13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

Data Preprocessing Analysis to Avoid Information Loss

Alexandra Laura Giczi[a], **Edith Alice Kovács**[a]

[a]Budapest University of Technology and Economics, Department of Analysis and Operations Research
giczi.alexandra@gmail.com
kovacsea@math.bme.hu

Abstract

Machine learning and data analytics methods are frequently used across many fields, such as healthcare, transportation, or manufacturing, where Industry 4.0 has revolutionized digitalization processes through several pillars, with a focus on Big Data. Production lines collect continuously increasing sensor data to improve efficiency, reduce costs and scrap rates, or support predictive maintenance. Across the traditional stages of these projects, of particular interest is the role of data processing before modelling as on the one hand, it can consume a significant portion of a project's timeline (sometimes even half of the invested time [3]), and on the other hand, it directly influences the quality of learned models and their predictive performance.

Despite its importance, many preprocessing steps are routinely applied in an automated fashion, for example, following a rule of thumb while handling empty fields – deleting rows if the dataset has only a certain number of them or imputing a value if this rate is higher. [2] If imputation is chosen, the next question of the value to be placed arises. Automatic data processing techniques based on statistics or underlying distribution were also introduced recently. [4] These solutions often lead to the inadvertent loss of potentially valuable information. In particular, data elements that do not conform to standard assumptions are frequently ignored or treated as noise. This causes a lot of turbulence and iteration in modern pipelines, since real-world data is rarely ideal and typically deviates from theoretical assumptions in a wide range of aspects.

Proposed approach

In our research, we propose a novel preprocessing analysis before conventional data transformations that systematically exploits information contained in such typically overlooked or discarded data. By incorporating this information into the learning process, the proposed approach aims to improve model robustness and performance across a range of machine learning tasks.

In the first step, we analyse all the key steps in data preparation: data cleaning, transformation of numerical and categorical values, and reduction by feature engineering, sampling. [1] We demonstrate that valuable information can be lost during data cleaning – in the *handling of duplicates*, *missing*, and *outlier values*.

Then we move on to the analysis part that consists of multiple steps, such as transforming the dataset with zeros and ones by scanning the places of missing values for each record and indicating this abnormality with the value one. Training a general machine learning model (like KNN, gradient boosting) on this dataset can analyse the effect of those problems on the output class. The new workflow, based on pre-analysis tasks, and this way containing the information loss module, is presented in Figure 1.

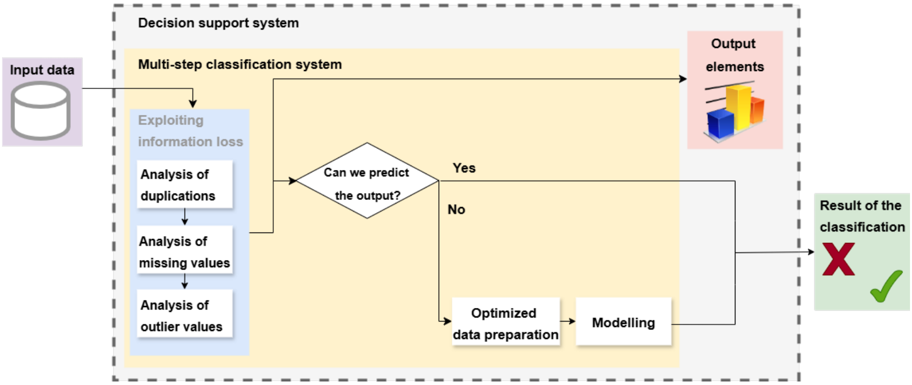
Decision support system
Multi-step classification system
Input data
Exploiting information loss
Analysis of duplications
Analysis of missing values
Analysis of outlier values
Can we predict the output?
Yes
No
Optimized data preparation
Modelling
Output elements
Result of the classification

Figure 1. Workflow with the preprocessing analysis.

After running these analyses, two outcomes can occur: the output classes can be defined by running only these steps before the traditional modelling part. If the classes cannot be predicted this way, then it does not matter how these data quality issues are managed, the user can choose an arbitrary method – for instance, no need to decide between imputation of the mean or maximum value – as the system indicates that this information does not influence the output classes. In this way, our proposed solution always provides result for the user by enabling faster data processing and saving time during the analysis.

The preprocessing methodology is complemented by an interactive visualization interface that supports decision makers by facilitating the exploration and interpretation of the extracted information. After loading the data and selecting the configuration that led to the expected results, the system can show the problems for each record in a highlighted manner. It also draws a chart based on the Shapley values for the attributes of the dataset. The usefulness of the proposed approach is demonstrated in this way through illustrative examples, highlighting the types of information that may remain undetected when the introduced methodology is not employed. These results underscore the potential of the proposed preprocessing step to enhance insight generation and improve downstream machine learning outcomes.

References

[1] J. Brownlee: *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*, Machine Learning Mastery, 2020.

[2] S. García, S. Ramírez-Gallego, J. Luengo, J. Benítez, F. Herrera: *Big data preprocessing: methods and prospects*, Big Data Analytics 1.1 (2016), pp. 1–22, doi: https://doi.org/10.1186/s41044-016-0014-0.

[3] K. Maharana, S. Mondal, B. Nemade: *A review: Data pre-processing and data augmentation techniques*, Global Transitions Proceedings 3.1 (2022), pp. 91–99, doi: https://doi.org/10.1016/j.gltp.2022.04.020.

[4] A. Mumuni, F. Mumuni: *Automated data processing and feature engineering for deep learning and big data applications: A survey*, Journal of Information and Intelligence 3.2 (2025), pp. 113–153, doi: https://doi.org/10.1016/j.jiixd.2024.01.002.