DIGITAL SYSTEMS
INDUSTRIAL SYSTEMS

A Guide to Data Quality Testing for AI Applications based on Standards

Nishat I Mowla

RISE Report :

2024

A Guide to Data Quality Testing for AI Applications based on Standards

Nishat I Mowla

Abstract

Data quality testing is critical for the development and deployment of Artificial Intelligence (AI) systems, particularly those used in decision-making processes. The integrity, accuracy, and reliability of data directly influence the performance and fairness of AI systems. Ensuring high-quality data is crucial as it not only helps in reducing biases but also in enhancing the overall effectiveness and trustworthiness of AI applications across various domains.

This report provides a detailed exploration of the necessary prerequisites for effective data quality testing, including the identification of key data attributes and the establishment of specific quality benchmarks. It discusses various data quality characteristics and metrics for assessing and improving the quality of data used in AI systems. In particular, the report discusses the relevant standards and guidelines that govern data quality testing, offering a structured framework for organizations to adhere to these practices.

By implementing rigorous data quality testing protocols, organizations can significantly mitigate risks associated with data-driven decisions, thereby ensuring that their AI systems operate within the desired scope of accuracy and fairness. This not only aligns with regulatory compliance but also enhances the credibility and reliability of AI applications in real-world scenarios.

Key words: Data quality, ISO/IEC standards, performance evaluation, artificial intelligence

RISE Research Institutes of Sweden AB

RISE Report :
ISBN:

Content

1 Introduction

The domain of machine learning (ML) is built on the foundation of data. Data is the raw material from which machine learning models learn and extract patterns. This document provides an introduction to data quality in machine learning, focusing on requirements, methods, tools and standards.

1.1 Machine learning Data

In machine learning, data is used to train models to make predictions or decisions without being explicitly programmed to perform the task. Data comes in various forms and structures, and the nature of the data often dictates the type of ML model that can be applied. Good machine learning data makes good machine learning models. Good machine learning data needs to be relevant, comprehensive, accurate, and prepared with minimal biases to allow the development of robust machine learning models. Data is collected from various sources and can be structured or unstructured. It undergoes several preprocessing steps such as cleaning, transformation, normalization, and feature extraction before it is used to train a model.

ISO/IEC dec[1] outlines that machine learning (ML) involves refining model parameters through computational methods so that the model accurately represents the data or experiences it is exposed to. Further expounded by ISO/IEC 23053[2], machine learning is identified as a subset of artificial intelligence that utilizes computational methods to allow systems to derive insights from data or experiences. ML is applicable to an array of tasks reliant on data and ML algorithms. Data within ML is differentiated into several types: training data, validation data, testing data, and production data. In the case of supervised ML, a model is developed through the training of an algorithm using training data. Validation and testing data are subsequently employed to confirm the model's operation within acceptable bounds. Following this, the model applies what it has learned to make predictions or decisions based on new, unseen production data. The efficacy of a trained ML model is tied to the data quality across all these categories. ISO/IEC 23053 outlines a variety of general ML algorithms, noting that each may be differently affected by the quality attributes of the data they process.

Example 1:

Representativeness is a crucial data quality attribute for machine learning. If the training data fails to adequately mirror the population seen in the production data, there's a heightened risk that the trained ML model will draw incorrect conclusions from that production data. This issue becomes particularly significant when the model's decisions impact people, potentially leading to biased outcomes against underrepresented groups.

Example 2:

Training an ML model is essentially a mathematical routine that repeatedly processes a set of training data, which reflects characteristics of a specific object or event. The accuracy of each data

[1] ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. Available at https://www.iso.org/standard/74296.html

[2] ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). Available at https://www.iso.org/standard/74438.html

sample in the training set directly affects the efficacy of the trained model. If a significant portion of the training data samples are inaccurate, the model is more likely to generate erroneous predictions or assessments based on the production data.

It is worth noting that the same dataset can serve multiple analytics or machine learning purposes. For instance, a data holder might distribute data to various users, both within and outside their organization. Similarly, a data user may be permitted to employ the data for several different tasks.

1.2 Data lifecycle

This section outlines a structured approach to managing the quality of data through various stages essential for analytics and machine learning (ML) projects as shown in Fig. 1.

