplit=5,minbucket = 7)
> plot(drzewo21a)
> text(drzewo21a,use.n=TRUE,all=TRUE,cex=1)
> drzewo21a
n= 85

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 85 20 nie (0.7647059 0.2352941)
   2) VEGF< 5 72 14 nie (0.8055556 0.1944444)
     4) HER< 5 46  6 nie (0.8695652 0.1304348) *
     5) HER>=5 26  8 nie (0.6923077 0.3076923)
      10) ER< 1.5 11  0 nie (1.0000000 0.0000000) *
      11) ER>=1.5 15  7 tak (0.4666667 0.5333333)
        22) HGFR< 2.5 8  3 nie (0.6250000 0.3750000) *
        23) HGFR>=2.5 7  2 tak (0.2857143 0.7142857) *
   3) VEGF>=5 13  6 nie (0.5384615 0.4615385) *
```

```
> drzewo21b <- rpart(nawrot.5lat~przerzuty.wchlonne+VEGF+HGFR+HER+PR+CK+ER+MRP2,
                     data=dane,minsplit=5,minbucket = 7)
> plot(drzewo21b)
> text(drzewo21b,use.n=TRUE,all=TRUE,cex=1)
> drzewo21b
n= 85

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 85 20 nie (0.7647059 0.2352941)
   2) przerzuty.wchlonne< 0.5 49  5 nie (0.8979592 0.1020408) *
   3) przerzuty.wchlonne>=0.5 36 15 nie (0.5833333 0.4166667)
     6) ER>=5 15  2 nie (0.8666667 0.1333333) *
     7) ER< 5 21  8 tak (0.3809524 0.6190476)
      14) HER< 4.5 13  6 nie (0.5384615 0.4615385) *
      15) HER>=4.5 8  1 tak (0.1250000 0.8750000) *
```

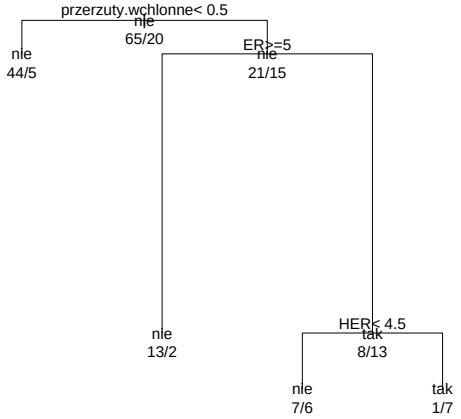
```
> drzewo22a <- ctree(nawrot.5lat~CK+PR+MRP2+przerzuty.wchlonne, data=dane,
                     controls=ctree_control(minsplit=2,minbucket = 1,
                     maxdepth = 10,mincriterion = 0.50))
> drzewo22b <- ctree(nawrot.5lat~CK+przerzuty.wchlonne, data=dane,
                     controls=ctree_control(minsplit=2,minbucket = 1,
                     maxdepth = 10,mincriterion = 0.50))
> plot(drzewo22a)
> plot(drzewo22b)
```

Wnioski

Na podstawie opisu wykresu `drzewo21a` badam kolejno zależność pomiędzy nawrotem a poziomem VEGF. Z węzłów 2 oraz 3 wynika, iż prawdopodobieństwo wystąpienia nawrotu jest większe

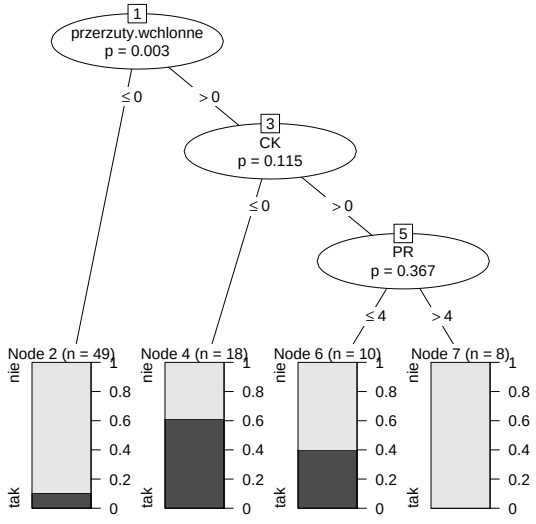


(a) Wykres `drzewo21a` zależności nawrotu choroby od parametrów VEGF, ER, HER i HGFR.

