A Data Pre-Processing Task is a data processing task that preceeds the Data Analysis Task.

- **AKA:** Data Preparation, Data Wrangling.
- **Context:**
  - Task Input: Raw Data.
  - **Task Output**: the final Training Set.
  - It can be solved by Data Pre-Processing System that implements a Data Pre-Processing Algorithm.
- **Example(s):**
  - Data Profiling Task,
  - Data Normalization Task,
  - Dimensionality Reduction Task,
  - Data Whitening Task,
  - Data Cleaning Task,
  - Instance Selection Task,
  - Feature Extraction Task,
  - Feature Selection Task,
  - Data Integration Task,
  - Data Transformation Task.
- **Counter-Example(s):**
  - Data Enrichment task,
  - Regularization Task,
  - Data Post-Processing,
  - Neural Network Weight Initialization Task,
  - Neural Network Training Task,
  - Neural Network Prediction Task,
- **See:** Data Mining Task, Principal Component Analysis, Outlier, Data Quality Measure, Overfitting.

---

# References

## 2018

- (Wikipedia, 2018) ⇒ https://en.wikipedia.org/wiki/Data_pre-processing Retrieved:2018-5-13.
  - **Data pre-processing** is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can

produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. [1] Often, data pre-processing is the most important phase of a machine learning project, especially in computational biology. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, Instance selection, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set. Kotsiantis et al. (2006) present a well-known algorithm for each step of data pre-processing. [2]

## 2017

- (Abdallah, Du and Webb, 2017) ⇒ Abdallah Z.S., Du L., Webb G.I. (2017) "Data Preparation". In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA
  - ABSTRACT: Before data can be analyzed, they must be organized into an appropriate form. Data preparation is the process of manipulating and organizing data prior to analysis. Data preparation is typically an iterative process of manipulating raw data, which is often unstructured and messy, into a more structured and useful form that is ready for further analysis. The whole preparation process consists of a series of major activities (or tasks) including data profiling, cleansing, integration, and transformation.

---

1. Pyle, D., 1999. *Data Preparation for Data Mining.* Morgan Kaufmann Publishers, Los Altos, California.

2. S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Preprocessing for Supervised Learning", *International Journal of Computer Science*, 2006, Vol 1 N. 2, pp 111–117.