STATISTICA
DANE ▸ WIEDZA ▸ SUKCES

# Krótki kurs data mining

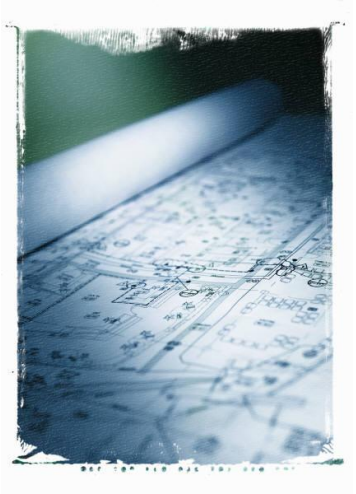
DANE ▸ WIEDZA ▸ SUKCES

StatSoft Polska, Kraszewskiego 36, 30-110 Kraków, www.StatSoft.pl, tel. 12 4284300. e-mail:info@statsoft.pl

# Informacje ogólne

- Co to jest data mining
- Wykorzystywane modele
- Metodyka
- Przegląd technik
- Podsumowanie

- Zastosowanie technik data mining staje się konieczne, gdy problemy, które się pojawiły, nie mają oczywistego rozwiązania:
  - Optymalizacja procesów wytwarzania produktu
  - Wykrywanie oszustw (fraud detection)
  - Ocena ryzyka
  - Segmentacja klientów



- Techniki data mining pomogą znaleźć potrzebne odpowiedzi.

# Czym jest data mining?

- Data mining jest procesem analitycznym zaprojektowanym do eksploracji dużych zbiorów danych w poszukiwaniu pewnych wzorców lub zależności między zmiennymi.
- Data mining jest procesem biznesowym, którego celem jest maksymalizacja wartości analitycznej zbieranych danych.

# Czym jest data mining?

- Wykorzystanie technik data mining
  - Wykrywanie prób oszustwa (bankowość, ubezpieczenia)
  - Wykrywanie wzorców zachowań
  - Modelowanie wzorców zachowań klientów pod kątem sprzedaży krzyżowej (cross-selling) i rozszerzonej (up-selling), a także pozyskiwania nowych klientów
  - Optymalizacja jakości i procesu wytwarzania produktu

- Data mining może być wykorzystany w każdej organizacji, dla której ujawnienie zależności ukrytych w danych przyniesie korzyści biznesowe

# Czym jest data mining?

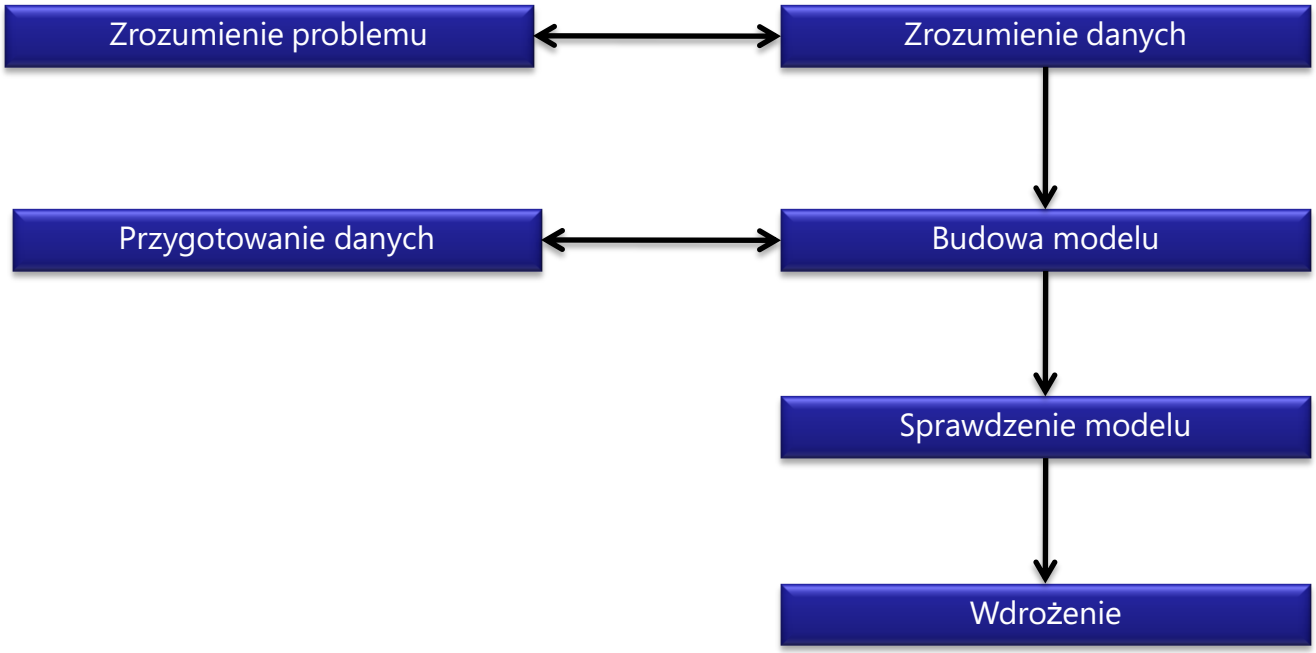
- Typowe cele projektów data mining
  - Identyfikacja grup, skupisk, warstw i wymiarów w danych, które nie wykazują się żadną wyraźną strukturą.
  - Identyfikacja czynników związanych z pewnym konkretnym wynikiem (np. przyznaniem kredytu).
  - Dokładne modelowanie zmiennych wynikowych (np. dla przyszłych klientów).