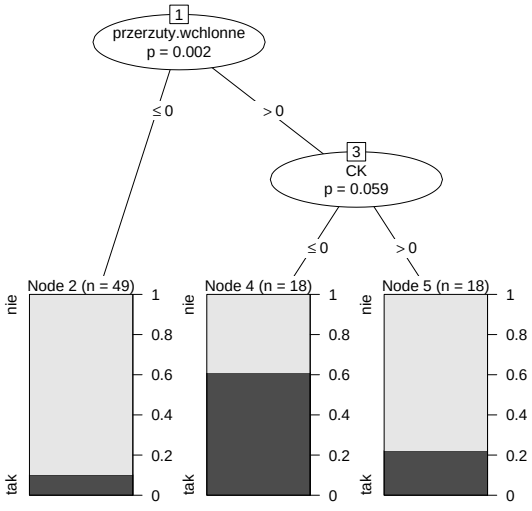


(b) Wykres `drzewo21b` zależności nawrotu choroby od przerzutów w węzłach chłonnych i zmiennych ER oraz HER.

Rysunek 3.4: Wykresy wykonane za pomocą `rpart()`.



(a) Wykres `drzewo22a` zależności nawrotu choroby od parametrów PR, CK oraz przerzutów w węzłach chłonnych.



(b) Wykres `drzewo22b` zależności nawrotu choroby od przerzutów w węzłach chłonnych i zmiennej CK.

Rysunek 3.5: Wykresy wykonane za pomocą `ctree()`.

dla osób, u których VEGF jest $\geqslant 5$ (p=0.4615385) niż u osób z VEGF < 5 (p=0.1944444). Przeprowadzając test proporcji dla tych węzłów otrzymuję p-value = 0.08287, co wzkazuję na brak zależności. Następnie na podstawie mojej próby dla HER $\geqslant 5$ oraz ER < 1.5 wynika, że jest całkowita szansa na brak nawrotu w ciągu 5 lat, która zdecydowanie zmniejsza się dla ER $\geqslant 1.5$. Wynik testu proporcji dla węzłów 4, 5 oraz 10, 11 to kolejno 0.1297, 0.0276. Biorąc pod uwagę parametr HGFR p-value dla testu proporcji dla węzłów 22, 23 wynosi 1. Można więc wnioskować wpływ poziomu ER, który przy odpowiednim zestawieniu zmiennych HER, VEGF weryfikuje, czy u osoby chorej wystąpi nawrót.
W przypadku zastawienia tych samych parametrów z przerzutami w węzłach chłonnych z rysunku 3.4b, widać pewną zależność między poziomem ER pod warunkiem, że wystąpił przerzut w węzłach chłonnych. Prawdopodobieństwo nawrotu zwiększa się z 0.1333333 do 0.6190476 dla kolejno ER $\geqslant 5$ i ER < 5. Jednakże stosując test proporcji dla węzłów 6, 7 otrzymuję p-value = 0.2084. Można stwierdzić, że zmienna ER nie ma wpływu na nawrót nowotworu. Co więcej, zmienna HER podobnie jak na rysunku 3.4b w odpowiednim złożeniu z parametrami ER i przerzuty w węzłach chłonnych, nie ma wpływu na nawrót (p-value = 0.2661).

