Data mining

Piotr Maliszewski

Plan

1. Czym jest data mining?
2. Czym nie jest data mining?
3. CRISP-DM
4. Przygotowanie danych + przykłady
5. Modelowanie danych + przykłady
    a. opis
    b. szacowanie i przewidywanie
    c. klasyfikacja
    d. grupowanie
    e. odkrywanie reguł asocjacyjnych

Czym jest data mining?

Data mining (po polsku zwykle - eksploracja danych) to interdyscyplinarna część informatyki, zajmująca się odkrywaniem wzorców w dużych zbiorach danych. Korzysta z narzędzi statystyki, uczenia maszynowego czy sztucznej inteligencji. Celem całego procesu jest wydobycie informacji ze zbioru danych i przetworzenie ich tak, by były zrozumiałe i nadawały się do późniejszego użycia. (Wikipedia)

Czym NIE jest data mining?

Metody eksploracji danych nie są żadnymi “magicznymi” (automatycznymi) sposobami wydobywania mądrości z danych. Niemożliwa jest eksploracja danych bez aktywnego działania człowieka.

Czym NIE jest data mining?

Eksploracja danych nie jest metodą “czarnoskrzynkową”. Nie ma oprogramowania, dzięki któremu każdy może prosto „wyklikać” rozwiązania problemów, nie zastanawiając się co tak naprawdę się dzieje z danymi.

Czym NIE jest data mining?

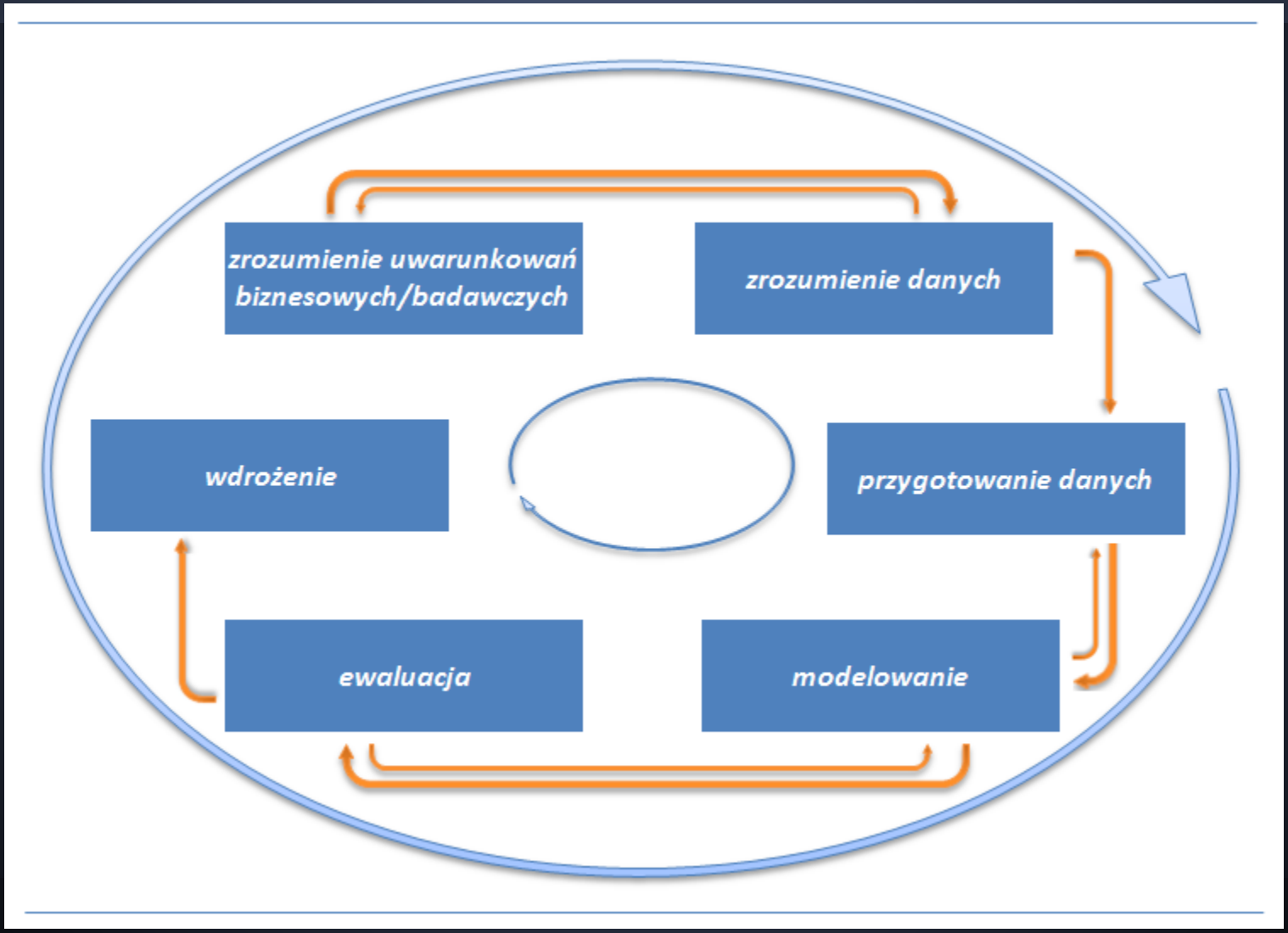
Eksploracja danych nie służy też do automatycznego czyszczenia / porządkowania baz danych. Procesy te są często bardzo pracochłonne - czasem bardziej, niż późniejsza analiza danych.

CRISP-DM: Cross Industry Standard Process for Data Mining

Utworzony w 1996 standard określający metodologię eksploracji danych. Jest uniwersalny dla wszystkich gałęzi przemysłu, biznesu i nauki.

Składa się z sześciu etapów:

CRISP-DM


zrozumienie uwarunkowań biznesowych/badawczych
zrozumienie danych
przygotowanie danych
modelowanie
ewaluacja
wdrożenie

1. Zrozumienie uwarunkowań biznesowych i/lub badawczych

Jeszcze przed rozpoczęciem jakichkolwiek czynności natury technicznej, takiej jak zbieranie danych, należy jasno określić: jakie cele stają przed badaczami? Jakie są wymagania projektu, a jakie jego ograniczenia? Po odpowiedzi na te pytania, należy stworzyć wstępny plan działania.

2. Zrozumienie danych

Surowe dane mogą być nieuporządkowane lub rozrzucone po wielu nośnikach (serwerach). Takie dane należy zebrać i wstępnie ocenić ich jakość oraz przydatność dla projektu. Może też okazać się, że dane są dla badaczy kompletnie niezrozumiałe – wtedy konieczne jest zasięgnięcie rady ekspertów w danej dziedzinie (mogą to być np. pracownicy firmy, dla której robiona jest analiza).

3. Przygotowanie danych

Bardzo ważny etap, który może okazać się niezwykle pracochłonny. Wstępnie przesiane pod kątem przydatności dane wciąż są w stanie surowym. Być może są w wielu niekompatybilnych formatach, są niekompletne lub zanieczyszczone zbędnymi informacjami. Mogły być nieużywane od wielu lat. Należy więc je oczyścić i poradzić sobie z brakującymi atrybutami, a następnie przygotować tak, by były gotowe do użycia przez narzędzia modelujące.

4. Modelowanie

Konieczny jest wybór jednej (lub kilku) technik modelujących dane – w zależności od ich rodzaju i efektów jakie chcemy uzyskać. Podstawowe zadania stojące przed eksploracją danych to opis (eksploracyjna analiza danych), szacowanie (estymacja), przewidywanie (predykacja), klasyfikacja, grupowanie i odkrywanie reguł asocjacyjnych.

5. Ewaluacja

Ocena zastosowanego modelu (modeli). Czy spełnił postawione na początku założenia biznesowe? Czy są jakieś cele biznesowe lub badawcze, które nie zostały uwzględnione? Czy wyniki dały odpowiedzi na pytania, czy są zadowalające? Jeśli nie, można wrócić do poprzedniego etapu.

6. Wdrożenie

Etap ten jest w podobnym stopniu etapem „biznesowym” (a nie „technicznym”), jak etap 1. Należy zastanowić się, jak i czy wykorzystać stworzone modele. Wiąże się to np. z napisaniem raportu skierowanego do osób decyzyjnych (zarządu firmy) czy też artykułu naukowego.

CRISP-DM

Ponieważ metodologia CRISP-DM jest iteracyjna, po etapie 6. może nastąpić powrót do etapu 1. Być może założenia biznesowe/naukowe należy sformułować inaczej. Z drugiej strony, badacze, bogatsi o zdobyte doświadczenia, mogą np. przeprojektować bazy danych tak, by bardziej nadawały się do odkrywania z nich danych.

CRISP-DM

Proces eksploracji danych zaczyna się już w momencie zbierania danych, co nieraz wymaga innego spojrzenia na to, jak powinna funkcjonować dana instytucja. Rzadko (w praktyce: nigdy) zdarza się tak, by już za pierwszym razem badacze odkryli w danych „świętego Graala”, który w magiczny sposób doprowadzi do przełomowych odkryć naukowych czy zwielokrotni dochody klienta.

Przygotowanie danych

Dlaczego należy obrabiać dane przed ich użyciem?

Baza danych może zawierać:

- Pola, które są przestarzałe lub zbędne
- Rekordy z brakującymi wartościami
- Punkty oddalone
- Dane znajdujące się w nieodpowiednim do eksploracji formacie
- Wartości niezgodne ze zdrowym rozsądkiem

Czyszczenie danych

Co z tymi danymi jest nie tak?...

ID	kod pocztowy	płeć	dochód	wiek	stan cywilny	kwota transakcji
1001	10048	M	75000	C	M	5000
1002	J2S7K7	K	-40000	40	W	4000
1003	90210		10000000	45	S	7000
1004	6269	M	50000	0	S	1000
1005	55101	K	99999	30	R	3000

Czyszczenie danych

Klienci 1002 i 1004 mają “dziwne” kody pocztowe. Pierwszy nie pochodzi z USA tylko z Kanady, a prawidłowy kod drugiego to “06269”, zapisany jako liczba a nie tekst i pozbawiony początkowego zera.