# Uwarunkowania

- Dane:
  - Przedsiębiorstwa gromadzą duże ilości danych np. na potrzeby sprawozdawczości, bilingów, prowadzenia procesów technologicznych itp.
  - Dostępne są narzędzia do zarządzania oraz udostępniania nawet bardzo dużych zbiorów danych
- Metody: istnieje wiele dopracowanych metod umożliwiających modelowanie i przewidywanie nawet bardzo złożonych zależności
- Oprogramowanie *STATISTICA Data Miner*
  - Kreatory prowadzące przez cały proces analizy (*Przepisy Data Miner*)
  - Wydajne narzędzia tworzenia i wdrażania modeli

# Modele data mining

- W literaturze poświęconej zagadnieniom data mining zaproponowano wiele sposobów podejścia do procesu zbierania i analizowania danych, wyciągania wniosków i wprowadzania w życie udoskonaleń.
- **CRISP** zaproponowany w latach 90 jako standardowy proces modelowania danych w technologii data mining



# Kroki w modelowaniu

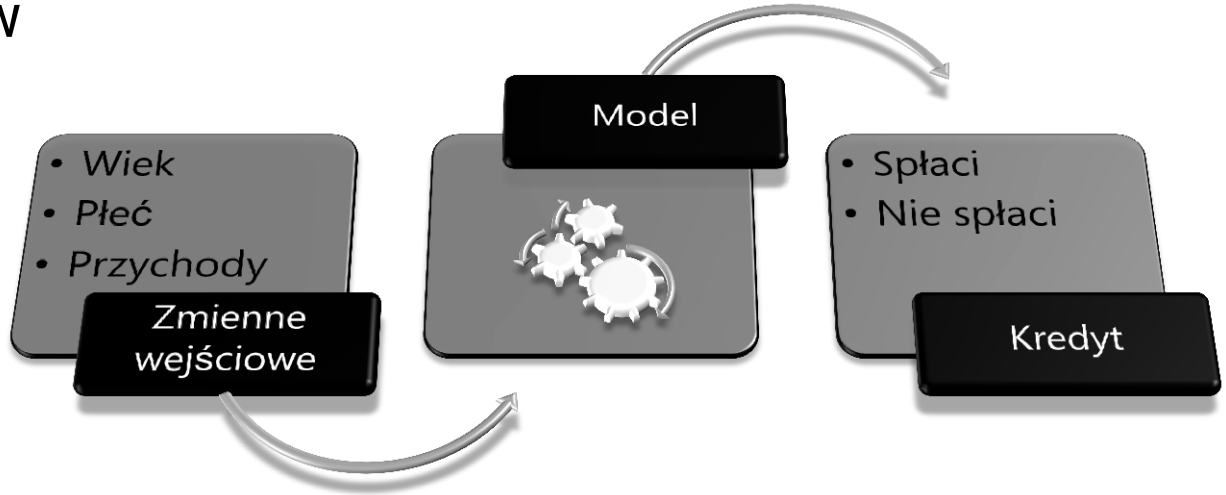
- Krok 0: Sprecyzuj istotę problemu
  - Przed uruchomieniem oprogramowania analitycznego i przeprowadzeniem analiz, wszystkie pytania, na które ma zostać znaleziona odpowiedź muszą być jasno sprecyzowane. Niedokładne sformułowanie problemu oznacza tylko stratę czasu i pieniędzy.
- Krok 1: Wstępna eksploracja
  - W tym kroku zwykle zaczyna się od przygotowania danych: uzupełnienia braków, usunięcia błędów, transformacji zmiennych, wydzielenia podzbiorów, a dla zbiorów z dużą liczbą cech – ich wstępnej eliminacji. Statystyki opisowe, korelacja i wykresy są podstawowymi narzędziami takiej analizy.
- Krok 2: Budowanie modeli i walidacja
  - W tym kroku budowane są rozmaite modele i wybierany jest ten z najlepszymi zdolnościami predykcyjnymi.
- Krok 3: Wdrożenie modelu
  - Celem projektu jest przewidywanie lub klasyfikacja (np. skoring kredytowy) nowych przypadków. Trzeci i ostatni krok zwykle wiąże się z wdrożeniem wybranego modelu lub modeli w celu otrzymania prognozy.

# Krok 1: Wstępna eksploracja

- Czyszczenie danych
  - Identyfikacja i usuwanie niepoprawnego kodowania danych
- Transformacje zmiennych
  - Logarytmowanie, transformacje Boxa – Coxa i inne
- Redukcja zbioru danych
  - Wybieranie do analizy tylko niektórych przypadków, a dla zbiorów z dużą liczbą zmiennych wstępna selekcja czynników
- Wizualizacja…
  - Korelacja, wykresy rozrzutu i inne
  - Statystyki opisowe informują nas o ważnych charakterystykach zmiennych, jak miara tendencji centralnej czy rozrzut
  - Zależności zwykle są lepiej widoczne na wykresie niż w tabelce z liczbami

# Krok 2: Budowanie modeli i walidacja

- Data mining ma na celu modelowanie rzeczywistości
- Z kilku zmiennych wejściowych chcemy otrzymać jeden lub kilka wyników




- Model może być łatwo interpretowalny, np. jako ciąg zdań „Jeżeli – To”, lub działać na zasadzie „czarnej skrzynki” (np. sieci neuronowe, gdzie zasady, na jakich została oparta predykcja, są bardzo trudne do interpretacji)

