Słów kilka o data mining (eksploracyjnej analizie danych) na przykładach w genetyce, medycynie i gospodarce

Przemyslaw.Biecek@gmail.com, MIMUW

Plan:

1. czym jest ten data mining?
2. skalowanie wielowymiarowe na przykładzie PCA i MDS,
3. analiza skupień na przykładzie dendrogramu,
4. klasyfikacja na przykładzie lasów losowych,
5. podsumowanie.

Co to ten Data Mining?

To paraphrase provocatively, 'data mining is statistics minus any checking of models and assumptions'.
Brian D. Ripley (zapytany o różnice pomiędzy DM a statystyką).

Rozwój technik komputerowych, baz danych oraz galopująca miniaturyzacja umożliwiają zbieranie coraz większych zbiorów danych. W wielu badaniach analizowane są tysiące zmiennych i setki milionów przypadków. Standardowe techniki statystyczne słabo nadają się do analizy takich zbiorów danych. Np. problem kontroli błędu pierwszego rodzaju traci na znaczeniu, co podważa stosowalność całej klasycznej teorii testowania w analizie naprawdę dużych zbiorów danych.

Terry Speed w swojej kolumnie w biuletynie IMS często podkreśla, że autorami nowych technik analizy danych coraz rzadziej są matematycy czy statystycy a coraz częściej informatycy, biolodzy, fizycy itp.
Nie rozwijają oni co prawda matematycznych, jednolitych teorii ale tworzą algorytmy rozwiązujące (z różnym skutkiem) konkretne problemy.

Co to ten Data Mining?

Nie ma (=nie znam) powszechnie przyjętej definicji dziedziny data mining.
To określenie pojawia się najczęściej w sytuacji gdy analizowane są duże zbiory danych. Roboczo przyjmijmy więc, że data mining to metody analizy dużych zbiorów danych.

Można wyróżnić trzy grupy metod, które często pojawiają się w podręcznikach czy szkoleniach poświęconych tematyce data mining.
Są to:

- Skalowanie wielowymiarowe, konstrukcja cech, wybór cech,
- Analiza skupień, podejście modelowe, kombinatoryczne, hierarchiczne i mieszane,
- Klasyfikacja, tak podejście parametryczne jak i nieparametryczne.

Skalowanie wielowymiarowe

Skalowanie wielowymiarowe jest techniką pozwalającą na transformacje, najczęściej redukcje, przestrzeni w której opisywane są obiekty. Zazwyczaj oznacza to redukcje przestrzeni cech do przestrzeni $\mathcal{R}^2$ lub $\mathcal{R}^3$.

Do tego celu może służyć wiele technik i algorytmów. Poniżej, w krótkich żołnierskich słowach, opiszemy techniki PCA i MDS.

Skalowanie wielowymiarowe ma wiele zastosowań, w tym:

- ułatwia wizualizacje, najczęściej przez redukcje danych do 2-3 wymiarów,
- etap pośredni w analizie skupień,
- ekstrakcja cech, w sytuacji gdy oryginalnych zmiennych jest zbyt dużo by móc je analizować (np. PCR).

Przykład: Psychotyczność pacjentów z projektu EDEN

Zastosowanie skalowania wielowymiarowego przedstawimy na przykładzie danych pochodzących z projektu EDEN. W ramach tego projektu przebadano pod kątem psychiatrycznym ponad 3 tysiące pacjentów z 6 europejskich ośrodków. Dla każdego pacjenta określono ponad tysiąc cech, między innymi początkową psychotyczność mierzoną w czterech obszarach oraz początkową jakość życia.

Pytanie: jak określać podobieństwo pomiędzy pacjentami, czy zbiór danych jest jednorodny ze względu na te cechy, czy w danych można wyróżnić typowe profile pacjentów, czy są wyraźne skupiska w danych?

	BPRS.Maniac	BPRS.Negative	BPRS.Positive	BPRS.Depression	MANSA
1	1.8	2.3	2.1	3.4	3.1
2	1.8	1.3	1.1	2.1	3.0
3	1.0	2.0	1.3	3.0	4.1
4	1.3	1.1	1.1	1.5	4.0
5	1.2	1.9	1.4	3.4	4.1
6	1.3	1.9	1.3	4.0	4.0

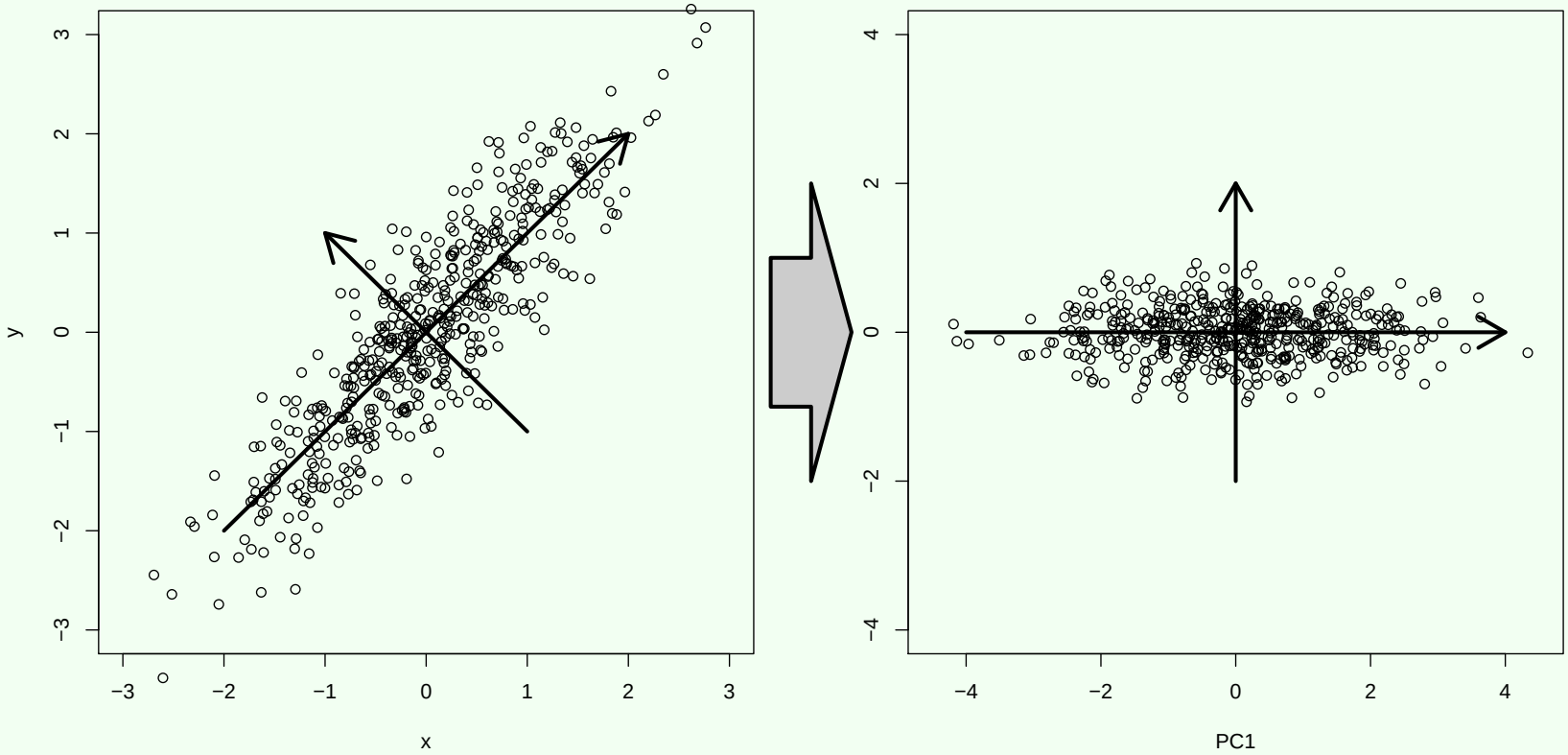
Analiza składowych głównych (PCA, ang. Principal Components Analysis)

