Assignment 2

Please follow the following rules when working on this assignment:

- This assignment is mandatory for all students.
- Work on this assignment in teams of three students. Each team has to prepare its own solution.
- Prepare a result document (PDF) with your solutions for all the tasks listed below.
    - You may write this document in English or German.
    - The PDF has to include the team number, name, study programme and matriculation number for each team member.
    - Document name: A2-Txx-Vy.pdf (where “xx” is the team number and “y” the version).
- Submit this document in Ilias no later than February 8, 10:00 am. This works as follows:
    - Go to ‘Submit Assignments’ and select Assignment 2.
    - One team member has to select 'Create Team' and afterwards 'Manage Team' to add your team members. Choose 'Add Users of Current Course' to add your team members.
    - Note that you can only select individual students and not directly the teams/groups we already have organized in Ilias. Anyway, the teams you are creating to submit your document have to match one of the available teams.
    - Finally click 'Hand in' to submit the prepared PDF.
    - Please make sure that you submit only one document per team and that you add all team members before submitting your document.
- Contact Holger Schwarz for any further questions or use the forum in Ilias.

Task 1:

Given set T of transactions. Each transaction is identified by a transaction ID and contains a set of items.

T:

ID	**items**
1	A, F
2	A, D, G
3	A, C, F
4	D, F
5	B, C, F
6	A, F
7	A, B, D
8	B, C, D, F
9	C, E, G, H
10	B, C, E, F

a)

Explain how the **support** of an itemset is defined and calculate the support of itemset {B, C, F}.

b)

Explain how the **confidence** of an association rule is defined and calculate the confidence of rule {B, C} → {F}.

c)

Use the **Apriori algorithm** to determine all frequent itemsets based on transactions T. The minimum support is 0,25. Provide for each iteration of the algorithm a separate table with the results (generated candidates, support of the itemsets, identified frequent itemsets) similar to the table below. When does the algorithm stop?

candidate itemset	support	frequent?
		☐ yes ☐ no
		☐ yes ☐ no
		☐ yes ☐ no
		☐ yes ☐ no

Task 2:

Remember the Mountain States Health Alliance (MSHA) from the exercises. The management has identified the following issues:

Descriptive
Association

a) To improve cost efficiency, the variety of drugs used in the MSHA should be reduced. To achieve this goal, it is necessary to identify typical combinations of drugs that doctors prescribe or that are used in certain departments.

Descriptive
Clustering

b) The MSHA management identified unusual high costs for drugs in certain departments. The goal is to do some root cause analysis.

Predictive
Classification

c) The MSHA offers various services to former patients as part of its aftercare program. Hence, the hospital would like to inform the patients about suitable services four weeks after they have been discharged. The goal is to provide such suggestions based on ~~services that were offered to patients or services patients made use of in the past~~.

Predictive
Regression

d) For planning purposes, the health alliance needs a good estimate for the time patients will stay at a hospital. The goal is to provide such an estimate based on data about previous patients that already have been discharged.

Discuss for these issues how data mining could provide relevant information that supports management in decision making.

Your result document (only two to three pages for the entire task) must include for each issue:

- a justification for the data mining technique you would use (Why descriptive or predictive? Which one is most suitable?)
- a list of attributes that you consider important as input to the data mining algorithm and the semantics of these attributes
- a description of the patterns that could be derived
- examples how these patterns could be used to address the mentioned issue