# Krok 2: Budowanie modeli i walidacja

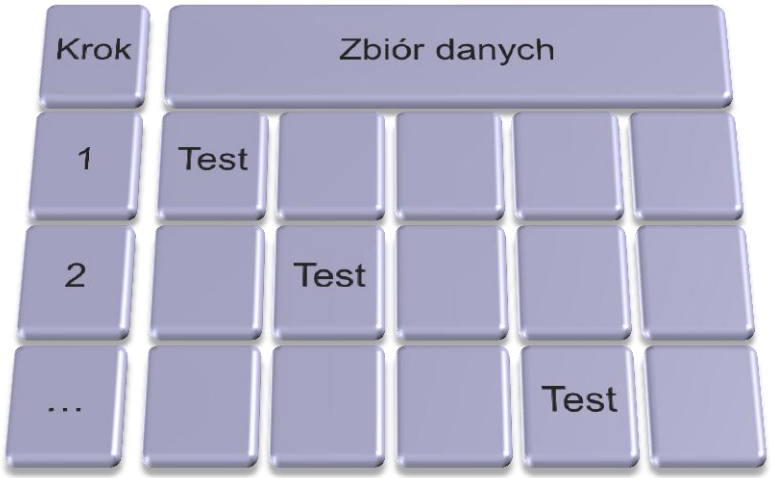
- Modele są zwykle oceniane w dwóch aspektach
  - Dokładność
  - Łatwość interpretacji
- Zdarza się, że nie da się uzyskać dokładnego modelu, który można łatwo interpretować
- Metody, takie jak drzewa decyzyjne czy regresja, są prostsze od modeli wykorzystujących drzewa wzmacniane czy sieci neuronowe i dlatego łatwo je interpretować, jednakże ceną za to może być obniżenie dokładności predykcji
- Modele data mining przybliżają rzeczywistość, nie są jej idealnym odbiciem i przy ich wdrażaniu i stosowaniu należy o tym pamiętać

# Krok 2: Budowanie modeli i walidacja

- Walidacja modeli wymaga, aby były one tworzone na jednym zbiorze (uczącym), a sprawdzane na innym (testowym)
- Typowe metody walidacji modeli:
  - Podział początkowego zbioru danych na dwa podzbiory – np. w stosunku 75% do 25%
  - Jeżeli danych jest za mało, aby je podzielić na podzbiory, należy wykorzystać v-krotny sprawdzian krzyżowy

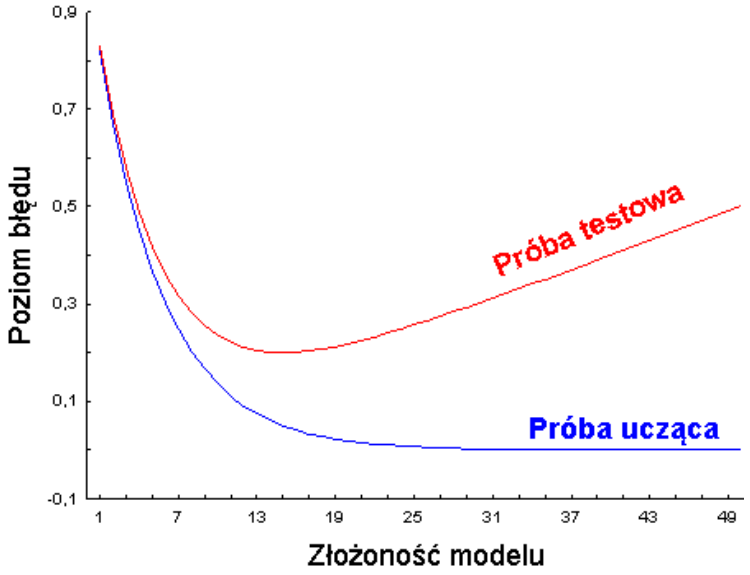
# Krok 2: Budowanie modeli i walidacja

- V-krotny sprawdzian krzyżowy




- Podziel zbiór danych na podzbiory
- Wybierz jeden z podzbiorów jako uczący, a na pozostałych zbuduj model
- Powtórz całą procedurę, wybierając inny podzbiór jako testowy

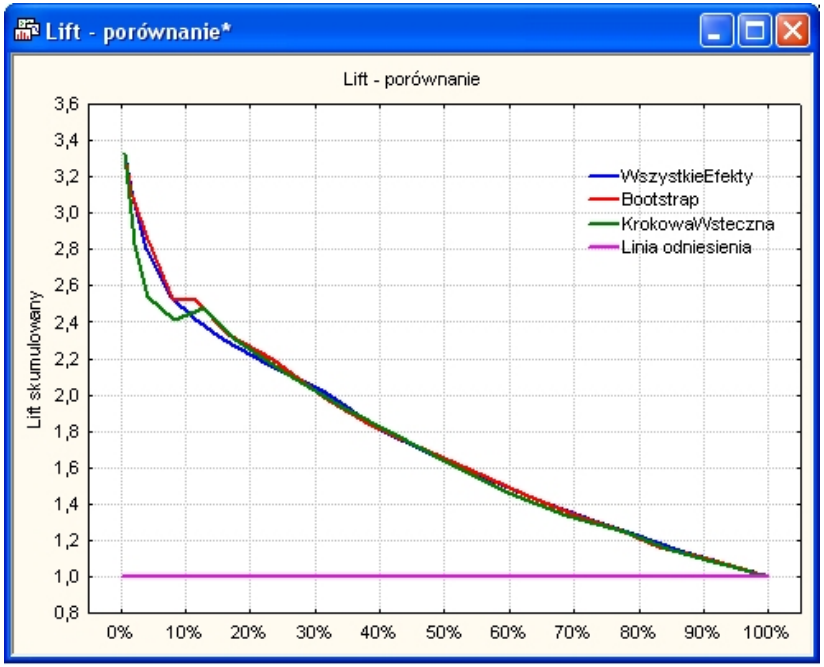
# Krok 2: Budowanie modeli i walidacja




- Ocena jakości predykcji:
  - Regresja – suma kwadratów reszt
  - Klasyfikacja – macierz pomyłek
- Poziom błędu na zbiorze uczącym nie jest dobrym wskaźnikiem zdolności modelu do generalizacji (przewidywania).
- Przeuczenie modelu – zbyt dobre dopasowanie modelu do zbioru uczącego prowadzi do zmniejszenia dokładności predykcji.