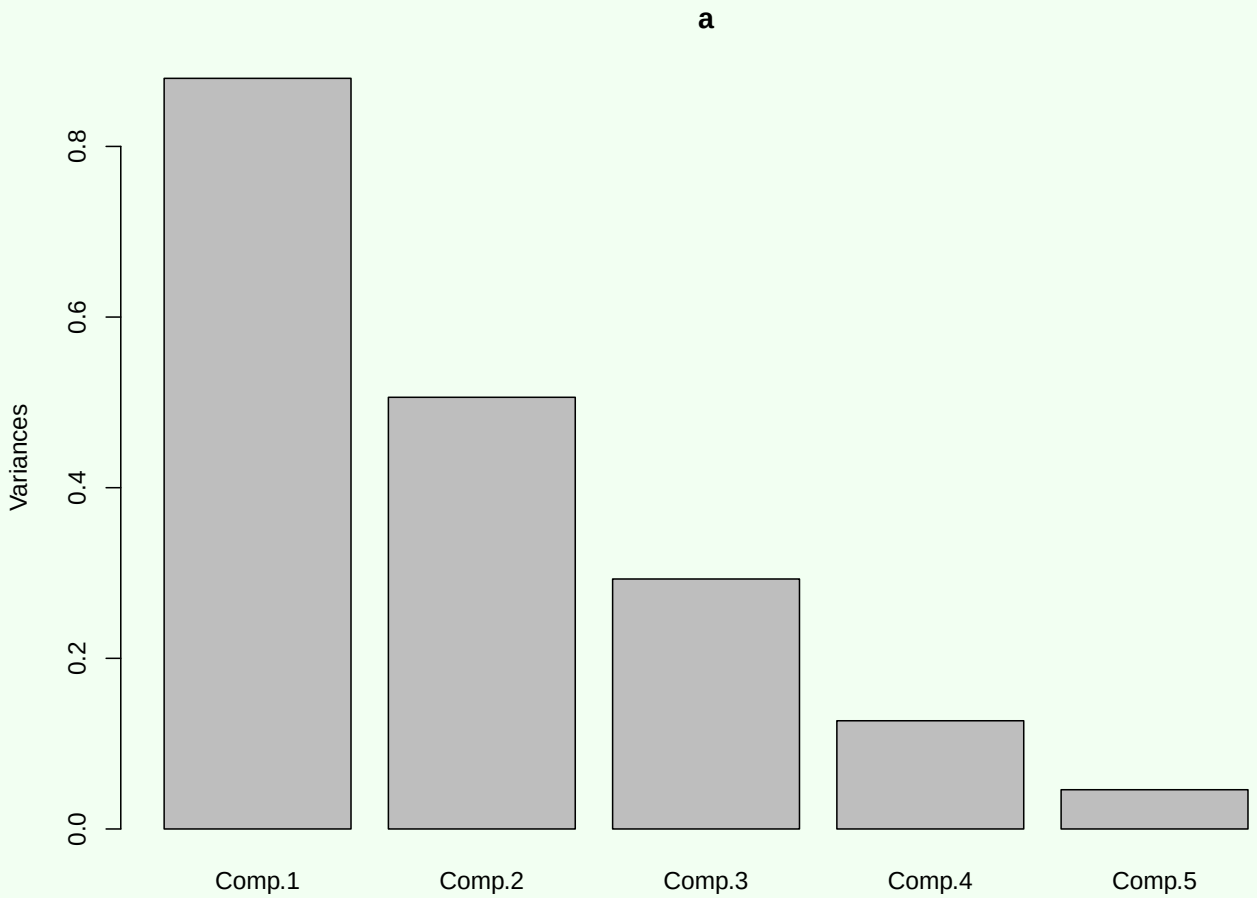
Metoda PCA pozwala na przekształcenie danych do przestrzeni k-wymiarowej. W opisanym przykładzie przekształcimy 5-wymiarową przestrzeń cech do dwuwymiarowej przestrzeni.

Wykorzystany jest następujący algorytm:

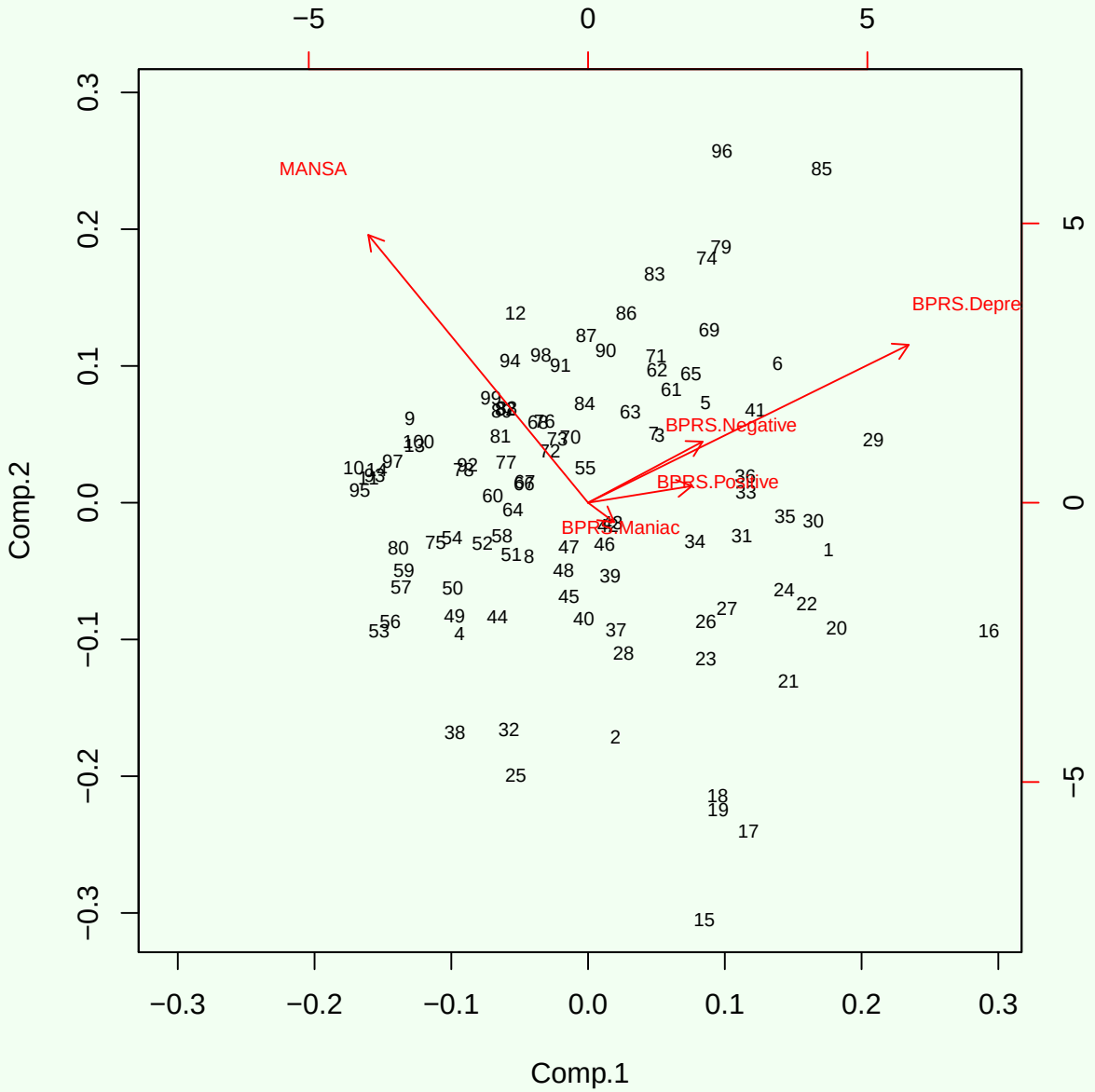
- wyznacz rozkład SVD dla analizowanej macierzy $X = UDV^T$. Macierz U opisuje bazę ortonormalną w przestrzeni kolumn macierzy X a macierz V opisuje bazę w przestrzeni wierszy,
- jako nowe współrzędne wybierz k pierwszych wektorów z macierzy V,
- wyznacz współrzędne obiektów w nowym (zredukowanym) układzie współrzędnych.

Analiza składowych głównych (PCA, ang. Principal Components Analysis)





Zmienność w danych wyjaśniona przez kolejne współrzędnej nowej bazy.