Klient 1003 najwyraźniej nie ma płci…

...oraz ma ponadprzeciętny dochód. Niezależnie od tego, czy jest to błąd czy prawda - taka wartość znacznie oddalona od pozostałych może zepsuć efekty działania niektórych technik statystycznych modelowania.

Czyszczenie danych

Klient 1002 ma ujemny dochód, a 1005 podany z dziwną dokładnością. Być może w jakieś bazie danych wartość “99999” oznaczała brakującą wartość.

Poza tym nie wiemy, w jakich jednostkach mierzony jest dochód. Być może klient z Kanady podał wartość w dolarach kanadyjskich a nie amerykańskich…

Pole wiek jest problematyczne - po pierwsze deaktualizuje się, po drugie czy “0” oznacza wiek nieznany, czy noworodka?

Stan cywilny w takiej formie jest nieoczywisty - bo czy “S” oznacza “Single” czy “Separated”?

Obsługa brakujących danych

zużycie paliwa	liczba cylindrów	pojemność silnika	konie mechaniczne
14,0	8	350,0	165,0
31,9	4	89,0	71,0
15,0		400,0	150,0
30,5			
23,0		350,0	125,0
11,0	8		215,0
25,4	5		77,0
37,7	4	89,0	62,0

Obsługa brakujących danych

zużycie paliwa	liczba cylindrów	pojemność silnika	konie mechaniczne
14,0	8	350,0	165,0
31,9	4	89,0	71,0
15,0	??	400,0	150,0
30,5	??	??	??
23,0	??	350,0	125,0
11,0	8	??	215,0
25,4	5	??	77,0
37,7	4	89,0	62,0

Obsługa brakujących danych

Możemy po prostu pominąć rekordy zawierające puste pola, ale może być ich za dużo… Lepsze rozwiązania to:

1. Zastępywanie brakujących wartości stałą, określoną przez analityka
2. Zastępywanie ich wartością średnią (np. pojemność silnika - w każdym pustym polu 200,65

Obsługa brakujących danych

3. Zastępywanie ich wartościami losowymi pochodzącymi z rozkładu zmiennej (np. pojemność silnika od góry w brakujących: 144,15; 323,45; 81,84)

Ostatnie rozwiązanie brzmi najbardziej sensownie z punktu widzenia “matematycznego”, ale otrzymaliśmy właśnie silnik o 5 cylindrach i pojemności 82. To chyba mało prawdopodobne...

Obsługa brakujących danych

W każdym takim przypadku musimy zastanowić się nad sensem danych które uzupełniamy.

Istnieją metody, które pytają: “Jaka wartość byłaby najbardziej prawdopodobna na miejscu brakującej, biorąc pod uwagę wszystkie pozostałe?”. Jedną z nich jest estymacja bayesowska.

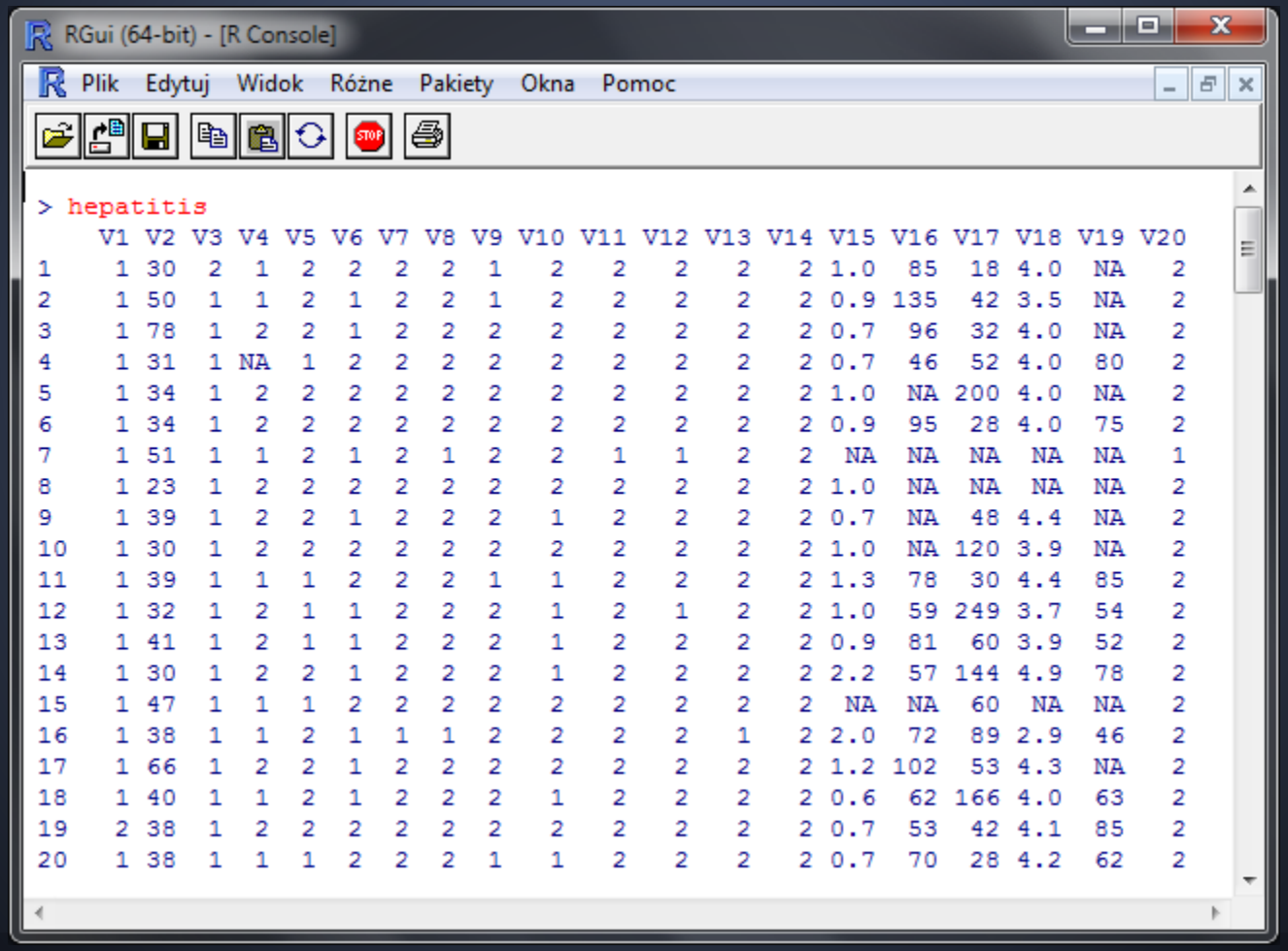
Obsługa brakujących danych

Przykład - zastosowanie środowiska R.

R to język programowania służący do obliczeń statystycznych i wizualizacji wyników.

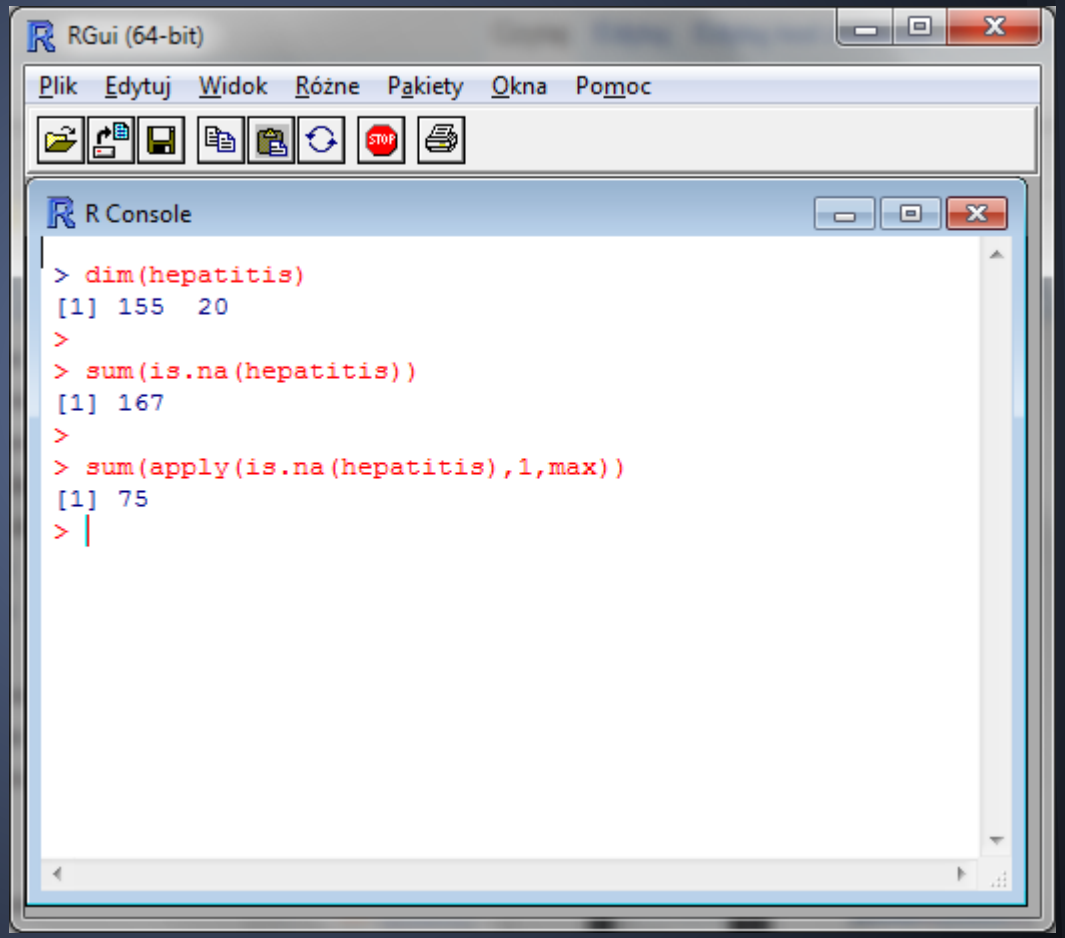
Skorzystamy ze zbioru danych *hepatitis*, zawierającego 155 rekordów po 20 atrybutów. Niektóre rekordy są niekompletne.