# Krok 2: Budowanie modeli i walidacja

- Miary dopasowania modelu:
  - Dokładność klasyfikacji
  - Całkowity Zysk/Koszt błędów
  - Krzywe Lift i Gain
  - Błędy prognoz liczbowych
  - Proporcja

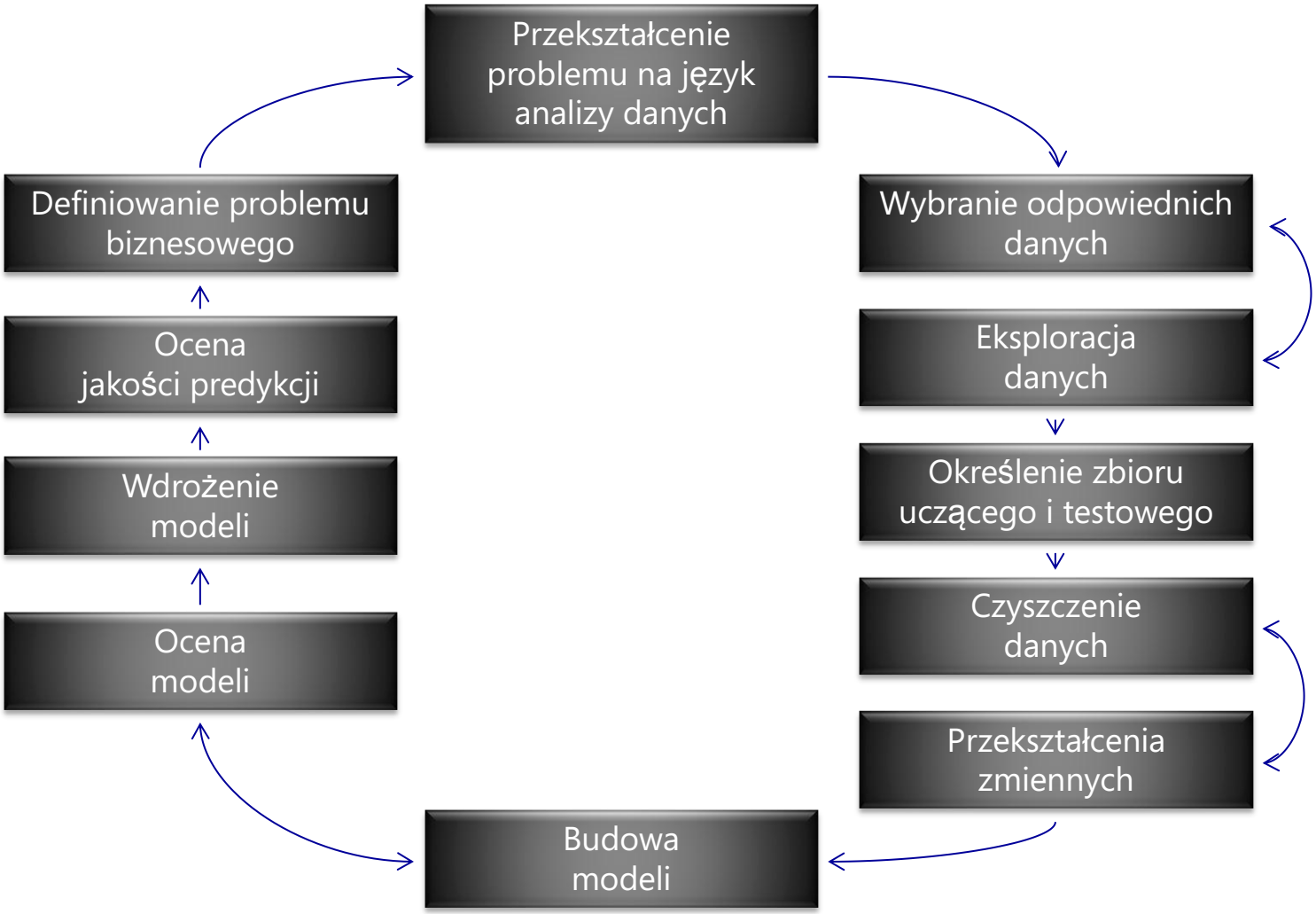


# Krok 3: Wdrożenie modelu

- Raz zbudowany model może być wykorzystywany przez dowolnie długi okres czasu

- Model powinien być łatwy do wdrożenia
  - Regresja liniowa – wystarczy znać współczynniki
  - Drzewa klasyfikacyjne – zestaw reguł Jeżeli – To

# Metodyka data mining



# Przegląd technik data mining

- Metody z nauczycielem:
  - Klasyfikacja – jakościowa zmienna zależna.
  - Regresja – ilościowa zmienna zależna.
  - Szeregi czasowe.
  - Optymalizacja.
- Metody bez nauczyciela:
  - Odnajdywanie ukrytych wymiarów i redukcja liczby zmiennych (analiza głównych składowych).
  - Segmentacja – grupowanie podobnych obiektów.
  - Analiza sekwencji, asocjacji i połączeń – pomaga wyjaśnić zależności między zmiennymi, jednoczesnego występowania zdarzeń (osoba kupująca młotek kupuje także gwoździe).

