Прізвище: **КИРИЛЮК**

Ім'я: **Дмитро**

Група: **ПП-22**

Варіант: **08**

Дата захисту: **24.03.2025р**.

Кафедра: **САПР**

Дисципліна: **Системи інтелектуального аналізу та візуалізації даних**

Перевірив: **Андрій КЕРНИЦЬКИЙ**

**ЗВІТ**

до лабораторної роботи №06

на тему “**Асоціативний аналіз. Методи побудови асоціативних правил.”**

**Мета роботи:** ознайомлення з методами пошуку асоціативних правил за допомогою алгоритмів Apriori та FPGrowth в середовищах Weka та Excel. Студенти мають набути навичок роботи з цими алгоритмами, виконавши певні тренувальні завдання в середовищі Weka, а потім застосувати ці навички для виконання індивідуальних завдань на власних наборах даних. У програмі Excel студенти повинні застосувати алгоритм Apriori до свого власного набору даних про покупки для знаходження двійкових частих наборів

**Вправа 1 - Датасет для голосування**

Завантажте набір даних `vote.arff'.

* Завдання A1: запустіть Apriori, використовуючи налаштування параметрів за замовчуванням. Достовірність правила 10 становить 0:96. Як було обчислено це значення впевненості? Запишіть пропорцію у вигляді ділення.
* Завдання A2: Скільки випадків підтримує правило 8?
* Завдання A3: Що означає «правило застосовується до певної кількості випадків»? Поясніть на прикладі правила номер 7. (Підказка: ви можете перевірити числа на панелі попередньої обробки.)
* Завдання A4: Що означає «кількість випадків, передбачених правильно»? Поясніть на прикладі правила номер 9.
* Завдання A5: Вивчіть опис параметрів для Apriori, натиснувши кнопку `More' у вікні, яке дозволяє вам змінити параметри для `Apriori'. Спробуйте змінити кількість правил, указаних у вихідних даних. Як ви думаєте, чи може кількість згенерованих правил перевищувати 100. Якщо так, то чому?
* Завдання A6: Що означає «найкращі правила»? Який критерій використовується для визначення найкращих правил?
* Завдання A7: Яке правило визначає, наскільки ймовірно, що якщо конгресмен не голосував за допомогу Сальвадору, він також голосував за допомогу нікарагуанським контрас?
* Завдання A8: 10 найкращих правил містять правила, які містять «Class=democrat» у правій частині. Чи говорить це щось про виборчі звички конгресменів-демократів?

vote.arff - Цей набір даних містить інформацію про те, як кожен із конгресменів Палати представників США голосував за 16 ключових законів. Один екземпляр представляє історію голосування одного конгресмена та його партій.

**Вправа 2 - Датасет для погоди**

Завантажте набір даних `weather.nominal.arff'.

* Завдання В1: Розгляньте правило: temperature=hot ==> humidity=normal. Чим підтверджується це правило? Скільки випадків застосовується до цього правила і яке значення достовірності? (Щоб відповісти на це запитання, відкрийте вікно «Перегляд» на панелі попередньої обробки.)
* Завдання B2: Розгляньте правило: temperature=hot humidity=high ==> windy=TRUE. Чим підтверджується це правило? Скільки випадків застосовується до цього правила і яке значення достовірності? Далі запишіть номери екземплярів, які підтримують правило, і кількість екземплярів, які застосовуються до цього правила.
* Завдання B3: Чи може правило мати перевірки двох (чи більше) атрибутів праворуч, як у прикладі нижче: outlook=sunny temperature=cool ==> humidity=normal play=yes

weather.nominal.arff - Це дуже маленький набір даних лише з номінальними атрибутами.

**Вправа 3. - Датасет для супермакету**

Завантажте набір даних `supermarket.arff'.Використовуйте Apriori для створення правил і використовуйте їх, щоб сказати щось прокупівельні звички клієнтів супермаркету. Згенеруйте близько 30 правил.Також може бути цікаво створити правила з одним конкретним атрибутом у правійчастині. Їх можна згенерувати, встановивши для першого параметра значення `true', а длядругого параметра — значення індексу атрибута (індекси атрибутів для цього параметрапочинаються з 0, а не з 1), яке ви хочете бачити в правій частині правил.

* Завдання C1: Вивчіть кілька згенерованих правил і опишіть одне спостереження, яке, на вашу думку, було зроблено щодо купівельних звичок клієнтів супермаркету. Також запишіть відповідні правила для цього спостереження.
* Завдання C2: Опишіть друге спостереження, яке, на вашу думку, було зроблено щодо купівельних звичок клієнтів супермаркету. Також запишіть відповідні правила для цього спостереження.
* Завдання C3: Чи пропонують спостереження, зроблені вами в завданнях C1 і C2, якісь напрямки дій для менеджера супермаркету? Якщо так, то якими вони можуть бути?

supermarket.arff - Цей набір даних описує купівельні звички покупців супермаркету. Більшість атрибутів позначають одну конкретну групу предметів. Значення є `t', якщо клієнт купив товар поза асортиментом товарів, а в іншому випадку відсутній. Є один екземпляр на клієнта. Набір даних не містить атрибут класу, оскільки це не потрібно для вивчення асоціаціативних правил.

**Індивідуальне завдання:**

**1. Виконайте наступні завдання для власного набору даних у Weka:**

• Запустіть алгоритм пошуку асоціативних правил Apriori.

• Яке значення для порогу підтримки було використано в побудованій моделі? Яке значення для порогу достовірності було використано?

• Запишіть 10 найкращих знайдених правил, вкажіть для них значення підтримки та достовірності.

• Що позначають числа ліворуч і праворуч від стрілки в знайдених асоціативних правилах?

**2. Виконайте наступні завдання для власного набору даних у Weka :**

• Запустіть алгоритм пошуку асоціативних правил FPGrowth.

• Порівняйте списки десяти найкращих правил, отриманих двома алгоритмами. Поясніть відмінність в роботі двох алгоритмів.

• Згенеруйте також правила, у правій частині яких буде знаходитися ваш цільовий атрибут.

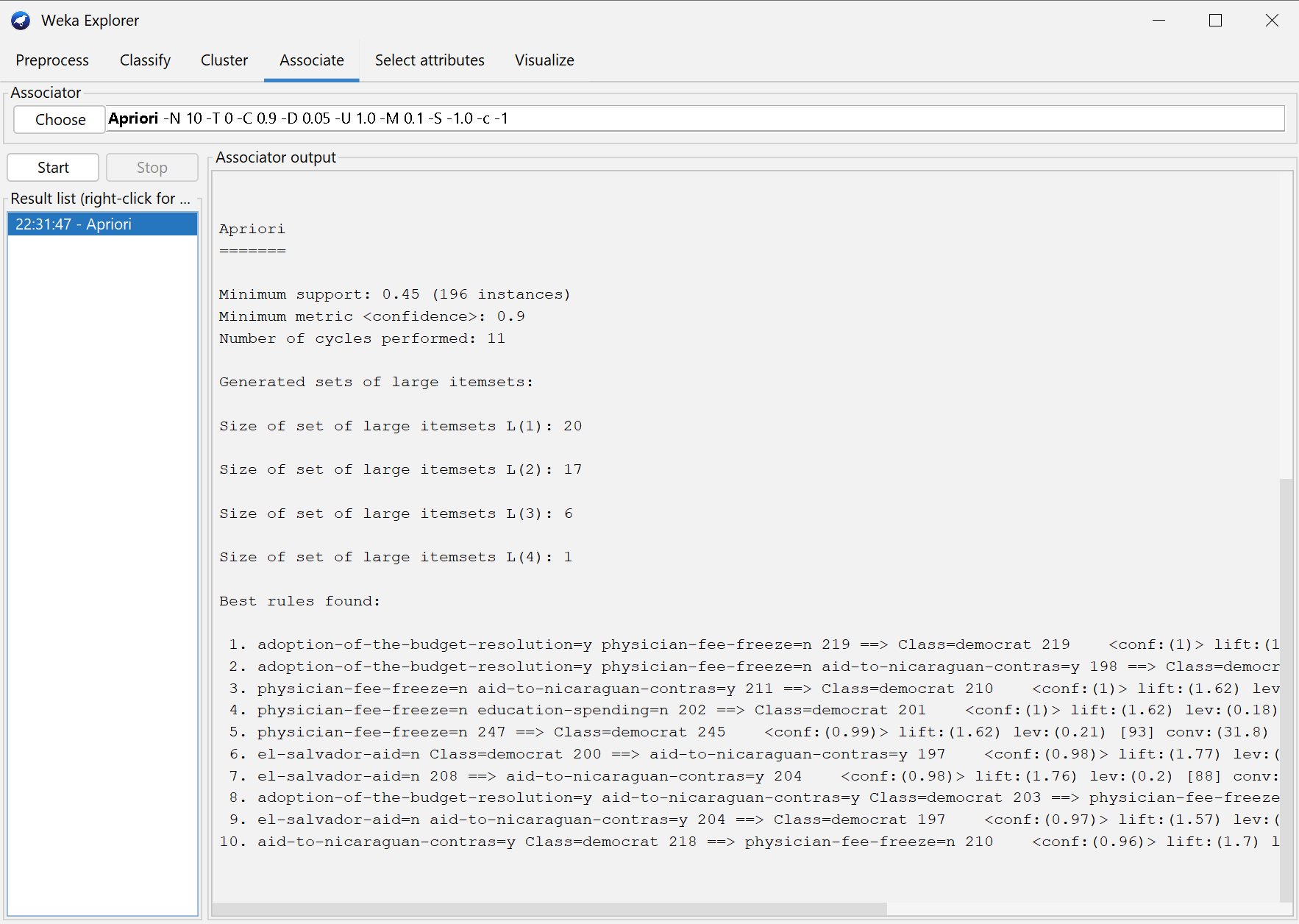
**3. Виконайте наступні завдання для власного набору даних в Excel** :

