Sprawozdanie

Temat: Reguły asocjacyjne

Przedmiot: Analiza Danych w Biznesie

Wykonał: Artur Kompała

1. Cel sprawozdania

Celem niniejszej pracy było wybranie zbioru danych transakcyjnych, wygenerowanie reguł asocjacyjnych z różnymi wartościami progowymi oraz porównanie i skomentowanie otrzymanych wyników. Do realizacji projektu wykorzystano język Python oraz bibliotekę mlxtend implementującą algorytm Apriori.

2. Wybór zbioru

Wybrany dataset: Wine Quality z repozytorium UCI Machine Learning Repository. Zbiór zawiera 6497 wina z 11 atrybutami reprezentującymi cechy chemiczne (kwasowość, gęstość, zawartość alkoholu, pH, siarczyny itp.) oraz ocenę jakości.

3. Przygotowanie danych

Dane zostały wstępnie przetworzone w następujący sposób:

- Dyskretyzacja atrybutów ciągłych – zmienne numeryczne (kwasowość, zawartość alkoholu, pH, itp.) zostały podzielone na przedziały kategoryczne (np. niski, średni, wysoki).
- Kodowanie binarne – każda kombinacja atrybutu i wartości została przekonwertowana na format binarny wymagany przez algorytm Apriori.
- Standaryzacja nazewnictwa – utworzono czytelne etykiety dla każdego itemsetu (np. alcohol_high, acidity_low).

4. Wartości progowe

Zestaw	Min Support	Min Confidence	Cel
1	5%	70%	Wysoka jakość, wysoką pewność
2	3%	50%	Równowaga między ilością a jakością
3	1%	30%	Eksploracja szeroka, włączając słabe powiązania

5. Porównanie i analiza wyników

Metryka	Zestaw 1	Zestaw 2	Zestaw 3
Liczba reguł	2735	28724	701042
Liczba reguł złożonych	2727	28633	700298
Procent reguł złożonych	99.7%	99.7%	99.9%
Średni support	7.04%	4.55%	1.79%
Średnia confidence	80.94%	67.75%	51.04%
Średni lift	2.60	2.71	3.06
Min support	0.50%	0.30%	0.10%
Max support	26.07%	26.07%	26.07%

ZESTAW 1 (Support 5%, Confidence 70%)

Charakterystyka:

- Liczba reguł: 2,735
- Średnia confidence: 80.94% – najwyższa spośród wszystkich zestawów
- Średni lift: 2.60

Zestaw 1 zawiera reguły o najwyższej wiarygodności. Średnia confidence 80.94% oznacza, że jeśli poprzednik reguły zaistnieje, to następnik pojawi się w średnio 81 na 100 przypadków. Lift 2.60 wskazuje, że powiązania są 2.6 razy silniejsze niż przypadkowe. To zestaw rekomendowany dla systemów decyzyjnych wymagających wysokiej pewności.

ZESTAW 2 (Support 3%, Confidence 50%)

Charakterystyka:

- Liczba reguł: 28,724 – **10x więcej** niż w Zestawie 1
- Średnia confidence: 67.75% – **kompromis** między ilością a jakością
- Średni lift: 2.71 – *nieznacznie wyższy* niż w Zestawie 1

Zestaw 2 reprezentuje złoty środek eksploracji. Obniżenie progu support z 5% do 3% wprowadza 10-krotny wzrost liczby reguł, ale średnia confidence pozostaje na rozsądnym poziomie 67.75%. Średni lift 2.71 jest porównywalny lub nawet nieznacznie wyższy, co sugeruje, że nowe reguły nie stanowią pogorszenia. Ten zestaw jest idealny dla analiz eksperymentalnych i odkrywczych.

ZESTAW 3 (Support 1%, Confidence 30%)

Charakterystyka:

- Liczba reguł: 701,042 – 256x więcej niż w Zestawie 1
- Średnia confidence: 51.04% – zaledwie 21 punktów powyżej progu minimalnego
- Średni lift: 3.06 – najwyższy, ale na tle ogromnej liczby reguł potencjalnie zafałszowany

Zestaw 3 to eksploracja ekstensywna. Prawie 1 milion reguł z confidence zaledwie o 21 punktów powyżej minimum (50% vs. 30%) sugeruje, że wiele z nich to prawie przypadkowe korelacje. Wysoki lift 3.06 może być artefaktem selekcji w zbiorze tak ogromnym, niektóre słabe powiązania mogą wyglądać statystycznie interesujące. Ten zestaw wymagałby dodatkowej filtracji i weryfikacji statystycznej.

6. Wnioski

Prawie wszystkie reguły są złożone (multi-itemset):

- ZESTAW 1: 2,727 z 2,735 (99.7%)
- ZESTAW 2: 28,633 z 28,724 (99.7%)
- ZESTAW 3: 700,298 z 701,042 (99.9%)

Oznacza to, że przeważnie kombinacje wielu atrybutów implikują inne atrybuty. Pojedyncze cechy nie decydują o wyniku.

Wysokie progi (Zestaw 1): Mniej reguł, ale każda jest zdrowo rozsądna i praktycznie użyteczna.

Niskie progi (Zestaw 3): Wiele reguł, ale większość to słabe powiązania, które mogą być artefaktami.

Wybór progów drastycznie wpływa na liczbę i jakość odkrywanych reguł, decyzja powinna wynikać z konkretnego celu analitycznego.