Obsługa brakujących danych



Obsługa brakujących danych

- 155 rekordów po 20 atrybutów
- 167 brakujących atrybutów w sumie
- 75 rekordów z co najmniej jednym brakującym atrybutem



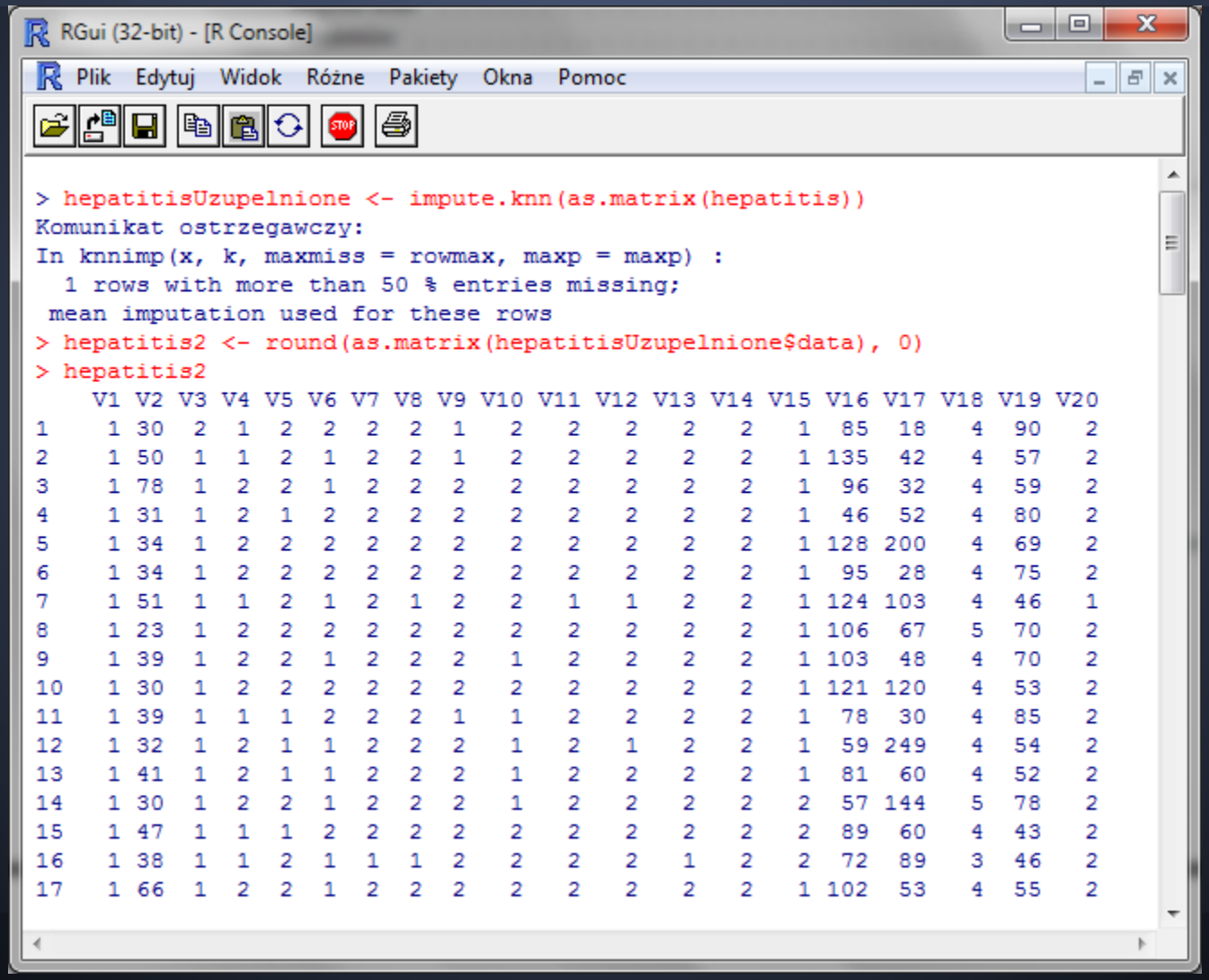
Obsługa brakujących danych

Zastosujemy funkcję *impute.knn* z repozytorium Bioconductor. Korzysta ona z algorytmu *k*-najbliższych sąsiadów (więcej o nim w części poświęconej klasyfikacji).

Domyślnie, korzysta ona z $k = 10$ pobliskich wartości, i generuje na ich podstawie wartości losowe, które są wstawiane w miejsce brakujących.

Jeśli w wierszu brakuje więcej niż 50% danych, funkcja korzysta ze średniej.

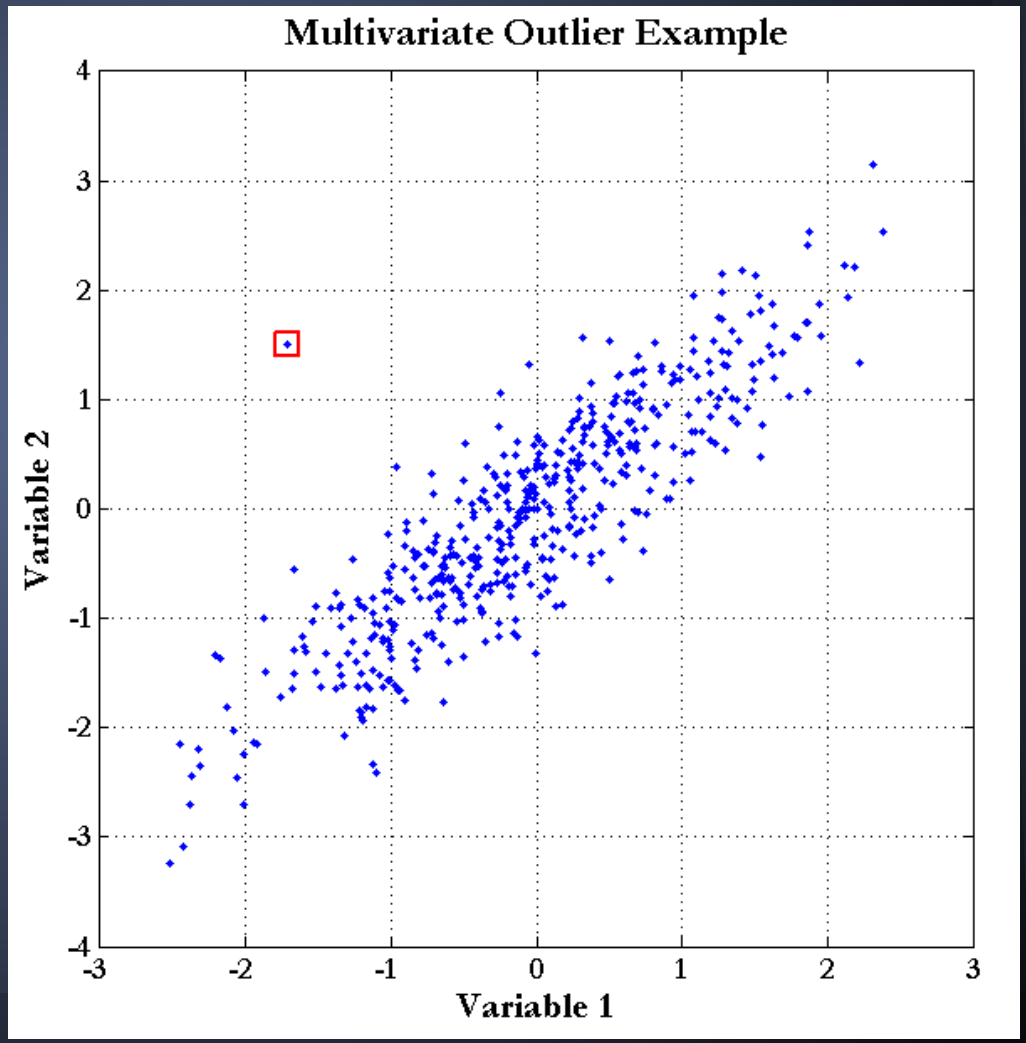
Obsługa brakujących danych



Graficzne metody identyfikacji punktów oddalonych

Punkty oddalone to skrajne wartości, które są sprzeczne z ogólnym trendem pozostałych danych.

Nawet jeśli taki punkt to nie błąd, niektóre metody (np. sieci neuronowe, algorytm k-najbliższych sąsiadów) są na to wrażliwe i mogą dać niestabilne wyniki.



Przekształcanie danych

Zmienne na ogół mają zakresy, które bardzo różnią się od siebie (np. X od 0 do 0,4; Y od 0 do 70). Dla niektórych algorytmów może powodować to, że zmienne o “większych” wartościach (tutaj: Y) będą miały nadmierny wpływ na wyniki.

W celu radzenia sobie z tym problemem, możemy skorzystać z normalizacji lub standaryzacji.

Normalizacja min-max

$$X^* = \frac{X - min(X)}{zakres(X)} = \frac{X - min(X)}{max(X) - min(X)}$$

Normalizując zmienną sprawdzamy, jak bardzo jej wartość jest większa od wartości minimalnej *min(X)*, i skalujemy tą różnicę przez zakres.

W nowym zakresie (od 0 do 1) minimalna wartość X to 0, a maksymalna - 1.

Standaryzacja

$$X^* = \frac{X - srednie(X)}{odchylenieStandardowe(X)}$$

Standaryzując zmienną, obliczamy różnicę pomiędzy daną i średnią wartością pola oraz skalujemy tą różnicę przez odchylenie standardowe wartości pól.

Po standaryzacji, wartości zmiennych należą z reguły do przedziału od -4 do 4, z wartością średnią równą 0.

Modelowanie danych

Nasze dane są już odpowiednio wyczyszczone, przekształcone i przygotowane do eksploracji. Nadszedł czas na wybór techniki odpowiedniej do założonych na początku celów.

Do podstawowych technik należą eksploracyjna analiza danych (opis), szacowanie, przewidywanie, klasyfikacja, grupowanie i odkrywanie reguł asocjacyjnych.