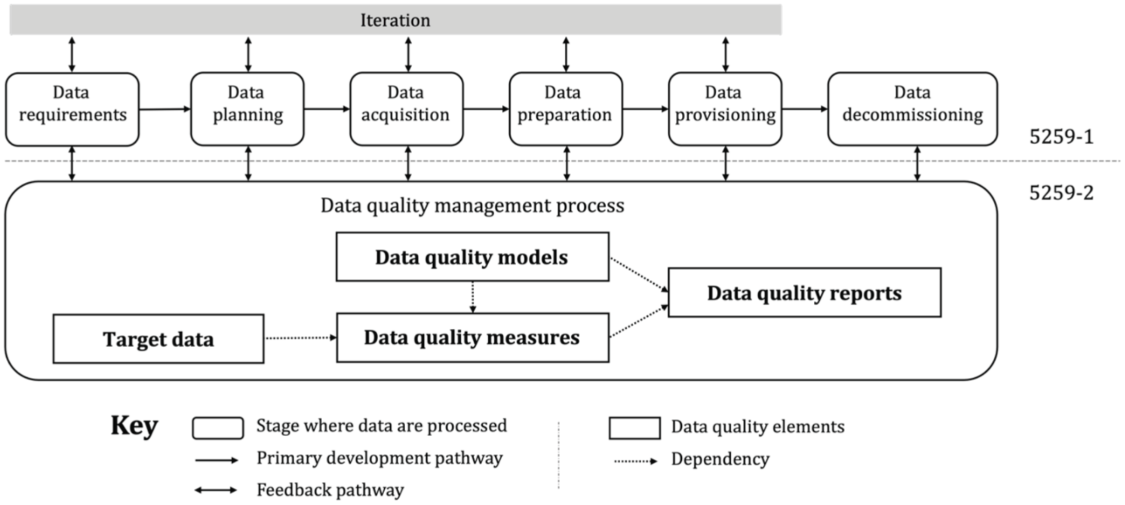

Fig. 1 Data quality elements in data life cycle for analytics and ML. ISO 5259-2[3].

- Stage 1: Data Requirements

Objective: Determine the necessary data for an analytics or ML project, assess availability, and identify relevant data quality characteristics.

- Stage 2: Data Planning

Objective: Ensure that the data meet the requirements identified in the previous stage and support the objectives of the analytics and ML projects. This includes designing data architecture, estimating efforts for data acquisition and preparation, and planning for data quality management.

- Stage 3: Data Acquisition

Objective: Collect data (both live and historic) identified in the planning stage. This involves:

- Protecting the privacy of data subjects and securing the data.

[3] ISO/IEC 5259-2 Artificial intelligence — Data quality for analytics and machine learning (ML). Available at https://www.iso.org/standard/81860.html

- Modifying data collection methods to include tests and improve data quality.

- Reducing risks of data inconsistencies from different transformations.

- Stage 4: Data Preparation

Objective: Process the collected data into a form suitable for input into analytics and ML models, ensuring data quality through:

- Transforming, validating, and cleaning data.

- Aggregating, sampling, and creating new features.

- Enriching data by linking diverse sources and annotating data for supervised learning tasks.

- Stage 5: Data Provisioning

Objective: Apply prepared data to analytics and ML projects and assess if they meet the performance requirements. If not, analyze potential data or algorithmic issues, communicate these issues for upstream quality improvement, and possibly repeat earlier stages to enhance data quality.

- Stage 6: Data Decommissioning

Objective: Manage the end-of-life of data by storing, archiving with metadata, or destroying it based on retention policies and project requirements. Ensure that archived data includes necessary context for future use.

These stages illustrate a comprehensive framework for handling data from its initial requirement gathering through to decommissioning, emphasizing continuous improvement in data quality to meet the specific needs of analytics and ML applications.

1.3 Data quality model

ISO/IEC 5259-1 outlines a data quality model, shown in Fig. 2, as a set of characteristics designed to help specify data quality requirements and evaluate data quality effectively. This model integrates data quality subjects (entities affected by data quality), data quality characteristics (categories of data quality attributes like accuracy, completeness, and precision), and data quality requirements (properties or attributes of data with specific acceptance criteria based on the data usage context). These elements are organized to align with the intended use of the data, particularly in analytics or machine learning tasks, such as training a neural network to predict product sales based on marketing strategy features. The model uses a UML diagram to illustrate the relationships among these elements, emphasizing the importance of context in defining and achieving target data quality. This framework allows organizations to select appropriate data quality characteristics and measures to achieve targeted quality requirements for specific data sets.

