**Generowanie reguł asocjacyjnych**

**Danologia, Laboratorium**

Do pracy nad zadaniem będzie potrzebna biblioteka mlxtend

https://rasbt.github.io/mlxtend/

Wszystkie interesujące nas funkcje należy znaleźć w mlxtend.frequent\_patterns i mlxtend.preprocessing, poza tym potrzebne będą operacje na tabelach pandas. Zadanie musi być oddane w formacie pdf-a ze sprawozdaniem, przy czym polecana forma pracy to pisanie kodu w notebooku jupyter, czytelne opisanie uzyskanych wyników i ewentualny eksport.

Na zajęciach będziemy działać na zbiorze supermarket. W zbiorze tym każdy rekord jest zbiorem wartości, oznaczających obecność w koszyku zakupów pewnych kategorie produktów (np. dairy) oraz ostateczną wartość rachunku (high/low). Tego rodzaju dane opisane są w dokumentacji mlxtend jako dane transakcyjne.

Potok przetwarzania będzie wymagał:

1. Wczytania danych
2. Zakodowania danych w formacie onehot
3. Wyciągnięcia częstych zbiorów np. algorytmem apriori
4. Wygenerowania reguł i przycięcia tablicy reguł do tych o konkretnych wartościach metryk

Metryki pokrótce opisane są w samej dokumentacji mlxtend, będą też omówione na początku laboratorium. Należy wykonać następujące zadania (dla każdego trzeba podać kod użyty do uzyskania odpowiedzi):

1. Podaj nazwy produktów i wsparcie dla najliczniejszego zbioru częstego spośród zbiorów częstych o minimalnym wsparciu równym 20%. Jeżeli istnieje wiele równolicznych zbiorów, podaj ten o największym wsparciu. Wyniki liczbowe zaokrąglij do 2 miejsc po przecinku. Zinterpretuj uzyskane wyniki.
2. Wyrysuj histogram występowania produktów w bazie. Ogranicz go do 10 najczęściej występujących produktów.
3. Wygeneruj zbiory często o minimalnym wsparciu 25%, na podstawie tych zbiorów wygeneruj reguły o minimalnej wartości ufności (confidence) równej 0.70. Podaj uzyskane reguły i zinterpretuj je. Czy patrząc na wsparcia poprzedników i następników implikacji, jak również na miarę lift można powiedzieć, że te reguły są wartościowe?
4. Dla jakich kombinacji produktów, wartość rachunku jest wysoka (jakie jest X z reguły X→ {high})? Zawęź poszukiwania poprzez wygenerowanie w pierwszej kolejności zbiorów częstych o minimalnym wsparciu 16%, a następnie, wygenerowanie reguł o ufności minimalnej 0.50. Dla tak otrzymanych reguł wybierz tylko te, które mają jako następnik implikacji jedynie element high. Podaj i zinterpretuj uzyskane wyniki. Jak oceniasz, czy te reguły są wartościowe?