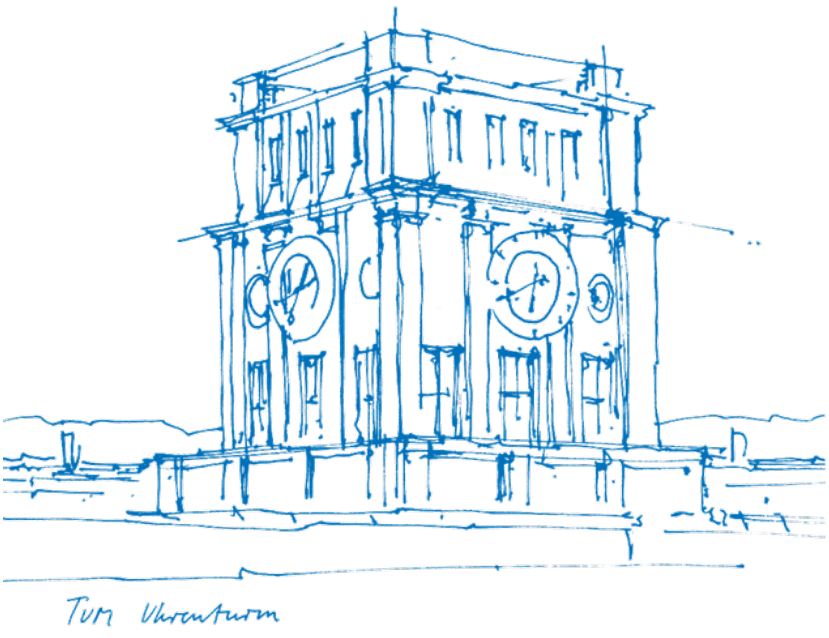
Chair of Information Systems and Business Process Management
TUM School of Computation, Information and Technology
Technical University of Munich

Technical Write-Up for Datasets in the Bachelor's Thesis: "Causation Integration in Root Cause Analysis for Business Process Violations"

Catherine Sai and Eduardo Breitembach Appio

Technical Write-Up on Dataset: Coffee Roasting Process



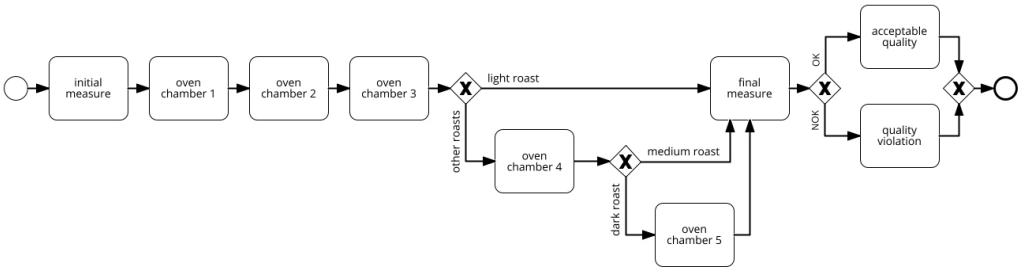
Coffee Roasting Process

The Coffee Roasting Process dataset [1] contains event logs related to the whole production cycle of a coffee product. In this guidebook, the logic, variables, and other metrics of the dataset will be elaborated.

Logic

The Coffee Roasting Process begins with the initial measures of evaluating the height and moisture of the coffee beans in a tray. This tray goes through several ovens and is then evaluated for good quality measures. An overview of the process can be seen in **Fig. 1**.

Figure 1
An overview of the coffee roasting process modeled using Signavio.



As seen in **Fig. 1** the process changes depending on the roasting degree of the desired product. Three main roasting degrees are distinguished: light roast, medium roast, and dark roast. In total, there are 5 roasting chambers that are used according to the desired roasting degree, each roasting chamber is equipped with 3 different temperature sensors.

Dataset and Event Parameters

The dataset originated from a Kaggle challenge related to a Coffee Roasting Process containing 29,184 traces, every trace contains information from event logs pertaining to a whole run through the model. The dataset was also adjusted to fit the needs of this thesis.

In total there are 18 parameters, unique parameters include: trace id, date time, height, moisture, roasting degree, five different roasting chambers with three different temperature sensors each,

violation, and case number. An exemplary trace can be seen in **Fig. 2**. Furthermore, every parameters has a different meaning and their definitions are specified in **Table 1**.