MANSA
BPRS.Depre
BPRS.Negative
BPRS.Positive
BPRS.Maniac
Comp.1
Comp.2
−5
0
5
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3

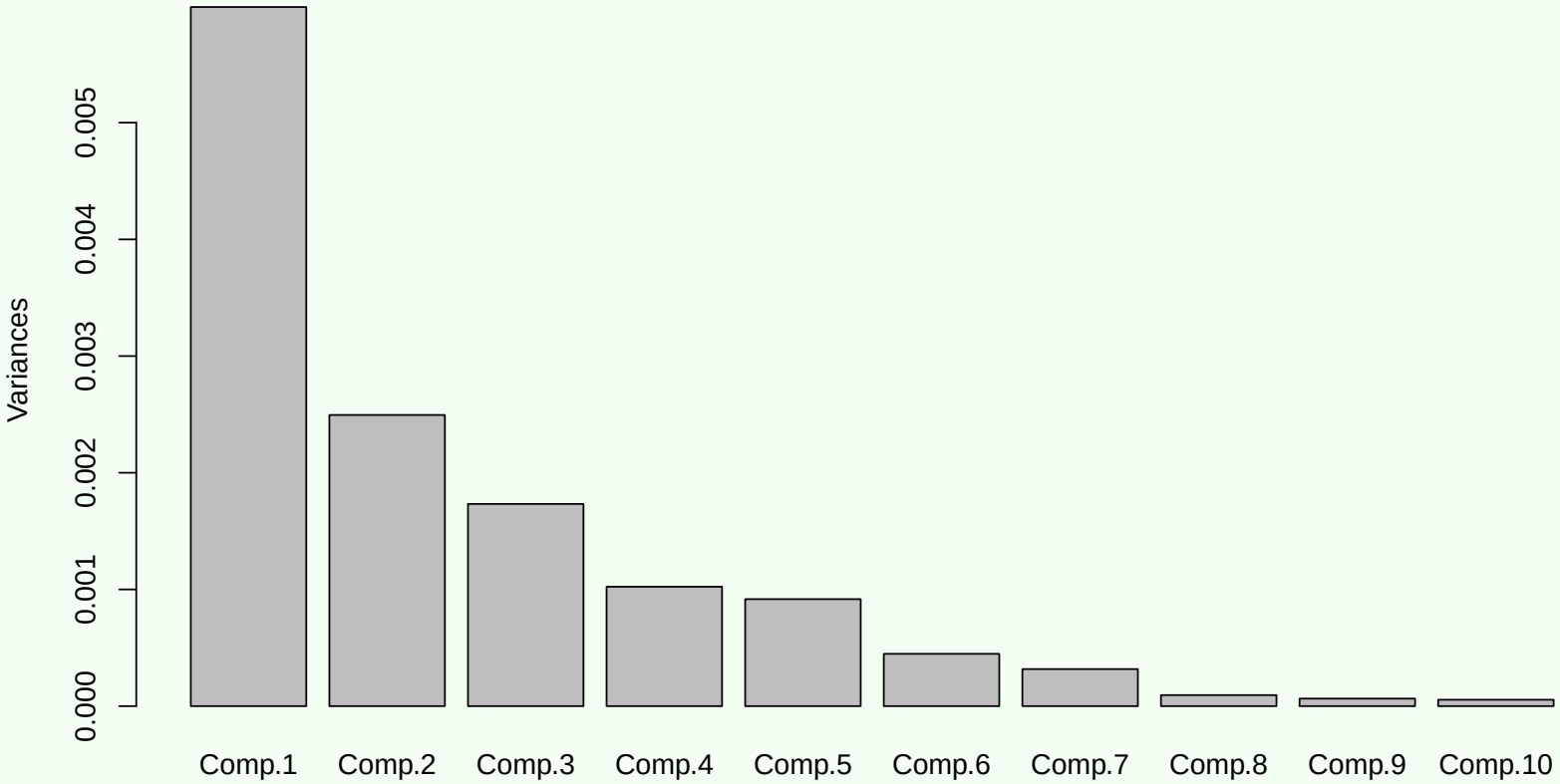
Dane GUS na temat studiowalności

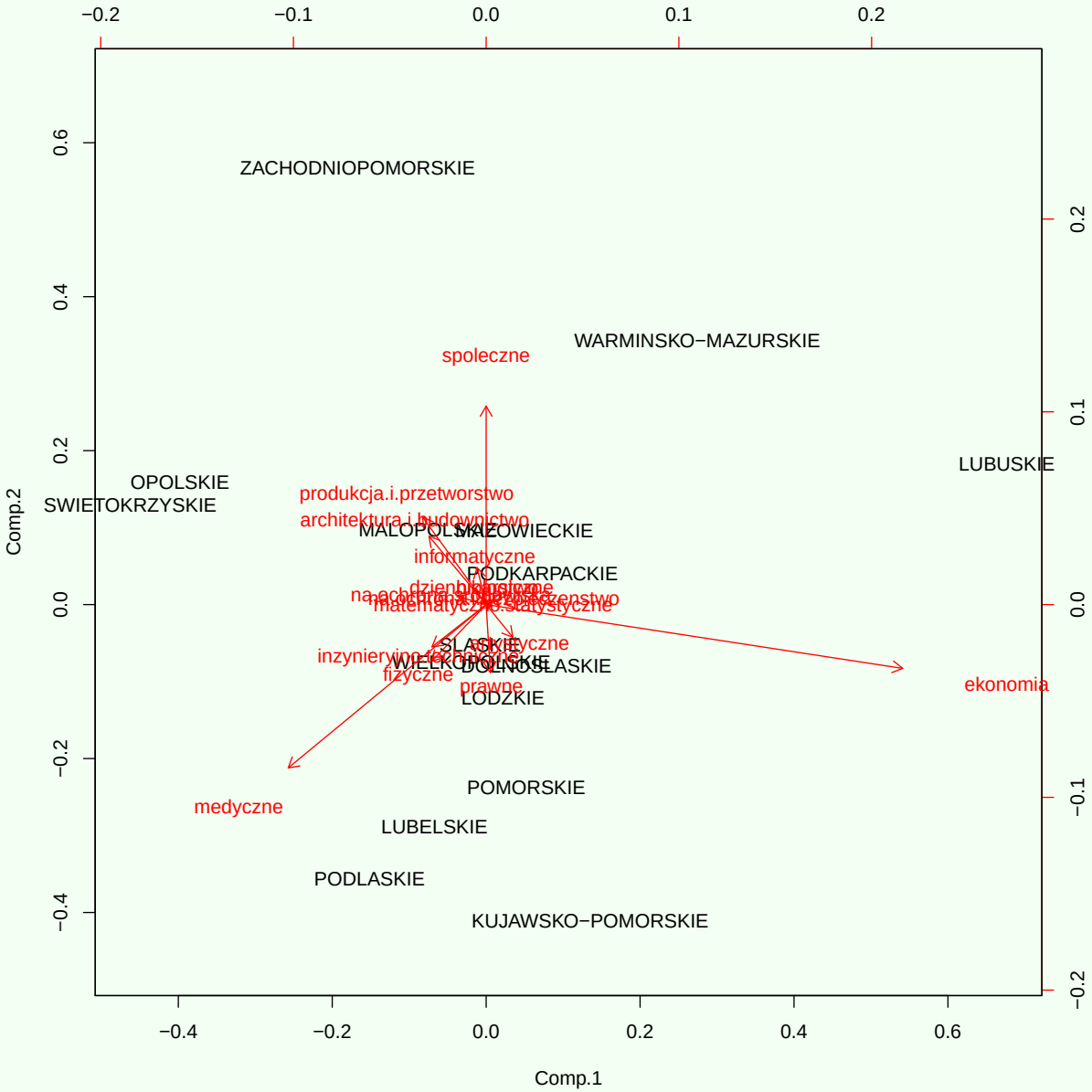
Zobaczmy teraz zbiór danych pobrany ze strony GUSu dotyczący liczby studentów studiujących w określonym województwie określony kierunek studiów.

Zbiór danych dotyczy 16 województw i 15 wyróżnionych kierunków.

Pytanie: Które województwa mają podobną strukturę kierunków studiów.

	studenci.artystyczne	studenci.spoleczne	studenci.ekonomia
DOLNOSLASKIE	2179	13313	23998
KUJAWSKO-POMORSKIE	1205	5097	10098
LODZKIE	2934	10242	13990
LUBELSKIE	341	5971	9655
LUBUSKIE	498	3474	7772
MALOPOLSKIE	2741	19528	23555
MAZOWIECKIE	3231	27381	30501


Variances
0.000
0.001
0.002
0.003
0.004
0.005
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Comp.7
Comp.8
Comp.9
Comp.10


ZACHODNIOPOMORSKIE
WARMINSKO−MAZURSKIE
spoleczne
LUBUSKIE
OPOLSKIE
SWIETOKRZYSKIE
produkcja.i.przetworstwo
architektura.i.budownictwo
MALOPOLSKIE
MAZOWIECKIE
informatyczne
PODKARPACKIE
matematyczne.statystyczne
ekonomia
fizyczne
prawne
LODZKIE
POMORSKIE
medyczne
LUBELSKIE
PODLASKIE
KUJAWSKO−POMORSKIE
Comp.1
Comp.2

Skalowanie wielowymiarowe Kruskalla (MDS, ang. Multidimensional Scaling)

Dla metody PCA skalowanie było przeprowadzane tak by w „docelowej" przestrzeni odległość pomiędzy obiektami była możliwie bliska odległości w przestrzeni oryginalnej. Odpowiada to minimalizacji tzw. stresu (kalka z ang. stress)

