{{design.mltask.name}}

{{design.visual\_analysis.name}}





|  |  |  |
| --- | --- | --- |
| **Version** | **Author** | **Date** |
| 1.0 | {{config.author.name}}{{config.author.email}} | {{config.generation\_date.name}} |

Table of contents

[I. Methodology 3](#_Toc60799681)

[A. Data Preparation 3](#_Toc60799682)

[B. Data Ingestion 4](#_Toc60799683)

[C. Model and Feature Tuning 5](#_Toc60799684)

[1. Pre-processings 5](#_Toc60799685)

[2. Pre-processings 5](#_Toc60799686)

[3. Tested Algorithms 6](#_Toc60799687)

[a) Selected Model 6](#_Toc60799688)

[b) Alternative Models 6](#_Toc60799689)

[4. Tested Algorithm 7](#_Toc60799690)

[II. Experiment Results 7](#_Toc60799691)

[A. Selected Model 7](#_Toc60799692)

[B. Alternative Models 8](#_Toc60799693)

[III. Selected Model Results 8](#_Toc60799694)

[A. Selected Model Metrics 8](#_Toc60799695)

[B. Sensitivity Testing and Analysis 9](#_Toc60799696)

[C. Diagnostics 9](#_Toc60799697)

[IV. Deployment and monitoring 9](#_Toc60799698)

[A. Implementation Details 9](#_Toc60799699)

[B. Version Control 10](#_Toc60799700)

# Methodology

This section deals with the methodological details used to build the clustering model:

{{if dataset.prepare\_steps.status != No}}

* *Data Preparation* steps may first be applied to generate the initial set of features used in the process and may also transform the target.

{{endif dataset.prepare\_steps.status }}

* *Data Ingestion* analyzes each feature in order to maximize its clustering potential.

{{if config.is\_saved\_model.value == Yes}}

* *Model and Feature Tuning* finds the best hyperparameter set for the selected algorithm.

{{endif config.is\_saved\_model.value}}

{{if config.is\_saved\_model.value != Yes}}

* *Model and Feature Tuning* describes the tested algorithms and the way to find the best hyperparameter set for each of them.

{{endif config.is\_saved\_model.value}}

## Data Preparation

When the original dataset is very big, the required computational resources may be too large compared to the expected benefit of training algorithms on it. As a result, the process may be performed on a subset of the dataset. The sampling method given in the table below defines how it is built.

{{design.training\_and\_testing\_strategy.table}}

|  |  |
| --- | --- |
|  |  |
|  |  |

{{/design.training\_and\_testing\_strategy.table}}

|  |
| --- |
| **Legend:*** *Sampling method:* A subset may have been extracted in order to limit the computational resources required by the evaluation and selection process. The *Record limit* gives its size.
	+ *No sampling (whole data)*: the complete dataset has been kept.
	+ *First records*: The first N rows of the dataset have been kept (or all the dataset if it has fewer rows. The current dataset has {{result.train\_set.sample\_rows\_count.value}} rows). It may result in a very biased view of the dataset.
	+ *Random (approx. ratio)*: Randomly selects approximately X% of the rows.
	+ *Random (approx. nb. records)*: Randomly selects approximately N rows.
	+ *Column values subset (approx. nb. records)*: Randomly selects a subset of values and chooses all rows with these values, in order to obtain approximately N rows. This is useful for selecting a subset of customers, for example.
	+ *Class rebalance (approx. nb. records)*: Randomly selects approximately N rows, trying to rebalance equally all modalities of a column. It does not oversample, only undersamples (so some rare modalities may remain under-represented). Rebalancing is not exact.
	+ *Class rebalance (approx. ratio)*: Randomly selects approximately X% of the rows, trying to rebalance equally all modalities of a column. It does not oversample, only undersamples (so some rare modalities may remain under-represented). Rebalancing is not exact.
 |

{{if dataset.prepare\_steps.status != No}}

The following data preparation steps are applied to the initial dataset in order to produce the Machine Learning dataset with {{design.features\_count.value}} features and {{result.train\_set.sample\_rows\_count.value}} samples:

{{dataset.prepare\_steps.table}}

|  |
| --- |
|  |
|  |

{{/dataset.prepare\_steps.table}}

{{endif dataset.prepare\_steps.status }}

## Data Ingestion

During the data-ingestion phase, the features are transformed into numerical features without missing values so as to be ingestible by the Machine Learning algorithm. The table below summarizes the processing applied to each of them.

{{design.input\_feature.table}}

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

{{/design.input\_feature.table}}

Legend

* *Feature name:* Name of the feature column
* *Feature status:* Input, Target or Rejected
* *Feature type:* Numeric, Category, Text, or Array
* *Processing:* Type of processing applied (Avg-std rescaling, dummy-encode…)

## Model and Feature Tuning

{{if design.outliers\_detection.status == Yes}}

### Pre-processings

The following outlier detection technique was used:

{{design.outliers\_detection.table}}

|  |  |
| --- | --- |
|  |  |
|  |  |

{{/design.outliers\_detection.table}}

{{endif design.outliers\_detection.status}}

{{if dataset.prepare\_steps.status != No}}

{{if design.outliers\_detection.status != Yes}}

### Pre-processings

{{endif design.outliers\_detection.status}}

The following dimensionality reduction technique was used:

{{design.dimensionality\_reduction.table}}

|  |  |
| --- | --- |
|  |  |
|  |  |

{{/design.dimensionality\_reduction.table}}

{{endif dataset.prepare\_steps.status}}

{{if config.is\_saved\_model.value != Yes}}

### Tested Algorithms

A selection of algorithms (candidate models) was then trained on the machine learning dataset, with various combinations of hyperparameters. The section below details the tested algorithms. It begins with the selected algorithm and its hyperparameter selection and continues with the other tested algorithms.