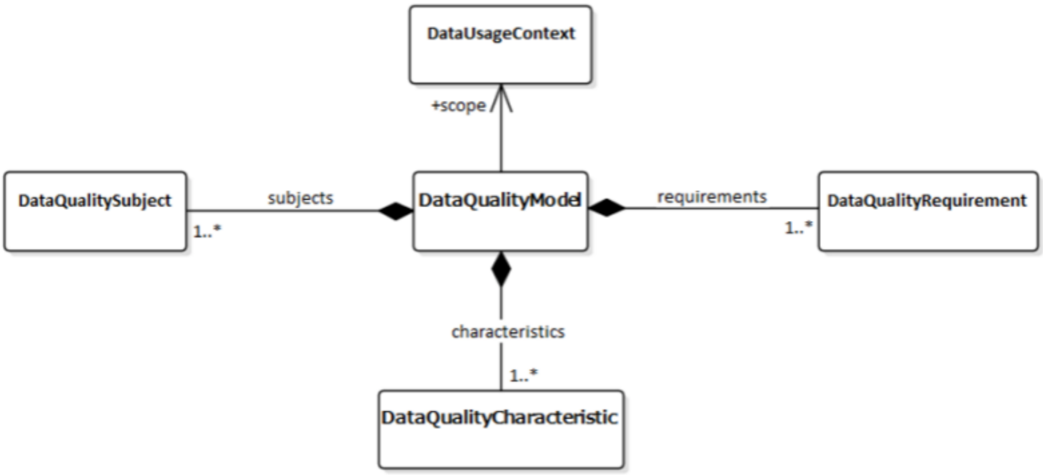

Fig. 2 Data quality model. ISO 5259-2.

Data will be utilized to train a deep neural network ML model that forecasts product sales, leveraging the attributes of a marketing strategy. This model will undergo training and deployment through cloud-based services. In this context, a "data quality subject" refers to any entity impacted by data quality. According to ISO/IEC 5259-1[4], a data quality characteristic encompasses a group of data quality attributes related to the overall data quality, such as accuracy, completeness, and precision. Data quality requirements detail the necessary properties or attributes of data, complete with acceptance criteria that are tailored to how the data will be used. These criteria might be quantitative, qualitative, or described in other terms.

1.4 Data readiness level

Data Readiness Levels (DRLs), as shown in Fig. 3, are a systematic method to assess the readiness of data for deployment, similar to Technology Readiness Levels (TRLs) used for evaluating technology maturity. This concept was developed by Amazon Research Cambridge and University of Sheffield, authored by Lawrence. The concept aims to quantify the overall quality and preparedness of data sets, which is crucial for project planning and development. It's noted that a significant portion of project time, up to 80%, is often devoted to pre-processing data, adhering to the Pareto principle where 80% of the effort might be spent on the last 20% of the work due to detailed adjustments and corrections.

[4] ISO/IEC 5259-1:2024 Artificial intelligence — Data quality for analytics and machine learning (ML). Available at https://www.iso.org/standard/81088.html

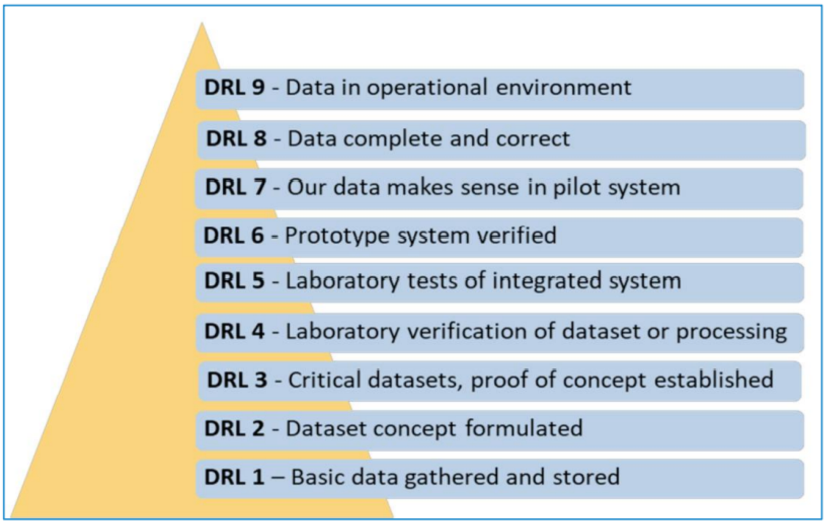
DRL 9 - Data in operational environment
DRL 8 - Data complete and correct
DRL 7 - Our data makes sense in pilot system
DRL 6 - Prototype system verified
DRL 5 - Laboratory tests of integrated system
DRL 4 - Laboratory verification of dataset or processing
DRL 3 - Critical datasets, proof of concept established
DRL 2 - Dataset concept formulated
DRL 1 – Basic data gathered and stored

Fig. 3 Data Readiness Level (DRL). Concepts from Amazon Research Cambridge and University of Sheffield and ROADVIEW Deliverable D4.5 on initial readiness assessment of specific datasets in ROADVIEW Robust Automated Driving in Extreme Weather by Ian Marsh[5].