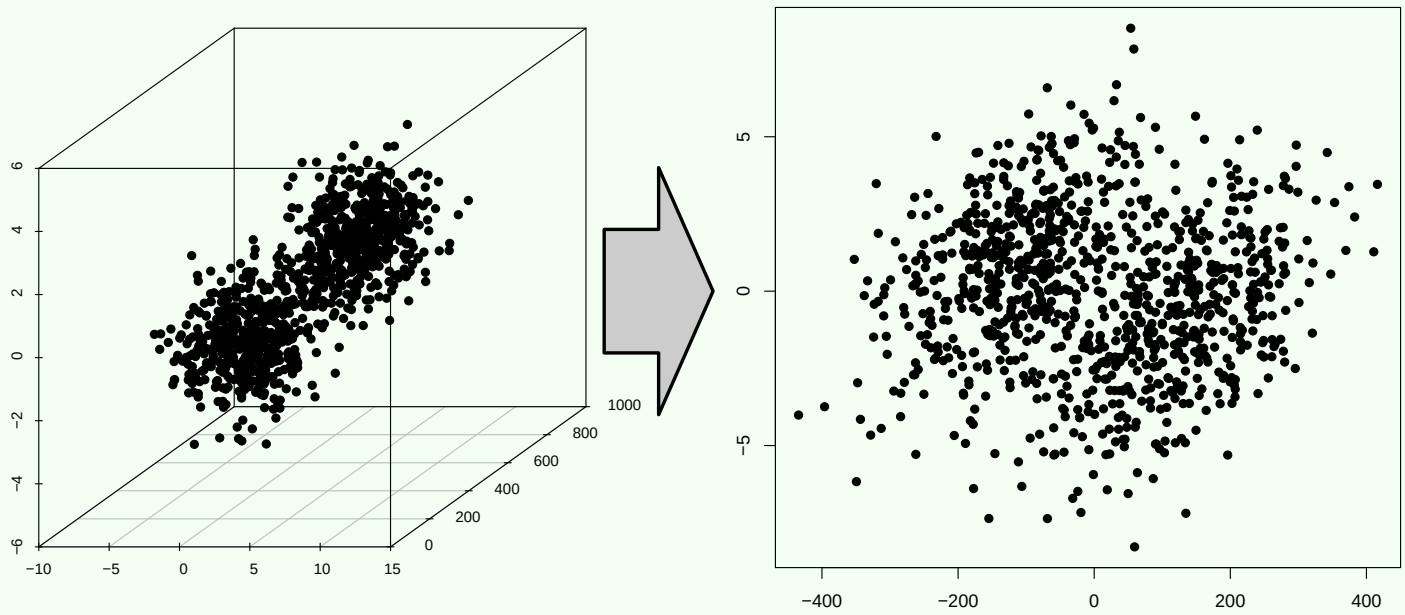
$$stress = \frac{\sum_{i,j}(d_{ij} - \tilde{d}_{ij})^2}{\sum_{i,j} d_{ij}}.$$

W przypadku metody MDS minimalizowana jest wartość

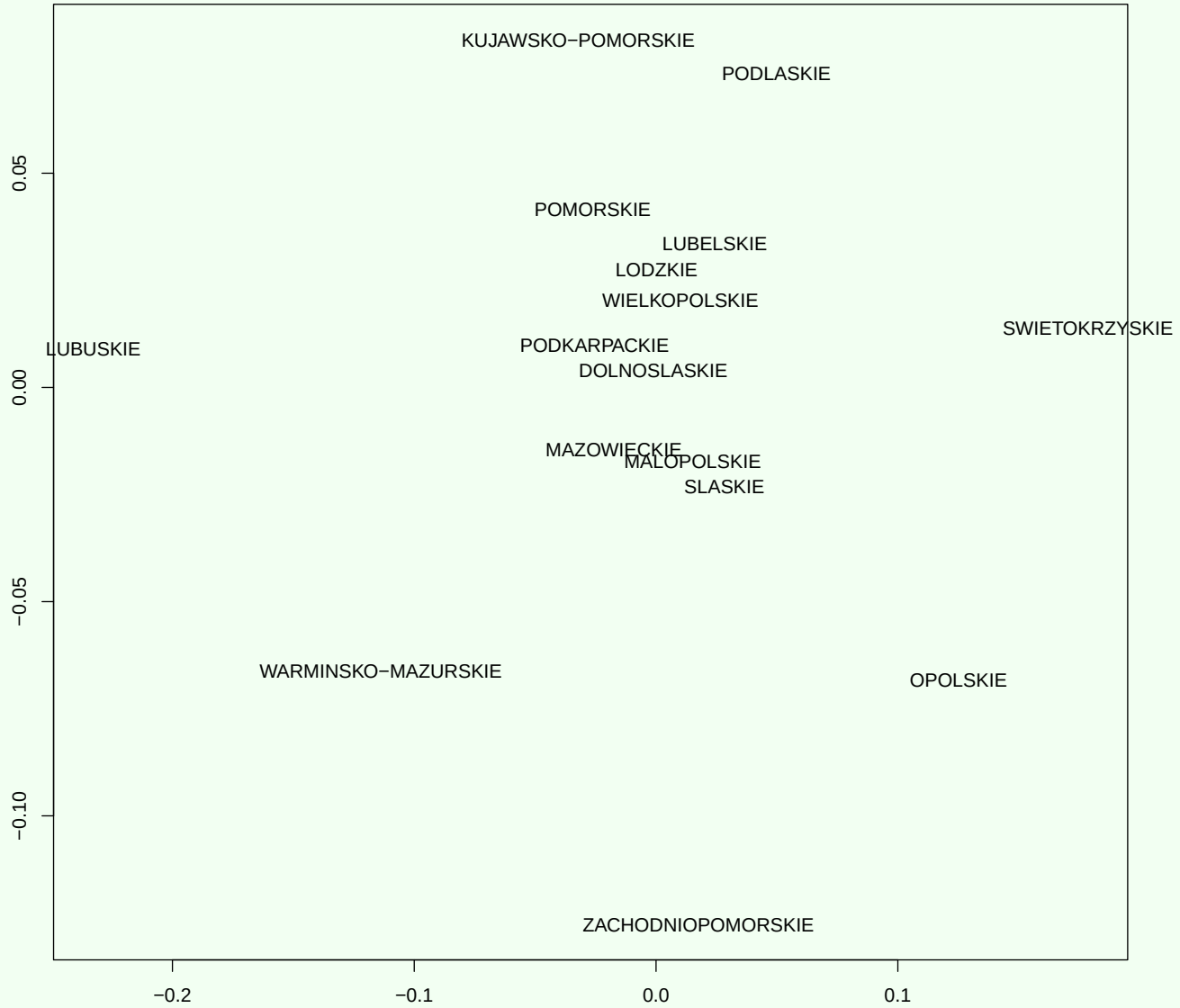
$$stress = \frac{\sum_{i,j}(f(d_{ij}) - \tilde{d}_{ij})^2}{\sum_{i,j} f(d_{ij})^2},$$

gdzie $\tilde{d}_{ij}$ to odległość pomiędzy obiektami i i j w nowej k-wymiarowej przestrzeni a d_{ij} to oryginalne odległości pomiędzy obiektami przekształcone przez pewną monotoniczną funkcję $f()$ (więc d_{ij} i $\tilde{d}_{ij}$ mogą być w różnych skalach!).

Skalowanie wielowymiarowe Kruskalla (MDS, ang. Multidimensional Scaling)



Skalowanie wielowymiarowe Kruskalla (MDS, ang. Multidimensional Scaling)



Analiza skupień

Analiza skupień pozwala na wyróżnienie zbiorów obserwacji (nazywanych skupieniami lub klastrami) podobnych do siebie. Proces szukania podziału na grupy, nazywany jest czasem klastrowaniem (kalka z ang. clustering). Jest to szybko rozwijany zbiór metod obejmujący metody kombinatoryczne, hierarchiczne, modelowe i inne.

Poniżej przedstawimy przykład analizy hierarchicznej aglomeracyjnej. Popularnie używane są również metody hierarchiczne dzielące, metoda k-średnich i metoda k-medoidów.

Analiza skupień ma wiele zastosowań, w tym:

- segmentacja klientów,
- identyfikacja wspólnie regulowanych genów,
- konstrukcja portfeli akcji lub innych instrumentów,
- filogenetyka i konstrukcja rodowodów gatunków.

Hierarchiczna analiza skupień, AGNES

AGglomerative NESting (AGNES) jest najpopularniejszą metodą aglomeracyjną.

Dendrogram opisujący skupiska tworzony jest w następujący sposób

1. Każdą obserwacje traktujemy jako osobne skupisko.
2. Znajdujemy dwa skupiska najbliższe sobie. Odległość pomiędzy skupiskami można wyznaczać na różne sposoby (tzw. szczegół implementacyjny).
3. Łączymy dwa najbliższe skupiska w jedno i przeliczamy odległości pomiędzy tym nowym skupiskiem a pozostałymi.
4. Jeżeli pozostało więcej niż jedno skupisko to wracamy do kroku 2.

