This file describes the analysis plan for the experiments of the journal paper "Learning by Sampling: Learning Behavioral Family Models from Software Product Lines" submitted to the Empirical Software Engineering Journal: special issue on "Configurable Systems".

This journal paper is an extended version of a study published in the 23rd International Systems and Software Product Line Conference - Volume A 2019.

The analysis of the data of this study is automated in the **data/wise2learn.sh** bash script. Originally, this script has been designed to run our analysis as parallel jobs submitted to our cluster. However, each step can also be performed in a standard computer by running each set of commands in the aforementioned script. The analysis plan of this study works as follows:

(1) FSM models are derived for all valid product of the SPLs "minepump" "aerouc5" "cpterminal".

```
- These SPLs are given as FTS models are used to derive FSM models
  given a product configuration.

- The FSM models are using the **learnFFSM.jar** program.

- To derive FSMs, you shall use the command:

    java -cp learnFFSM.jar uk.le.ac.fts.FsmFromFTS

    with the following parameters:

        -fts <fts> with the path to a file describing
                   a Featured Transition System, and

        -conf <cnfg> with the path to a file containing a
                     configuration file to assist the
                     FSM derivation process.

- These refer to lines 4-8 of the data/wise2learn.sh script

Obs.: The "agm" "vm" "ws" SPLs do not need this conversion.
      They are already described as (F)FSM models
```

(2) Once the FSM models are generated, FFSM models shall be learned from each sampled subset of valid products.

```
- The sampling process can be made using the FeatureIDE toolkit.

- The sampled configurations shall be organized as in
  the **products_*.prtz** files. These are found in the folders
  named as products_[1-4]wise and products_all in the data directory.
```

- The FFSM learning process can be run using the emse_prtz.py script.
  This script performs the learning process following the order set
  in the .prtz file. Two log files are generated:

    - report.log -> It indicates a set of statistics of each learning
                     process, such as numbers of transitions/states,
                     landmarks/state mapping, precision and recall

    - report_fmeasure_l.log -> It essentially indicates the precision
                                of the FFSM models learned compared to all valid
                                product FSMs. This process is done using the class
                                uk.le.ac.compare.CompareStructure.

- These steps are indicated in the emse_prtz.py file.
  The emse_prtz.py script is called for each SPL
  in lines 8-12 and 18-21 of the wise2learn.sh script.

(3) Once the FFSM model recovery process is done, we start the FFSM learning from all pairs of FSMs. This step is indicated in lines 26-28 of the wise2learn.sh script. For each SPL, there is an specific python script that runs this process. These scripts are named as emse_pairs_.py

(4) Once the FFSM models are learned from all pairs, we calculate the similarity for all models learned. This step is indicated in lines 30-32 of the wise2learn.sh script. For each SPL, there is an specific python script that runs this process. These scripts are named as emse_dissim_.py

(5) Once the learning processes are done, we have to tabulate the log files so we can run our statistical analysis. The log files are tabulated using the grep/sed commands so that only the required statistics are collected. These are indicated in lines 37-74 of the wise2learn.sh script. Two tab files are generated for each sampled subset of all SPLs:

    - report.tab -> Tabulated version of report.log

    - report_fmeasure_l.tab -> Tabulated version of report_fmeasure_l.log

(6) The data/exp_emse/learningFFSMs.Rproj file is an RStudio project file that supports the statistical analysis of the experiments. As result, the data/exp_emse/script.r file generates a few plots and tables:

- boxplot_pairs_<states|transitions>_size.pdf -> These indicate the size
    of the FSM models learned from all pairs of FSM models of the SPLS

- correlation_* -> These depict the correlation between multiple parameters:
    configuration similarity/ratio of features vs. ratio states/transitions

- histogram_* -> Histograms for config. similarity, ratio of features/states/transitions

```
- Precision_* -> Boxplots indicating the precision of the FFSM models
    learned from each sampled subset of product configuration

- twise_sizes -> Bar plots indicating the size of each sampled subset.
```

(7) The files generated in our experiment are found in the zip file data/wise2learn.zip and the data/exp_emse folder contains all plots and tables that supported the analysis of results in the journal paper.

If you have any questions, feel free to contact me via damascenodiego@alumni.usp.br