• Сформуйте власний набір транзакцій покупок, як показано у прикладі

• Знайдіть двоелементні групи частих наборів.

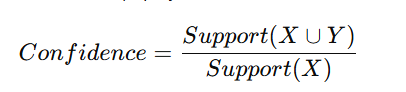
• Організуйте ваші обчислення на кількох аркушах (мінімум одниаркуш із вхідними даними і один аркуш для аналізу.)

**Вправа 1**

****

****

1) Впевненість (confidence) обчислюється за формулою:

****

Тобто це відношення випадків, коли зустрічаються **X і Y разом**, до випадків, коли зустрічається тільки **X.** У нашому випадку це 210/218 = 0.96.

2) Правило 8 підтримується у 203 випадках. Достовірність (conf): 0.98.

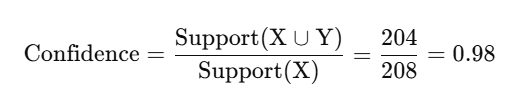


3) Це означає, що в наборі даних є конкретна кількість випадків (рядків), які відповідають умові правила. Іншими словами, це кількість разів, коли ліва частина правила виконується в даних.

Приклад для правила №7:

* 208 конгресменів не голосували за допомогу Сальвадору (el-salvador-aid = n).
* З них 204 також голосували за допомогу Нікарагуа (aid-to-nicaraguan-contras = y).
* Впевненість правила (confidence) 0.98 означає, що 204 з 208 випадків підтверджують це правило.

Формула:



Отже, це правило застосовується до 208 випадків, тобто в 208 рядках датасету конгресмен не голосував за допомогу Сальвадору. І з них у 204 випадках він голосував за допомогу Нікарагуа, що робить правило дуже надійним (98%).

4) Правило говорить:

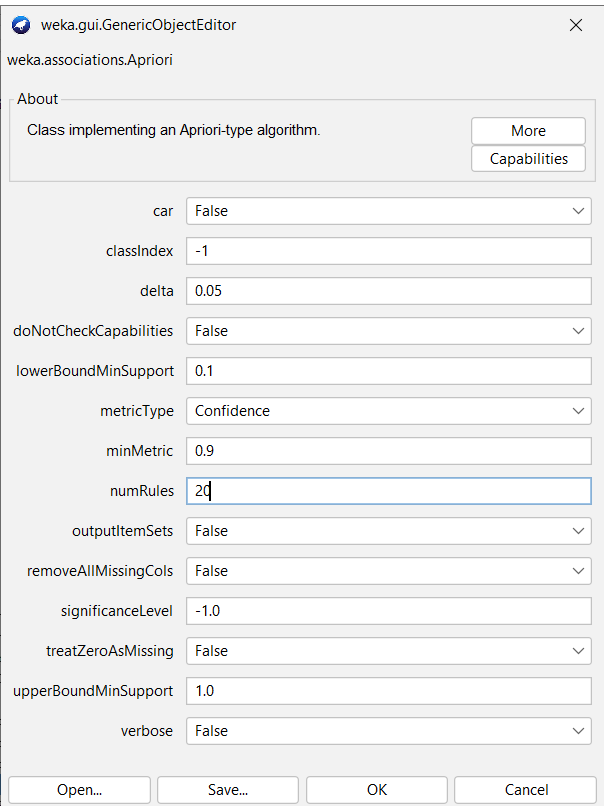
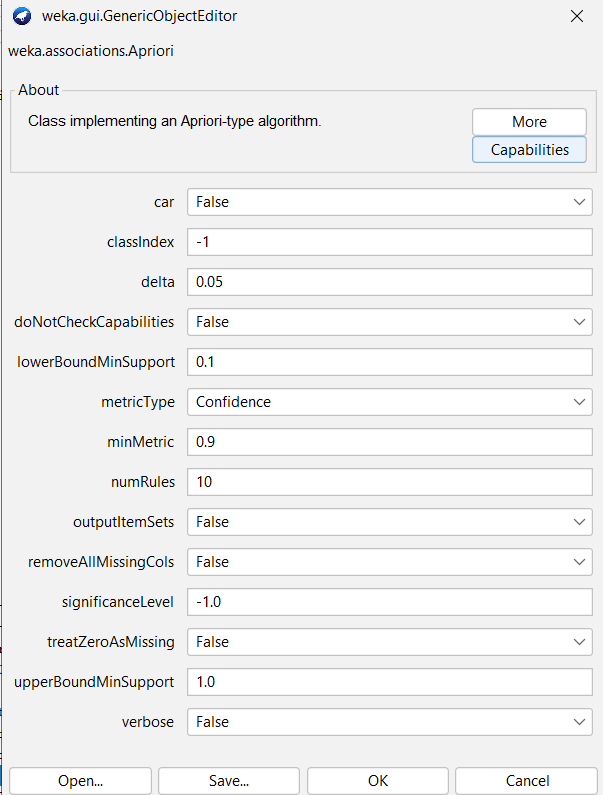
"Якщо проголосував за законопроєкт A, то проголосував за законопроєкт B".

Правило застосовується до 204 випадків у датасеті.

З них у 197 випадках передбачення правильне, тобто ці конгресмени дійсно є демократами.

Це означає, що правило має високу точність (97%) і є надійним для прогнозування партійної приналежності.

5)