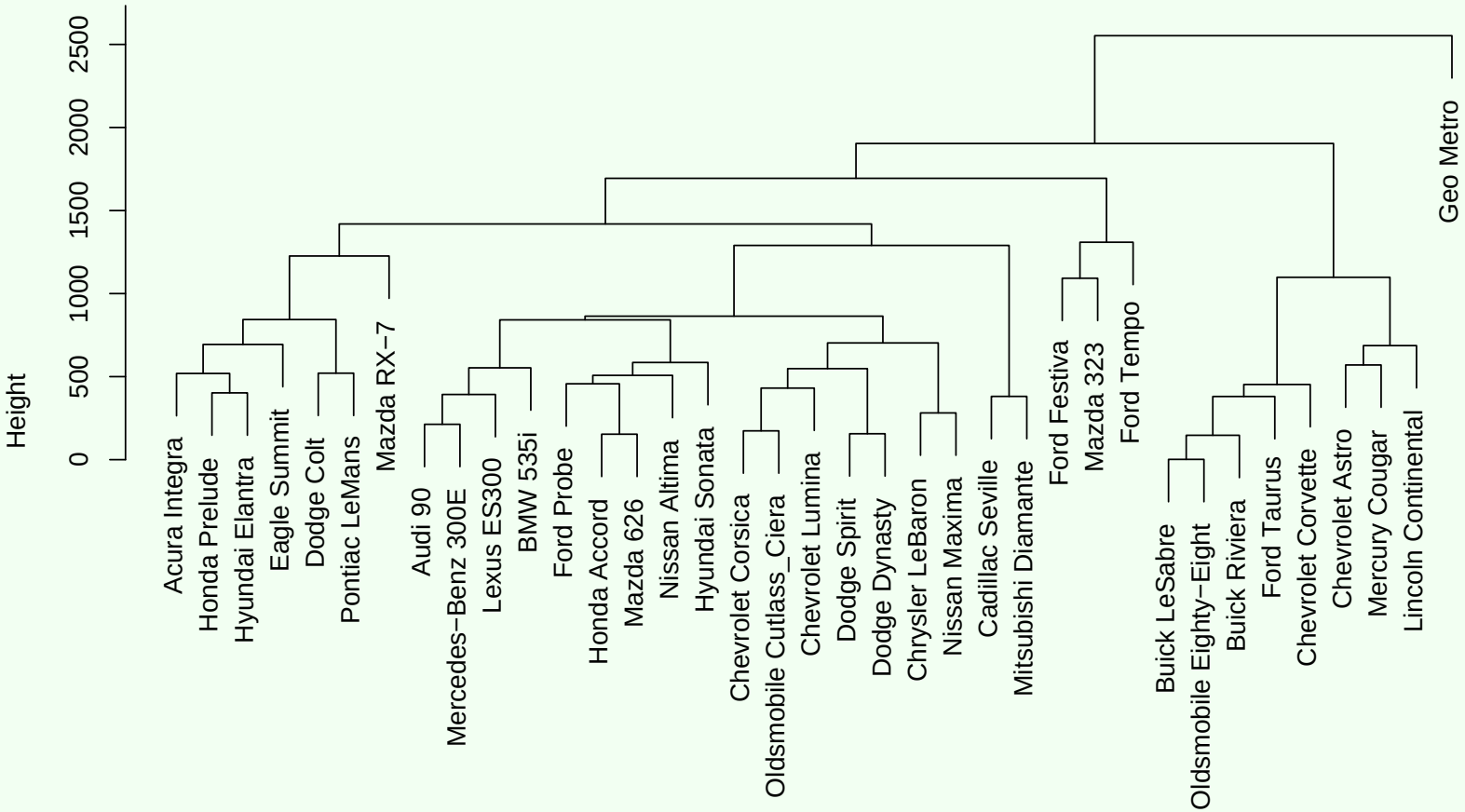
Przykład: samochody

Przedstawimy analizę skupień na przykładzie zbioru danych dotyczącego samochodów.

Dla 93 różnych modeli samochodów zebrano informacje o 27 cechach, takich jak moc silnika, pojemność, rozmiar, liczba drzwi, liczba poduszek, liczba cylindrów, liczba koni mechanicznych, maksymalne obroty, spalanie w trasie, mieście i średnie, wielkość bagażnika, masa itp.

Na podstawie tych danych chcemy ocenić, które samochody są do siebie podobne i które z nich można ewentualnie połączyć w grupy.

Przykład: samochody



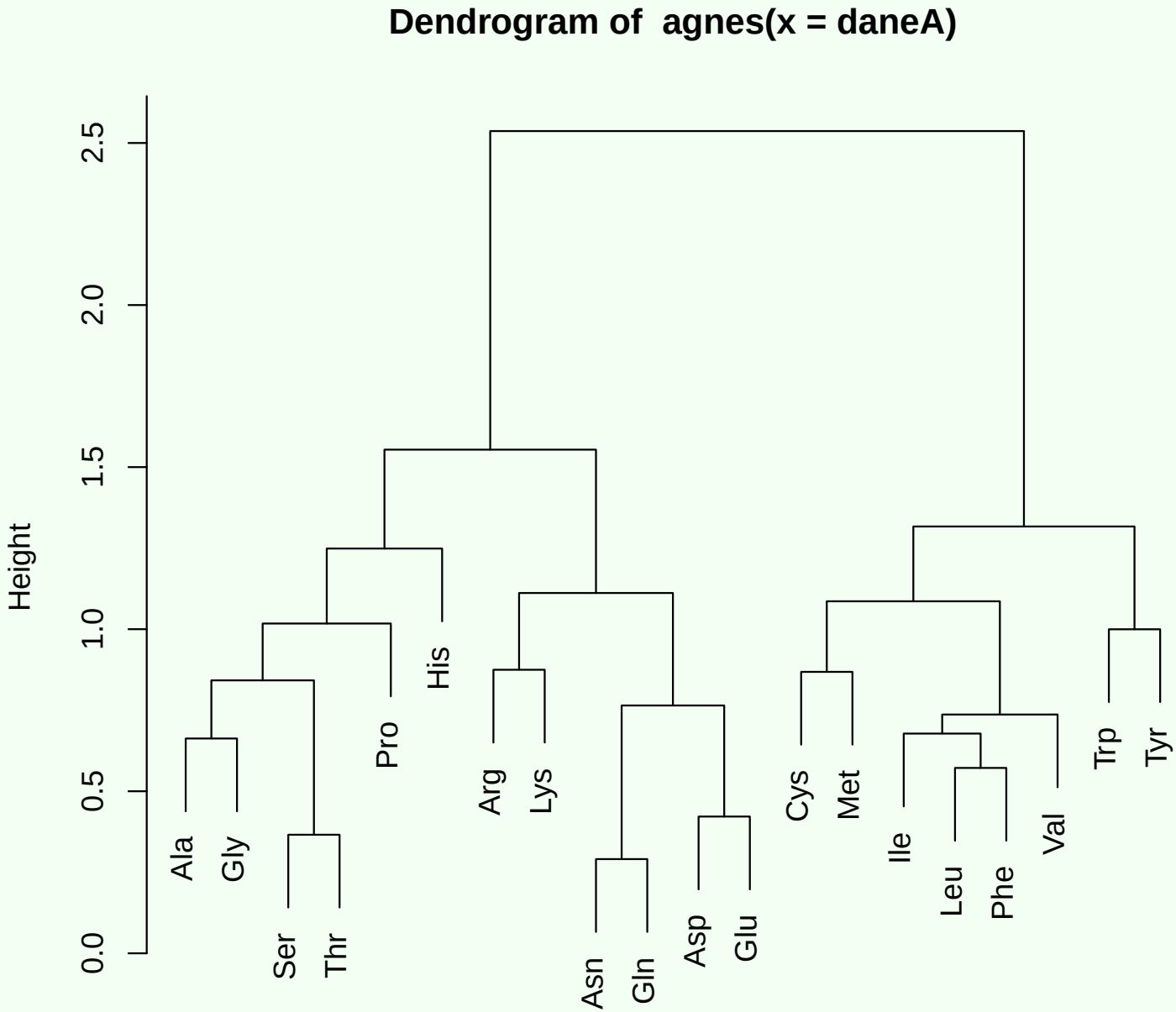
Agglomerative Coefficient = 0.8

Przykład: aminokwasy

Białka wszystkich żywych organizmów składają się z aminokwasów. Rożnych aminokwasów jest jedynie 20, każdy o innych właściwościach fizyko-chemicznych (zasadowość, polarność, rozmiar, homofilność, itd.).

