**Mod CRISP-DM** to standard proces eksplor dan w biznes.Najpierw identyfi się cele biznes i analiz dostępne dane pod kątem ich struktury i jakości. Nast dane są przygotowy poprzez transfor i czyszczeni. Kolejn krok to bud model eksploracyj, które nast są ocenian pod kątem zgodnoś z założeniam. Ostat etap jest wdroż odkryt wiedzy w praktyce biznes.Def **podobieństwa string** określa różnice między teks. Metryk Hamminga(równej dł.) liczy różn w znakach, np. „pies” → „piec” (1 zmiana), a odl Levenshteina(różnej dł.)mierzy operac potrzebne do przekształc tekst, np. „pies” → „piasek” (3 oper).Wyszukiw wzorca w string polega na znalezi wzorca w tekst, np. „wys” w „wsiadło”.Grupow tekst klasyfiku dokumenty według treści, np. „sport”, „nauka”. Wymaga analiz cech i algor jak k-means.Przetwarzani jęz naturalnego (NLP) pozwala maszynom analiz tekst, dzielić go na słowa, rozpoznawa części mowy i interpret znaczenie w kontekście. **Algo naiwny wyszukiw wzorca w łańcuchu tekst: Wej:** s - łańcuch znakowy, p – łańcuch wzorca, **Wyj:** Wszystkie pozycje wzorca p w łańcuchu s.**Elem pomoc:** i : pozycja okna, i ∈ N; n : dł łańcucha s; n ∈ N; m : dł wzorca p; m ∈ N. **List kroków:** 1. n ← |s| ; obliczamy dł łańcucha s. 2. m ← |p| ; obliczamy dł wzorca p. 3. Dla i = 0, 1, ..., n - m wykonuj krok 4: 4. Jeśli p = s[i : i + m] to wypisz i ; znalez wzorzec na pozycji i. 5. Zakończ.Metryk Hamminga mierzy różn między dwoma ciągami o tej samej dł, określając liczb pozycji, na których się różnią. Przykład: dist(„pies”, „piec”) = 1.Odl Levenshteina określa min liczbę operacji (wstawienie, usunięcie, zamiana znaku) potrzebnych do przekształcenia jednego tekstu w drugi. Przykład: dist(„pies”, „piasek”) = 3.Algo Knutha-Morrisa-Pratta (**KMP**) efektywna, nienaiwna metod wyszukiwania wzorca w tekście, eliminująca zbędne porównania. Działa w czas liniowym O(n + m), n-dł tekstu,m-dł wzorca.Kluczowy elementem to tablica "przesunięć" (prefix table), przechowująca informacje o powtarzających się prefiksach i sufiksach wzorca.Kroki:Przygotowanie tablicy prefiksów– określa, jak daleko można przesunąć wzorzec w razie niedopasowania. Dopasowan wzorca do tekstu–algo porównuje znaki, w przypadku niedopasowania przeskakuje wzorzec zgodnie z tablicą przesunięć, unikając zbędnych porównań. Np.Tekst:"abababcabababc" Wzorzec:"ababc"Tablica przesunięć:[0,0,1,2,0]Po pierwsz liter. a brak przesun; po ab najd prefiks to a; po aba najd prefiks to ab; itd.Algorytm porównuje znaki i przesuwa wzorzec zgodnie z tablicą, pomijając powtórne sprawdzanie pasujących liter.