The main challenges in implementing DRLs include:

1. Assigning a singular readiness level (DRL 1-9) to large and complex datasets that often contain imperfections such as missing values, inaccuracies, or incomplete readings.
2. The context-sensitive nature of data readiness which may imply different things to different users.
3. Variability in the methods available to users for handling data imperfections, which can affect the accuracy of the assigned DRL.
4. The continuous production of sensor data from sources like real cars, which are essential yet challenging to quality control.

The DRL system used in ROADVIEW project assigns a scale from 1 to 9 to describe the effort, time, and cost required to address or rectify data issues:

- Lower DRL values indicate simpler fixes.
- Higher values signify more complex problems that are harder to correct and might impact further stages of data processing.

Lawrence’s framework initial concept divides data readiness into three bands:

- A (Utility): How useful the data is for a specific objective.
- B (Validity): The accuracy and reliability of the data.
- C (Accessibility): The ease of accessing and using the data.

These bands are detailed in ROADVIEW project in a structure analogous to the 9-level scale used in TRLs, providing a comprehensive measure of data readiness and highlighting the importance of understanding and preparing data thoroughly to ensure successful project outcomes.

[5] Initial readiness assessment of specific datasets. Available at https://roadview-project.eu/wp-content/uploads/sites/59/2024/05/ROADVIEW_Deliverable-4.5_v04.pdf

1.5 Data quality requirements in the AI act

Article 10 addresses data and governance for high-risk AI systems, stipulating that such systems must be built using training, validation, and testing datasets that adhere to quality standards specified in subsequent paragraphs. The datasets are required to undergo rigorous data governance and management, focusing on aspects such as (a) design choices; (b) data collection; (c) relevant data preparation processing operations, such as **annotation, labelling, cleaning, enrichment and aggregation**; (d) the formulation of relevant assumptions, notably with respect to the information that the data are supposed to measure and represent; (e) a prior assessment of the availability, quantity and suitability of the data sets that are needed; (f) examination in view of possible biases; (g) the identification of any possible data gap. It's essential to assess the datasets for **relevance, availability, potential biases,** and **any gap**s that might impact their effectiveness.

The datasets must also be **relevant, error-free, representative, complete**, and possess **statistical properties suitable** for the target demographic or application environment. Moreover, they should reflect specific characteristics related to the geographical or functional contexts in which the AI system will operate. If necessary for **bias monitoring** and correction, AI system providers may process sensitive personal data under strict conditions to ensure **privacy** and **security**, employing techniques like pseudonymization or encryption where needed.

Lastly, rigorous data governance practices are mandated for all **high-risk AI systems** to ensure compliance with established **data quality requirements**, regardless of whether they involve training models or other methodologies.

2 Standardized data quality

AI systems require high-quality data to function effectively. The purpose of data quality testing is to verify the accuracy, completeness, and relevance of data used for AI decision-making. Ensuring data integrity is paramount in avoiding biases and making fair decisions that do not disadvantage any group.

Various analytics and machine learning tasks may have distinct data quality needs. These differing requirements can influence the selection of a data quality model, along with the corresponding data quality measures and evaluation criteria.

The AI Act proposal outlines the requirements of data quality in Article 10 on data governance regulation with stringent guidelines for the development of high-risk AI systems, emphasizing the necessity of utilizing high-quality training, validation, and testing datasets. These datasets must adhere to rigorous data governance and management practices as discussed in section 1.6. These practices are essential to ensure the reliability and fairness of high-risk AI systems. The primary requirements and characteristics for data quality in AI can be mapped to major data quality standards such as ISO/IEC 5259 series, as shown in Fig. 4, and ISO/IEC 24027[6].