Dysponujemy opisem każdego z 20 aminokwasów przez 23 różne współczynniki. Interesuje nas określenie podobieństwa pomiędzy aminokwasami na podstawie tych współczynników.

	Sweet	Kyte.and.Doolittle	Abraham.and.Leo	Bull.and.Breese	Guy
Ala	0.2817337	0.7000000	0.576846307	0.8625954	0.5508685
Arg	0.2229102	0.0000000	0.005988024	0.8931298	1.0000000
Asn	0.1207430	0.1111111	0.225548902	0.9694656	0.6451613
Asp	0.0000000	0.1111111	0.427145709	0.8625954	0.7196030
Cys	0.4582043	0.7777778	0.604790419	0.7671756	0.1736973
Gln	0.1238390	0.1111111	0.347305389	1.0000000	0.7617866


Dendrogram of agnes(x = daneA)
Height
0.0
0.5
1.0
1.5
2.0
2.5
Ala
Gly
Ser
Thr
Pro
His
Arg
Lys
Asn
Gln
Asp
Glu
Cys
Met
Ile
Leu
Phe
Val
Trp
Tyr

Klasyfikacja

Klasyfikacja to zagadnienie w którym naszym celem jest budowa algorytmu (klasyfikatora), który na podstawie określonych cech przypisze obiekt do jednej z g kategorii.

Również ten dział szybko się rozwija. Począwszy od lat dwudziestych XX wieku (np. Fisher) budowane są nowe algorytmy budowy klasyfikatorów. Do popularniejszych należą: regresja logistyczna, drzewa decyzyjne, metoda wektorów podpierających, sieci neuronowe i wiele innych.

Klasyfikacja ma wiele zastosowań, w tym:

- klasyfikacja do grupy zdrowy-chory,
- identyfikacja z jakim rodzajem nowotworu mamy do czynienia,
- rozpoznawanie numerów rejestracyjnych ze zdjęć tablic samochodowych,
- scoring i rating dla klientów banku.

Przykład: ryzyko wznowy wśród pacjentek chorych na raka piersi

Przykładowy klasyfikator przedstawimy na przykładzie danych zebranych od pacjentek chorych na raka piersi. Dane te zostały zebrane w Dolnośląskim Centrum Onkologii i dotyczą ponad 300 pacjentek leczonych w tym centrum.

Dla każdej z pacjentek zbadano wiele cech, np. informacje o przerzutach do węzłów chłonnych, menopauzie, wieku, poziomie rozmaitych substancji, markerów genetycznych HER, BCRP, Wimentyna i szeregu innych cech. Część pacjentek w analizowanym okresie doświadczyła nawrotu choroby a część pozostała zdrowa (analizujemy 5 letni DFS).

Pytanie na które chcemy odpowiedzieć to jakie czynniki mogą świadczyć o ryzyku wznowy oraz jak dla danej pacjentki ocenić ryzyko nawrotu choroby.

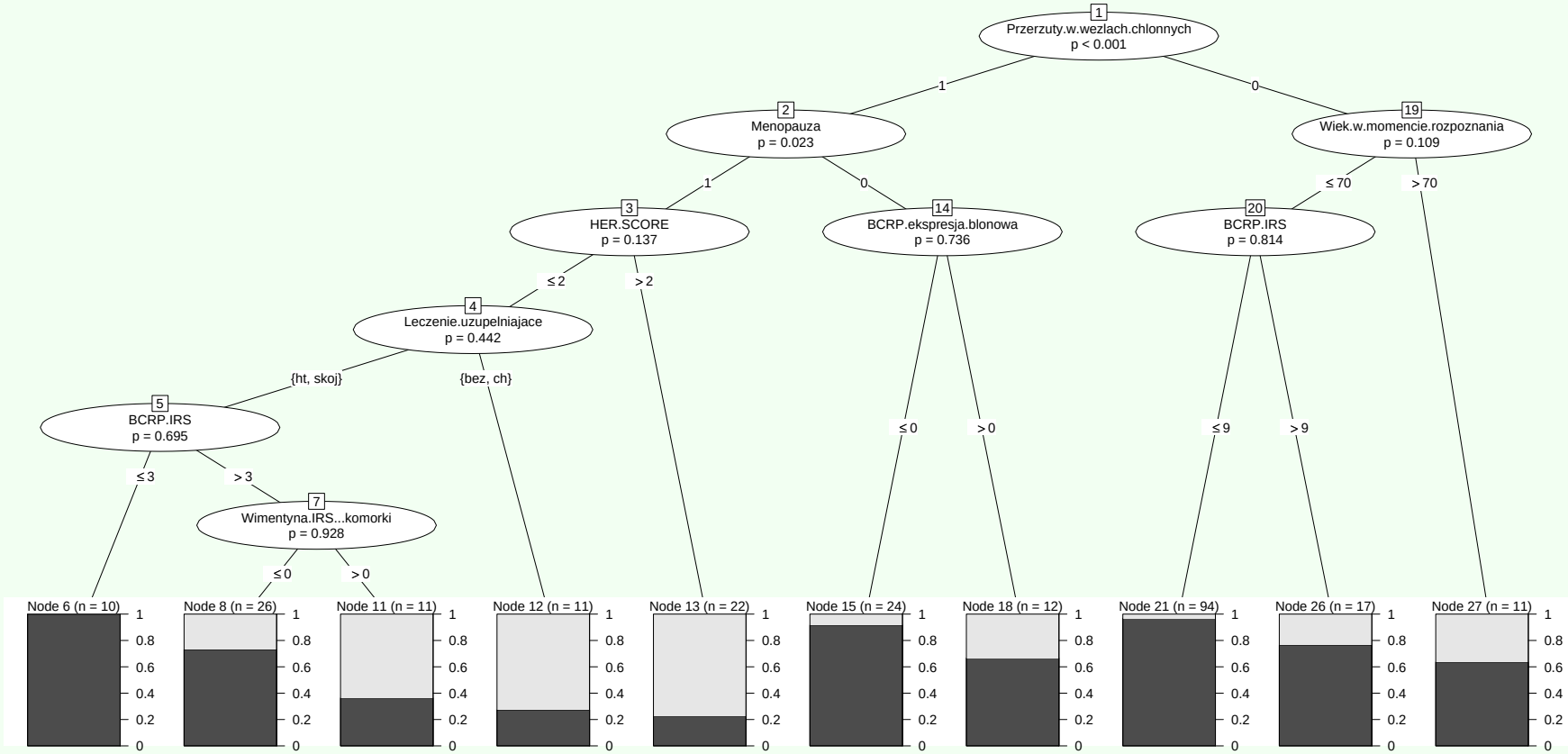
Drzewa decyzyjne

Drzewa decyzyjne to technika konstrukcji klasyfikatora, który można przestawić w postaci drzewa. Liście takiego drzewa odpowiadają klasom a węzły odpowiadają pewnym testom logicznym.

Drzewo konstruowane jest następująco:

- dla obserwacji znajdujących się w danym węźle identyfikujemy cechę, która najbardziej różnicuje badane klasy (używając współczynnika Giniego, entropii lub innej miary),
- jeżeli dana cecha jest ciągła to szukamy punktu podziału, który najsilniej różnicuje badane klasy,
- tworzymy dwa podwęzły, w jednym umieszczamy obserwacje o wartości wybranej cechy mniejszej niż wartość progowa, w drugim pozostałe,
- jeżeli różnorodność w danym podwęźle jest większa niż pewna zadana stała to wracamy do kroku 1.

Drzewo decyzyjne



Klasyfikacja z użyciem lasów losowych

Drzewa decyzyjne są metodą mało stabilną, której często towarzyszy problem tzw. przeuczenia. Aby poprawić właściwości tego klasyfikatora używa się tzw. metody boosting. Bazuje ona na metodzie bootstrap i pozwala na poprawę stabilności klasyfikatora.

Algorytm konstrukcji i użycia lasu losowego wygląda następująco:

- z próby danych generujemy B (np. 1000) podzbiorów danych, losując ze zwracaniem wiersze oraz bez zwracania cechy,
- na każdym z uzyskanych podzbiorów danych budujemy drzewo decyzyjne, ponieważ podzbiory zawierają średnio tylko około 35% wspólnych obserwacji, uzyskane drzewa różnią się między sobą,
- tak zbudowany las klasyfikuje nowe obserwacje do klasy najczęściej wybieranej przez drzewa z których jest zbudowany.

Klasyfikacja z użyciem lasów losowych

Po co konstruować las losowy?

Okazuje się że taki komitet klasyfikatorów ma bardzo dobre właściwości.

