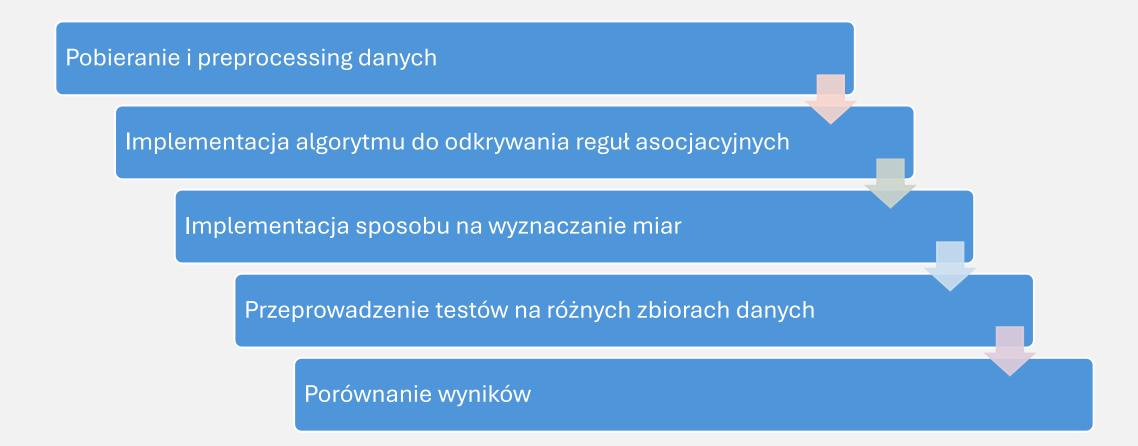
Implementacja algorytmu do odkrywania reguł asocjacyjnych

Mikołaj Bańkowski

Mateusz Kołacz

1. Omówienie problemu



2. Wybór algorytmu

Algorytm Apriori

Algorytm AprioriRuleGen

Algorytm Eclat

Algorytm Partition

3. Wybrany algorytm

Do zadania wybrany został algorytm Apriori.

Uzasadnienie:

- **Uniwersalność** działa dobrze na różnorodnych zbiorach danych (małych, średnich i większych), może być stosowany zarówno do danych dyskretnych, jak i o bardziej złożonej strukturze
- **Przystosowanie do analiz reguł asocjacyjnych** pozwala wygenerować nie tylko częste zbiory, ale i same reguły asocjacyjne,
- **Latwy w implementacji** bogata literatura oraz biblioteki, które ułatwiają jego implementacje

4. Dane

Dane będą pobierane z **UCI Machine Learning Repository** (https://archive.ics.uci.edu) za pośrednictwem biblioteki ucimlrepo. Zostanie stworzona klasa udostępniająca interfejs pozwalający na pobranie jednego z wybranych zbiorów danych.

Z początku wybrane zostaną dyskretne zbiory danych, odpowiednie do problemu. W dalszej części rozwiązania, jeżeli będzie to wymagane przez eksperymenty, stworzone zostanie rozwiązanie pozwalające na dyskretyzację dowolnych danych.

5. Wybrane zbiory danych

Car Evaluation - Zbiór danych dotyczący oceny samochodów. Zawiera informacje na temat cech samochodu (np. cena, liczba drzwi, pojemność bagażnika, bezpieczeństwo) oraz decyzji dotyczących oceny samochodu (klasyfikacja jako: unacceptable, acceptable, good, very good).

Zbiór ten zawiera 1728 rekordów, z czego każdy opisuje pojedynczy egzemplarz samochodu.

Dane są wyraźnie zdefiniowane i sklasyfikowane, co ułatwia analizę i zrozumienie danych.

5. Wybrane zbiory danych cd.

Nursery - Zbiór danych dotyczący oceny przedszkoli. Zawiera informacje o cechach przedszkola (np. ilość dzieci, ilość osób dorosłych, poziom opieki) oraz decyzji dotyczących oceny przedszkola (klasyfikacja jako: recommend, priority, not recommend, very-recom).

Zbiór ten zawiera 12960 rekordów, z czego każdy opisuje pojedyncze przedszkole.

Dane są zróżnicowane i uwzględniają różne aspekty funkcjonowania przedszkoli, co pozwala na wszechstronną analizę.

5. Wybrane zbiory danych cd.

<u>Tic-Tac-Toe Endgame</u> - Zbiór danych dotyczący końcówki gry w kółko i krzyżyk. Zawiera informacje o ułożeniu planszy gry (np. pozycje poszczególnych pól: top-left, top-middle, top-right, itd.) oraz decyzji dotyczących oceny końcowego stanu gry (klasyfikacja jako: win, loss, draw).

Zbiór ten zawiera 958 rekordów, z czego każdy opisuje pojedynczą sytuację w końcówce gry w kółko i krzyżyk.

Dane są w pełni dyskretne i reprezentują wszystkie możliwe kombinacje ułożenia pól na planszy w końcowej fazie gry.

6. Wybrane miary

- Współczynnik podniesienia (lift),
- · Współczynnik pewności (certanity factor),
- Jaacard / coherence,
- iloraz szans (odds-ratio)

7. Testy

Celem testów projektu jest porównanie zachowania się wybranych miar w odniesieniu do wartości współczynnika podniesienia.

Algorytm zostanie uruchomiony na 3 wybranych zbiorach danych, z których wybrane zostaną reguły asocjacyjne o różnych (zbliżonych) wartościach wsp. podniesienia - lift.

Następnie zbadana zostanie zależność między wartościami wybranych miar a wartością wybranego wsp. podniesienia lift. Wnioski zostaną dokładnie opisane i zestawione w sprawozdaniu końcowym.

8. Technologie

Projekt zostanie w całości zrealizowany w środowisku Python. Stworzony zostanie skrypt konsolowy, na podstawie którego przeprowadzone zostaną badania.

Przy implementacji rozwiązania zachowane zostaną wszystkie najnowocześniejsze i specyficzne dla języka Python techniki i paradygmaty programowania.