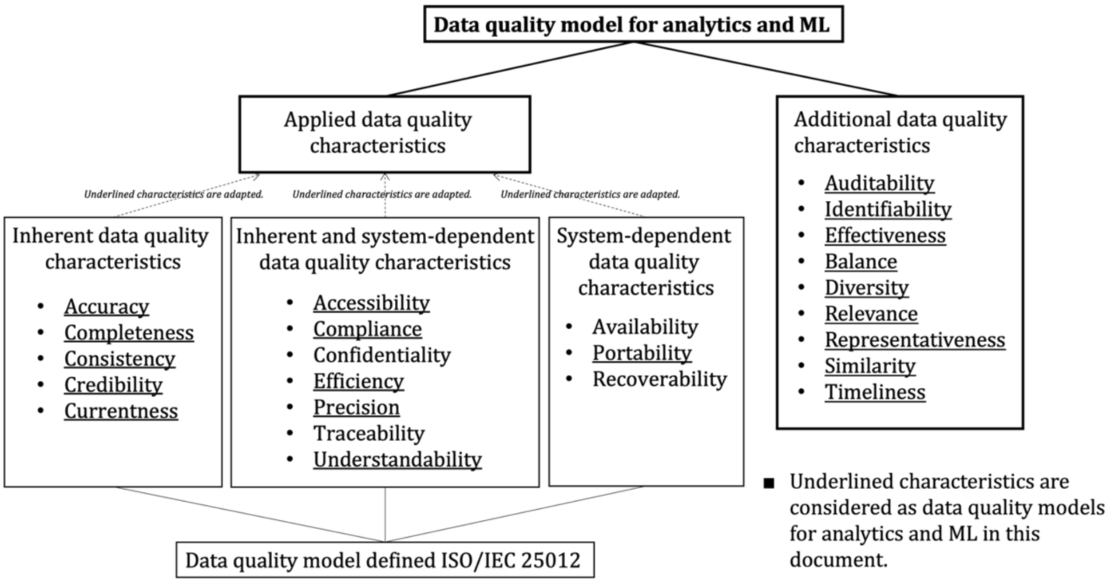
Data quality model for analytics and ML
Applied data quality characteristics
Additional data quality characteristics
Auditability
Identifiability
Effectiveness
Balance
Diversity
Relevance
Representativeness
Similarity
Timeliness
Underlined characteristics are adapted.
Underlined characteristics are adapted.
Underlined characteristics are adapted.
Inherent data quality characteristics
Accuracy
Completeness
Consistency
Credibility
Currentness
Inherent and system-dependent data quality characteristics
Accessibility
Compliance
Confidentiality
Efficiency
Precision
Traceability
Understandability
System-dependent data quality characteristics
Availability
Portability
Recoverability
Underlined characteristics are considered as data quality models for analytics and ML in this document.
Data quality model defined ISO/IEC 25012

Fig. 4 Data quality characteristics for analytics and ML in ISO/IEC 5259-2.

2.1 Applied data quality characteristics (ISO/IEC 5259 series, 25012, 25024)

2.1.1 Accuracy

The accuracy of a dataset refers to the extent to which each data item in the dataset holds the correct data value. ISO/IEC 25012[7] defines accuracy as the degree to which data values accurately reflect the true nature of the attributes they are intended to represent. ISO 5259-2 elaborates on accuracy by dividing it into two aspects:

- Syntactic accuracy: This involves the extent to which data values conform to a set of syntactically correct values within a specific domain.
- Semantic accuracy: This pertains to how closely data values align with a set of semantically correct values within a relevant domain.

A data item is considered syntactically correct when its data value matches its explicit data type, and semantically correct when the data value aligns with expected values useful for the machine learning (ML) task at hand. Given that ML models are based on mathematical frameworks, low syntactic or semantic accuracy in the training, validation, testing, or production datasets can lead to inaccuracies in the model or the conclusions it draws.

[6] ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making. Available at https://www.iso.org/standard/77607.html

[7] ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. Available at https://www.iso.org/standard/35736.html

In the context of a supervised learning classification system, the precision of the label sequence is critical to the model's inference accuracy. Factors to assess the accuracy of labeling include:

- Correctness of label names: Ensuring that labels are correctly named according to what they signify.
- Correctness of labeled tags: Verifying that tags attached to labels are accurate.
- Correctness of label sequence contents: Ensuring the sequence of labels is correctly ordered and appropriate for the dataset.

Example 1:

If the phrase "lazy dog" is entered as "lzy dg" an ML-based natural language understanding system can fail to correctly interpret the phrase.