Figure 2
An exemplary trace for the Coffee Roasting Process.

case_id	event_id	date_time	activity	activity_key	activity_value
18648	30000	2017-02-19 00:05:00	Height Measure	height	173.52
18648	30001	2017-02-19 00:05:30	Moisture Measure	moisture	6.69
18648	30002	2017-02-19 00:06:00	Roast Degree Selection	roasting_degree	0
18648	30003	2017-02-19 00:06:30	Roasting Chamber 1	RC1_Sensor_1	246
18648	30004	2017-02-19 00:11:30	Roasting Chamber 1	RC1_Sensor_2	206
18648	30005	2017-02-19 00:16:30	Roasting Chamber 1	RC1_Sensor_3	242
18648	30006	2017-02-19 00:21:30	Roasting Chamber 2	RC2_Sensor_1	317
18648	30007	2017-02-19 00:26:30	Roasting Chamber 2	RC2_Sensor_2	316
18648	30008	2017-02-19 00:31:30	Roasting Chamber 2	RC2_Sensor_3	318
18648	30009	2017-02-19 00:36:30	Roasting Chamber 3	RC3_Sensor_1	416
18648	30010	2017-02-19 00:41:30	Roasting Chamber 3	RC3_Sensor_2	390
18648	30011	2017-02-19 00:46:30	Roasting Chamber 3	RC3_Sensor_3	425
18648	30012	2017-02-19 00:51:30	Final Measure	violation	0
18648	30013	2017-02-19 00:52:00	Quality Assessment	case_no	1

It is important to note that there are two types of datasets, a **categorical** and a **binary** one. The only difference is that the binary dataset does not distinguish violations, it just simply indicates based on binary values whether one happened or not.

Table 1
Explanation on event variables.

height	Maximum height of the coffee beans in the tray. Number value rounded to 2 decimals and measured in mm.
moisture	Moisture level recorded from the coffee beans. Number value rounded to 2 decimals.
roasting degree	0 = light roast. 1 = medium roast. 2 = dark roast.
temperature sensor	Recorded temperature from the sensor in °F.
violation	0 = no violation 1 & 2 = light roast violation. 3 & 4 = medium roast violation. 5 & 6 = dark roast violation.
case number	0 = violation 1 & 2 = light roast; no violation. 3 & 4 = medium roast; no violation. 5 & 6 = dark roast; no violation.

Textual Descriptions

As part of the thesis, it was important to fabricate a synthetic textual description that would relate to the dataset and represent all the process constraints that would explain the violations that occur in the dataset. A gold standard can visualize all the constraints, it does however not reflect the reality of only having constraints defined inside textual documents.

The dataset did not include any textual descriptions and in order to make it as realistic as possible, several real-world handbooks [2][3][4] of coffee roasting businesses were chosen as guidance for textual composition. The textual descriptions includes a lot of superfluous text that is irrelevant to the thesis' use-case, however, to study how well the specifically relevant text formats would perform, it was decided that three different textual constraint definition was going to be written within the synthetic text. An overview as well as examples can be seen in **Table 2**.

Table 2
Overview and example of textual descriptions.

purely textual	If beans roast much hotter than 780°, the coffee will start to taste more and more of charcoal and will not pass the final quality check. This roasting degree is the only one that requires the use of roasting oven 5.
semi formal	Roasting oven 1 should follow temperature rule [140°, 420°]. Afterwards, it is not allowed for roasting oven 2 to go above 520° or below 240°.
formal	Coffee tray height of at most 170mm: ->Temperatures of roasting 1 should follow [170°,450°]. ->Temperatures of roasting 2 should follow [270°,550°].