**Problem rekomend**-sugerow użytk produktów, usług lub treści, które mogą ich zainteres, na podst dostęp danych i preferencji.Stosuje się w e-commerce (Amazon), serwisach streamingowych (Netflix) i mediach społecz.Item-to-item–rekomendow produktów podobnych do tych, które użytk już wybrał.Np.Kupiłeś X, może zainteres Cię także Y.User-to-user-rekomend na podst preferencji innych klient o podobn wyborach.Np.Os, które kupiły X, często kup także Y.Hybrydowe-łączy oba podejść.**Baza trans**-zbiór transak składających się z różn elementów(np. produktów).Zbiór częst-podzbiór elementów, który pojawia się w transakc co najm określoną liczb razy (min\_sup). Reguła asoc-zależność w bazie transak w formie „jeżeli T’, to T”, gdzie T i T’ to rozłącz podzbiory. Odkrywa się je poprzez analiz zbiorów częst i przekształc ich w reguły. Apriori-znajdowanie częst zbior elemen w baz transak.Dane wejśc: baz tran nad zbior T, min\_sup (min wsparcie).Dane wyjś: Frequentitemsets–maks zbior częst.Kroki:1.Wyznaczenie zbiorów 1-elem–znajdujemy poj elem spełniające próg min\_sup.2. Generow większych zbior–łączym częst k-elem zbior w k+1-elem.3.Filtrowanie – usuw zbiory niespełniaj min\_sup.4. Aktualizacja – dodajemy nowe częst zbior i usuw mniejsz podzbior.5.Powtarzanie-zwiększamy k i kontynuujemy, aż nie można utworz większ zbior.**Algor indukc reguł asocja** służ do wyodrębniania reguł z częst zbior transak.Dane wej: baz trans(B) i zbiór częst(ZC) spełniający próg min\_sup.Dane wyj: zbiór reguł asocja(R) spełniających min\_sup i min\_conf.Kroki:1.Inicjalizacja– ustawie początk liczb elem(z = 2) i pust zbior reguł.2.Generow reguł–tworz reguł z częst zbiorów, obliczenie confidence i odrzucenie niespełniających min\_conf.3.Uogólnianie–scalanie reguł o wspóln poprzednik i usuw mniej ogóln.4.Powtarz procesu dla dłuższ zbior częst.5.Zakończen–zwrócenie zbioru maks ogólnych reguł. **Standary**-przeskalowanie wart atrybutów tak,by miały śred 0 i odchyl standard 1, co ułatwia analiz danych.Normaliz- przekształc wart atrybutów do przedzi[0,1]lub innego zakresu, co pozwala na porównyw różnych cech.Dyskretyz-zamienia wart atrybutów na przedzi, zmniejszając liczb unikaln wart i upraszczając analizę.**Dyskretyz nienadzorowana**–metoda przedz równ dł dzieli wartości atrybutu na n równ przedz.Dane wej: przedz wartości atrybutu[m, M]oraz liczba przedz n.Dan wyj: zbiór D zawierający podział przedz[m, M]na równ częś.Kroki:1.Inicjali pust zbior D.2.Oblicz dł przedz d=(M-m)/n.3.Ustawien końca przedz k=m.4.Iterac od 1 do n-1:-Ustalen początk przedz p=k.-Przesunięc końca przedz k=k+d.-Dodan przedz[p, k]do zbior D.5.Dodan ostatnieg przedz[k, M]do zbior D.Wynikiem algor jest podział zakre wartości atrybutu na n równych przedz.

**Nadzor**–metoda podz wart atrybutów tak,by obiekty z różn decyzj miały różne wart.Dan wej:system decyz z atryb a, b oraz etykiet decyzyjnymi.Dan wyj: system z dyskretyzowanymi wart atrybutów.Kroki:1.Sort wart atrybutu a.2.Identyfi miejsc, gdzie wart decyzyjne się zmieniaj.3.Wyznaczeni granic przedz, by różne decyz miały różne zakres wart.4. Przypis wartościom atrybutu odpowied przedz.5.Powtórz proces dla atryb b. Wynik jest podział wart atrybutów na przedz, które minimalizują powiązan między różn decyz w tych samych zakresach wart.