Pattern Mining		Summer Semester 2024
Student number	Surname	First name

100 points **Deadline: 15.04.2024 12:00**

Please write short, succinct answers, that are to the point. Please submit PDF reports, scripts, and KNIME workflows (if applicable).

Please indicate if you work on the assignment together with someone else. You are allowed to work in groups of at most **two** individuals. Please note, however, that you should not partition the work but rather work together.

Please disclose any help you got from artificial intelligence (AI) tools such as ChatGPT. Carefully evaluate the answers that you get from such tools and critically discuss those answers. If you use AI tools, be sure to understand your submission and be able to explain and possibly modify it during the assignment review.

In order to be awarded any points for your submission you have to be able to explain the submitted solutions.

In general, for the following tasks, you may use R and Python.

Task 1: Product Orders (50 points)

You have tabular data about store sales. The `Order.ID` determines which rows belong to the same order, i.e., which products were bought together.

File: `store.csv`

a. Data Preparation (10 Points)

Familiarize yourself with the provided dataset for product orders. Document and eliminate any data quality issues that you may find (if you find any), i.e., perform the necessary data cleaning.

Choose a suitable package for association rule mining for the platform of your choice. Transform the dataset into the format required for association rule mining.

For this task, you may also use KNIME for the data preparation (besides R or Python).

b. Mining Association Rules (20 Points)

You should now discover potentially interesting association rules in the data.

Determine appropriate support and confidence levels. Justify your choices.

Determine the frequent itemsets using an algorithm (package) of your choice.

Try to derive at least one interesting association rule, for which you conduct a detailed evaluation, including an analysis of the rule's contingency table as well as computation of the rule's confidence, support, lift, Kulczynski coefficent, and imbalance ratio, if applicable.

c. Mining Multilevel Association Rules (20 Points)

The `Category` and `Sub.Category` attributes define for each product the corresponding category and subcategory, respectively. Use those columns to discover potentially interesting multilevel association rules. A completely arbitrary example of a multilevel association rule would be `orders('Hamilton Beach Toaster, Black')` ⇒ `orders('Furnishing')`. Check for redundancy of the discovered association rules. Justify why the discovered rules are interesting.

Hint: The `arules` package for R supports hierarchies. Check out the manual on Moodle.

Task 2: Health Data (50 points)

The provided health dataset consists of seven attributes: `ID`, `patientID`, `date`, `reportID`, `type` ('S' = symptom, 'C' = condition, 'T' = treatment, 'N' = note, 'W' = weather, 'F' = food), a `name` and a `value`, the type and range of which is dependent on the type, e.g., if the type is 'T' (treatment), the value is the prescribed quantity of the medication. The `reportid` attribute determines which rows belong to the same report. Be aware that labels for symptoms, conditions, treatments, etc. are not necessarily disjoint, i.e., the same label may denote a condition and a symptom, which you should take into account as it may affect the working of the employed algorithm.

File: `data.csv`

a. Data Preparation (10 Points)

Familiarize yourself with the provided health dataset. Document and eliminate any data quality issues that you may find (if you find any), i.e., perform the necessary data cleaning.

Choose a suitable package for association rule mining for the platform of your choice. Transform the dataset into the format required for association rule mining.

For this task, you may also use KNIME for the data preparation (besides R or Python).

b. Mining Association Rules (20 Points)

Look for associations between different symptoms, different conditions, and between different treatments, respectively, i.e., look at which symptoms frequently occur together, which conditions frequently occur together, and which treatments frequently occur together. Determine appropriate support and confidence levels. Justify your choices.

Try to derive at least one interesting association rule for symptoms, conditions, and treatments, for which you conduct a detailed evaluation, including an analysis of the rule's contingency table as well as computation of the rule's confidence, support, lift, Kulczynski coefficent, and imbalance ratio, if applicable.

c. Mining Multidimensional Association Rules (20 Points)

Look for associations between treatments and conditions, weather and conditions, and food and conditions. Determine appropriate support and confidence levels. Justify your choices.