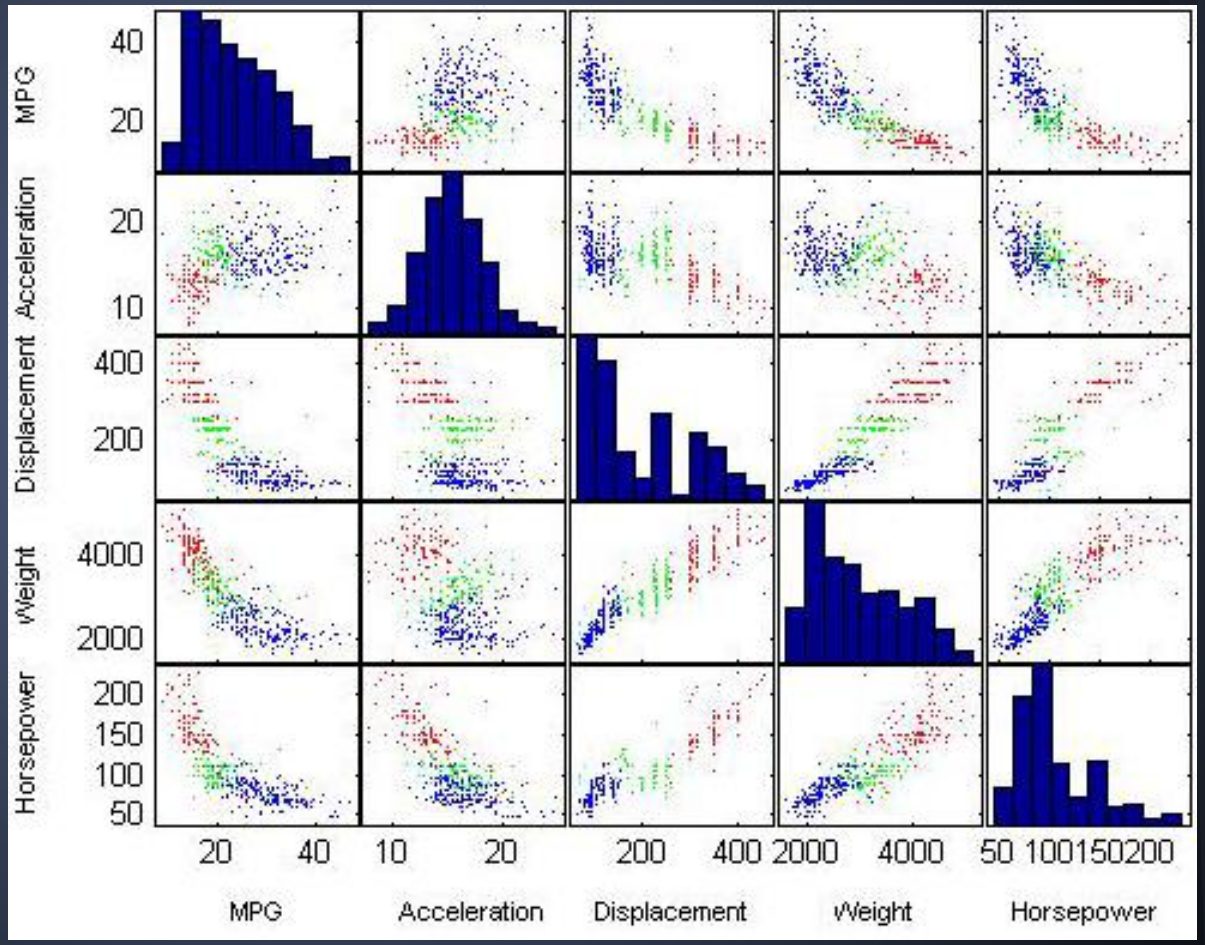
Eksploracyjna analiza danych

EDA, dzięki wizualizacji danych, pozwala na sprawdzenie wzajemnych relacji między atrybutami i identyfikację ciekawych podzbiorów obserwacji.

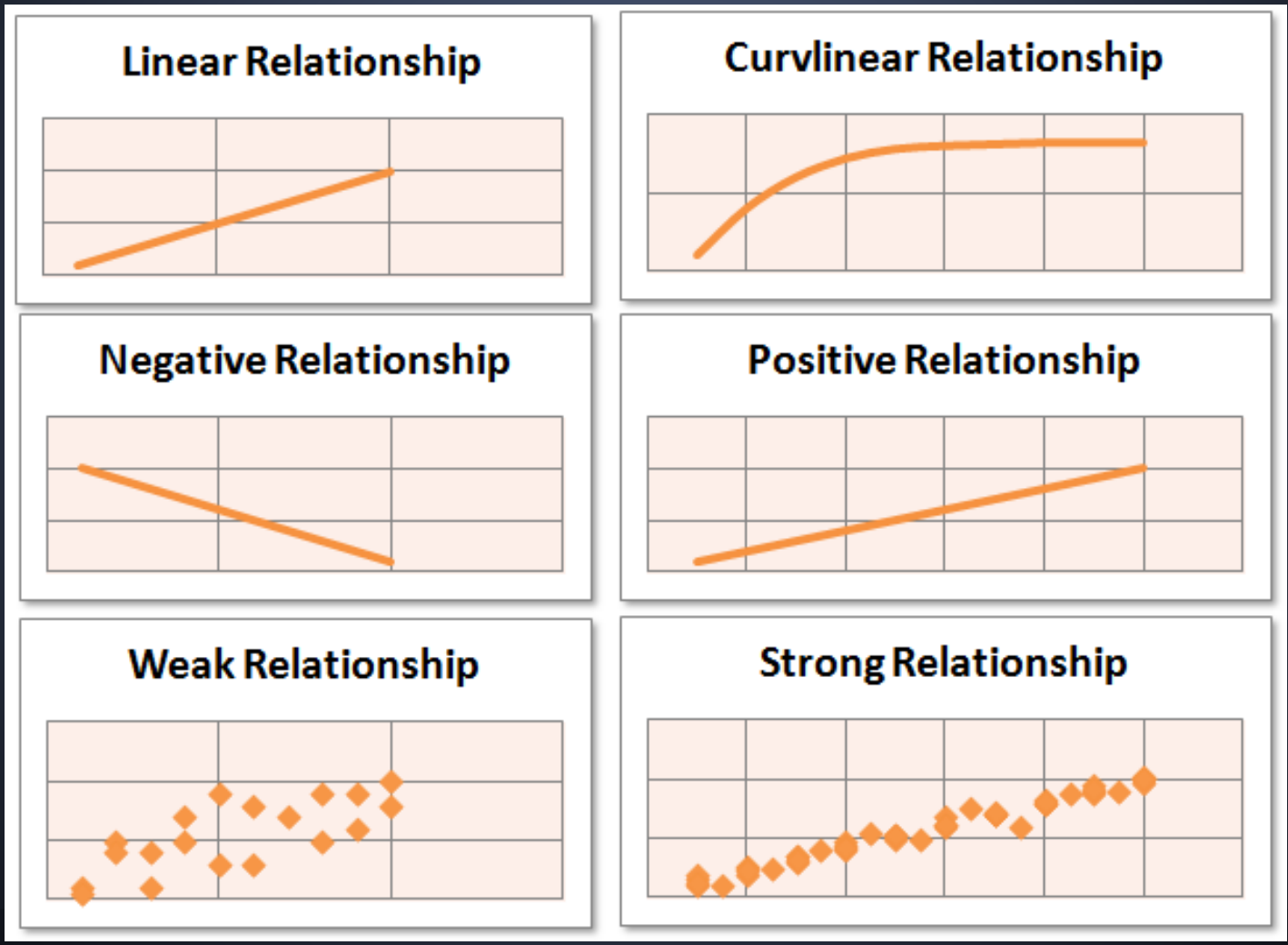
Proste wykresy, rzuty i tabele często odkrywają ważne relacje, które mogą wskazać interesujący podzbiór danych do dalszych badań.

Eksploracyjna analiza danych

Już podczas najprostszej wizualizacji dwuwymiarowej, możemy starać się zauważyć pewne zależności.



Eksploracyjna analiza danych



Eksploracyjna analiza danych

Na przykład, jeśli zauważymy gdzieś zależność idealnie liniową, oznacza to, że jedna z tych zmiennych jest po prostu funkcją drugiej.

Rozsądne jest wtedy wyrzucenie jednej z tych zmiennych ze zbioru danych, w celu zmniejszenia rozmiaru zadania oraz uniknięcia wielokrotnej regresji.

Eksploracyjna analiza danych

Inną ważną techniką EDA jest dyskretyzacja zmiennych.

Np. możemy podzielić cały zbiór danych o klientach za pomocą zmiennej wieku na trzy części: 0-18, 19-39 i 40-100. Sprawdzając ponownie zależności dla każdego z tych przedziałów i porównując je ze sobą, możemy zauważyć coś interesującego.

Dyskretyzacja jest czynnością równie przygotowującą dane, jak i eksplorującą.

Szacowanie i przewidywanie

Szacowanie i przewidywanie należą do dziedziny wnioskowania statystycznego. Są stosowane przez statystyków od ponad stu lat.

Wnioskowanie statystyczne składa się z metod szacowania i testowania hipotez o cechach populacji (całego zbioru danych), na podstawie informacji zawartych w próbce (reprezentatywny podzbiór populacji).

Szacowanie i przewidywanie

Estymując nieznane wartości dla całego zbioru danych, nie uda nam się zrobić tego dokładnie. Możemy co najwyżej oszacować prawdopodobieństwo, z jakim wartość zawiera się w jakimś przedziale.