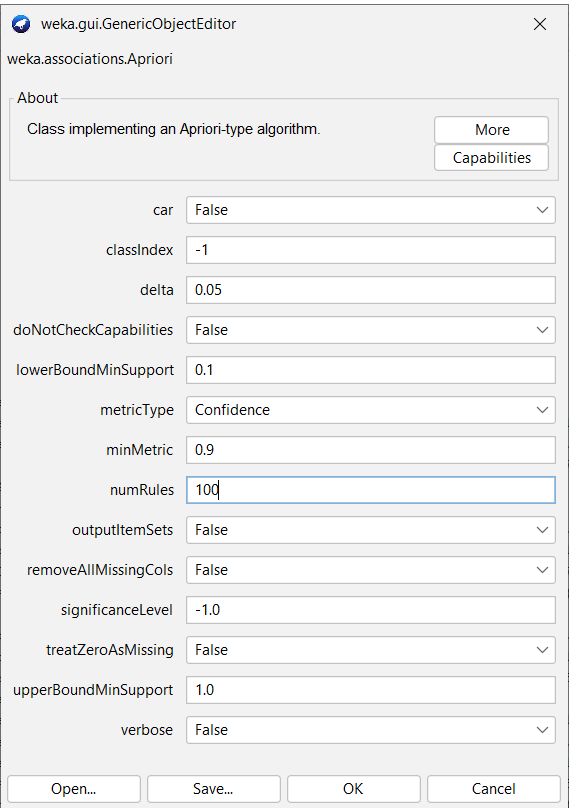
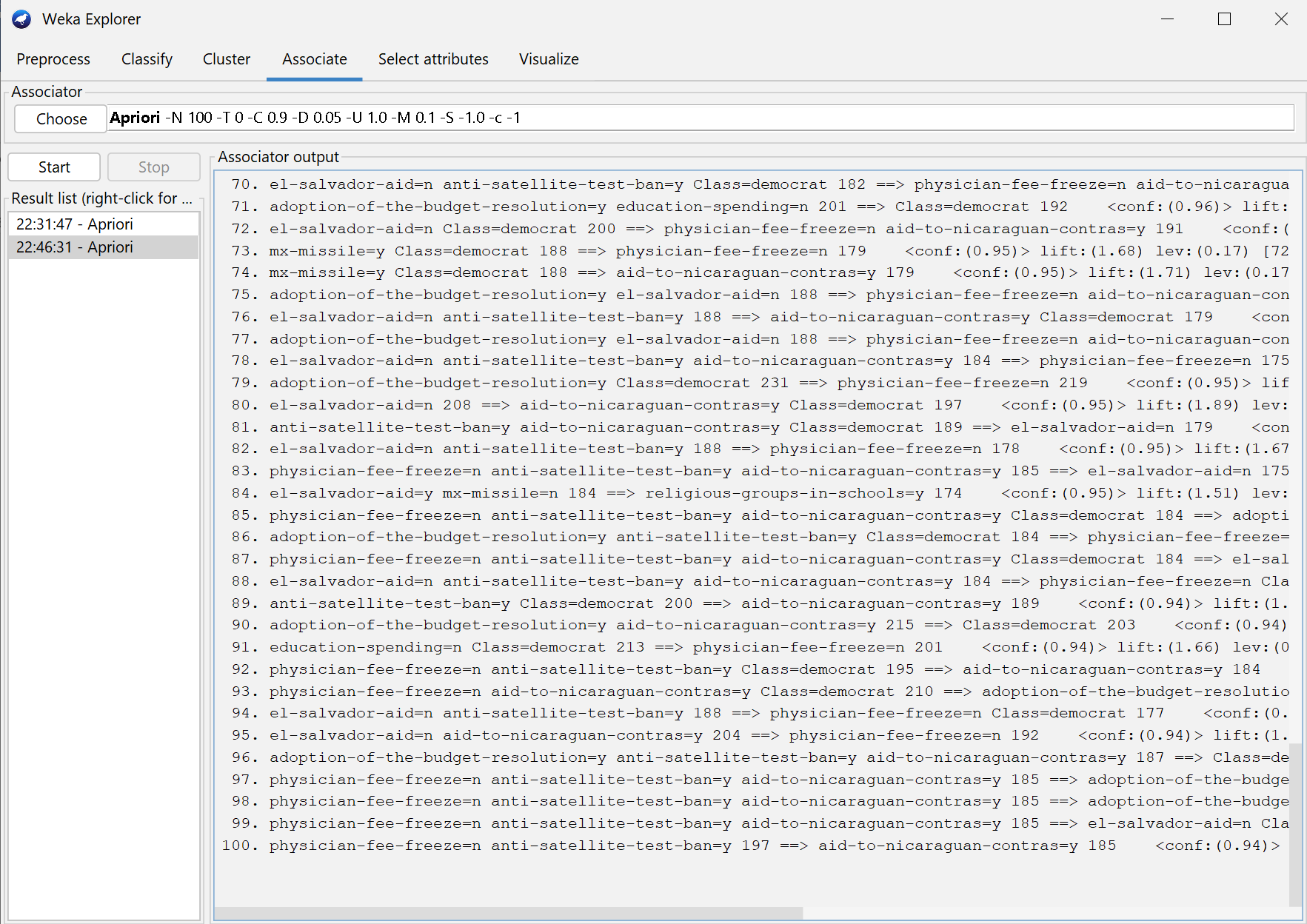


**Чи може кількість згенерованих правил перевищувати 100? Чому?**

Так, може, оскільки кількість правил залежить від:

* Мінімальної підтримки (min support) – якщо встановити дуже низьке значення, буде багато правил.
* Мінімальної довіри (min confidence) – якщо значення занадто низьке, генерується більше правил.
* Максимальної кількості правил (numRules) – цей параметр можна змінити.

Якщо параметр numRules збільшити, Weka може згенерувати більше ніж 100 правил.

6)Що означає «найкращі правила»?

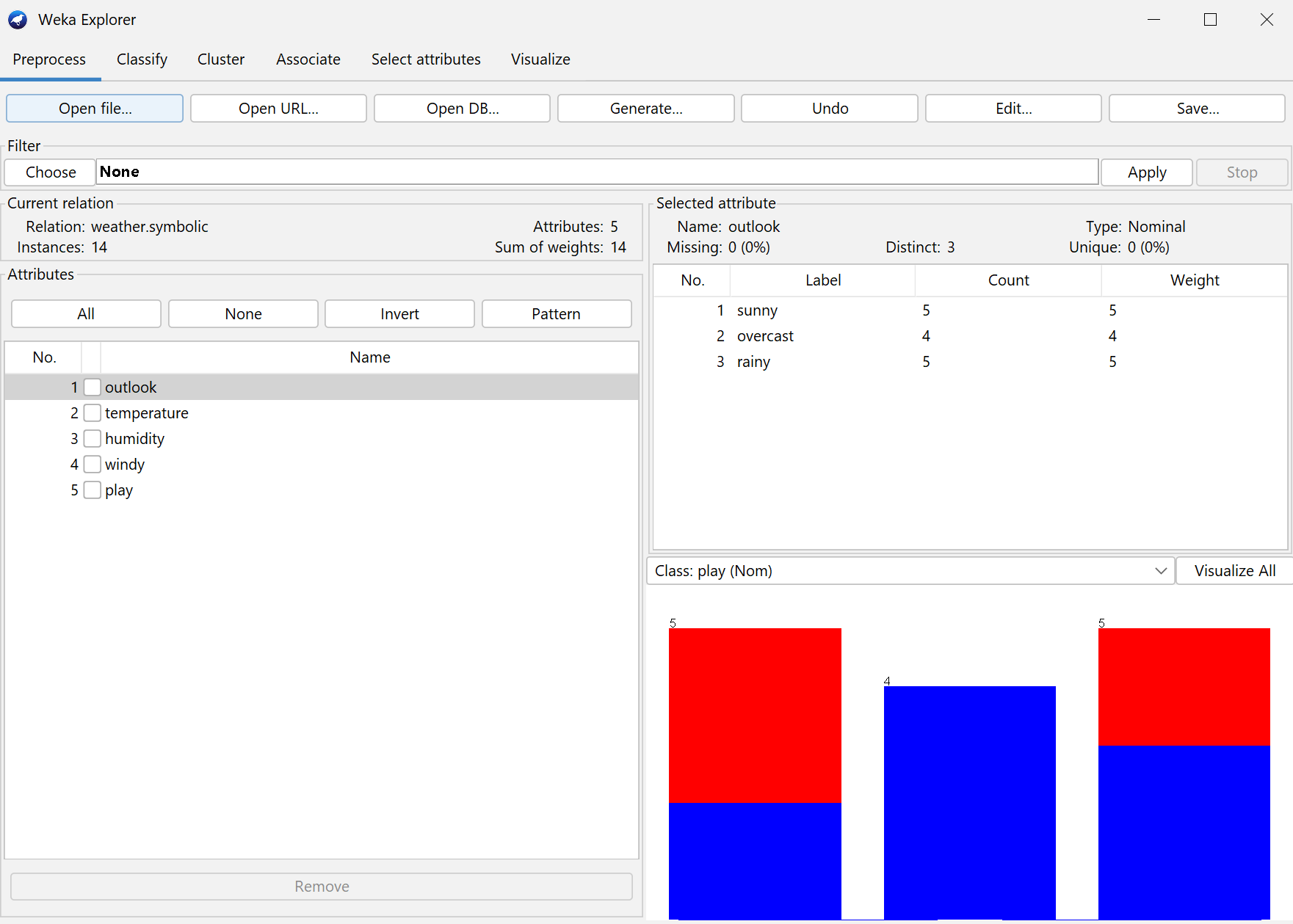
Weka вибирає найкращі правила за впевненістю (confidence) та підтримкою (support). Тобто:

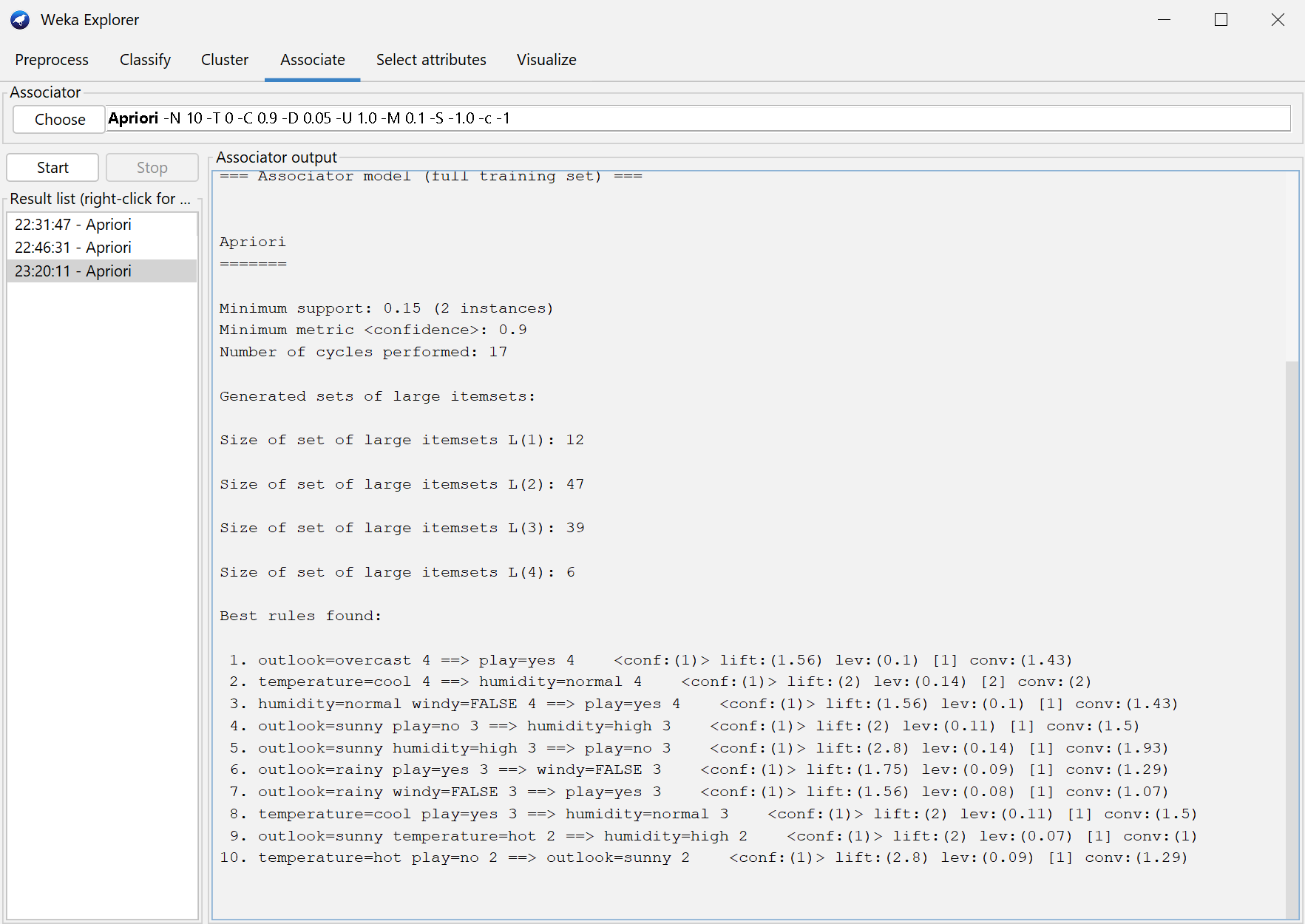
* Чим вища підтримка, тим частіше зустрічається правило.
* Чим вища впевненість, тим точніше правило передбачає результат.

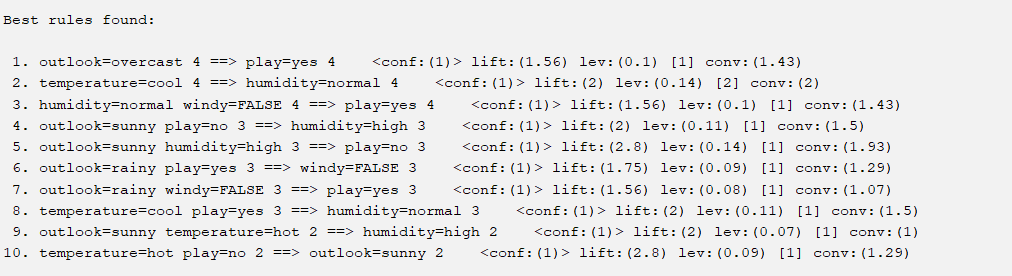
7) 

8) У 10 найкращих правилах часто зустрічається Class=democrat у правій частині. Це свідчить про те, що багато голосувань демократів мають певні закономірності, і їхні голоси можна передбачити з високою точністю. Це означає, що демократи частіше голосували за певні законопроєкти або ухвалювали рішення згідно з партійною лінією.

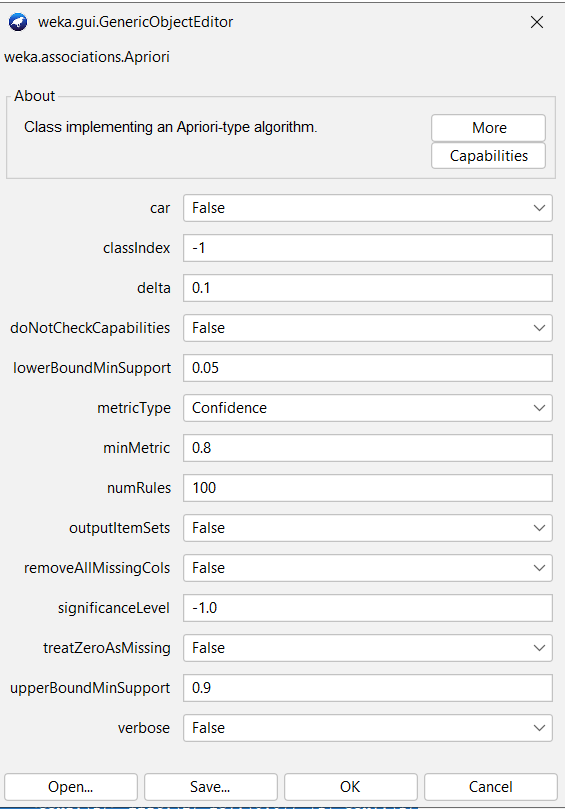
**Вправа 2**

****

****

****

Я розширив набір правил, зробивши аналіз глибшим, але ціною того, що частина правил має низьку підтримку (з'являються тільки в 1-2 випадках). Якщо мета – знайти стабільніші закономірності, варто підняти підтримку назад. Якщо мені треба знайти всі можливі залежності, то залишаю так, як є.

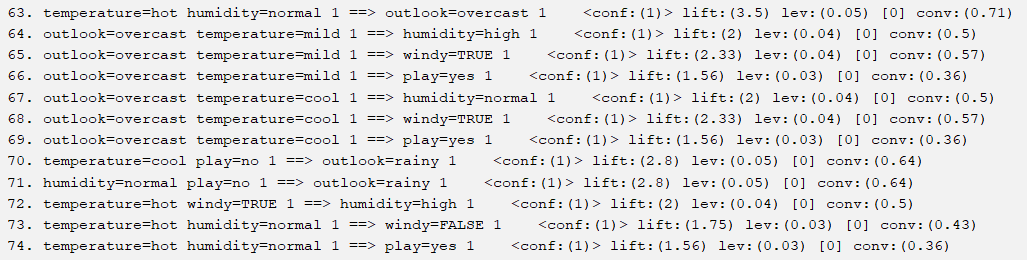
****

1)

Перевіряємо список правил – такого немає.  
Натомість є:

* temperature=hot humidity=normal 1 ==> outlook=overcast 1 (№63)
* temperature=hot humidity=normal 1 ==> windy=FALSE 1 (№73)
* temperature=hot humidity=normal 1 ==> play=yes 1 (№74)

Це означає, що temperature=hot не веде до humidity=normal у цьому наборі даних. Ймовірно, у вибірці humidity=high є більш типовим значенням для temperature=hot.