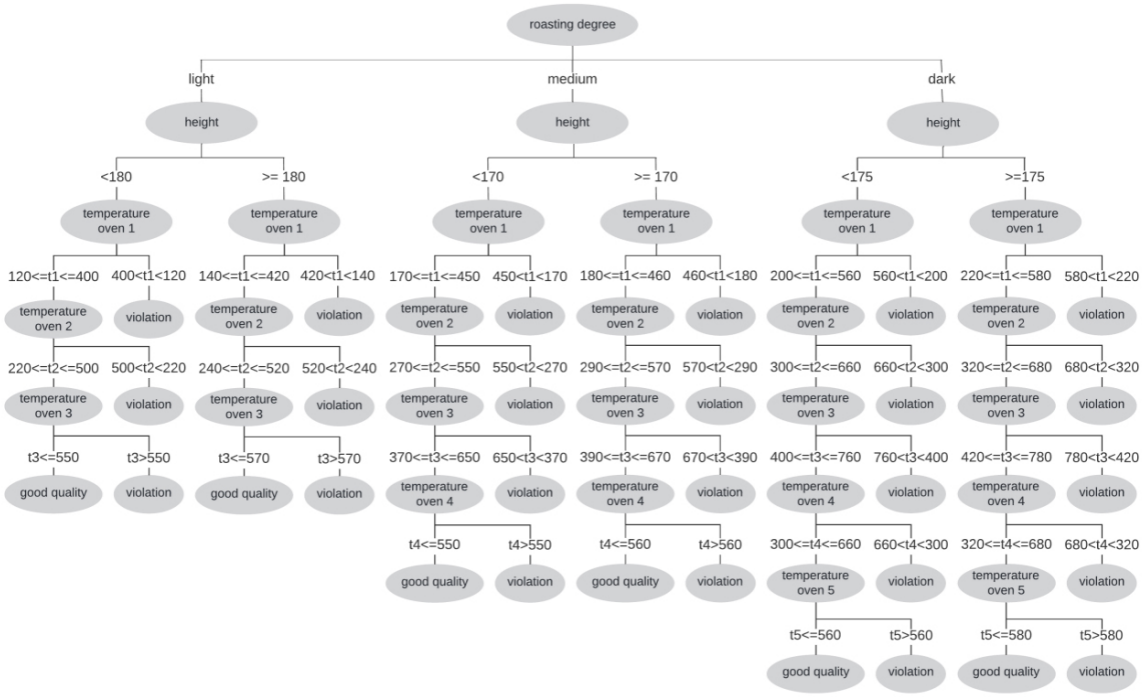
It is worth noting that formal and semi-formal formats can include rules that are also defined within the textual descriptions.

Gold Standard

To better visualize the proposal of the thesis and understand constraints in the context of the Coffee Roasting Process dataset, it is important to have a gold standard realized by a decision tree for the Coffee Roasting Process which is visible in **Fig. 3**.

The gold standard was created manually and represents the ideal solution as the accuracy and recall are 100%. In total, it displays a depth of 7, with a total of 57 nodes (conditional and leaf nodes).

Figure 3
Multivariate Coffee Roasting Process decision tree gold standard (modeled using Lucidchart).



Constraints

With the gold standard and textual descriptions at hand it is now possible to define the constraints regarding the Coffee Roasting Process. For this, the notations presented in [5] were used and extended in order to fit the parameters of the thesis. An exemplary constraint can defined such as:

c = ({Chamber 3}, {Chamber 4}, {directly follows}, {roasting degree = medium OR roasting degree == dark})

The goal of these constraints is to define a decisive behavior that the first activity must follow in order to successfully transition into the second activity without violating any business process rules described in the formal textual document.

Recorded Sources

In addition to the sources listed in the bibliography, the document sources utilized for this dataset are also available in PDF format. These can be accessed in the dataset repository.

Bibliography

[1] PODSYP, *Dataset: Product quality produced by a roasting machine*, Jan. 2015. [Online]. Available: https://www.kaggle.com/datasets/podsyp/production-quality.

[2] C. Five, "Coffeehouse five employee handbook, version 2-13-19," pp. 1–18, Feb. 2019. [Online]. Available: https://www.coffeehousefive.com/uploads/9/0/1/6/9016048/employee_handbook.pdf.

[3] C. M. Coffee, *What is the difference between light, medium, and dark roast coffee?* Feb. 2020. [Online]. Available: https://www.coppermooncoffee.com/blogs/newsroom/what-is-the-difference-between-light-medium-and-dark-roast-coffee.

[4] Q. C. Roasters, "Quest coffee roaster handbook, amended may 2021," pp. 1–22, Apr. 2021. [Online]. Available: https://library.sweetmarias.com/wp-content/uploads/2021/05/QuestHandbook-May-2021.pdf.

[5] Q. Chen, K. Winter, and S. Rinderle-Ma, "Predicting unseen process behavior based on context information from compliance constraints," in *Business Process Management Forum*, C. Di Francescomarino, A. Burattin, C. Janiesch, and S. Sadiq, Eds., Cham: Springer Nature Switzerland, 2023, pp. 127–144, ISBN: 978-3-031-41623-1.