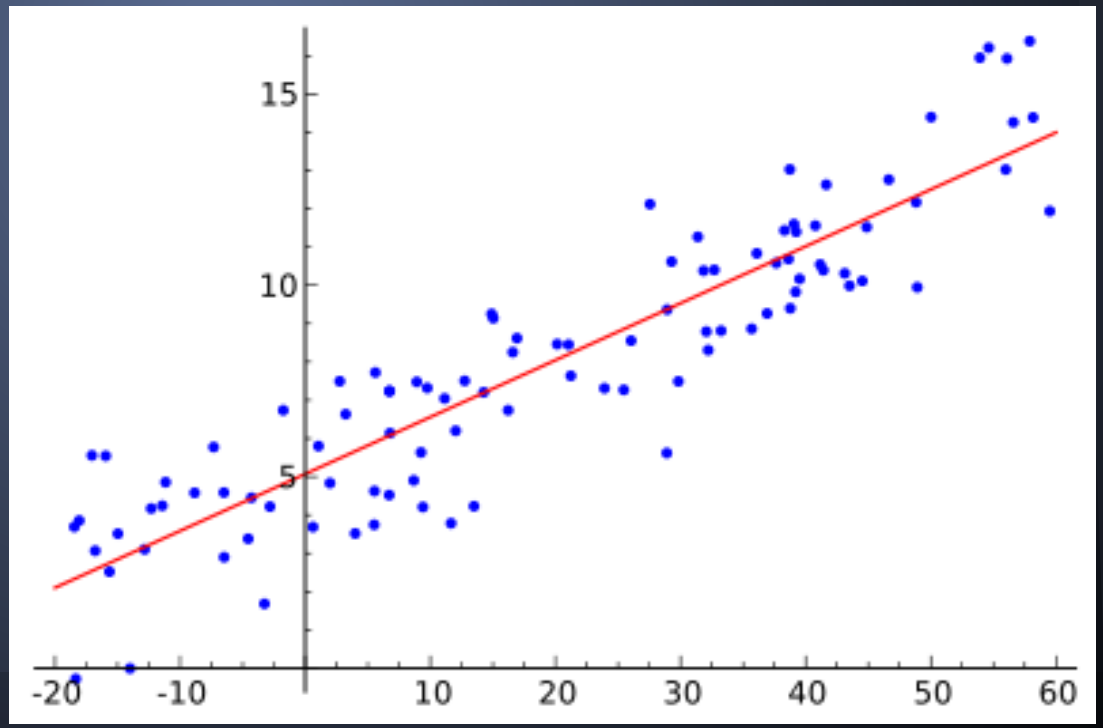
Np. dla pewnej próbki składającej się z 3333 rekordów, średnia wartość parametru wynosi 1,563. Możemy wyliczyć, że z prawdopodobieństwem 95% wartość dla całego zbioru danych zawiera się w przedziale (1,518, 1,608) - margines błędu wynosi tu 0,045.

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Szacowanie i przewidywanie

Poprzednie metody służyły do szacowania wartości jednej zmiennej. Możemy też spróbować przewidzieć wartość jednej zmiennej na podstawie drugiej, biorąc za dane znane korelacje.

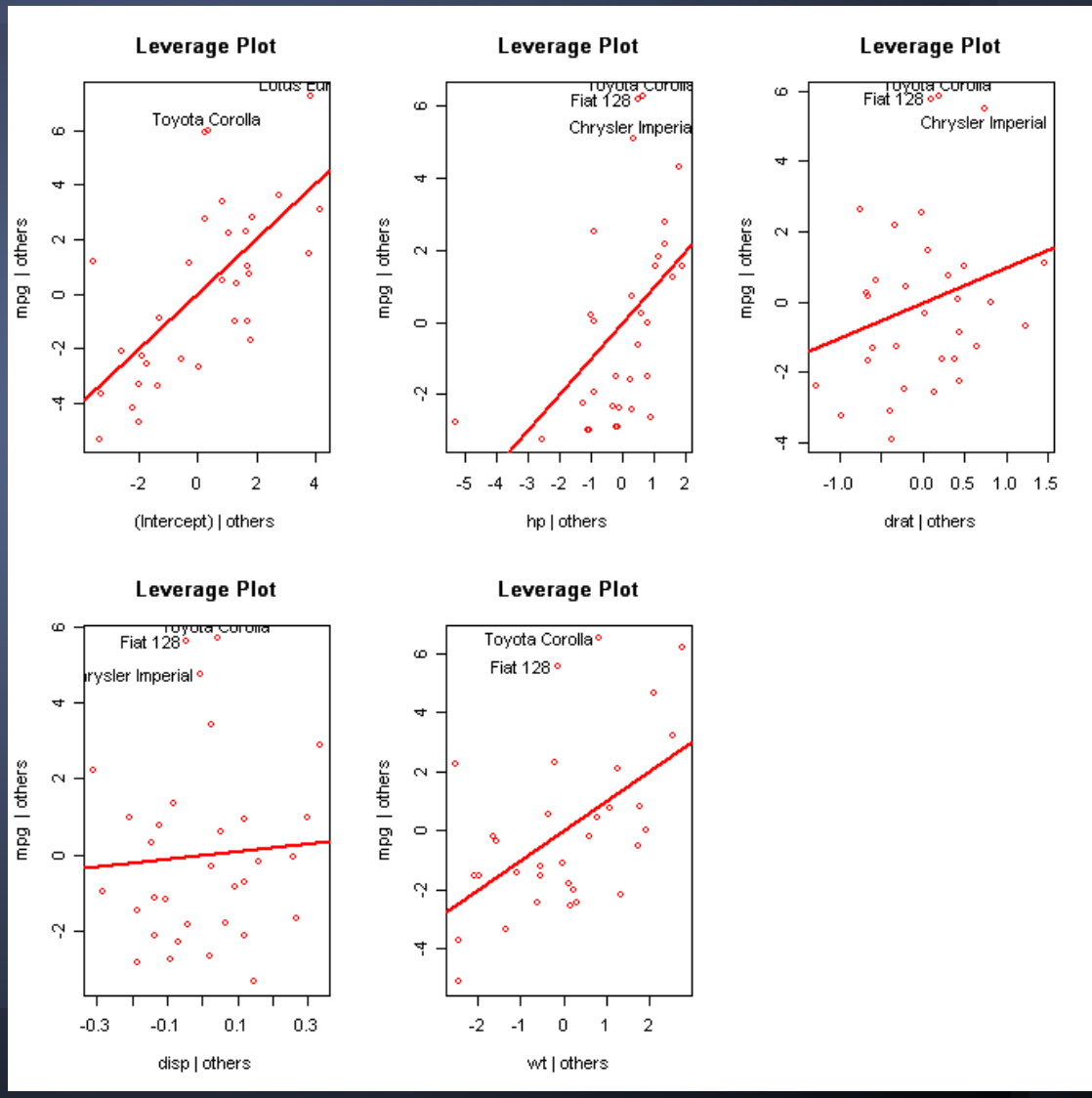
Metoda ta nazywa się regresja liniowa (tu: dwuwymiarowa, jednokrotna).



Szacowanie i przewidywanie

Regresja wielokrotna służy do zobrazowania zależności większej ilości zmiennych (osie X) od zmiennej celu (wynikowej - oś Y).

Ważne jest, by eliminować zmienne współliniowe, aby nie doprowadzać do niestabilności w rozwiązaniach.



Klasyfikacja

Klasyfikacja jest algorytmem, starającym się przyporządkować dane do kategorii (klas) na podstawie otrzymanego zbioru uczącego.

Np. mamy bazę danych osób, zawierającą ich wiek, płeć, zawód oraz zmienną celu - grupę dochodu (niski, średni, wysoki). Za pomocą klasyfikacji, możemy “nauczyć” algorytm rozpoznawać wzorce w tych danych tak, że gdy otrzymamy nowe dane - tym razem bez grupy dochodu - będzie on w stanie przyporządkować je do odpowiednich klas.

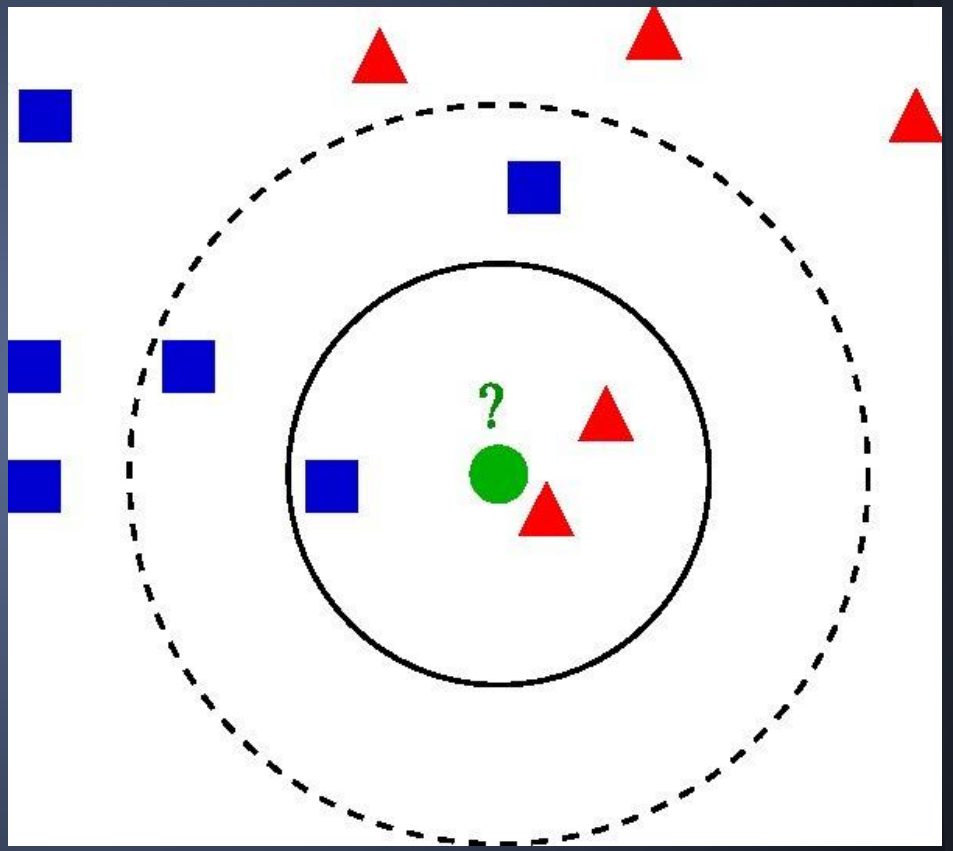
Algorytm *k*-najbliższych sąsiadów

Algorytm *k*-nn (*k*-nearest neighbor) jest przykładem uczenia leniwego. Cały zbiór uczący jest zapamiętywany, tak, że klasyfikacja nowych niesklasyfikowanych rekordów może zostać dokonana przez porównywanie z najbardziej podobnymi rekordami ze zbioru uczącego.