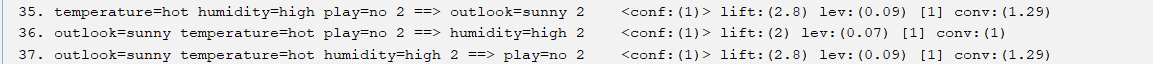
2)

Правило: temperature=hot humidity=high ==> windy=TRUE  
Шукаємо в списку – такого немає.  
Натомість є схожі:

* temperature=hot humidity=high play=no 2 ==> outlook=sunny 2 (№35)
* temperature=hot humidity=high windy=TRUE 1 ==> outlook=sunny 1 (№80)
* outlook=sunny temperature=hot humidity=high 2 ==> play=no 2 (№37)

У випадку №80 правило справджується лише в одному екземплярі (support = 1). Це означає, що підтримка дуже низька.

Тобто, правило temperature=hot humidity=high ==> windy=TRUE не має значної підтримки в наборі.





3)

Чи можливе правило з декількома атрибутами праворуч?

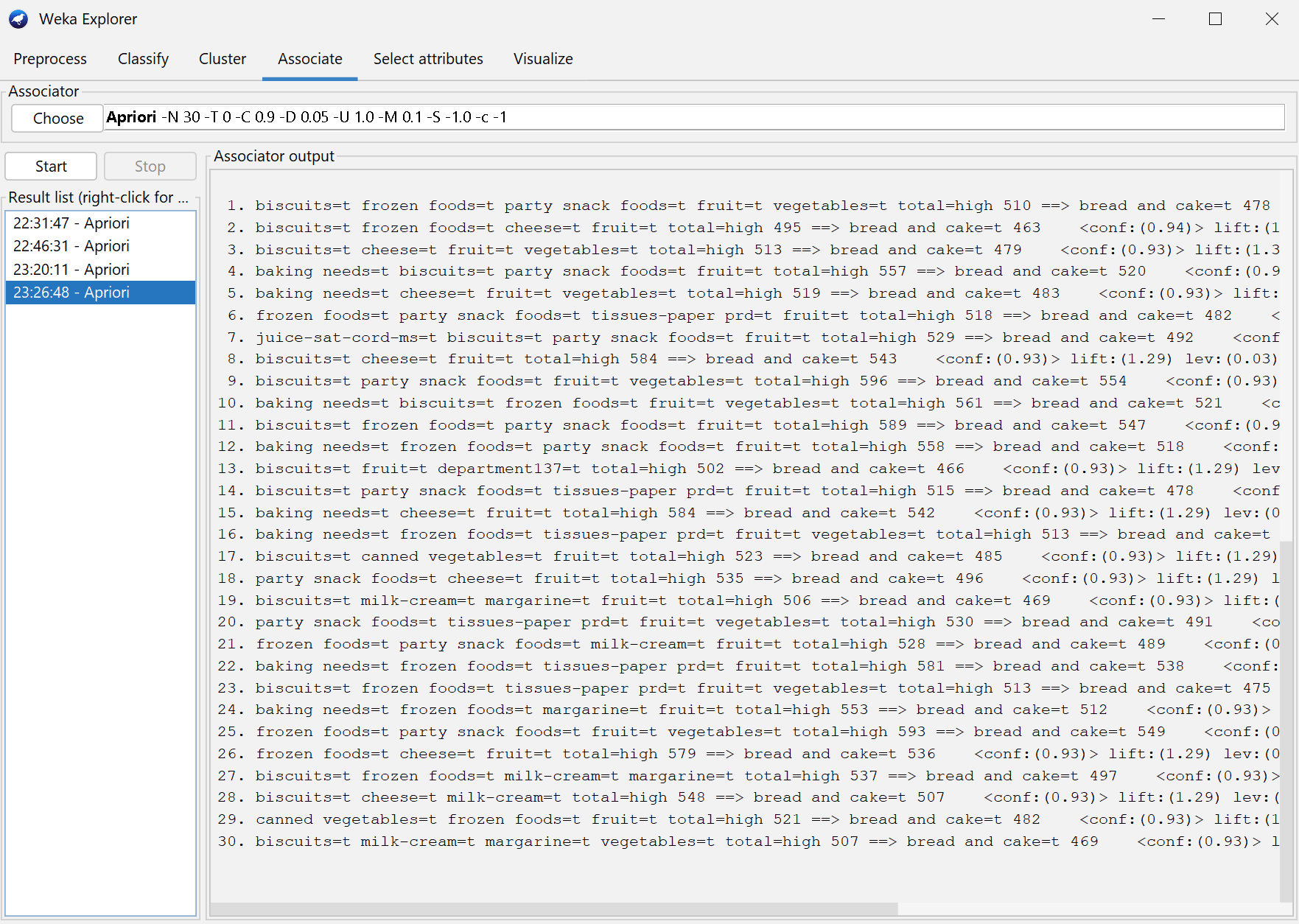
Так, у цьому виводі з Weka є такі правила! Наприклад:

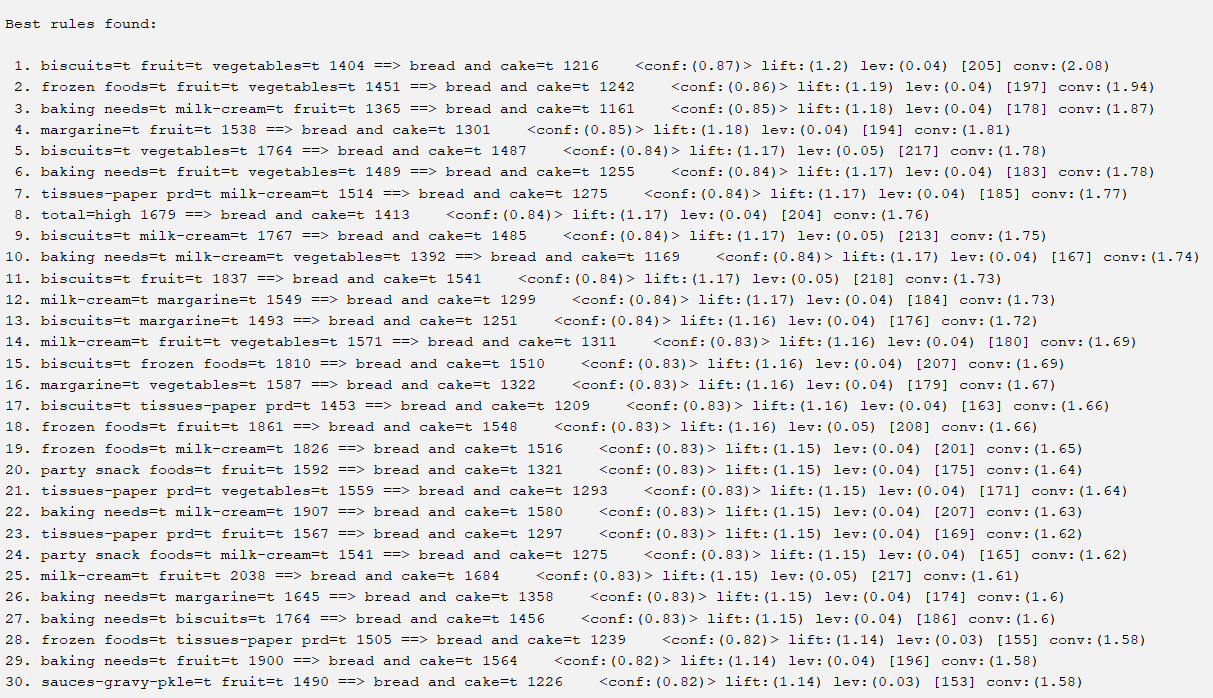
* temperature=hot play=no 2 ==> outlook=sunny humidity=high 2 (№38)
* outlook=sunny temperature=hot 2 ==> humidity=high play=no 2 (№39)

Отже, Weka може створювати правила з двома атрибутами праворуч, але такі випадки рідкісні. Це залежить від підтримки в наборі даних.