Rysunki 3.5a oraz 3.5b wykonałam w celu zbadania parametrów CK oraz PR (które nie pojawiły się na wyżej omówionych wykresach). Dla zerowego CK szansa na brak nawrotu raka jest około trzy razy mniejsza niż dla CK > 0. Dodatkowo, przy CK > 0 zmienna PR rozdziela obserwacje na dwie grupy: PR $\leqslant 4$, gdzie prawdopodobieństwo nawrotu to około 0.4 oraz PR > 4, gdzie prawdopodobieństwo to jest równe 0. Postanowiłam nie odrzucać zmiennej CK, ponieważ p-value = 0.059 dla zmiennej CK (rysunek 3.5b), które jest w przybliżeniu równe 0.05, może świadczyć o istotności zmiennej CK podczas badania występowania nawrotów u chorego (być może dla większej próby ta zmienna okazałaby się istotna).

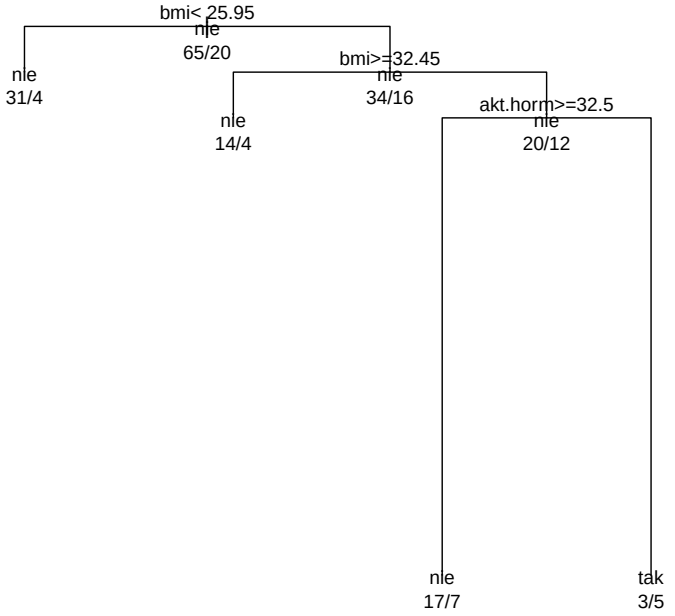
Podsumowując, spośród zmiennych ER, PR, HER, CK, VEGF, HGFR oraz MRP2 ważna wydaje się być przede wszystkim zmienna ER, wpływ zmiennej CK jest niejednoznaczny.

Punkt 3 Ten paragraf dotyczy pozostałych zmiennych tzn. ilości porodów oraz poronień, poziomu BMI, oraz długości okresu aktywności hormonalnej.

```
> drzewo31 <- rpart(nawrot.5lat~akt.horm+porody+poronienia+bmi, data=dan,
                    minsplit=5,minbucket = 7)
> plot(drzewo31)
> text(drzewo31,use.n=TRUE,all=TRUE,cex=0.75)
```

```
> drzewo31
n= 85

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 85 20 nie (0.7647059 0.2352941)
   2) bmi< 25.95 35  4 nie (0.8857143 0.1142857) *
   3) bmi>=25.95 50 16 nie (0.6800000 0.3200000)
     6) bmi>=32.45 18  4 nie (0.7777778 0.2222222) *
     7) bmi< 32.45 32 12 nie (0.6250000 0.3750000)
      14) akt.horm>=32.5 24  7 nie (0.7083333 0.2916667) *
      15) akt.horm< 32.5 8  3 tak (0.3750000 0.6250000) *
```

```
> drzewo32a <- ctree(nawrot.5lat~akt.horm+przerzuty.wchlonne, data=dane,
                     controls=ctree_control(minsplit=2,minbucket = 1,
                     maxdepth = 10,mincriterion = 0.50))
> drzewo32b <- ctree(nawrot.5lat~przerzuty.wchlonne+poronienia, data=dane,
                     controls=ctree_control(minsplit=2,minbucket = 1,
                     maxdepth = 10,mincriterion = 0.50))
> plot(drzewo32a)
> plot(drzewo32b)
```

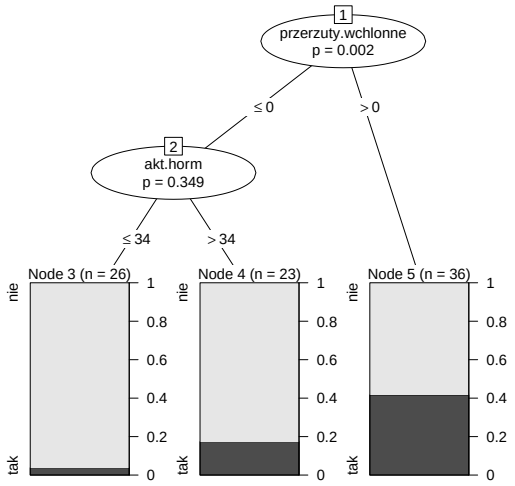


Rysunek 3.6: Wykres `drzewo31` zależności nawrotu choroby od poziomu bmi oraz długości aktywności hormonalnej wykonany za pomocą `rpart()`.