Klasyfikacja dokonywana jest w oparciu o odległość nowego punktu od *k* punktów sąsiednich.

Algorytm *k*-najbliższych sąsiadów

Przykład:
klasyfikujemy zielony punkt. Zależnie od wyboru *k*, zostanie on sklasyfikowany jako: czerwony (*k* = 1, 2, 3), czerwony lub niebieski (*k* = 4), niebieski (*k* = 5).



Odległość pomiędzy punktami to odległość euklidesowa, a wszystkie punkty powinny być znormalizowane.

Algorytm *k*-najbliższych sąsiadów

Ulepszeniem algorytmu może być wprowadzenie głosowania ważonego, tzn. dla $k > 1$ bliżsi sąsiedzi mają większy wpływ na decyzję niż bardziej oddaleni.

Ułatwia to postępowanie w sytuacji jak na poprzednim slajdzie, gdy dla $k = 4$ w prostym głosowaniu był remis. W głosowaniu ważonym, punkt zostałby wtedy sklasyfikowany jako czerwony.

Algorytm *k*-najbliższych sąsiadów

Problemem jest wybór najbardziej odpowiedniego *k*. Dla małego *k*, np. *k* = 1, algorytm zwróci po prostu wartość zmiennej celu najbliższej obserwacji, a to może prowadzić do przeuczenia (zapamiętania zbioru uczącego kosztem porządanej umiejętności generalizowania).

Z drugiej strony, zbyt duże *k* może spowodować, że lokalnie ciekawe zachowanie może zostać przeoczone.

Drzewa decyzyjne

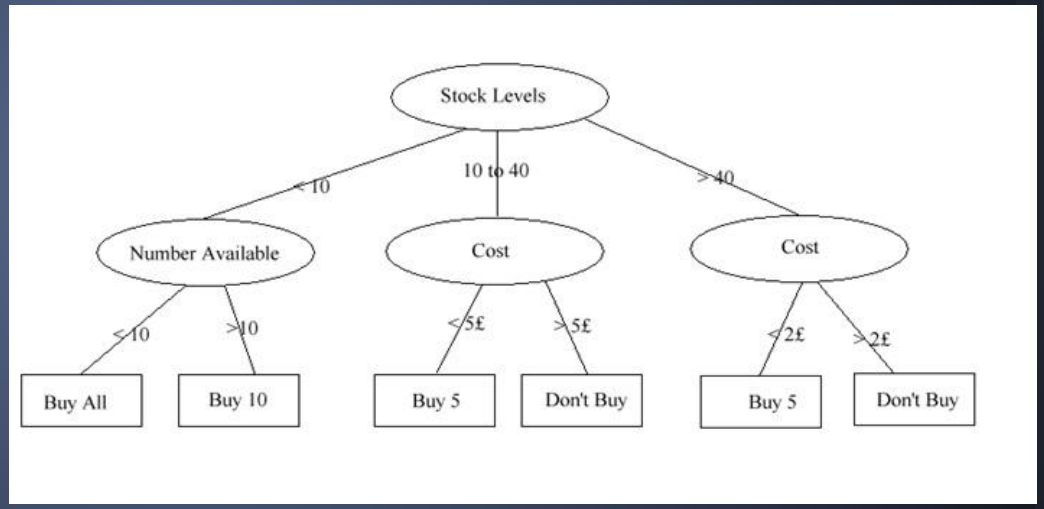
Kolejną metodą klasyfikacji są drzewa decyzyjne. Jest to metoda graficzna, wspomagająca proces decyzyjny.

Zbiór uczący powinien był bogaty i różnorodny, aby klasyfikacja każdego podzbioru była możliwa.

Klasy zmiennej celu muszą być dyskretne (a nie ciągłe). Przynależność do danej klasy musi być jasno określona.

Drzewa decyzyjne

Liście drzewa to zmienna celu - odpowiedź na pytanie zadawane przez badacza.



Przykład: popularny zbiór danych *golf*, na temat warunków pogodowych do grania w golfa, zawierający informacje o pogodzie oraz o opinii graczy, czy ten dzień “nadawał się” do gry (tak / nie).

Drzewa decyzyjne

Atrybuty kolejno:

pogoda, temperatura, wilgotność powietrza, wiatr (tak/nie), czy ten dzień był dobry na golfa

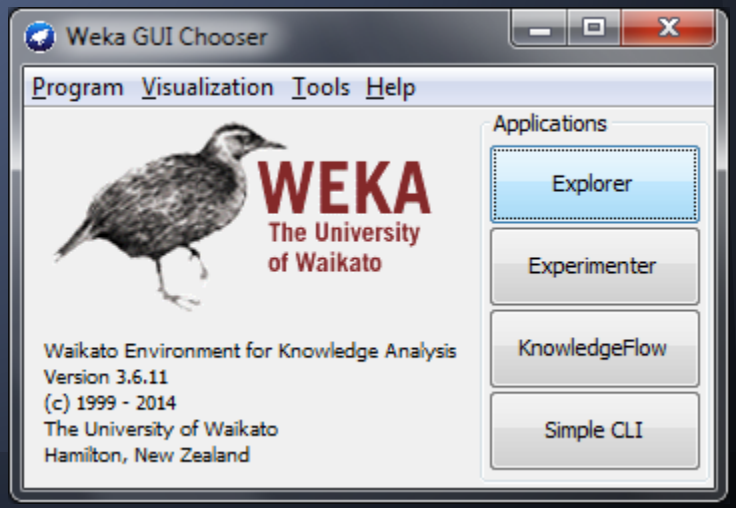
```
overcast, 83, 86, false, yes
overcast, 64, 65, true, yes
overcast, 72, 90, true, yes
overcast, 81, 75, false, yes
rainy, 70, 96, false, yes
rainy, 68, 80, false, yes
rainy, 65, 70, true, no
rainy, 75, 80, false, yes
rainy, 71, 91, true, no
sunny, 85, 85, false, no
sunny, 80, 90, true, no
sunny, 72, 95, false, no
sunny, 69, 70, false, yes
sunny, 75, 70, true, yes
```

Drzewa decyzyjne

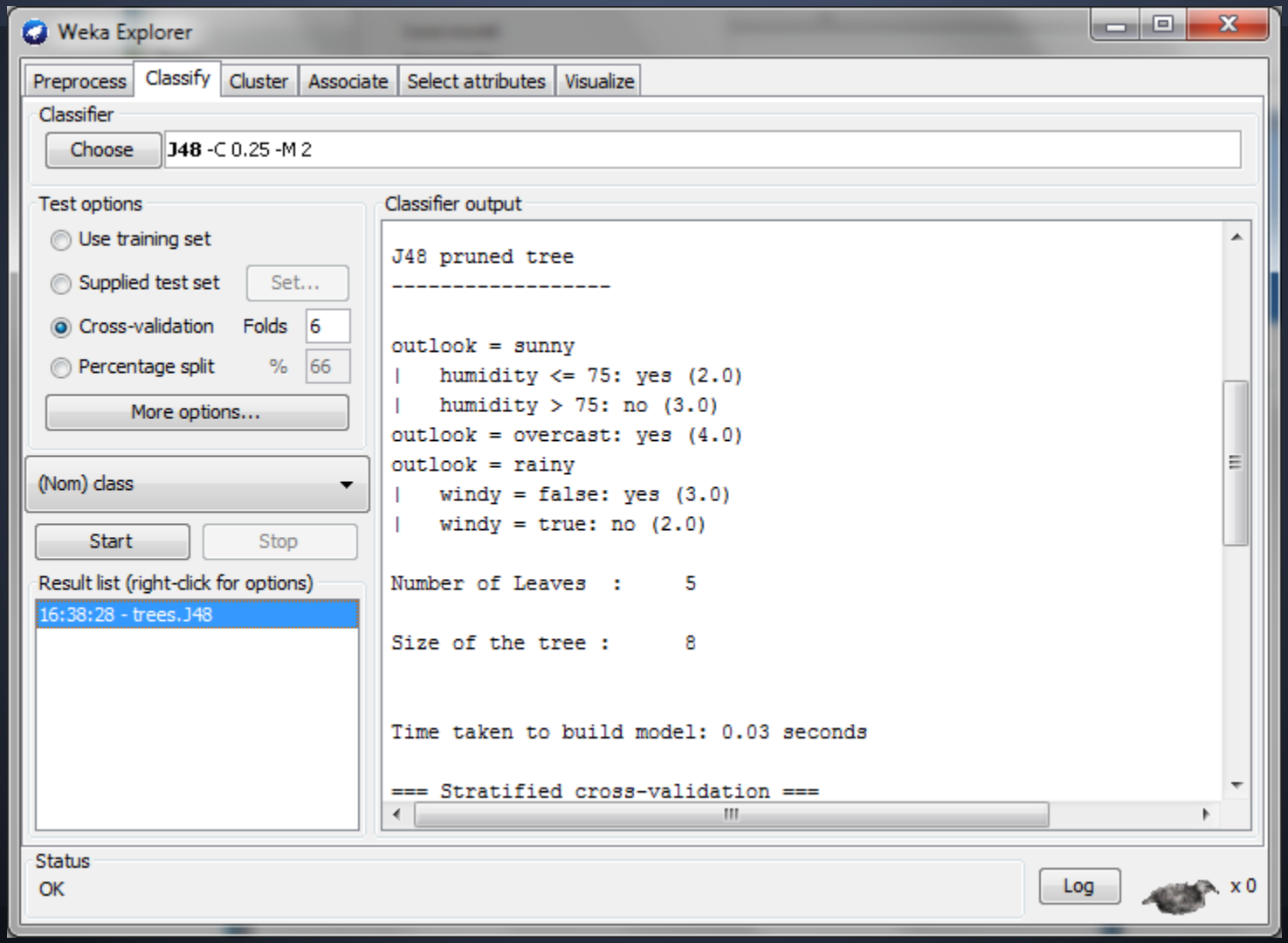
Zastosowanie środowiska Weka.

Weka to darmowy zestaw narzędzi, służący do analizy danych. Posiada graficzny interfejs, może być więc stosowany praktycznie przez każdego.