Example 2:

If the number 100 is entered as 1000 in training data, a regression model can fail to correctly calculate the weight of the related feature and if the entry was made in the production data, inferences can be incorrect.

Syntactic data accuracy: Ratio of closeness of the data values to a set of values defined in a domain:

$$\frac{number\ of\ data\ items\ which\ have\ related\ values\ syntactically\ accurate}{total\ number\ of\ data\ items\ for\ which\ syntactic\ accuracy\ is\ required}$$

Concerns all Data Life-Cycle except data design, data file, data item, and data value

Symantic data accuracy: Ratio of how accurate the data values in terms of semantics in a specific context are:

$$\frac{number\ of\ data\ values\ semantically\ accurate}{total\ number\ of\ data\ values\ for\ which\ symantic\ accuracy\ is\ required}$$

Concerns all Data Life-Cycle except data design, data file, and data value

Data accuracy assurance: Ratio of measurement coverage for accurate data:

$$\frac{number\ of\ data\ items\ measured\ for\ accuracy}{total\ number\ of\ data\ items\ for\ which\ measurement\ is\ required\ for\ accuracy}$$

Concerns all Data Life-Cycle except data design, data file, and data item

Risk of dataset inaccuracy: Number of outliers in values is indicating a risk of inaccuracy for data values in a dataset:

$$\frac{number\ of\ data\ values\ that\ are\ outliers}{total\ number\ of\ data\ values\ to\ be\ considered\ in\ a\ dataset}$$

Concerns all Data Life-Cycle except data design, data file, and data value

Data model accuracy: Data model describes the system with the required accuracy:

$$\frac{number\ of\ elements\ of\ the\ data\ model\ that\ accurately\ describe\ the\ system}{\begin{matrix} number\ of\ elements\ of\ data\ model\ that\ describe\ required\ accuracy\ within\ requirement \\ specification\ of\ system \end{matrix}}$$

Concerns data design, data models and elements

Data accuracy range: Are data values included in the required interval?:

$$\frac{\begin{matrix} number\ of\ data\ items\ having\ a\ value\ included\ in\ a\ specified\ interval \\ (range\ from\ minimum\ to\ maximum) \end{matrix}}{number\ of\ data\ items\ for\ which\ can\ be\ defined\ a\ required\ interval\ of\ values}$$

Concerns all Data Life-Cycle except data design, data file, data item, and data value

2.1.2 Completeness

ISO/IEC 25012 defines data completeness as having values for all required attributes and entity instances. ML algorithms might experience failure when they come across any empty data entries in training, validation, or testing datasets. Similarly, trained ML models might also malfunction when they encounter null values in production data. Completeness measures are critical for ML practitioners to ensure their data meets necessary standards, and they provide guidance on whether to implement additional data imputation methods as outlined in ISO/IEC 5259-4[8]. The concept of data completeness varies across different scenarios and must be evaluated within the context of its specific application. Criteria for assessing data completeness might include: For ML-based image classification, it is important to check for unlabeled samples that are unsuitable for use in supervised learning. For ML-based object detection, one must evaluate any incompleteness in the labeling of bounding boxes around objects. In practice, it is common to encounter samples containing multiple objects across various categories, making it challenging to obtain images with a single isolated object dominating the frame. Thus, when evaluating the completeness of a dataset for ML-based image recognition, considerations should include: The presence of any intended object within a sample, the categorization of all intended objects, and the labeling of all detected objects with bounding boxes or other identification methods.

Example 1:

A completeness assessment shows that over half of the data values for the zip code feature are missing. Considering that the zip code is not essential for their classification task, the data scientist opts to exclude this feature from the training, validation, testing, and production datasets.

Example 2:

A completeness analysis for a dataset used in an ML regression task reveals that 1% of the values for a critical predictive feature are missing, with the remainder of the data

[8] ISO/IEC 5259-4:2024 Artificial intelligence — Data quality for analytics and machine learning (ML). Available at https://www.iso.org/standard/81093.html

following a normal distribution. The data scientist decides to impute these missing values with the statistical mean of the available data.

Example 3:

In a dataset used for an ML clustering task, a completeness check finds a few records with empty data items. The data scientist chooses to remove these incomplete records from the training dataset.

Example 4:

For an ML classification task assessing plant images across the United States, a completeness measure is used to evaluate the proportion of missing data relative to the expected number of data items for proper dataset fidelity. For example, if the dataset is missing ten plant types from the northeastern U.S., this would be noted in the completeness evaluation.