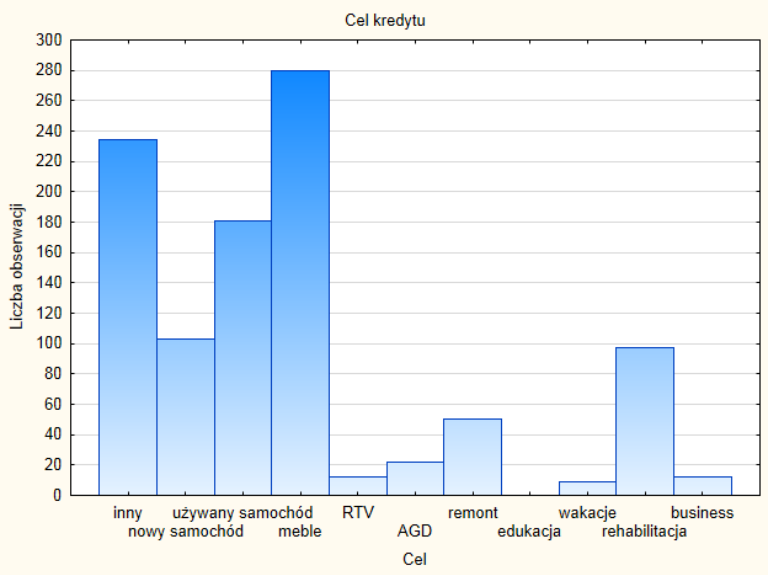
# Przegląd technik data mining

- Na następnych kliku slajdach zostaną krótko omówione techniki data mining
  - Statystyki opisowe
  - Regresja liniowa i logistyczna
  - Analiza dyskryminacyjna
  - Drzewa decyzyjne
  - Techniki segmentacji (k-średnich & EM)
  - Sieci neuronowe
  - Analiza asocjacji i połączeń

# Statystyki opisowe

- W klasycznym problemie data mining ilość zgromadzonych danych jest zbyt duża, aby można było na ich podstawie wyciągać wnioski. Pierwszym krokiem jest utworzenie dużo mniejszego zestawu cech, charakteryzującego zbiór danych.
- Do tego celu zwykle wykorzystuje się dwie miary charakteryzujące daną cechę
  - Miary tendencji centralnej – średnia arytmetyczna, moda, mediana.
  - Miary rozrzutu – odchylenie standardowe, wariancja.
- Przedstawienie danych w postaci graficznej jest jednym z obowiązkowych kroków, większość zależności o wiele łatwiej zobaczyć na wykresie niż odczytać z tabelki z liczbami

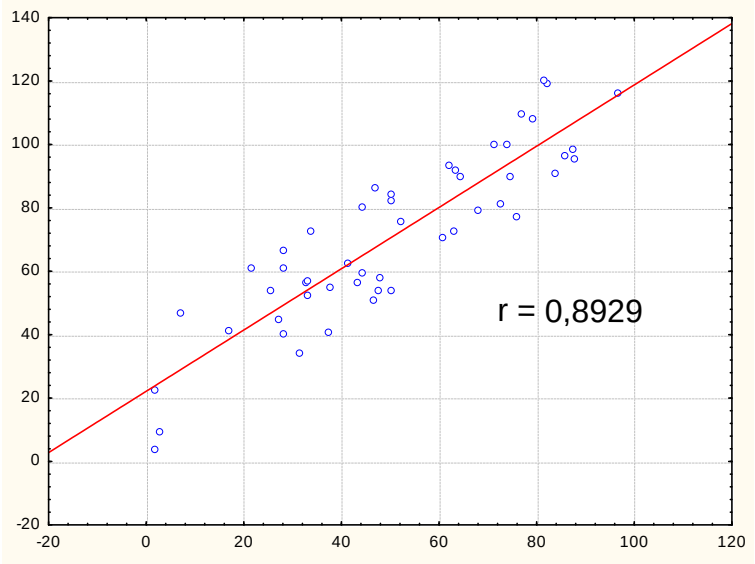
- Histogram jest prostym i efektywnym sposobem wizualizacji informacji zawartych w danej zmiennej.
- Z tabelki szybko odczytamy średnią, wariancję czy najmniejszą i największą wartość danej cechy.

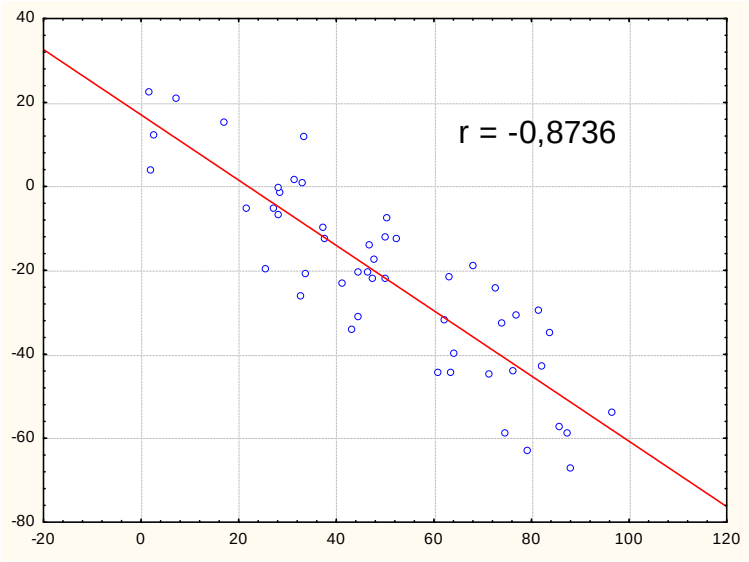



| Zmienna | Statystyki opisowe (CreditScoring.sta) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Średnia | Mediana | Moda | Liczność Mody | Minimum | Maksimum | Wariancja | Odch.std |
| Kwota | 4579,7 | 3247,3 | Wielokr. | 3 | 350,0 | 25793,6 | 15617138 | 3951,9 |

