|  |  |  |  |
| --- | --- | --- | --- |
|  | 17/12/2023 |  | |
| |  |  | | --- | --- | |  |  | |  |  | |  |  | |  |  | | |  | |
| Automatic lithoclassification of unconsolidated sediments, Becerra-Shaller  *Peer Review* | | | |
|  |  | | |  |  |  | | --- | --- | --- | |  |  |  | |  |  |  | |  |  |  | |  |  |  | |  |  |  | |  |  |  | |
|  |  | | Lara Nonis, M4 CAS in applied data science, University of Bern |

# Objectives and background information

The goal of the machine learning project under review, namely automatic lithoclassification of unconsolidated sediments from Patricio Becerra and Sebastian Shaller, is to classify rock samples in lithoclastic types using selected physical properties of the samples collected and without the need of a complex field classification.

The current method necessitates the extraction of rock samples through core drilling—a intricate and costly process involving the use of specific equipment in the targeted area. The extracted sample, referred to as a core or drilling core, must meet high-quality standards to enable visual classification. Field experts conduct this classification, along with other analyses, and the entire process can span up to two years for a comprehensive stratigraphic profile. Becerra and Shaller's ongoing project aims to streamline this classification process by eliminating the necessity for the core to meet the highest quality standards. This innovative approach not only significantly reduces drilling costs but also shortens the time required for classification, making the overall procedure more efficient.

The declared targets of the project, from the author’s presentation, are the following:

* Visualise the data in such a way that the visually-classified clusters are distinguishable
* Attempt to find patterns in the data independent of the visual classification through unsupervised clustering algorithms by two approaches:
  + Direct clustering by several algorithms such as K-Means, and HDBSCAN
  + PCA-analysis for dimensionality reduction followed by K-Means and HDBSCAN clustering
* Compare the visual-based classification of the sediments (classic geological route) with the semi-automated data-based clustering of the logs.

The project, initiated by Shaller in the context of the PhD program (further information in "Drilling into a deep buried valley (ICDP DOVE): a 252 m long sediment succession from a glacial overdeepening in northwestern Switzerland, Shaller et al., 2023"), continued with an initial statistical analysis of the available data in the M2 project, which is beyond the scope of the present report. The implementation of the machine learning algorithm, along with the corresponding data preparation for its execution, involved the creation of a Jupyter notebook and a project presentation, both of which will be comprehensively discussed in the forthcoming sections of this report.

# Notebook description

### 2.1 Import and pre-processing

At the time of the review, the notebook encompasses a total of 112 cells. It kicks off with a concise introduction to the project and offers key insights from the M2 project, setting the stage for the subsequent analysis conducted within the notebook. The initial steps involve importing the necessary libraries, categorized for data handling, visualization, numeric and statistical operations, machine learning, and other functionalities. This is followed by the loading and cleaning of the dataset.

The cleaning process unfolds in multiple stages. In the first pass, only selected features are retained from the entire dataset, accomplished through indexed selection and subsequent decimal adjustments. The resulting dataset is then subjected to a classification in a second step, wherein the litho class is categorized into four main types, each appropriately labeled. The third and final cleaning routine involves the exclusion of heavily disturbed data (indicative of poor core sample quality) and NaN values from the core quality column. The resultant dataset undergoes verification through the info function, revealing 40276 rows and 9 columns, all populated with non-null numeric entries.

The dataset is subsequently displayed through pie-plots where the different litho types and the core quality data are displayed before and after data cleaning. This presentation aims to illustrate the impact of the pre-processing itself and visually demonstrate the maintenance of the relative distribution of the different litho-types within the dataset, showcasing the representativity of the cleaned dataset with respect to the original. The dataset is then normalized to be used for the first target mentioned above, namely the visualization.

### 2.2 Visual classification

Visual classification unfolds through three distinct plotting strategies, progressively increasing in complexity: scatter plots, triangular plots, and 3D plots. Density, Magnetic Susceptibility, and Natural Gamma are meticulously charted in all-vs-all combinations for the four litho-types. These combinations are presented in an iterative fashion using Matplotlib for the initial scatter plots. Further scatterplots delve into the individual litho-classes, employing the same plotting variables, while consistently maintaining the color scheme associated with each class across all sets of plots.

In Section 3.2.3, triangular plots are showcased, portraying the 4 litho classes within the area delimited by the selected three features (Density, Magnetic Susceptibility, and Natural Gamma), each at a vertex of the triangle. Following a methodology akin to the scatter plots, the assessment is initially conducted collectively and subsequently for each litho-class in isolation.

The exploration extends to 3D plots, where the three features align along the axes, and the different litho-classes reveal their distribution both collectively and then individually in four interactive 3D charts.

### 2.3 Unsupervised learning

# X. Could be improved

- Not possible to run code w/o dataset:2.1.2 works only if dataset already available in computer, if not the path is not found and the dataset not imported

- First cleaning: column to be used subsequentially are selected via column index, where the columns would have another order or one is n/a in a new dataframe the whole analysis would not work 🡪 better use column name

- Classification performed by index (-2) same issue as before 🡪 classification main step for analysis