**wart brakujące1**.Usuw brakujących danych–metoda polega na usun wierszy lub kolumn z brakującymi wart. Stosow, gdy braki stanowią niewielki % danych.2.Zastępow braków wybranymi wart–polega na uzupełnianiu brakujących danych na kilka sposobów: Użyc najczęstszej wart lub średn;Zastąp wartościami z tej samej klasy decyz;Wykorzyst algor 1NN, który szuka najbardz podob obiektów.3.Omijanie braków–podział danych na podzbior bez brakujących wart, analiz ich osobno, a nast scalanie wyników.Metod te pozwal na minimalizowanie utrat informacji i popraw jakośc analiz danych. **Wart nieocz(anomalie**)-dane odbiegające od normy, niespój z reszt zbior. Mogą przyjmo formę nieoczekiw wart atrybutu (np. wart spoza dozwolon zakres) lub wartości odstając, które znacząco różnią się od większoś danych.Identyf wart nieoczekiw odbyw się poprze analiz spójnoś danych w sys decyz:Sys determini–jeśl 2 obiek mają te sam wart cech, powin mieć tę samą decyz. Różnic wskazuj na anomal.Sys niedeterminist–dopuszcz różn dec dla identycz cech, co może oznaczać wart nieoczekiwane.Sys niespójny monotonicznie–obiekt lepszy we wszyst atrybut powin mieć co najm taką samą war decyz jak gorszy. Jeśl jest inaczej, występ anomal. **Wart odstające** to dan nietypow, odbiegając od reszt zb. Mogą wynikać z błęd lub istot anomal.Sposob identyfi:1.Podejścia statys–analiz częstotli wart, zakresów kwartyli lub odchyleń standard.2.Meto odległościowe–wykryw pkt oddalon od sąsiadów.3.Ekspercka wiedz–iden na podst reguł biznesow. Algor Las Izo:1.Bud drzew izolac–los wybór atryb i wart podziału.2.Izolowanie anomal–pkt odstaj mają krótsze ścież w drzewach.3.Oblicz wyniku–śred dł ścież decyduje, czy pkt to anomal.4.Klasyfikac–wartości poniż ustalonego progu są uznawan za odstaj.**Reg dec**-wyrażeni „jeżel warunk, to dec”, gdzie warunki mogą obejmo wszyst lub część atryb.Pros reg dec ma post koniunkcji wart atrybutów, np. „a₁ = v₁ ∧ a₂ = v₂ ⇒ d”.Pewna reg dec-reg,która jednoznacznie klasyfikuje obiekty w sys dec.Min pros reg dec-reg,która nie może być skrócona bez utraty poprawności.

**3 probl wielokryt wspomag dec i ich algo:** 1.**Wybór najlepszego wariantu** – polega na wskazaniu jednej, optymalnej opcji spośród dostępnych wariantów. **Algorytm**: **ELECTRE** **I**– metoda eliminacji przez ograniczenia, wykorzystująca relacje przewagi między wariantami. 2.**Ranking wariantów** – porządkowanie opcji od najlepszej do najgorszej na podstawie wielu kryteriów. **Algorytm**: **TOPSIS** – metoda klasyfikacji według podobieństwa do rozwiązania idealnego, obliczająca odległość od ideału i antyideału. 3.**Sortowanie** – przypisanie wariantów do wcześniej zdefiniowanych kategorii decyzyjnych (np. klasyfikacja jakości).**Algorytm**: **Drzewo decyzyjne** – model hierarchiczny, który klasyfikuje warianty poprzez iteracyjne podziały danych według kryteriów. Każdy węzeł odpowiada pytaniu o wartość kryterium, a gałęzie reprezentują możliwe odpowiedzi. Proces kończy się w liściach drzewa, które wskazują ostateczne kategorie decyzyjne wariantów.**Algo indukcji wszystkich minimalnych prostych pewnych reg dec:**1.Inicjali puste zbior reg.2.Dla każdego obiektu sprawdzenie, czy istnieje obiekt niedeterminist.3.Bud wyrażeń boolowskich odróżniających obiekt od innych w sys.4. Tworz koniunkcji atryb warunkowych jako prost reg dec.5.Doda reg do zbior wynikowego.Algo LEM2 indukcji niektórych min prost pewnych reg dec:1.Wybór podzbiór obiektów z identyczną decyzją.2.Iteracyjne tworz reg dec:Wybór deskryptora odróżniającego najwięcej obiektów;Doda deskryptora do reguły;Powtarz, aż reg obejmie cały podzbiór. 