Value completeness: Ratio of data items of no presence of null data values in a dataset:

$$\frac{number\ of\ data\ items\ whose\ value\ is\ not\ null}{total\ number\ of\ data\ items\ in\ the\ dataset}$$

Value occurrence completeness: Ratio of the number of occurrences of a given data value to the expected number of value occurrences in data items with the same domain in a dataset.

$$\frac{number\ of\ occurances\ of\ the\ data\ value\ in\ the\ data\ items}{expected\ number\ of\ occurances\ of\ that\ data\ value\ in\ data\ items\ with\ the\ same\ domain\ in\ the\ dataset}$$

Feature completeness: Ratio of data items of no presence of null data values for a given feature in a dataset.

$$\frac{number\ of\ data\ items\ associated\ with\ the\ given\ feature\ with\ an\ associated\ data\ value\ not\ null}{number\ of\ data\ items\ associated\ with\ the\ given\ feature\ in\ the\ dataset}$$

Record completeness: Ratio of data records of no presence of empty data items in a dataset.

$$\frac{number\ of\ data\ records\ in\ the\ dataset\ not\ having\ any\ empty\ data\ item}{total\ number\ of\ data\ records\ in\ the\ dataset}$$

Label completeness: *Ratio of unlabelled or incompletely labelled samples in a dataset.*

$$1 - \frac{number\ of\ unlabelled\ or\ incompletely\ labelled\ samples}{number\ of\ all\ samples\ in\ the\ dataset}$$

2.1.3 Consistency:

ISO/IEC 25012 defines consistency in terms of data being in agreement with other data and lacking contradictions. Consistency is crucial for machine learning because the features utilized in the training data need to collectively support a model that can accurately make predictions on production data. Machine learning models tend to interpret data values literally, and as such, repeated records could lead to an overemphasis on certain features. Conflicting data within the training set may result in a model performing inadequately against its specified requirements. Furthermore, the distribution of data across features is often used as a criterion for assessing consistency. For example, certain ML models might need data that is normally distributed to achieve expected performance levels.

Data record consistency: Ratio of duplicate records in the dataset

$$\frac{number\ of\ duplicate\ records\ in\ the\ dataset}{total\ number\ of\ data\ records\ in\ the\ dataset}$$

Distribution of data values: Statistical distribution of data values for a given feature in the dataset.

An appropriate distribution measure and measurement function should be determined according to the ML task).

Data format consistency: Consistency of data format of the same data item (according to ISO 25024).

$$\frac{number\ of\ data\ items\ where\ the\ format\ of\ all\ properties\ is\ consistent\ in\ different\ data\ files}{total\ number\ of\ data\ items\ for\ which\ format\ consistency\ can\ be\ defined}$$

Semantic consistency: Degree to which semantic rules are respected (according to ISO 25024).

$$\frac{number\ of\ data\ items\ where\ values\ are\ semantically\ correct\ in\ the\ data\ files}{total\ number\ of\ data\ items\ for\ which\ semantic\ rules\ are\ defined}$$

2.1.4 Credibility:

ISO/IEC 25012 defines credibility as the extent to which data attributes are considered believable by users within a specific usage context. This applies to individual data items, related items within a data record, and entire datasets. The context in which the data is used can affect its perceived accuracy and trustworthiness. Data may be altered during processes such as transit, storage, or computation, either by authorized or unauthorized parties. A particular concern in machine learning is the risk of unauthorized parties tampering with training, validation, testing, and production data, potentially rendering trained models ineffective or influencing the outcomes they produce.

Data preparation methods, such as normalization, imputation, or the splitting and combining of features, can modify data without altering its underlying significance, thereby preserving its credibility.

Example 1:

A dataset intended for training, validating, and testing an ML model does not perform as expected on production data. A security audit reveals that unauthorized changes were made to the data in the training set by an intruder.

Example 2:

A training dataset features numerical data with significantly different ranges. To achieve uniformity, a data scientist decides to normalize these data values. While normalization alters the data values, their credibility remains intact within the machine learning context as their underlying meaning is preserved.