**Вправа 3**

****

****

1)Спостереження:  
Клієнти, які купують фрукти, овочі та печиво, часто також купують хліб і випічку.

Правила, що це підтверджують:



Висновок:  
Це свідчить про те, що клієнти, які купують корисні продукти, також купують хлібобулочні вироби, можливо, для збалансованого харчування або приготування страв.

2) Спостереження:  
Клієнти, які купують молочні продукти (молоко, вершки, маргарин), часто купують хліб і випічку.

Правила, що це підтверджують:





Висновок:  
Це може свідчити про те, що покупці купують молочні продукти разом із випічкою або для приготування домашніх десертів.

3) Оптимізація розташування товарів

* Розмістити хліб і випічку поруч із фруктами та овочами, а також біля печива, щоб підвищити ймовірність купівлі.
* Виділити окрему полицю для молочних продуктів (молоко, вершки, маргарин) поруч із хлібом і випічкою.

Спеціальні пропозиції та акції

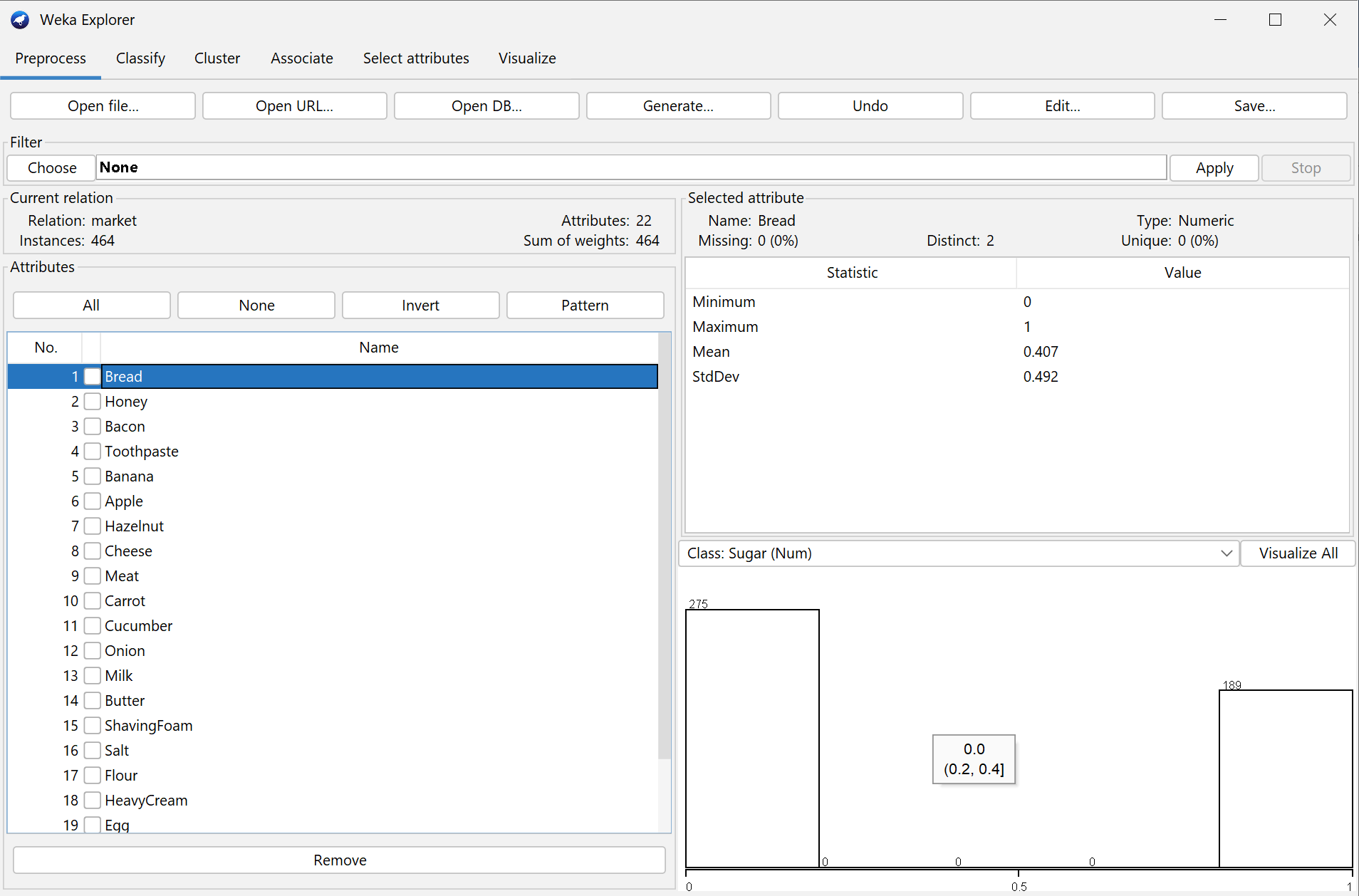
* Зробити знижку "Купіть печиво + овочі + хліб зі знижкою 10%".
* Запропонувати комплекти для випічки (борошно, маргарин, молочні продукти + хліб) за зниженою ціною.

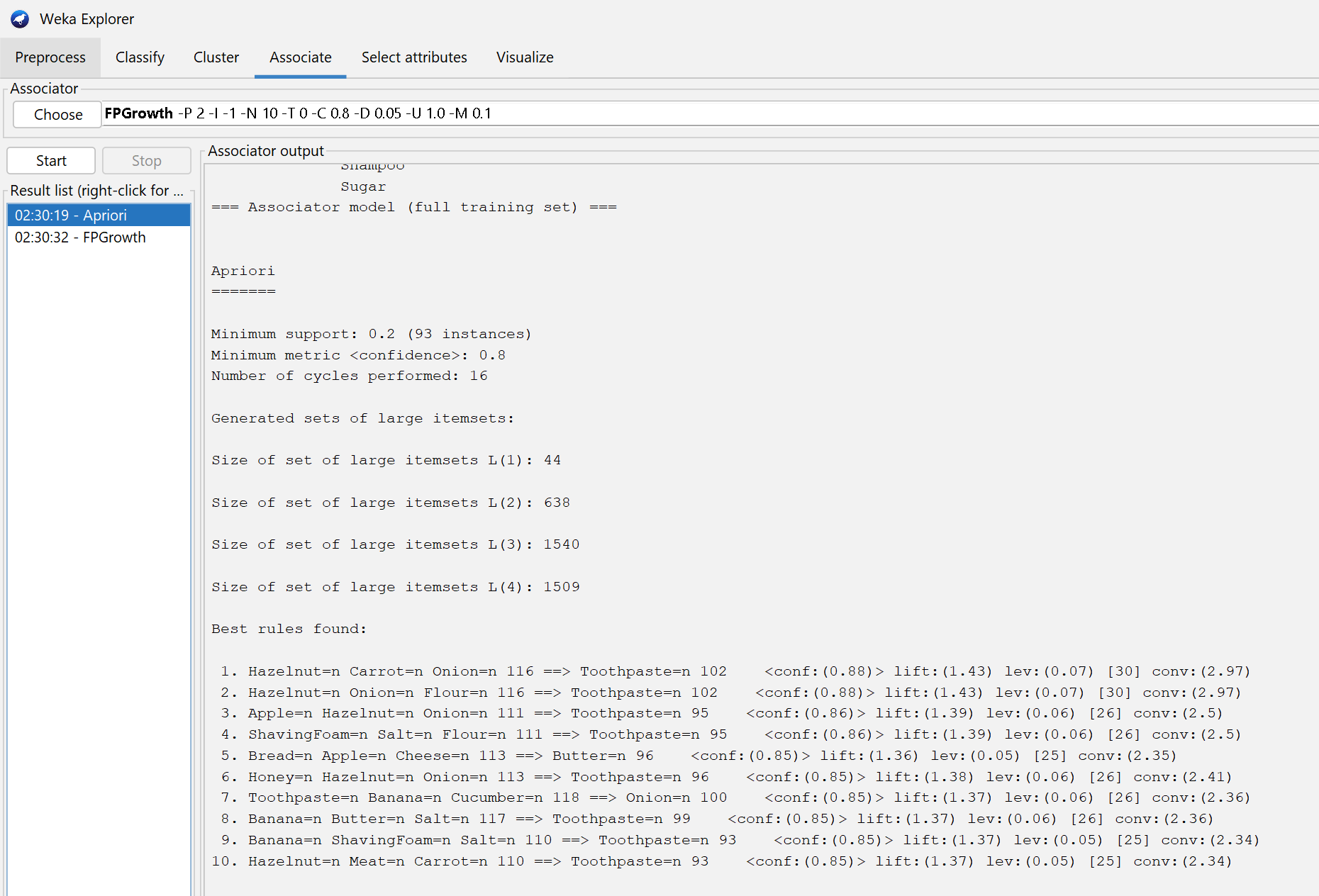
Персоналізовані рекомендації

* Для покупців, які часто купують молочні продукти, пропонувати знижки на хлібобулочні вироби через програму лояльності.