# Statystyki opisowe

- Korelacja mówi o liniowym związku między zmiennymi



# Regresja liniowa i logistyczna

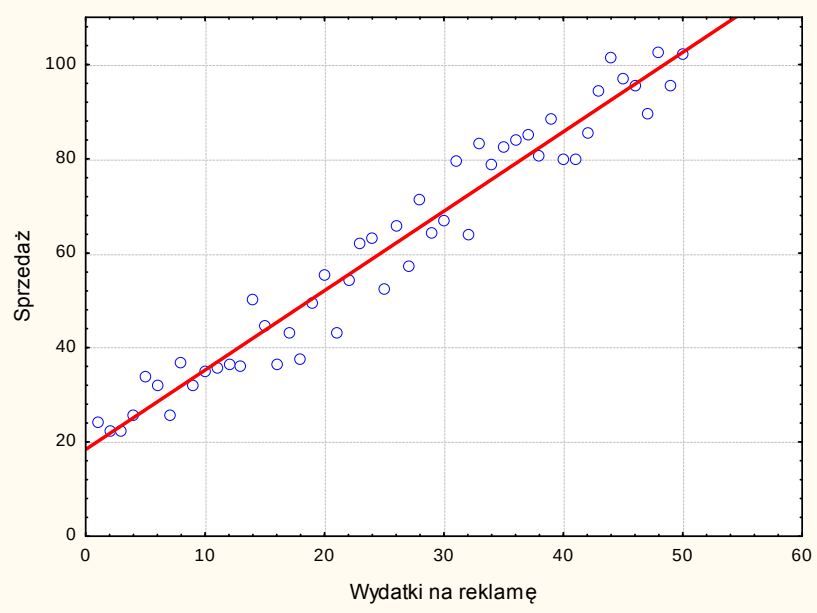
- Analiza regresji to zestaw metod statystycznych, które badając zależności między dwiema lub większą liczbą zmiennych, pozwalają przewidzieć wartości danej zmiennej, gdy znane są pozostałe
- Rodzaje regresji
  - Liniowa
  - Logistyczna
- Przykłady zastosowania
  - Przewidywanie sprzedaży produktu z wykorzystaniem zależności Sprzedaż <=> Wydatki na reklamę
  - Przewidywanie wydajności pracownika na podstawie informacji o jego wykształceniu i wyników testów badających różne zdolności
  - Przewidywanie liczby słów, jaką będzie się posługiwało dziecko w danym wieku

# Regresja liniowa

- Najprostszy model regresji zawierający tylko jedną zmienną objaśniającą jest postaci:

$$Y = \beta_0 + \beta_1 X$$

- Wartości współczynnika nachylenia i przecięcia z osią oY są tak dobierane, aby suma kwadratów odległości od otrzymanej prostej była jak najmniejsza.



# Regresja liniowa

- Niektóre działy personalne używają regresji wielorakiej, aby określić poziom wynagrodzenia. Analitycy przeprowadzają ankietę na temat wysokości zarobków, wykształcenia i umiejętności w porównywalnych firmach i wśród określonych grup zawodowych. Zebrane informacje mogą posłużyć do zbudowania modelu, np.:

$$Pensja = 0.5 \cdot (Ilość\ odpowiedzialności) + 0.8 \cdot (Liczba\ podwładnych)$$

- Nie wszystkie zależności mają charakter liniowy. Co wtedy zrobić?
  - Zwiększyć liczbę zmiennych zależnych.
  - Wykonać transformację danych – często udaje się sprowadzić problem do postaci liniowej.
  - Podnieść niektóre zmienne do wyższej potęgi, lub dodać interakcje między nimi.
- Podejście zaproponowane w ostatnim punkcie wymaga jednak sporej wiedzy i doświadczenia; techniki, które przedstawimy, wykonają sporą część pracy za nas.

# Regresja logistyczna

- Często zdarza się tak, że zmienna zależna, która nas interesuje, przyjmuje tylko dwie wartości, np. Tak/Nie – 1/0. Do jej modelowania moglibyśmy wykorzystać regresję liniową, ale rodzi to kilka problemów.
  - Przewidywane wielkości mogą przyjmować wartości większe od 1 i mniejsze niż 0.
  - Rozrzut zmiennej zależnej jest różny dla różnych wartości zmiennej niezależnej
  - Test istotności współczynników regresji korzysta z założenia normalności reszt, ale skoro zmienna zależna przyjmuje tylko dwie wartości 0 i 1, nawet asymptotycznie to założenie jest nie spełnione, a co za tym idzie istotność współczynników regresji jest mocno niepewna.
- Regresja logistyczna umożliwia zbudowanie poprawnego modelu, który jednocześnie może być interpretowany podobnie jak regresja liniowa