- Na zbiorach obserwacji, które nie znalazły się w zreplikowanych próbach (tzw. OOB, ang. out of bag) można liczyć nieobciążoną ocenę błędu predykcji.
- Badając jakość drzew w lesie można konstruować ranking ważności zmiennych.
- Można uzupełniać brakujące dane stosując tzw. imputacje.
- Można identyfikować obserwacje odstające i wpływowe.
- Otrzymuje się predyktor o małej wariancji.

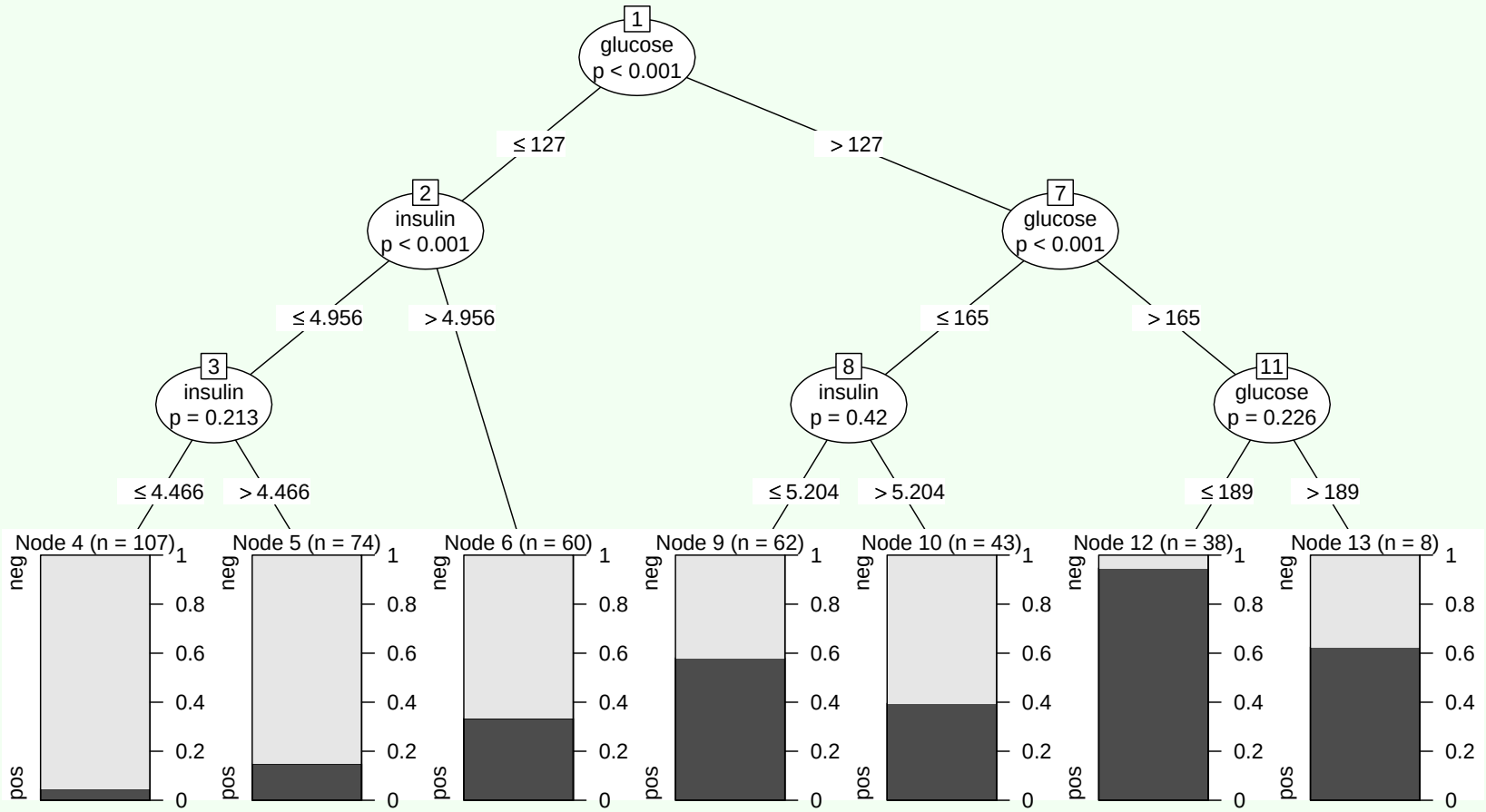
Przykład: Indianki z plemienia Pima

Kolejny przykład dotyczący klasyfikacji dotyczyć będzie badania cukrzycy u Indianek z plemienia Pima.

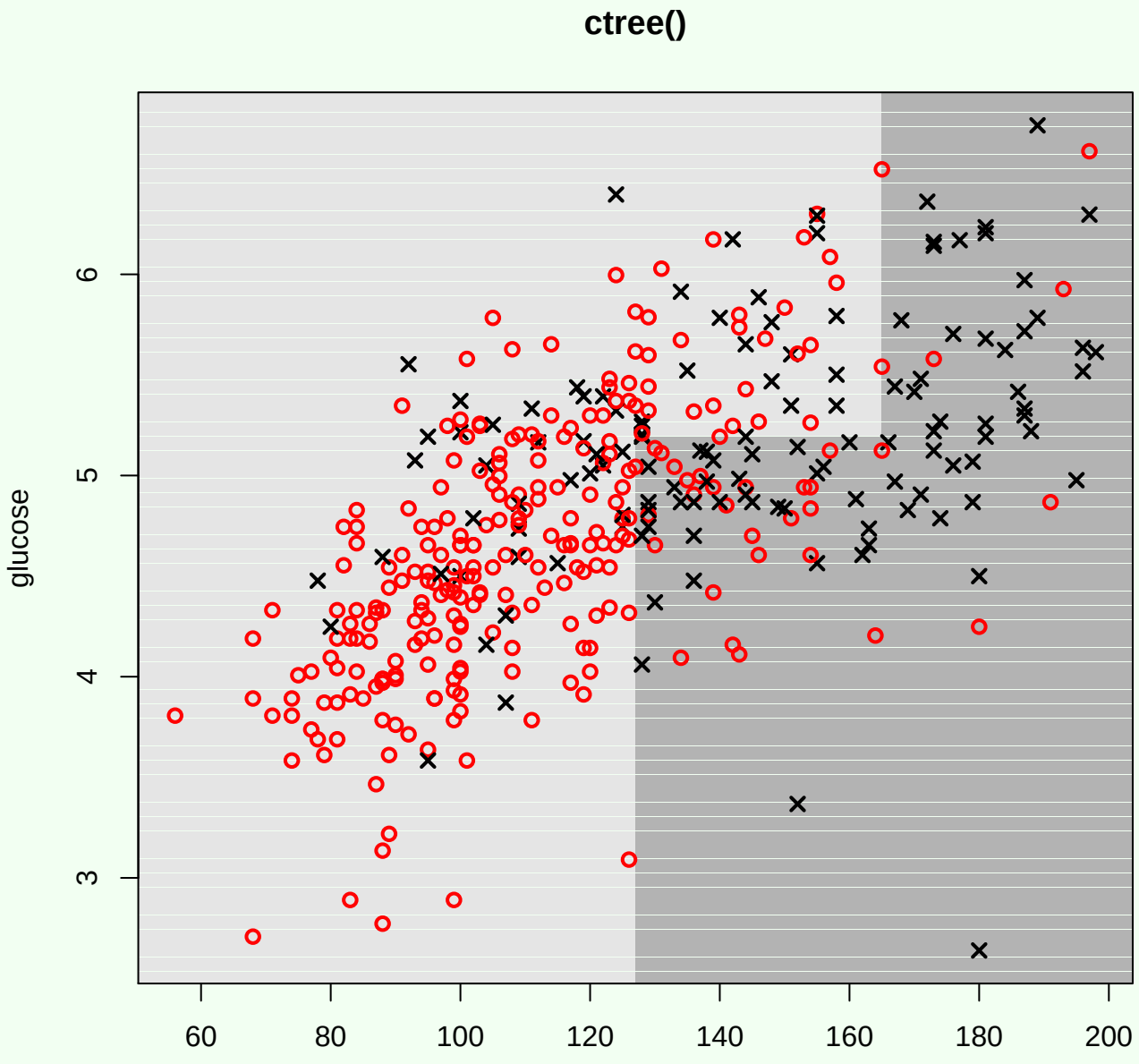
Pytanie: Która cecha lub zestaw cech pozwoli na skuteczną wstępną diagnozę cukrzycy?

	npreg	glu	bp	skin	bmi	ped	age	type
1	5	86	68	28	30.2	0.364	24	No
2	7	195	70	33	25.1	0.163	55	Yes
3	5	77	82	41	35.8	0.156	35	No
4	0	165	76	43	47.9	0.259	26	No
5	0	107	60	25	26.4	0.133	23	No
6	5	97	76	27	35.6	0.378	52	Yes