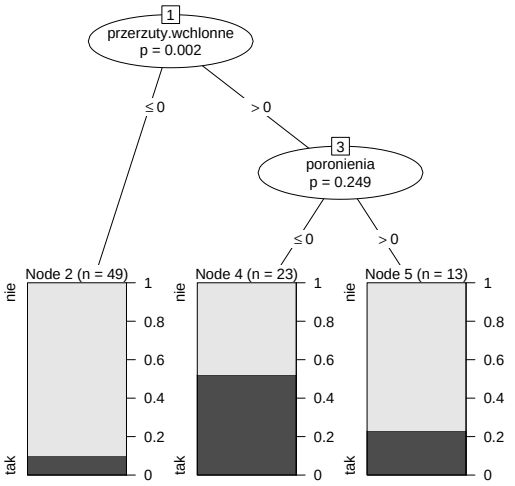
Wnioski

Opierając się na opisie `drzewo31` oraz węzłach 2, 7 przeprowadzam test proporcji. Wynikiem jest p-value = 0.02688. Można zatem stwierdzić, że zmienna `bmi` ma wpływ na prawdopodobieństwo nawrotu choroby w przeciągu 5 lat. Ważne jest czy pacjent znajdzie się w grupie gdzie BMI < 25.95 czy $[25.96, 32.45)$ (nadwaga, I stopień otyłości) (prawdopodobieństwo to zwiększa się). Zmienna długość okresu aktywności hormonalnej nie wykazuje istotności. P-value dla testu proporcji dla węzłów 14, 15 wynosi 1 (co wydaje się być sprzeczne z poszechną wiedzą na temat zachorowalności na raka piersi).

W oparciu o rysunek 3.7a, pod warunkiem braku przerzutów w węzłach chłonnych, zmienna `akt.horm` wykazuje niewielki wpływ na nawrót oraz w tym przypadku zwiększenie prawdopodobieństwa dla `akt.horm` > 34. Zmienna `poronienia` na podstawie rysunku 3.7b wydaje się wpływać na `nawrot.5lat`. Oznaczałoby to, że jeżeli kobieta poroniła szansa uniknięcia nowotworu zwiększa się o ponad połowę. Jednakże p-value w tym przypadku wynosi 0,249. Nie należy



(a) Wykres `drzewo32a` zależności nawrotu choroby od parametrów aktywność hormonalna oraz przerzuty w węzłach chłonnych.



(b) Wykres `drzewo32b` zależności nawrotu choroby od zmiennych poronienia oraz przerzuty w węzłach chłonnych.

Rysunek 3.7: Wykresy wykonane za pomocą `ctree()`.

więc uznawać tej zmiennej za istotną. Zmienna porody nie wykazuje zależności.
W konkluzji, istotną zmienną wydaje się być zmienna `bmi`. W przypadku pozostałych zmiennych, takich jak `porody`, `poronienia` oraz `akt.horm` nie widać wyraźnej zależności z nawrotem nowotworu w okresie 5 lat.

Zakończenie

W mojej pracy przedstawiłam wstęp do teorii drzew decyzyjnych oraz jej wykorzystywanie za pomocą pakietu R. W pierwszym rozdziale omówione zostały zagadnienia takie jak budowa drzewa oraz algorytm lasów losowych. Drugi oraz trzeci rozdział poświęciłam zastosowaniom.

