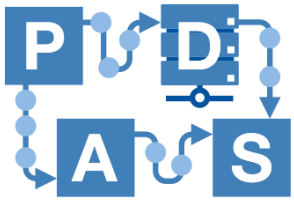
Advanced Process Mining

Assignment Part 1 - Description

RWTH Aachen – Summer Semester 2020

Tobias Brockhoff MSc
Lisa Mannel MSc
Sebastiaan J. van Zelst MSc PhD



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Introduction

In this assignment you will deal with a hypothetical process of large-scale infection testing and potential treatments. The healthcare institution has extracted a log in csv format that is provided as "log.csv" for further analysis using PM4Py. As we all want to become good data/process scientists, we will carry out our analysis in a python notebook using JupyterLab. In addition, you are supposed to deliver a written report.

When answering the questions, document what you did, as well as carefully describe and explain your results. In particular, especially for questions 3-5, explain how you derived your results, on which facts you base your claims, and motivate the methods you used.

It is reccomended to work on the tasks from top to bottom, since some questions may make use of preceding results.

Clearly indicate, which answer belongs to which question. Note, that the questions separate the presentation of results, their description and their discussion. This should also apply for the presentation of your analysis in the notebook.

Results from previous questions can (and should) be referred to, to improve your discussion and explanation.

Total Number of Points Obtainable

100 Points (20 % of final grade)

Group Size

Work in groups of size 2-3. Clearly indicate all group members on your report.

Input

JupyterLab Template, Event log, Report Template

Optional Resources

Social network mining

- W.M.P. van der Aalst and M. Song. Discovering Social Networks from Event Logs. BETA Working Paper Series, WP 116, Eindhoven University of Technology, Eindhoven, 2004.

Decision trees

- Process mining MOOC lectures '1.4: Learning Decision Trees' and '1.5: Applying Decision Trees'. Note that the MOOC also contains lectures on process model quality, social networks, etc

PM4Py

- New Doc: https://pm4py.fit.fraunhofer.de/documentation#discovery
- Old Documentation (some additional information about e.g. decision trees): http://pm4py.pads.rwth-aachen.de/documentation/
- Source Code: https://github.com/pm4py/pm4py-source/tree/release/pm4py

Deliverables

The assignment should be submitted via RWTHMoodle. Upload a PDF report based on the provided template, describing your methods and the results you obtained. The document should not contain more than **20** A4 pages, including title page (which shows all student numbers of your group), excluding appendices. If you cannot use the provided template, create a document mimicking the template by a method of your choice.
Moreover, upload an analysis notebook which satisfies the following requirements:

- The cells can be run top to bottom in order to reproduce your results
- Minimal number of additional dependencies (not counting PM4Py or modules that are in a basic anaconda installation)
    - Additional dependencies have to be free and installable from standard online repositories using `pip install` or `conda install` (NO further dependencies)

- All larger output figures are also saved as pdf or png in the provided figures folder

Do not re-upload the event log

Notebooks that intentionally access files or "play outside" of the notebook directory will lead to failing the exam

The Data

The provided artificial event has the following columns:

- *Patient: The patient unique id*
- *Activity: Activity in the process*
- *Resource: Resource involved in the given activity*
- *PatientName: Name of the patient*
- *Age: Age of the patient*
- *Insurance:* statutory health insurance or private health insurance
- *Start_timestampe: Start timestamp of the activity*
- *TimeStamp: End timestamp of the activity*
- *@@duration: Duration of the activity*

Setup

1. Install PM4Py (https://pm4py.fit.fraunhofer.de/install)
2. Install JupyterLab (https://jupyter.org/install)
3. Download the template notebook and directory
4. Start JupyterLab (see 1.)
5. Browse to the template notebook
6. Have fun ☺

Tasks (90 Points + 10 Points for style)

Q1. Inductive Miner (25 pts)

In this question you should discover a model for the given event log with a special focus on the Inductive Miner implemented in PM4Py.

a) Apply the Inductive Miner implemented in PM4Py to the given event log and describe the process. Furthermore, give and reason about the fitness and precision results, respectively. On a high level, describe the potential problems of the model and reason how they were caused by the algorithm and the log. (7 pts)

b) From the process owner we know that patients are called in order to control the quarantine and that there are two potential quarantine phases, i.e., before and after a positive test. Implement a function that resolves the duplicate activity *Control Call* by context sensitive renaming. Discuss the impact on the discovered model. (5 pts)

(*Hint*: The *Test* activity is not affected by noise)

c) The log has considerable data quality issues induced by errors during the event logging. Apply the IM to a DFG filtered for noise. Describe your results and explain why the IM mines a different model. Which type of noise is prominent in the log? (5 pts)

d) Investigate the DFG of the log after applying the preceding steps. Which activities might be filtered out in order to obtain an improved model that explains most of the process more precisely? Why might this yield better results when applying the IM? Implement a filter and apply the IM to the filtered log. (3 pts)

(*Hint*: Have a look at the log utility in the sources of the PM4Py project)

e) Consider the process model for the patients who were prescript the special medication. What do you observe? How is this behavior captured by the complete model in d)? (2pt)

f) Apply additional miners to the log and compare the results. Which model is the best model? (3pts)

Q2. Social Network Analysis (12 pts)

Discover the organizational perspective of the process. For each of the following networks, try to find a clear organizational structure and discuss the structure obtained. If no clear structure is to be found, explain why this is the case.

a) Handover-of-Work Social Network (3 pts)

b) Subcontracting Social Network (3 pts)

c) Working-Together Social Network (3 pts)

d) Joint-Activity Network (3 pts)

Q3. Performance Analysis (20 pts)

Which parts of the process have the biggest influence on the total case duration?

a) Provide and briefly describe results of your performance analysis. Remember to also consider your current results which may give you a good entry point for a deeper analysis. (15 pts)

b) Discuss insights obtained from you analysis, for example identify bottlenecks, and discuss their impact. (5 pts)

Q4. Decision Points (20 pts)

Investigate how patients are referred for further treatment by means of a decision tree. Describe the factors that you observe.

a) Create a decision tree of reasonable complexity using the available attributes in the log. (7 pts)

b) Since it is likely that the resources at the treatment facilities are limited, implement a function that assigns a(n) (estimate) of the number of patients at each facility to each event. To this end, you have to decide which event occurs at which facility based on your analysis in question 2. Create a decision tree of reasonable complexity using this derived attribute. (13 pts)

(*Hint*: There is an easy pattern that might help you to find a proper facility assignment)

Q5. Process Improvement Suggestions (13 pts)

Based on the information that has been obtained for the previous four questions, are there any opportunities for improving the care process? For each of the above questions can you mention any improvement opportunity? If yes, indicate how the process can be improved. If not, explain why.