Podstawowymi algorytmami budującymi drzewa decyzyjne, są algorytmy CART i C4.5. Skorzystamy z tego drugiego, bo w przeciwieństwie do CART, nie buduje wyłącznie drzew binarnych.



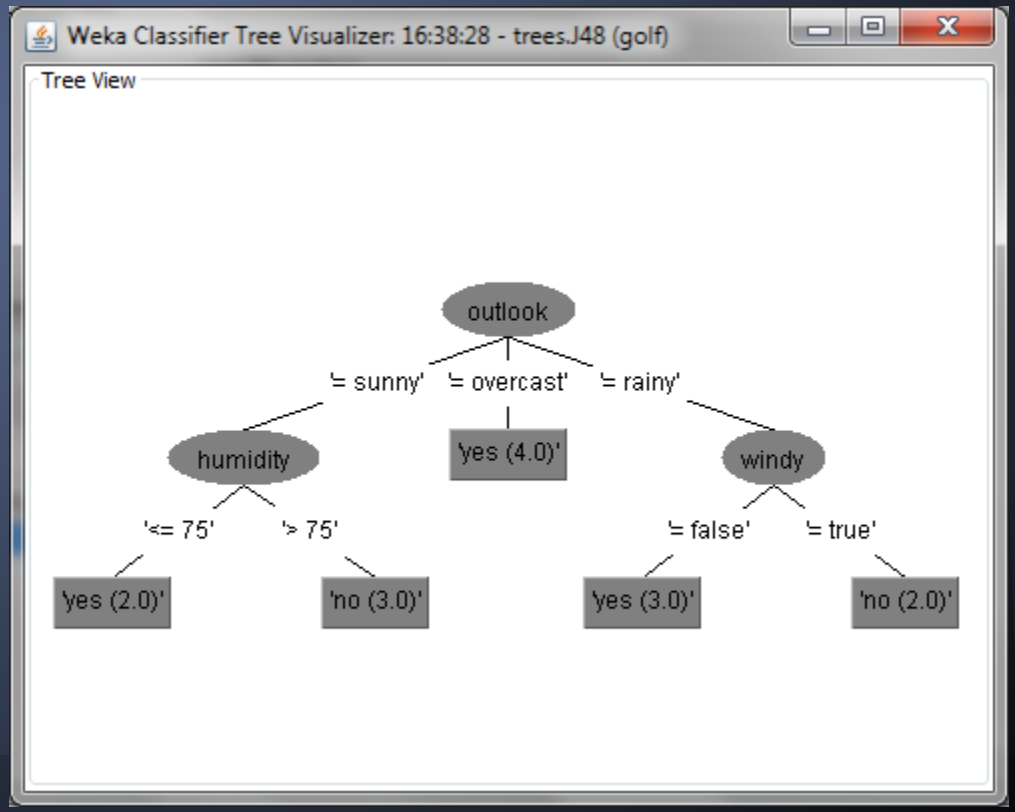
Drzewa decyzyjne



Drzewa decyzyjne

Wybraliśmy drzewo J48, które jest realizacją algorytmu C4.5 w środowisku Weka.

Zostało wygenerowane drzewo decyzyjne dla zbioru danych *golf*.



Grupowanie

Grupowanie (inaczej: klastrowanie) oznacza grupowanie danych w klasy podobnych obiektów.

Różni się od klasyfikacji tym, że nie ma zmiennej celu. Grupowanie nie próbuje klasyfikować ani przewidywać wartości zmiennej celu - zamiast tego, algorytm próbuje podzielić cały zbiór danych na stosunkowo zgodne grupy.

Algorytm *k*-średnich

Algorytm *k*-średnich (*k*-means) jest prostym i efektywnym algorytmem znajdującym grupy w zbiorze danych. Działa następująco:

1. Wybierz *k* (np. *k* = 3)
2. Wybierz losowo *k* początkowych środków grup (centroidów)
3. Dla każdego rekordu, znajdź który z *k* centroidów jest mu najbliższy i przypisz go do niego
4. Dla każdej z grup, znajdź nowy centroid (punkt będący najbliżej środka każdego z jej elementów)
5. Powtarzaj kroki 3-5 do zbieżności lub zakończenia

Algorytm *k*-średnich

Odległość między punktami, podobnie jak dla klasyfikacji, może być odległością euklidesową, a dane powinny zostać znormalizowane przed użyciem algorytmu.

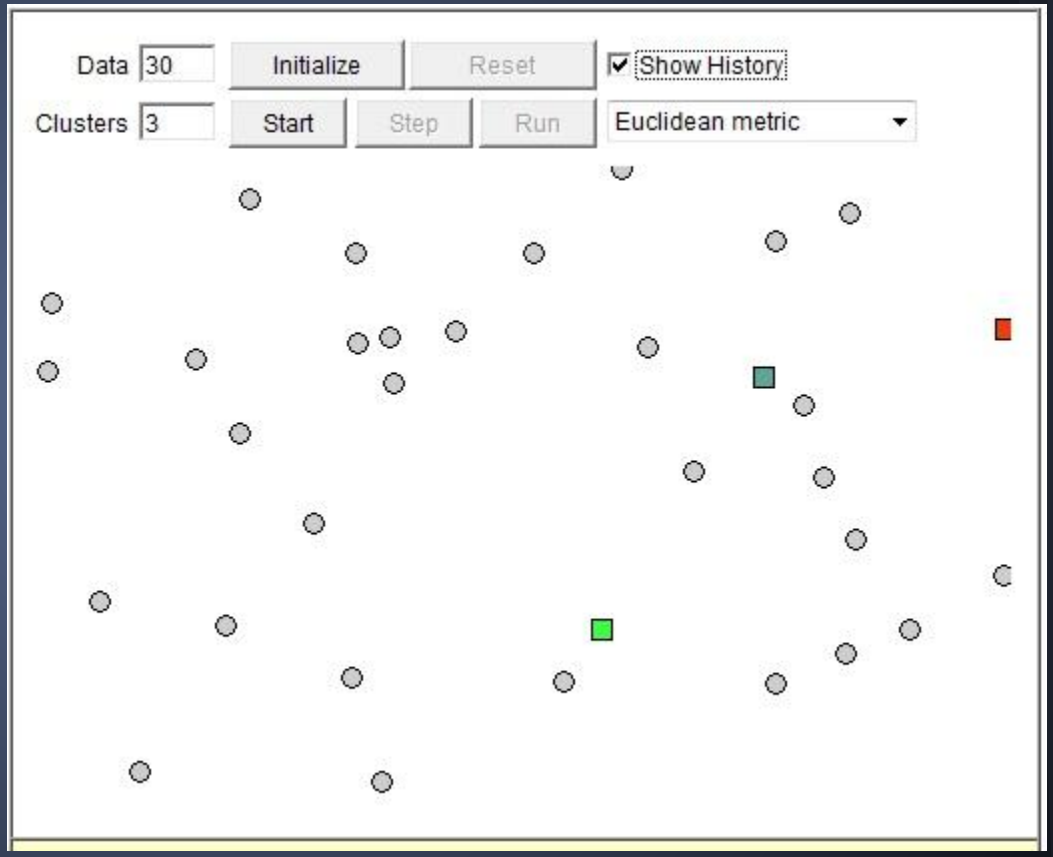
W kroku 4 algorytmu oblicza się środek ciężkości punktów grupy. Przykładowo, dla punktów (1,1,1), (1,2,1), (1,3,1), (2,1,1) nowym centroidem jest punkt (1,23, 1,75, 1).

Algorytm *k*-średnich

Przykład:

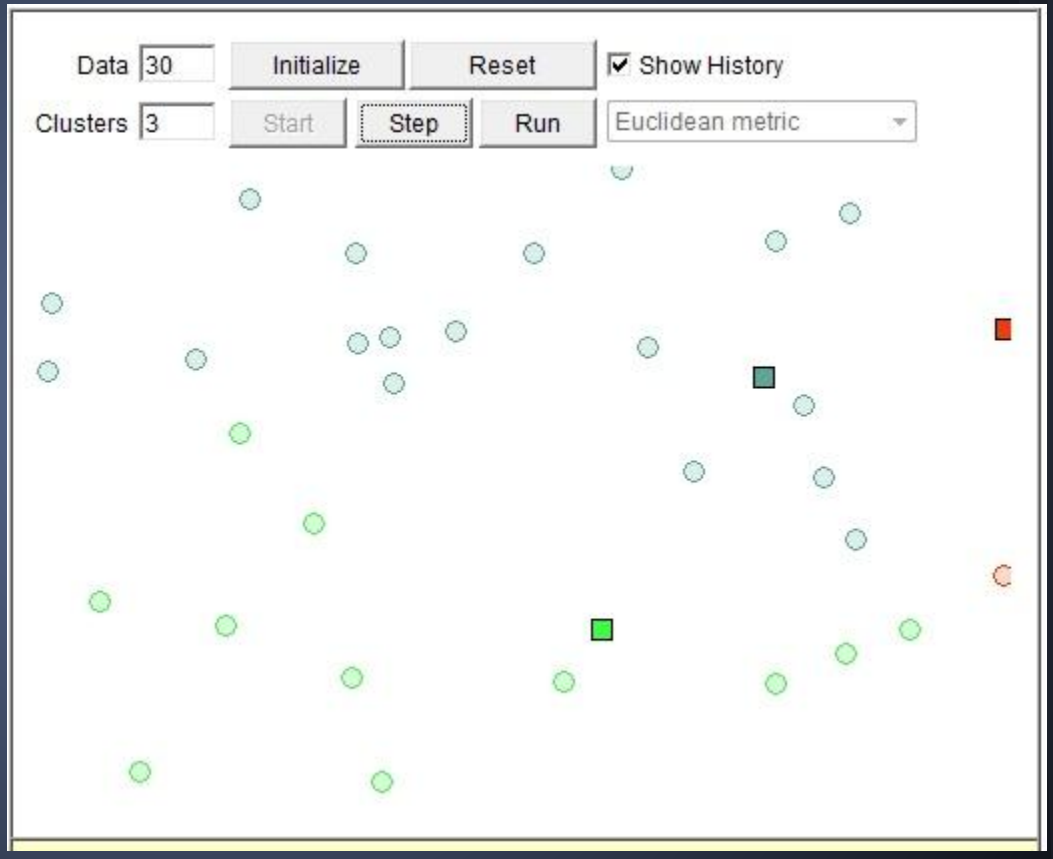
zbiór danych (szare kropki) składa się z 30 elementów. Każdy element posiada dwa parametry (oś X oraz Y).