#### Selected Model

The {{result.chosen\_algorithm.name}} algorithm has been selected.

|  |
| --- |
| {{design.chosen\_algorithm\_search\_strategy.text}} |

The settings for this algorithm are given below. By definitions, hyperparameters are settings with multiple values. They are defined by lists or ranges in the table below:

{{design.chosen\_algorithm\_search\_strategy.table}}

|  |  |
| --- | --- |
|  |  |
|  |  |

{{/design.chosen\_algorithm\_search\_strategy.table}}

#### Alternative Models

Other algorithms are also tested. They are listed below, along with their settings:

{{design.other\_algorithms\_search\_strategy.table}}

|  |  |
| --- | --- |
|  |  |
|  |  |

{{/design.other\_algorithms\_search\_strategy.table}}

{{endif config.is\_saved\_model.value}}

{{if config.is\_saved\_model.value == Yes}}

### Tested Algorithm

The {{result.chosen\_algorithm.name}} algorithm has been tested.

|  |
| --- |
| {{design.chosen\_algorithm\_search\_strategy.text}} |

The settings for this algorithm are given below. For hyperparameters, the possible values or ranges are listed:

{{design.chosen\_algorithm\_search\_strategy.table}}

|  |  |
| --- | --- |
|  |  |
|  |  |

{{/design.chosen\_algorithm\_search\_strategy.table}}

{{endif config.is\_saved\_model.value}}

# Experiment Results

The methodology detailed in the previous section has been run. The obtained results are presented in this section.

{{if config.is\_saved\_model.value != Yes}}

## Selected Model

{{result.chosen\_algorithm.name}} was finally selected by the user with the set of hyperparameters given in the table below:

{{endif config.is\_saved\_model.value}}

{{if config.is\_saved\_model.value == Yes}}

The set of hyperparameters for the selected algorithm {{result.chosen\_algorithm.name}} is given in the table below:

{{endif config.is\_saved\_model.value}}

{{result.chosen\_algorithm\_details.image}}

See section II.D.2.a) for detailed explanations on the algorithm and its hyperparameters.

{{if config.is\_saved\_model.value != Yes}}

## Alternative Models

For all other algorithms and hyperparameters, the table below gives the performance obtained for the {{design.test\_metrics.name}} metrics:

{{result.leaderboard.image}}

Performance results obtained with the inertia metric are given below:

{{result.leaderboard.inertia.image}}

Finally, the number of clusters obtained for each algorithm is given in the table below. Depending on the algorithm, it may be the optimal number of clusters or the one defined in the settings.

{{result.leaderboard.clusters.image}}

{{endif config.is\_saved\_model.value}}

# Selected Model Results

## Selected Model Metrics

The selected clustering algorithm gave the following clusters:

{{result.cluster\_summary.image}}

The following rules describe each of them:

{{result.cluster\_description.image}}

The heatmap below shows how important each feature is to characterize a cluster.

{{result.cluster\_heat\_map.plot}}

|  |  |
| --- | --- |
| Legend:  |  |

{{if result.chosen\_algorithm.name == Interactive clustering }}

The clusters are organized with the following hierarchy:

{{result.hierarchy.plot}}

{{endif result.chosen\_algorithm.name}}

The details of the obtained evaluation metrics obtained is given below.

{{result.detailed\_metrics.table}}

|  |  |
| --- | --- |
|  |  |
|  |  |

{{/result.detailed\_metrics.table}}

{{if result.feature\_importance.status != No}}

## Sensitivity Testing and Analysis

Feature importance values that assess which features have a significant impact on the algorithm results have been computed:

{{result.feature\_importance.plot}}

{{endif result.feature\_importance.status}}

## Diagnostics

ML Diagnostics are designed to identify and help troubleshoot potential problems and suggest possible improvements at different stages of training and building machine learning models.

{{result.diagnostics.table}}

|  |  |
| --- | --- |
|  |  |
|  |  |

{{/result.diagnostics.table}}

# Deployment and monitoring

## Implementation Details

* The backend used by the model is: {{design.backend.name}}
* The model can be found at: {{config.project.link}}
* The name of the generated file is: {{config.output\_file.name}}
* The timing of the training was the following:

{{result.timings.table}}

|  |  |
| --- | --- |
|  |  |
|  |  |

{{/result.timings.table}}

## Version Control

* The model was trained at {{result.training\_date.name}} (In the DSS server time zone).
* The model was trained with the following version of DSS: {{config.dss.version}}
* With the following code environment: {{config.environment.name}}