3. Minimalizacja reg–usunięcie zbędnych deskryptorów.4.Usunięcie reg, które nie są konieczne.5.Doda finalnych reg do zbioru wynikowego.LEM2 tworzy mniej reg, eliminując nadmiarowe, podczas gdy metoda wyczerpująca generuje wszystkie możliwe reguły. **Model akademicki** (Fayyad i inni) koncentruje się na eksploracji danych jako procesie badawczym. Obejmuje kroki od zrozumienia obszaru zastosowania, poprzez przygotowanie i redukcję danych, wybór celu i algorytmu eksploracji, aż po analizę wyników i ich interpretację. Model ten kładzie nacisk na techniczne aspekty eksploracji danych oraz na dokładność analizy odkrytych zależności.**Model biznesowy** (CRISP-DM) to podejście stosowane w przemyśle, które strukturyzuje proces analizy danych w sposób ułatwiający praktyczne wdrożenie. Proces rozpoczyna się od zrozumienia problemu biznesowego i danych, następnie obejmuje przygotowanie danych, modelowanie i ocenę wyników, a kończy na wdrożeniu rozwiązania. Model ten jest szeroko stosowany w praktyce biznesowej, gdyż pozwala na iteracyjne doskonalenie modeli.**Model hybrydowy** (Cios i inni) łączy podejścia akademickie i biznesowe, uwzględniając zarówno aspekt techniczny, jak i praktyczne zastosowanie odkrytej wiedzy. Proces zaczyna się od zrozumienia problemu i danych, przechodzi przez ich eksplorację i ocenę odkrytej wiedzy, a kończy na jej praktycznym zastosowaniu. Model ten jest elastyczny i nadaje się zarówno do badań naukowych, jak i do wdrażania systemów eksploracji danych w organizacjach.

**Apriori Wejście:** Zbiór transakcji T, próg minimalnego wsparcia min\_sup **Wyjście:** Zbiór częstych zbiorów elementów; Znajdź wszystkie częste pojedyncze elementy, Generuj kandydujące zbiory elementów o długości k+1 z częstych zbiorów długości k, Oblicz wsparcie dla kandydatów, odrzucając te, które nie spełniają min\_sup, Powtarzaj kroki 2-3 aż nie będzie więcej częstych zbiorów, Zwróć zbiór wszystkich częstych zbiorów elementów **Knuth-Morris-Pratt (KMP) Wejście:** Łańcuch znakowy s, wzorzec p **Wyjście**: Wszystkie pozycje wystąpień p w s, Oblicz tablicę prefiksów dla p,Przeglądaj s i p, dopasowując kolejne znaki, W przypadku niedopasowania przesuwaj wzorzec zgodnie z tablicą prefiksów, Jeśli dopasowanie, zapisz pozycję i kontynuuj, Zakończ i zwróć znalezione pozycje **Algorytm sortowania wariantów  
Wejście:** Zbiór wariantów V z wartościami atrybutów **Wyjście**: Posortowane warianty; Określ kryteria sortowania (np. malejąco lub rosnąco dla każdego atrybutu), Normalizuj wartości atrybutów jeśli wymagane, Posortuj warianty według pierwszego kluczowego atrybutu, W przypadku remisów, użyj kolejnych atrybutów jako tie-breaker, Zwróć posortowaną listę wariantów **TOPSIS Wejście:** Macierz decyzyjna z wartościami wariantów i wagami kryteriów **Wyjście:** Ranking wariantów; Normalizuj macierz decyzyjną, Pomnóż przez wagi kryteriów, Wyznacz idealne rozwiązanie pozytywne i negatywne, Oblicz odległość każdego wariantu od rozwiązań idealnych, Oblicz współczynnik bliskości, Posortuj warianty malejąco według współczynnika bliskości, Zwróć ranking wariantów **Algorytm AHP (Analytic Hierarchy Process) Wejście:** Cel decyzji**,** Kryteria oceny**,** Alternatywy do wyboru**,** Macierze porównań parowych (oceny ważności kryteriów i alternatyw)**, Działanie: Tworzenie hierarchii** – cel na górze, kryteria w środku, alternatywy na dole. **Porównania parowe** – ocenianie ważności kryteriów i alternatyw względem siebie (np. na skali 1–9). **Obliczanie wag** – wyliczenie priorytetów dla kryteriów i alternatyw (np. metodą wartości własnych). **Ocena i wybór** – suma ważonych ocen pozwala wybrać najlepszą opcję. **Wyjście:** Najlepsza alternatywa zgodna z podanymi kryteriami.