Losujemy położenie *k* = 3 początkowych centroidów (kolorowe kwadraty).



Algorytm *k*-średnich

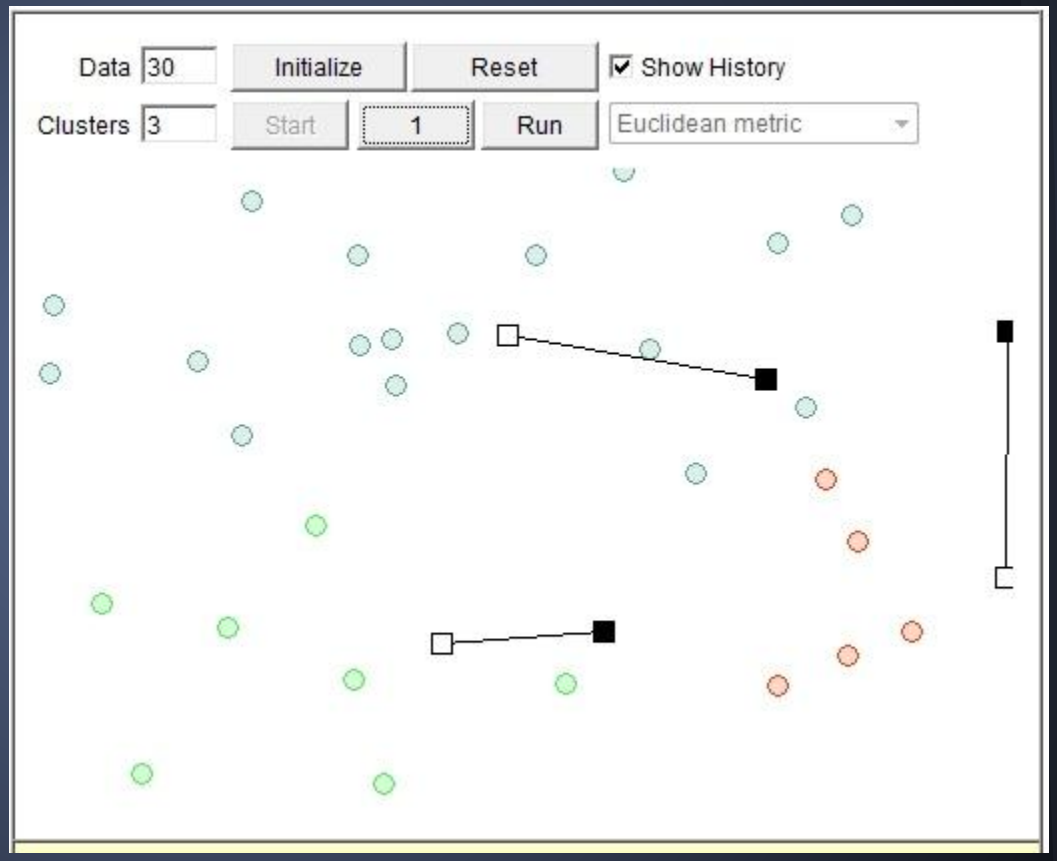
W pierwszym kroku przyporządkowujemy każdej kropce najbliższy centroid. Jak widać, grupa “czerwona” składa się na razie zaledwie z jednego elementu.



Algorytm *k*-średnich

Zgodnie z przewidywaniem, czerwony centroid “przykrywa” swój jedyny dotychczasowy punkt.

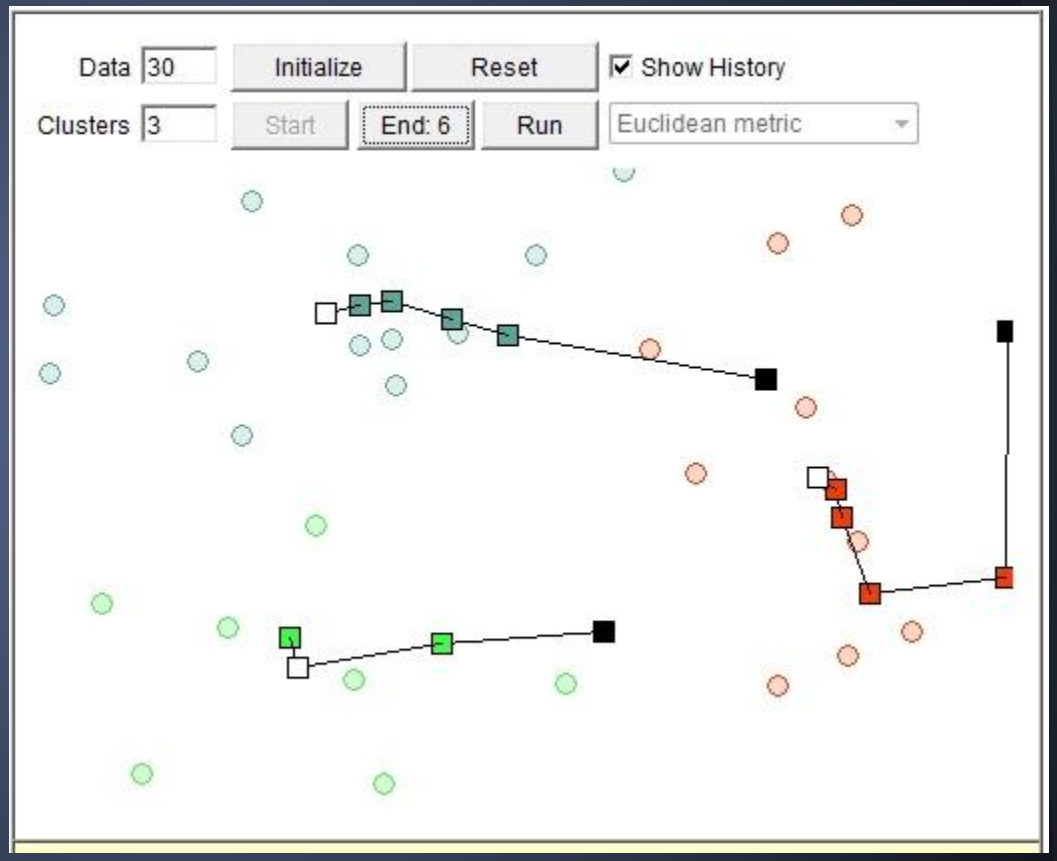
Jednakże, dzięki przesunięciu się środków grupy niebieskiej i zielonej, grupa czerwona “zyskuje” kolejne punkty.



Algorytm *k*-średnich

Po sześciu krokach, algorytm zakończył działanie. Ostateczne położenie centroidów symbolizują białe kwadraty.

Otrzymaliśmy trzy grupy, których elementy są do siebie podobne.



Algorytm *k*-średnich

Największym problemem w stosowaniu *k*-means jest wybór punktów początkowych. Z tego powodu, analityk powinien kilkakrotnie uruchomić algorytm z różnymi początkowymi centroidami. Można też, po wylosowaniu pierwszego położenia centroidu, starać się, by kolejne znajdowały się jak najdalej od niego.

Tak samo jak w algorytmie klasyfikacji *k*-nn, problemem jest też sam wybór *k*.

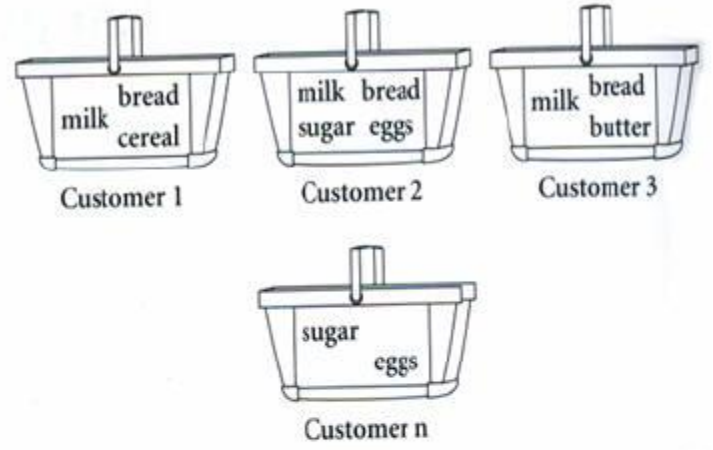
Odkrywanie reguł asocjacyjnych

Reguły asocjacyjne przypominają nieco drzewa decyzyjne. Tym razem jednak, algorytm nie ma z góry określonej prawidłowej odpowiedzi. Zamiast tego ma opisać wewnętrzne zależności między atrybutami.

Przykładowo, wiemy że na 1000 klientów 200 kupiło chleb, a z tych 200 - 50 kupiło piwo. Zatem odkryta reguła byłaby następująca: “Jeśli kupuje chleb, to kupuje piwo” ze wsparciem 200/1000 = 20% i ufnością 50/200 = 25%.

Odkrywanie reguł asocjacyjnych

Typowym przykładem zastosowaniem reguł asocjacyjnych jest analiza koszyka sklepowego. Sprawdzamy, jak wiele klientów którzy kupili towar A, kupili też towar B.



Tid	Towary
1	Bread, Milk
2	Beer, Diaper, Bread, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Bread, Diaper, Milk

$\Longrightarrow$

Przykłady reguł asocjacyjnych:

$$\{Coke\} \Longrightarrow \{Diaper\}$$
$$\{Diaper, Milk\} \Longrightarrow \{Coke\}$$
$$\{Milk\} \Longrightarrow \{Bread, Beer\}$$

Źródła

- Daniel T. Larose - *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*. PWN, Warszawa 2006.
- Przykład *k*-means: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- Reguły asocjacyjne: http://edu.pjwstk.edu.pl/wyklady/adn/scb/wyklad12/w12.htm
- Internet