# master\TOPSIS.py

Contains the TOPSIS program that accesses the alternative-criteria matrix master\Results.csv and calculates the scores for each alternative.

**Parameters**

fleet\_type. Elimination criteria that selects a subset of the data based on the fleet type, that should be used for further calculation. Possible values: **'Identical'**, **'Homogeneous'**, **'Heterogeneous'**

fleet\_feature\_type. Elimination criteria that selects a subset of the data based on the fleet feature type, that should be used for further calculation. Possible values: **'Numerical'**, **'Categorical'**, **'Semantics'**, **'None'**

output\_type.Elimination criteria that selects a subset of the data based on the output type, that should be used for further calculation. Possible values: **'Point-estimate'**, **'Interval'**, **'Distribution'**

w. Criteria weighting vector of ‘Missing Data’, ‘Noise’, ‘Low Sample Size’, ‘General Accuracy’, ‘Robustness’, ‘Time Complexity’, ‘Space Complexity’, ‘Explainability’, ‘Parameter Handling’.

The set values were used for the Master thesis.

# master\backblaze\data\_exploration.py and master\backblaze\data\_exploration.r

The Python script contains the explorative data analysis, the R program contains explorative analysis and visualizations of the results.

# master\backblaze\load\_data.py

This script contains helper functions to retrieve data from the csv files that are published by Backblaze. The script is used in master\backblaze\main.py.

# master\backblaze\load\_from\_r.py

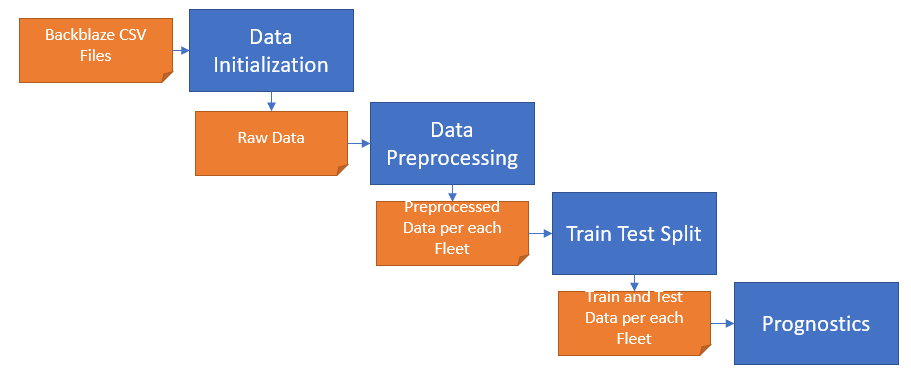
This script allows to return a pickle in dataframe format that can be used in an R script via the reticulate package.

# master\backblaze\PlottingHelpers.py and master\backblaze\ProcessingHelpers.py

Contains helper functions for processing and visualization only used in master\backblaze\data\_exploration.py.

# master\backblaze\main.py

Main program that comprises the 4 steps data initialization, preprocessing, train and test split and proportional hazard model prognostics (master\backblaze\Main.r contains the prognostics for the random survival forest). The following diagram shows the steps and their inputs/outputs:



**Parameters**

General.

* data\_path: The path where all inputs are retrieved from and all outputs are saved to.
* cols\_no\_data: Contains the columns that do not contain any SMART measurements. Should be kept fixed.
* filtered\_models: Contains the models that should be processed. To process manufacturers, insert the initial letters of the manufacturers (T for Toshiba, H for Hitachi and HGST, W for Western Digital, Sa for Samsung and ST for Seagate)

Data Initialization.

* initialize\_data\_set: Set true to execute the initialization script. Only necessary if you have not generated the further data (i.e. raw, preprocessed or train and test data).
* start\_date: Lower bound of csv file that should be extracted.
* end\_date: Upper bound of csv file that should be extracted.

Data Preprocessing.

* preprocess\_data\_set: Set true to execute the preprocessing script. Only necessary if you have not generated the further data (i.e. preprocessed or train and test data).
* sample\_size: Allows decreasing of sample size for testing purposes. 1 does not decrease the data.
* measurements\_in\_last\_n\_days: Removes HDDs that do not have measurements of critical SMART values in the last n days.
* threshold\_nan: Deletes columns if they have a share of n NaN values.
* elbow\_point\_detection: Boolean value to enable/disable the elbow point detection.
* rolling\_window: Number of days for the feature extraction.

Train Test Split.

* train\_test\_split: Set true to execute the train test split script. Only necessary if you have not generated the further data (i.e. train and test data).
* test\_size: Size of the test data. Rest is training data.

Prognostics.

* svm: Obsolete. Set to false.
* cox: Set to true to enable the proportional hazard model
* k\_cv: Obsolete.

# master\backblaze\main.r

Main program that contains the prognostics for the random survival forest. It is required that all steps of master\backblaze\main.py have been executed until the train test split, because the R script uses this as an input.

**Parameters**

topimpv: The number of most important variables that are used for prediction.

ntree: The number of trees used for the random survival forest.

data\_path: The absolute data path for output and input files.

sample\_size: Downsampling rate used for variable importance calculation.

models: Vector of models that should be trained and predicted.