BPI Assignment 1

This assignment is to be executed individually. (Copying solutions from each other is considered to be fraud.) The grade for this assignment counts for 20% of the final grade. The deadline for the assignment is Sunday, 20/05/2018, 23:59, and it is a hard deadline. Do not submit at the very last minute. Last minute technical problems when uploading are your own responsibility and the submission facility will automatically reject later upload attempts.

Participation in the assignments is required for participation in the final test. Only the final test can be retaken; assignments can only be redone in the next academic year.

Q1 Clustering (1.5 points) Using the K-Means algorithm, cluster the following instances into three clusters. For this question just consider the attributes "AMOUNT_REQ", "Case duration", and "Total Activities".

Process Instance ID	AMOUNT_REQ	Case duration	A_ACCEPTED	A_APPROVED	A_DECLINED	O_ACCEPTED	O_CANCELLED	O_DECLINED	Total Activities
174574	7000	0,29309	0	0	1	0	0	0	6
174577	15000	28,43964	1	1	0	1	0	0	74
174599	23112	0,000532	0	0	1	0	0	0	3
174602	6000	0,048206	1	1	0	1	0	0	25
174605	10000	12,95023	1	1	0	1	0	0	26
174608	2500	0,021134	0	0	1	0	0	0	7
174935	19000	19,78213	1	1	0	1	1	0	45
174938	5000	29,51885	1	1	0	1	0	0	46
174941	13000	0,515625	0	0	1	0	0	0	10
174944	35000	19,74352	1	1	0	1	1	0	88
174947	5000	7,612419	1	0	0	0	1	0	25
174950	7500	0,489861	0	0	0	0	0	0	11
174953	6000	6,503808	1	1	0	1	0	0	22
175284	6500	0,002049	0	0	1	0	0	0	6
175287	25000	21,14506	1	0	1	0	0	1	41
175290	10000	0,000799	0	0	1	0	0	0	3
175293	7800	19,1099	1	1	0	1	0	0	52
175296	8500	0,000486	0	0	1	0	0	0	3
175299	17000	0,01375	0	0	0	0	0	0	12
191449	3000	6,955382	0	0	0	1	0	0	15

You need to use the initial values for centroids mentioned in the following table.

Centroids	AMOUNT_REQ	Case duration	Total Activities
K1	50000	1.5	6
K2	10000	0.7	10
K3	30000	4	26

Please note that you just need to repeat the algorithm (find the distances) three times. In each step find the distance of instances with centroids and specify the cluster of each instance. You should do this task with Excel, manually, or using your own program. It is not required to normalize the data beforehand. You are not allowed to use RapidMiner (or similar data mining tools) for this task and need to show the intermediate steps.

Q2 Association rules (1 point) Based on the data used in the previous question discover all the association rules with a minimum support of 0.3 using RapidMiner. Then compute confidence and lift of all discovered rules (just take a screenshot of the results in the RapidMiner). For this question only consider the attributes "A_ACCEPTED"," A_APPROVED", "A_DECLINED", "O_ACCEPTED", "O_DECLINED" and "O_DECLINED". Select two of the best rules and discuss why you selected them.

Q3 Decision Tree (2 points) Consider again the data used in Question 1. Let "Total Activities" be the response variable, and learn a decision tree for these data. Instances where "Total Activities" is higher than 40 should be mapped to class "High" and others should be mapped to class "Low". The maximum depth of the tree should be set to three. After discovering the tree explain the result by explaining the resulting rules and interpreting the confusion matrix (this task should be done with RapidMiner).

Q4 Alpha Algorithm (2.5 points) For each of the following event logs apply the Alpha algorithm (show the footprint and the model), then discuss whether the discovered model is a sound process model or not. This task should be done manually.

L1= [⟨a,b,e,f⟩,⟨a,b,e,c,d,b,f⟩,⟨a,b,c,e,d,b ,f⟩,⟨a,b,c,d,e,b,f⟩,⟨a,e,b,c,d,b,f⟩]

L2= [⟨a,b,c,d⟩,⟨a,c,b,d⟩,⟨ a,e,f,d ⟩,⟨a,e,g,d⟩,⟨ a,e,h,d ⟩]

L3= [⟨d,c,b,e,f⟩,⟨a,e,f⟩,⟨ d,b,b,c,ef⟩,⟨a,b,c,d,e,f⟩,⟨ b,d,a,c,f ⟩]

Q5 Using Process Mining tools (3 points)

The data set for this part of the assignment can be downloaded via the Moodle system.

Using Disco and ProM answer the following questions:

a) How many process instances and events are in this event log? What is the median number of events in each trace and what is the average duration of them?
b) What will be the Disco process model for this event log when you set *Activities* slider to 80% and *Paths* slider to 10%? Interpret the (self-loop) edge from "Payment" to itself. How many times and for how many process instances this behavior happens?
c) Using Disco, analyze the time distribution of events over the time covered by this log? What is your interpretation of the distribution of events over the time? Do you see any remarkable patterns in the distribution of events (drifts, repeating behavior, etc.)?
d) How many variants are in this event log? What is the size (number of process instances) of the third most frequent variant? Also, explain what happens in this variant (report the sequence of activities).
e) Filter the event log using Disco while keeping 50% of the most common variants of process instances that finished until 01.01.2012 12: 12. What happens to the median and the average of case duration compared to the whole event log. Please also explain the difference between the median and the average (mean) of case duration.
f) Discover the dotted chart view of the event log and interpret it using ProM. Adjust the Dotted chart settings in a way to answer question *c*. Are there any interesting (or odd) patterns in the dotted chart view that could explain the patterns found when answering question *c*?
g) Filter the event log using "filter log using simple heuristics" and then apply the Alpha algorithm (*Alpha ++*) on it. Filter the log appropriately. Show some insights that can only be seen after filtering.
h) Use the filtered event log in the previous question and this time use it as input for Disco. Which parts of the process are time consuming for most of the process instances? Answer this by interpreting the Disco model.

Report requirements

The assignment should be uploaded to the Moodle system as a single zip file named **StudentNumber_BPI_assignment_1.zip** (StudentNumber is your student number). Your zip file must contain following files:

1. **Process models for RapidMiner and the files you used as input there.**
2. **Report_ StudentNumber.docx or Report_ StudentNumber.pdf**: A word or pdf file containing the following information
    - Succinct but very clear answers of the questions. About being verbose while not answering the question.

- Describe the analysis performed to answer each question, including the details about the filters applied, etc.
- Provide the findings, accompanied with screenshots of different model variants, statistics, decision trees, association rules.
- Provide screenshots of ProM and Disco to answer the process-related questions.
- Provide conclusions based on your analysis and include suggestions for possible improvements of the process.

The report should not exceed 15 pages of length. Do not forget to put your Name and your Student ID on the title page of the report!