**Індивідуальне завдання**

Посилання на датасет: <https://www.kaggle.com/datasets/rukenmissonnier/real-market-data?resource=download&select=market.csv>

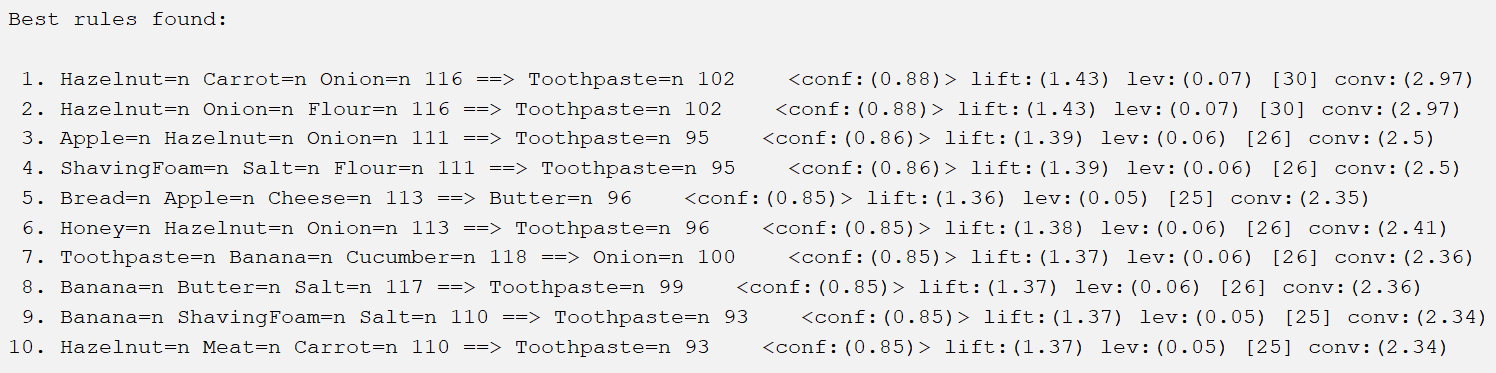
****

****

1)

Поріг підтримки (Support):  
Параметр lowerBoundMinSupport встановлений на значення 0.1, що означає, що мінімальна підтримка для правил дорівнює 10%.

Поріг достовірності (Confidence):  
Параметр minMetric встановлений на 0.8, що означає, що мінімальний поріг достовірності (confidence) для правил дорівнює 80%.



Support (Підтримка): Кількість транзакцій, у яких виконується дане правило

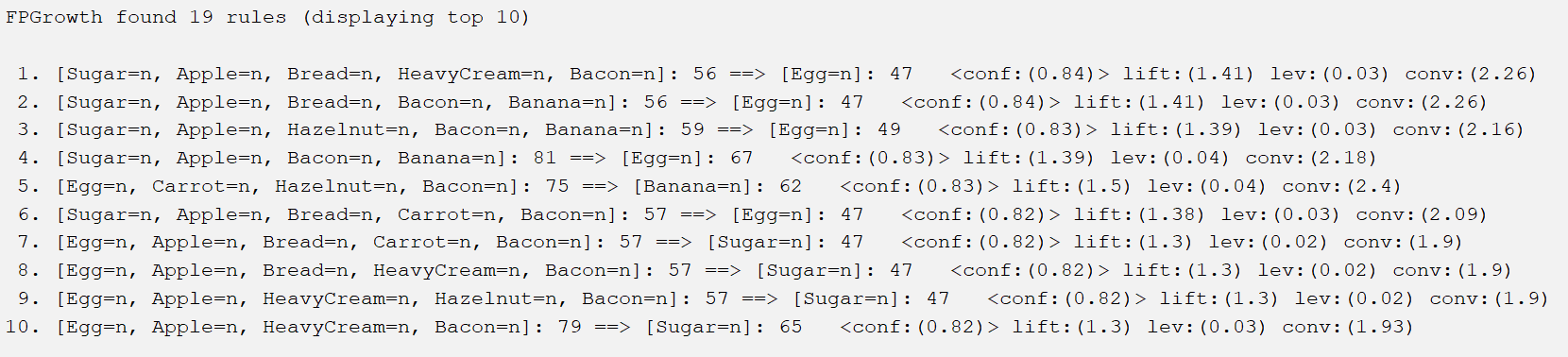
Confidence (Достовірність): Ймовірність того, що права частина правила виконається, якщо виконані всі умови з лівої частини.

Lift: Показує, наскільки часто продукти в правилі трапляються разом порівняно з їх випадковим співпадінням. Значення Lift > 1 означає позитивну асоціацію (сильний зв'язок).

Conviction (Conv): Чим вище значення Conviction, тим сильніша впевненість, що правило не є випадковим.

* Ліва частина (до стрілки) – це умова або передумова (Antecedent). Вона вказує на набір продуктів чи атрибутів, які одночасно зустрічаються в певній кількості транзакцій.
* Права частина (після стрілки) – це наслідок або результат (Consequent). Вона показує, який товар або атрибут буде зустрічатися з високою ймовірністю за виконання умов лівої частини правила.

2)