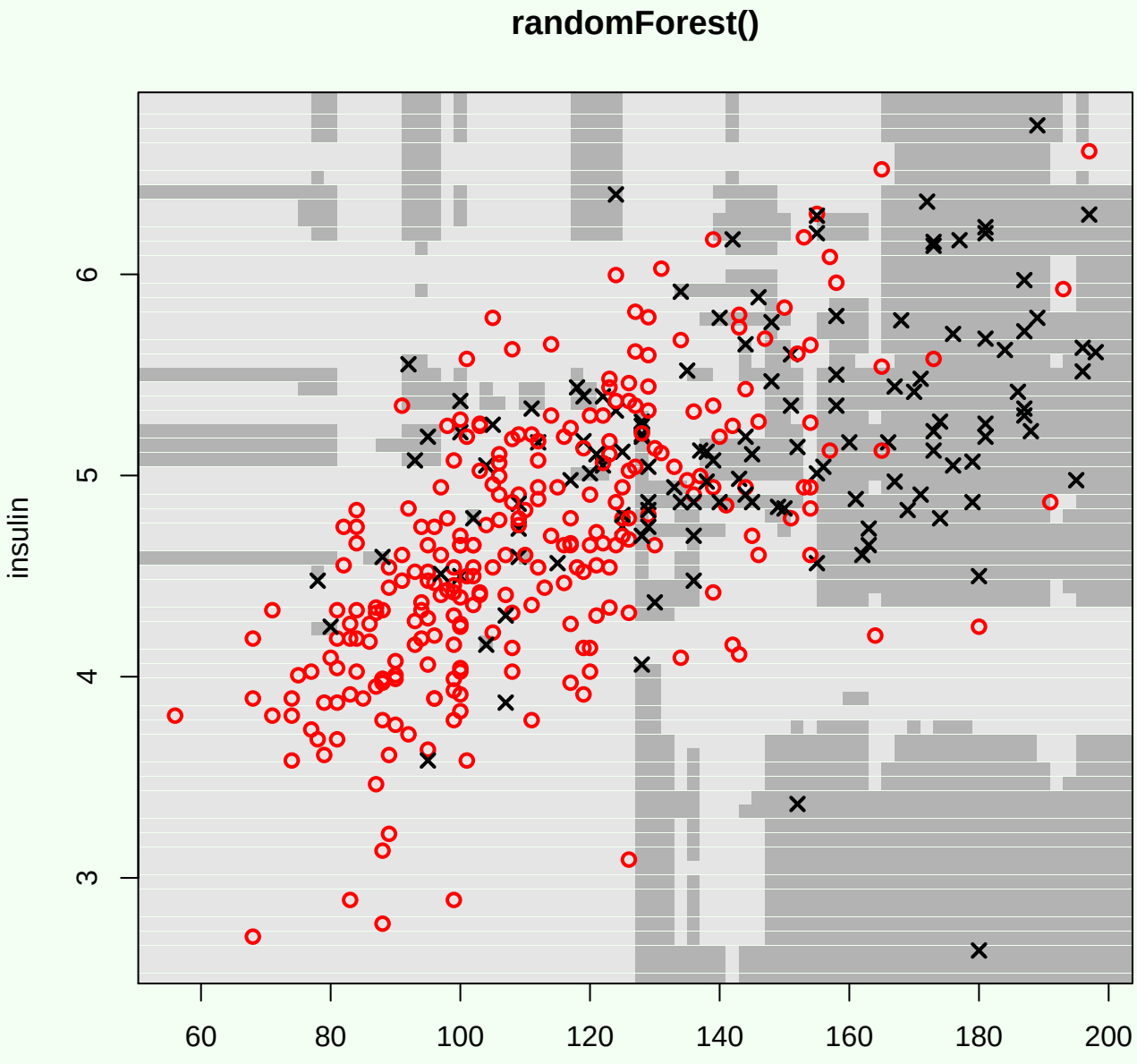
Przykład: Indianki z plemienia Pima



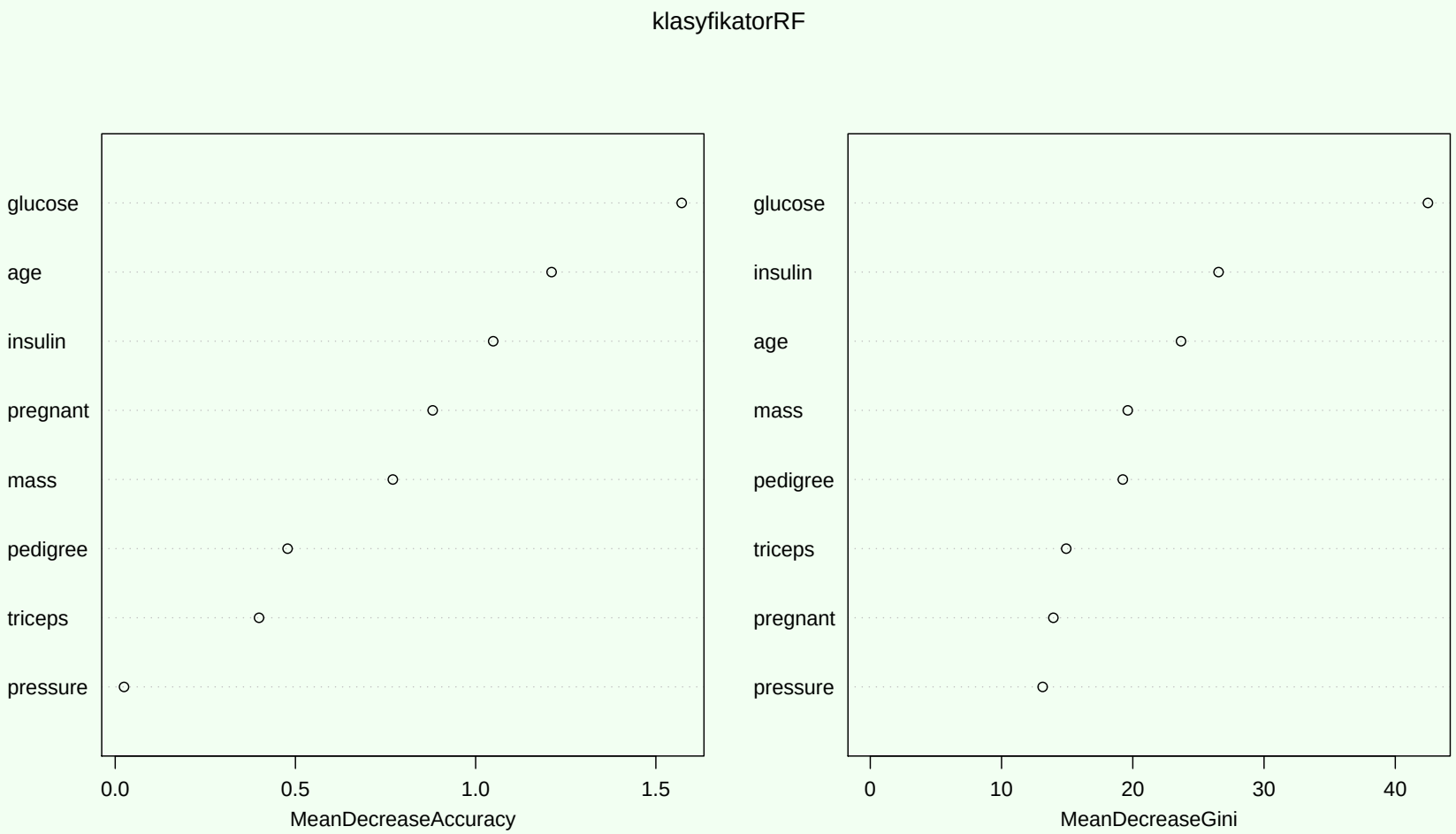
Przykład: Indianki z plemienia Pima, obszar decyzyjny drzewa decyzyjnego



Przykład: Indianki z plemienia Pima, obszar decyzyjny lasu losowego



Przykład: Indianki z plemienia Pima



Podsumowanie

- Data mining = statystyka - założenia,
  ale
  data mining + założenia $\neq$ statystyka.
- Przedstawione metody to algorytmy, które niczym w książce kucharskiej zostały zaprojektowane do rozwiązania określonych problemów.
- Algorytmom tym towarzyszą metody badające ich właściwości. Np. dla metod budowy klasyfikatorów potrzebne są metody do oceny błędu klasyfikacji, stabilności algorytmu, badania przeuczenia, diagnostyki modelu itp. W pewnym sensie ta diagnostyka odpowiada twierdzeniom z klasycznej statystyki matematycznej.
- Obecnie data mining jest znacznie szybciej rozwijany na wydziałach informatycznych niż przez „statystyków matematycznych". Czy to dobrze?