Project Title

Big Data – Curation, cleansing, and wrangling of big and large datasets

Introduction

During data gathering—whether for sophisticated Artificial Intelligence (AI) methods or more basic ones—data must first be curated. This involves sourcing, integrating, and cataloguing data, all of which should be carefully documented. Next, data cleansing must be carried out. Depending on the type and characteristics of the data, this includes handling missing values, removing duplicates, and standardizing the data. Various strategies can be applied, and it is essential to determine when and how to implement them. Further data wrangling is required to prepare the data for future use. Key steps in this process include data transformation, filtering, sampling, and handling outliers. These can often be harmonized based on the types of variables present in the dataset.

Lastly, data quality assurance is critical. This involves validating and verifying the data, profiling its characteristics, and ensuring proper governance. Procedures should be established to manage not only the quality but also the security and compliance of the data throughout its lifecycle.

Scope

Effective curation, cleansing, and wrangling of big and large datasets are crucial for ensuring data quality, reliability, and suitability for downstream analytics, machine learning, or business intelligence tasks. These processes help organizations derive meaningful insights, make informed decisions, and maintain competitive advantages in today's data-driven landscape. Thus, the proposed standard project includes data curation, cleansing, wrangling, and quality assurance of big and large datasets."

Keyworks

Metadata management; Cataloguing; Standardization; Formatting; Anonymization; Integration; Accuracy, Completeness, Consistency, Timeliness