Homework 4: Predictive Process Monitoring

This task uses the Turnaround process event log that we used in Practice 8 of the course. The link has been provided in Slack.

This goal of this homework is to train and use predictive process monitoring techniques to predict the outcome of a process from a log of events. Make the necessary modifications to the framework proposed by Taineema et al. (link[1]) reviewed in class to meet this goal.

Tasks

1. (1 point) As part of the log preprocessing, it is necessary to categorize the process traces as deviant or regular. This log contains a column called SLA. it is a "case attribute," which indicates how many minutes each case must complete. You must create a new column in the log that contains a case attribute called *label*, which contains a value of 1 for deviant cases or 0 for regular ones. This column's value is 0 if the duration of the case (in minutes) is less than or equal to the SLA; otherwise, this column's value must be 1 (the SLA has not been met). NB! If there are cases that do not have SLA, categorize them as 0.
2. (2 points) Add a column to the event log that captures the WIP of the process at the moment where the last event in the prefix occurs. Remember that the WIP refers to the number of active cases, meaning the number of cases that have started but not yet completed.
3. (4 points) Currently, the work proposed by Taineema et al. performs the extraction of the prefixes of the traces registered in the log to train the classification models. For large logs, this approach leads to an increase in the dimensionality of the models' input (too many features) without necessarily improving its precision, especially in cases in which the event traces are very long. You must modify this technique to extract subsequences of size n (*n*-grams), where *n* is a user-defined parameter, instead of encoding entire prefixes. An n-gram is a contiguous sequence of n items from a given trace. For example, in the following trace of a process, all possible n-grams of size three were extracted:

Trace									
Event number	**1**	**2**	**3**	**4**	**5**		**6**	**7**	**label**
Activity label	ER Registration	Leucocytes	CRP	LacticAcid	ER Triage	ER Sepsis Triage	IV Liquid	IV Antibiotics	1

Ngram number	**Ngram**								**label**
1	ER Registration	Leucocytes	CRP						1
2		Leucocytes	CRP	LacticAcid					1
3			CRP	LacticAcid	ER Triage				1
4				LacticAcid	ER Triage	ER Sepsis Triage			1
5					ER Triage	ER Sepsis Triage	IV Liquid		1
6						ER Sepsis Triage	IV Liquid	IV Antibiotics	1

NB! Consider using the n-grams module from the nltk library for python (link[2]).

4. (3 points) Test the results of your modifications with the Turnaround process event log using cluster bucketing, index encoding, and the XGboost model.

[1] https://github.com/Mcamargo85/predictive-monitoring-benchmark.git

[2] https://stackoverflow.com/questions/17531684/n-grams-in-python-four-five-six-grams

What to submit?

A report in PDF format containing the explanation of the modifications made to the approach, the evaluation of the changes, and the link to the repository containing the modifications made. You must submit this document via the Submit link on the course website.