NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

SC4020 Data Analytics and Mining

Academic Year 2025/2026

Semester 1

Group 18

LUNBERRY NOAH IWATA (N2503869H)

PIKERINGA ANTONINA DAILA (N2504101A)

RAHLFS FREDERIC MAURITZ (N2504096K)

SARAH EMILY ONG XIN WEI (U2440124G)

Table of Contents

1 Task 1: Analysis of Symptom Co-occurrence Patterns

In this task, we analysed the co-occurrence pattern of different symptoms within disease profiles using the Apriori algorithm.

1.1 Data Preprocessing

The dataset is loaded from a CSV file from the Disease Symptom Prediction Dataset. It was then cleaned to remove null values. Each row represents a transaction. We then extract the symptoms (items) associated with each case from every row (transaction).

The symptoms are then convereted to lowercase and we remove any whitespace This ensures that when we store the symptoms, there are no duplicate symptoms within transactions. Using the `TransactionEncoder` from the `mlxtend` library, the transactions are transformed via one-hot encoding. This facilitates the implementation of the Apriori algorithm.

1.2 Implementation of Apriori Algorithm

The Apriori algorithm is implemented using the `mlxtend` library, with the goal of mining frequent itemsets from the preprocessed transaction data. The minimum support threshold is set to 0.1, meaning that an itemset must appear in at least 10% of the transactions to be considered frequent.

Following the discovery of frequent itemsets, the `association_rules` function is used to generate association rules with a minimum confidence threshold of 0.6. This ensures that only rules with a confidence greater than or equal to 60% are considered valid.

The key parameters used are support and confidence, which are essential for filtering out weak rules and itemsets.

1.3 Results

The frequent itemsets and association rules discovered using the Apriori algorithm are shown in the table below. The frequent itemsets include individual symptoms that appear together in the dataset with a support greater than or equal to 0.1. The association rules highlight the relationships between symptoms, where the antecedents (conditions) are linked to the consequents (outcomes), with each rule being evaluated based on its support and confidence.

Antecedents	**Consequents**	**Support**	**Confidence**
dark_urine	abdominal_pain	0.110976	0.957895
abdominal_pain	loss_of_appetite	0.132927	0.633721
abdominal_pain	vomiting	0.176829	0.843023
yellowing_of_eyes	abdominal_pain	0.114634	0.691176
yellowish_skin	abdominal_pain	0.154878	0.835526

Table 1: Association Rules with Support and Confidence

1.4 Second Place Monologue

There is nothing wrong with second place. Your best effort is all that anyone's asking for. And, if you give your best and you come in second, you come in third, you come in last, it's not about winning or losing. It's about giving it everything you've got. Now, Sam has built a monument to devilry, and chaos. I, deserve second place, I came in second. The only crime that's been committed here is that Oscar and Ally deserve first. We should be applauding them for getting more points. But in this, sick rodeo, this, bizarre, *fucked-up* clown festival that Sam has put together, we're here celebrating, what I can only describe as the sickness at the core of America. And uh, I'm gonna get him, I'm gonna get Sam

2 Task 2: Mining Cancer Feature Patterns

3 Task 3: Open Advanced Task

4 Conclusion

Appendix A - Breakdown of Contributions

LUNBERRY NOAH IWATA (N2503869H)
Coding, analysis, and write-up for Task 2.

PIKERINGA ANTONINA DAILA (N2504101A)
Coding, analysis, and write-up for Task 2.

RAHLFS FREDERIC MAURITZ (N2504096K)
Coding, analysis, and write-up for Task 3.

SARAH EMILY ONG XIN WEI (U2440124G)
Coding, analysis, and write-up for Task 1.