Rozdział trzeci bazuje na danych rzeczywistych dotyczących 85 kobiet chorych na raka piersi. Poświęciłam go zbadaniu wpływu różnych czynników na wystąpienie nawrotu w ciągu 5 lat od rozpoznania choroby. Po weryfikacji zmiennych oraz obliczeniu ich istotności ograniczyłam się do 12 zmiennych. Istotność badałam poprzez obliczenie średniej zmiany indeksu Giniego dla każdej zmiennej według algorytmu lasów losowych. Wybrane przeze mnie zmienne podzieliłam na podstawie kontekstu medycznego na trzy grupy. W każdej z nich tworzyłam drzewo za pomocą funkcji `rpart` uwzględniając wszystkie zmienne znajdujące się w danej grupie. Postanowiłam także tak dobrać liczbę obserwacji w węźle, który ulega podziałowi oraz minimalną liczbę obserwacji w liściu drzewa, aby uniknąć analizowania węzłów o zbyt małej ilości elementów. Dopełnieniem było użycie funkcji `ctree` uwzględniając zmienne, które nie wykazały większej istotności po użyciu funkcji `rpart` oraz często łącząc je ze zmienną `przerzuty.wchlonne`, która wykazywała dużą istotność. Kotrolowałam wówczas graniczny poziom istotności dla hipotezy zerowej niezależności pomiędzy zmienną `nawrot.5lat` a pozostałymi zmiennymi.

Przeprowadzona przeze mnie analiza za pomocą drzew decyzyjnych, dotyczącą pojawienia się nawrotu choroby w ciągu 5 lat od rozpoznania, wskazuje przede wszystkim na istotny wpływ zmiennych takich jak pojawienie się przerzutów w węzłach chłonnych, ER oraz BMI. Trudno określić wpływ zmiennej CK, która być może dla większej próby okazałaby się ważna.

Na podstawie medycznej literatury wymienię niektóre czynniki o dobrze udokumentowanym wpływie na raka piersi. Ryzyko wystąpienia raka sutka u nieródek jest około 3 razy większe niż u kobiet, które rodziły. Ryzyko to wzrasta także z wiekiem rozpoznania choroby począwszy od 30 roku życia. Po klimakterium utrzymuje się na mniej więcej stałym poziomie. Długość okresu aktywności hormonalnej wiąże się z nieznacznym wzrostem ryzyka zachorowania na raka sutka (prawdopodobnie zależność ta ma związek z przedłużonym działaniem estrogenów na tkanki organizmu). Ciekawy wydaje się więc brak wyraźnej zależności pomiędzy porodami, długością okresu aktywności hormonalnej, wiekiem rozpoznania choroby oraz nawrotem nowotworu w przeciągu 5 lat.

Bibliografia

[1] Przemysław Biecek, *Na przełaj przez Data Mining*, `http://www.biecek.pl/R/naPrzelajPrzezDM.pdf`.

[2] Jacek Koronacki, Jan Ćwik, *Statystyczne systemy uczące się*, Warszawa 2005, Wydawnictwo Naukowo-Techniczne, rozdział 4.

[3] Jan Ćwik, Jan Mielniczuk, *Statystyczne systemy uczące się – ćwiczenia w oparciu o pakiet R*, Warszawa 2009, oficyna wydawnicza Politechniki Warszawskiej.

[4] Marek Walesiak, Eugeniusz Gatnar, *Statystyczna analiza danych z wykorzystaniem programu R*, Warszawa 2009, Wydawnictwo Naukowe PWN.

[5] Terry Therneau, Elizabeth Atkinson, *An Introduction to Recursive Partitioning Using the RPART Routines*, 3 września, 1997

[6] Agnieszka Nowak-Brzezińska, *Drzewa klasyfikacyjne. Konspekt do zajęć : Statystyczne metody analizy danych*, 11 stycznia, 2010

[7] Jacek Jassem, *Rak sutka. Podręcznik dla studentów i lekarzy*, Warszawa 1998, Springer PWN.