

To prepare the data for our experiments, we employed basic tactics for preprocessing. First, we split the log into a training and a test set. The test set will act as our primary source for evaluating factuals entirely unknown to the model. We split the training set into a training set and a validation set. This procedure is a common tactic to employ model selection techniques. In other words, Each dataset is split into 25% Test and 75 remaining, and from the remaining, we take 25% validation and 75% training data.

First, we filter out every case whose sequence length exceeds 25. We keep this maximum threshold for most experiments focusing on the evolutionary algorithm. The reason is the polynomial computation time of the viability measure. The similarity and sparsity components of the proposed viability measure have a run time complexity of at least  $N^2$ . Hence, limiting the sequence length saves a substantial amount of temporal resources.

Next, we extract time variables if they are provided in the log. Then, we normalise the values. For a time format, we encode all information from seconds to a year. If the complete log occurs within one time unit only, e.g. every event that happened within a year, drop the extracted column—afterwards, we standard scale all remaining time features.

Each categorical variable is converted using binary encoding. Binary encoding is very similar to one-hot encoding. However, it is still distinct. The binary encoding uses a binary representation for each class encoded. This representation saves a lot of space as binary encoded variables are less sparse than one-hot encoded variables.

We also add an offset of 1 to binary and categorical columns to introduce a symbol which represents padding in the sequence. All numerical columns have a zero mean and a standard deviation of 1.

We omit the case id, the activity and the label column from this preprocessing procedure for reasons explained in ???. The activity is label-encoded. Hence, every category is assigned to a unique integer. The label column is binary encoded, as we focus on outcome prediction.