# Regresja logistyczna

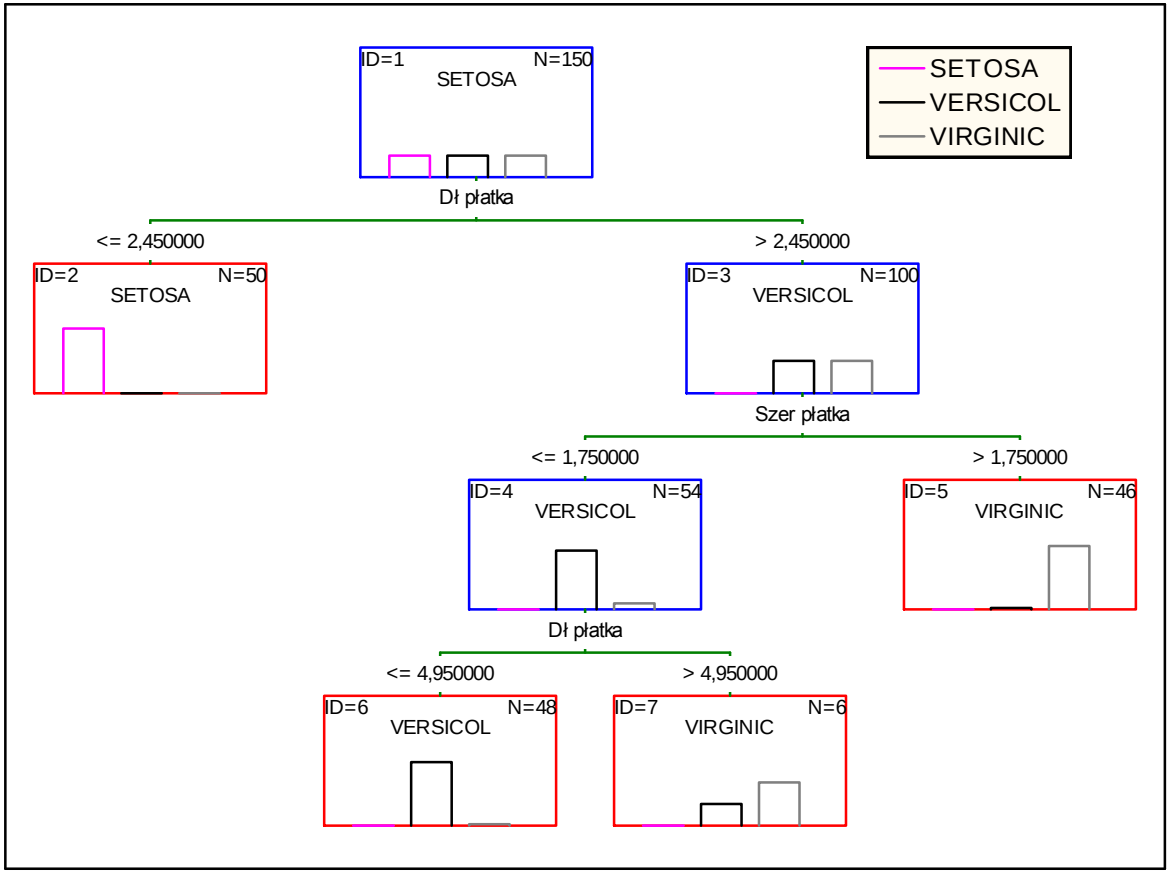
- Badanie skłonności klienta do niespłacenia kredytu (skoring kredytowy) w zależności od wieku, przychodów, stanu konta i innych czynników.
- Badanie skłonności klienta do rezygnacji (churn) w zależności od częstotliwości korzystania z usług, stażu klienta, liczby usług, z których korzysta.
- Badanie czynników wpływających na wystąpienie choroby wieńcowej
- Znajdowanie klientów, którzy najprawdopodobniej odpowiedzą na ofertę (cross-selling, up-selling)

# Analiza dyskryminacyjna

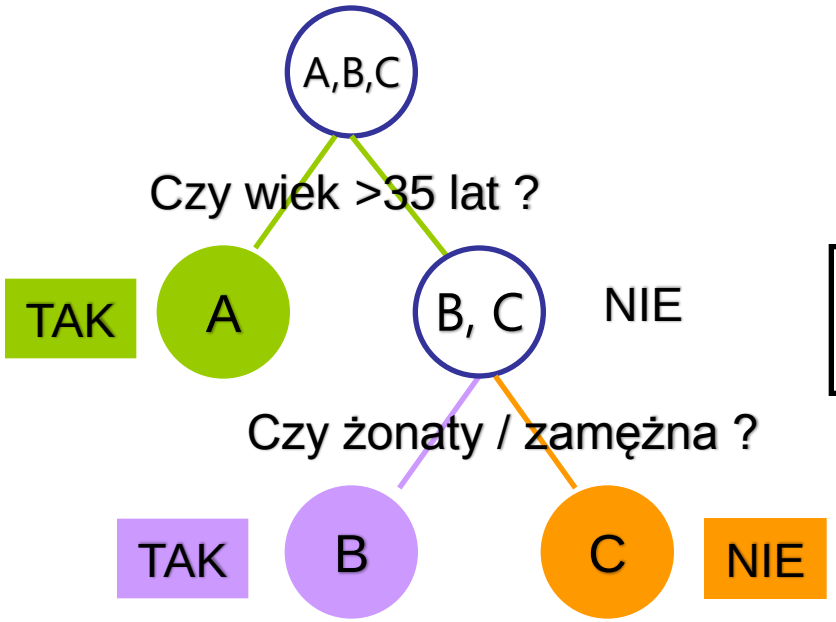
- Analiza dyskryminacyjna służy do przewidywania, do której klasy trafi obiekt o pewnych cechach, np. możemy mieć *Niskie*, *Średnie* albo *Wysokie* ryzyko kredytowe.
- Model uzyskany metodą analizy dyskryminacyjnej jest łatwy w interpretacji i do zastosowania dla nowych obiektów.
- W wyniku analizy dyskryminacyjnej uzyskujemy model liniowy: jest to szybka metoda, ale nie jest w stanie uwzględnić bardziej skomplikowanych, nieliniowych zależności.
- W przypadku zadania z dwoma klasami (np. *Tak / Nie*) na ogół lepsze wyniki daje regresja logistyczna.

# Drzewa decyzyjne

- Modele budowane za pomocą drzew decyzyjnych mogą służyć do przewidywania zarówno zmiennych jakościowych, jak i ilościowych.
- Drzewa klasyfikacyjne są łatwe do interpretacji – otrzymujemy zestaw reguł „Jeżeli – To”.
- Nie dokonujemy założeń co do natury związku łączącego predyktory ze zmienną zależną.