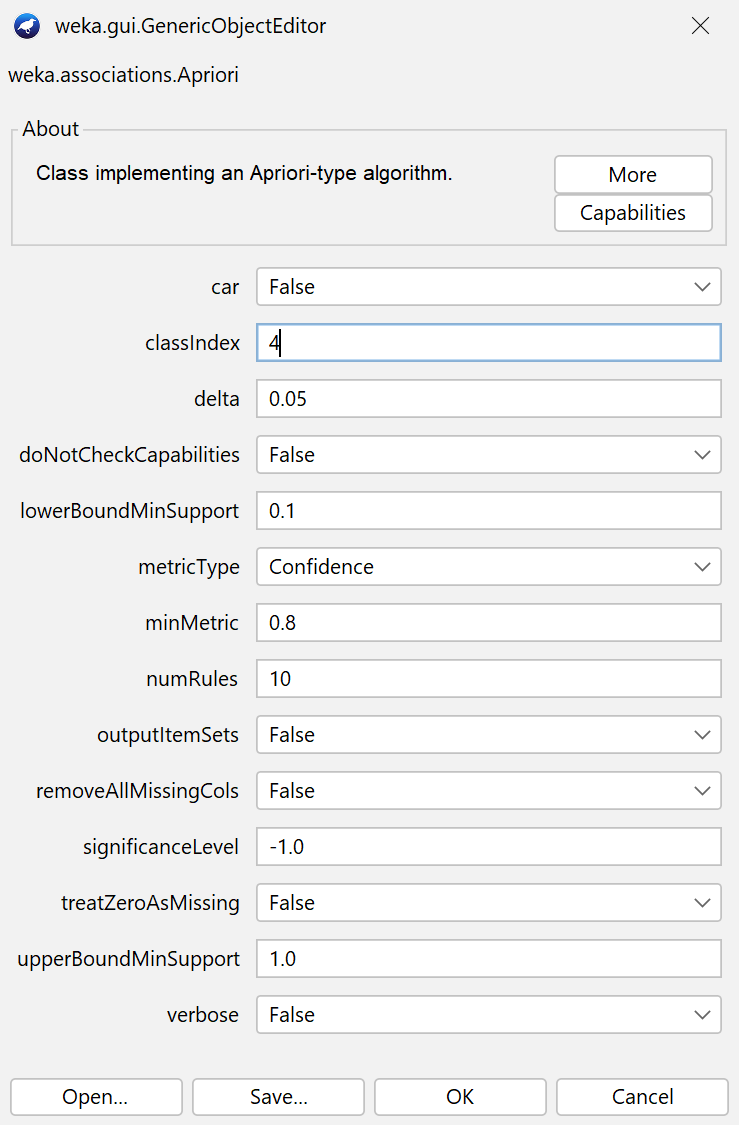


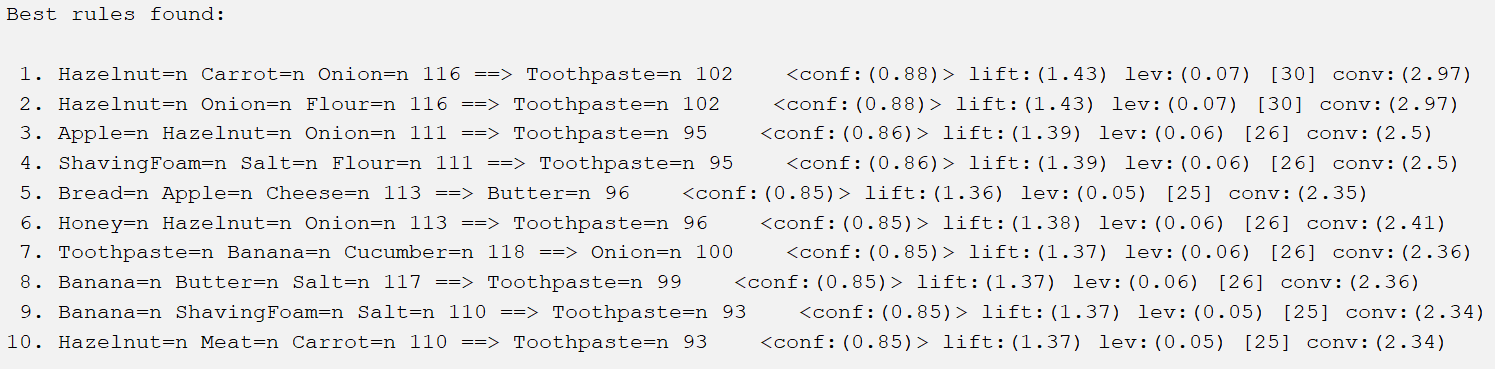
Таблиця порівняння правил(за результатами):



Відмінності:

* Apriori частіше пов’язує Toothpaste з різними комбінаціями продуктів (Hazelnut, Onion, Flour тощо).
* FP-Growth частіше знаходить правила, що включають Egg та Sugar разом з іншими продуктами (Apple, Bread, Bacon, Banana).
* Apriori використовує окремі товари як основу для правил, а FP-Growth частіше включає набори з більшою кількістю товарів.

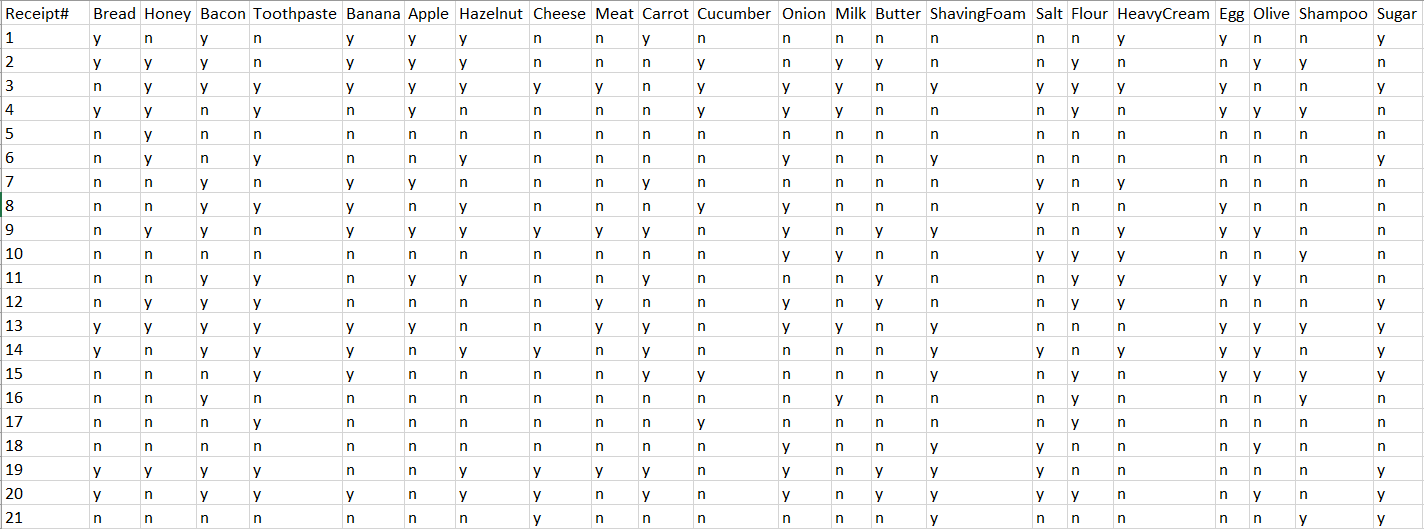




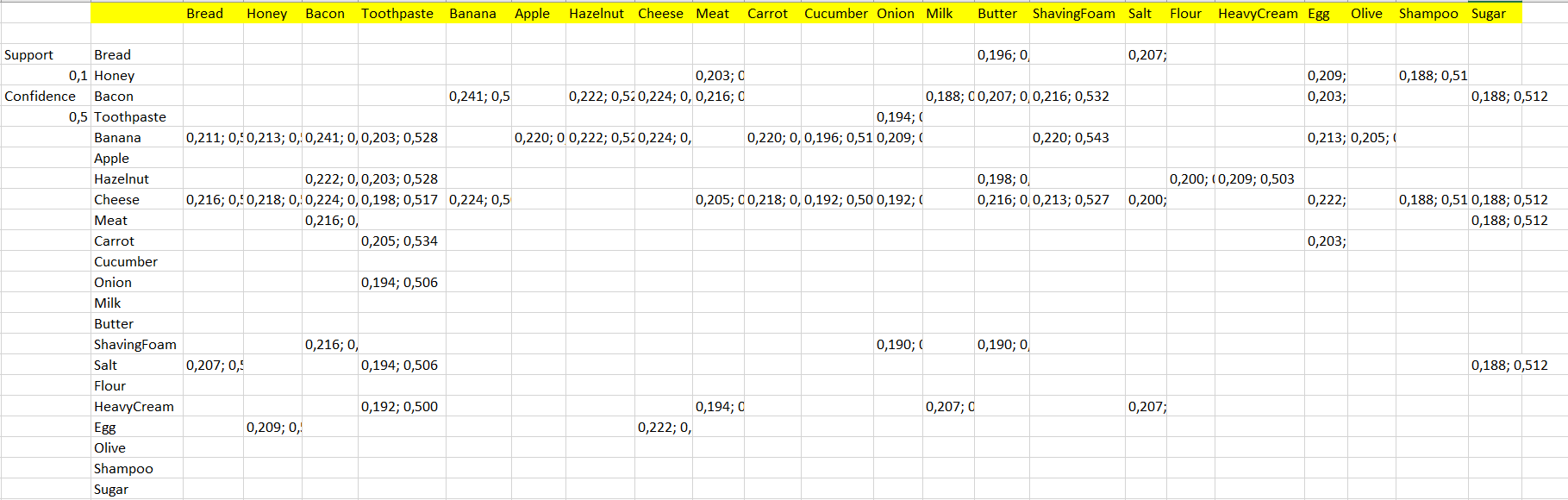
Генерація таких правил дозволяє аналізувати фактори, які впливають на купівлю чи відсутність купівлі цільового атрибуту (в даному випадку **Toothpaste**). Подібні правила можуть бути корисними для створення рекомендаційних систем чи оптимізації маркетингових стратегій.

3)

Сформував власний набір транзакцій покупок:



Знайшов двоелементні групи частих наборів:



**Висновок:** У цій лабораторній роботі досліджено алгоритми Apriori та FP-Growth для пошуку асоціативних правил у наборах даних. У Weka було знайдено правила з високою достовірністю та підтримкою. Алгоритм Apriori виявив часті стабільні правила, тоді як FP-Growth працював швидше й знайшов рідкісні комбінації. В Excel створено набір транзакцій і знайдено двоелементні часті набори. Отримані результати можна застосувати для оптимізації викладки товарів та персоналізованих рекомендацій у супермаркетах.