# Drzewa – interpretacja liści




Jeżeli wiek > 35 lat, to klasa A

Jeżeli wiek 35 albo mniej i żonaty, to klasa B

Jeżeli wiek 35 albo mniej i nieżonaty, to klasa C

# Analiza skupień

- Pojęcie analizy skupień (ang. *cluster analysis*, termin wprowadzony w pracy Tryon, 1939) obejmuje faktycznie kilka różnych algorytmów klasyfikacji.
- Ogólny problem badaczy wielu dyscyplin polega na organizowaniu obserwowanych danych w sensowne struktury lub grupowaniu danych. Na przykład biolodzy zanim będą mogli sensownie opisywać różnice między zwierzętami, muszą klasyfikować je ze względu na gatunki.
- Analiza skupień ma olbrzymie znaczenie w badaniach marketingowych, w których zachowanie klienta jest przewidywane ma podstawie pewnego zestawu cech.

# Segmentacja

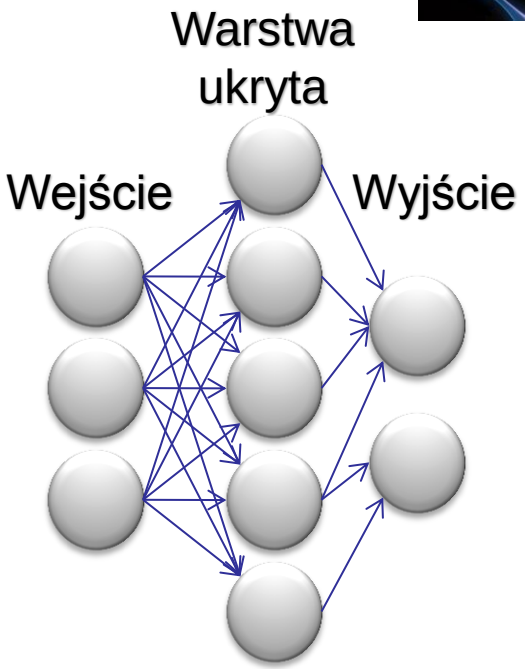
- Zastosowania segmentacji:
  - Tworzenie kampanii marketingowych, usług i produktów skierowanych do konkretnej grupy ludzi.
  - Znajdowanie podobnych, konkurujących ze sobą produktów (np. modeli samochodów).
  - Identyfikacja podobnych cykli procesów produkcyjnych.

# Sieci neuronowe

- Umożliwiają modelowanie praktycznie dowolnych zależności.
- Szerokie możliwości zastosowania:
  - Regresja – sieć MLP (perceptron wielowarstwowy) i RBF (o radialnych funkcjach bazowych).
  - Klasyfikacja – MLP, RBF.
  - Analiza skupień – sieć SOM Kohonena.

# Sieci neuronowe

- Wadą są bardzo ograniczone możliwości interpretacji otrzymanych wyników; nie otrzymamy zestawu prostych i jasnych reguł postępowania, jak to ma miejsce w drzewach decyzyjnych.



# Analiza sekwencji asocjacji i połączeń

- "Problem koszyka" oznacza następującą sytuację: jest bardzo dużo produktów, które mogą być kupowane przez klientów w pojedynczych transakcjach lub w sekwencji kolejnych transakcji rozłożonych w czasie. Mogą to być na przykład towary na półkach w sklepie wielkopowierzchniowym oferującym asortyment od pietruszki po sprzęt elektroniczny lub pakiety ubezpieczeniowe dostępne w ofercie towarzystwa ubezpieczeniowego, itp. Klienci wkładają do koszyka bardzo mały ułamek tego, co jest dostępne w ofercie.
- Reguły asocjacji. W bazie transakcji można wyszukać reguły asocjacji mówiące, które artykuły kupowane są często równocześnie. Na przykład może się okazać, że kupno latarki często łączy się z zakupem baterii, tzn. baterie i latarki są często w jednym koszyku. Jeśli każda transakcja ma swój stempel czasowy, analityk może też szukać typowych sekwencji zakupów.

Podsumowanie reguł asocjacji (Fastfood.sta)
Min: wsparcie= 20,0%, zaufanie = 10,0%
Maks. liczność zestawu = 10

| | Poprzednik | ==> | Następnik | Wsparcie% | Zaufanie(%) | Przyrost |
|---|---|---|---|---|---|---|
| 1 | HAMBURGR | ==> | PIZZA | 39,0 | 68,42 | 0,9 |
| 2 | PIZZA | ==> | HAMBURGR | 39,0 | 56,52 | 0,9 |
| 3 | HAMBURGR | ==> | PIZZA, PŁEĆ==MĘSKA | 32,5 | 57,02 | 0,9 |
| 4 | PŁEĆ==MĘSKA | ==> | PIZZA, HAMBURGR | 32,5 | 39,63 | 1,0 |
| 5 | PŁEĆ==MĘSKA, HAMBURGR | ==> | PIZZA | 32,5 | 69,15 | 1,0 |
| 6 | PIZZA | ==> | PŁEĆ==MĘSKA, HAMBURGR | 32,5 | 47,10 | 1,0 |

# Podsumowanie

- Dzięki technikom data mining możliwe jest wykrycie zależności ukrytych w danych (często zbieranych w innych celach) i przełożenie ich na konkretne decyzje biznesowe, badawcze lub technologiczne.
- Tam gdzie zawodzą tradycyjne metody analizy danych, często sprawdza się data mining
- Metodyki data mining to przepis na przeprowadzenie udanego projektu
- Dostępne jest wydajne oprogramowanie umożliwiające praktyczną realizację data mining
- Data mining to narzędzie, nie czarodziejska różdżka
- Trzeba dobrze rozumieć co znaczą dane i skąd się wzięły, żeby poprawnie interpretować wyniki
- Więcej informacji na portalu data mining:

http://statsoft.pl/datamining.html