Value credibility: Degree to which information items are regarded as true, real and credible (according to ISO/IEC 25024[9]).

$$\frac{\textit{number of information items where values are validated/certified by a specific process}}{\textit{total number of information items to be validated/certified}}$$

Source credibility: Degree to which values are provided by a qualified organization (according to ISO/IEC 25024).

$$\frac{\textit{number of data values provided or validated/certified by a qualified organization}}{\textit{total number of data values for whic source credibility can be defined}}$$

Data dictionary credibility: Degree to which data dictionary provides credible information (according to ISO/IEC 25024).

$$\frac{\textit{number of information items in data dictionary for which values are validated/certified by specific process}}{\textit{total number of information items in the data dictionary}}$$

Data model credibility: Degree to which data model provides credible information (according to ISO/IEC 25024).

$$\frac{\textit{number of elements of a data model with appropriate definition validated/certified by specific process}}{\textit{total number of elements of a data model}}$$

2.1.5 Currentness

Data currentness is the time difference (ΔT) between the time a data sample is recorded and the time it is used. It ensures that the data is of the correct age relative to its intended usage. For machine learning, currentness may relate to an appropriate age range for the ML task. For instance, data concerning demographic groups may be outdated due to changes in regulations and societal norms. Similarly, economic data spanning several decades may lead to inaccurate ML models if not adjusted for inflation, exchange rates, and other time-sensitive factors. These variances, often referred to as data-drift, impact the data used in production compared to that used in training and testing phases. This can be mitigated by maintaining data currentness. The concept of dataset currentness

[9] ISO/IEC 25024:2015 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality. Available at https://www.iso.org/standard/35749.html

might include the total time span covered by the dataset (such as data collected from 2010 to 2021), the time elapsed since the last data entry (e.g., 8 months), and the frequency of updates (e.g., every 6 months). Currentness should thus be evaluated as a composite metric that incorporates these aspects.

Feature currentness: Ratio of data items for a feature in the dataset that fall within the required age range.

$$\frac{\textit{number of data items for a feature that fall within the required age range}}{\textit{total number of data items for the feature}}$$

Record currentness: Ratio of data records in the dataset where all data items in the record fall within the required age range.

$$\frac{\textit{number of data records that fall within the required age range}}{\textit{total number of data records in the dataset}}$$

Another related concept is the **Age of Information** as will be further explored in Section 3.

2.1.6 Accessibility

Accessibility involves the extent to which data is reachable in a given usage context, especially for individuals requiring assistive technologies or special setups due to disabilities. Furthermore, it is essential that datasets are readily accessible and seamlessly deployable through suitable tools for analytics and machine learning applications.

User accessibility: Degree to which data values are considered accessbile by intended users

$$\frac{\begin{array}{c}\textit{number of data items relevant to the user' task within a specific context of use having values accessible}\\ \textit{by intended users}\end{array}}{\begin{array}{c}\textit{total number of data items that are relevant to the user' task within the context of use having values}\\ \textit{that are required to be accessbile in conformance to specification}\end{array}}$$

Data format accessibility: Degree to which data or information are not accessible by the intended users due to a specific format

$$\frac{\textit{number of data items not accessible due to its format}}{\textit{total number of data items for which format accessibility can be defined}}$$

Data accessibility: Ratio of accessible records in the dataset.

$$\frac{\textit{number of accessbile records in the dataset}}{\textit{total number of data records in the dataset}}$$

2.1.7 Compliance

Compliance refers to the data meeting regulations, standards, conventions or other rules. For instance, personal data used for analytics/ML can be subject to legal and regulatory requirements. Data users can have their own compliance requirements and certification schemes can have compliance requirements.

Data item compliance: *Degree to which data items meet compliance requirements*

$$\frac{\text{number of data items that meet compliance requirements}}{\text{total number of data items in the dataset}}$$

2.1.8 Confidentiality

Confidentiality is the degree to which data has attributes that ensure that access and interpretation are restricted to authorized users within a specific usage context. Confidentiality can be evaluated from both inherent and system dependant perspective.

Encryption usage: *Degree to which data values are fulfilling the requirement of encryption*

$$\frac{\text{number of data values correctly and successfully encrypted and decrypted}}{\text{total number of data values with encryption and decryption requirement}}$$

Non vulnerability: *Degree to which data item defined as confidential can be accessed by authorized users only*

$$1-\frac{\begin{array}{c}\text{number of accesses successfully performed during formal peneration attempts by unauthorized}\\ \text{users to reach target data item in a specific period of time}\end{array}}{\begin{array}{c}\text{number of accesses attempted by unauthorized users to target data}\\ \text{item in a specific period of time}\end{array}}$$

2.1.9 Efficiency

Efficiency is the degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.

Data format efficiency: Unnecessary space occupied rate due to data format definition.

$$1-\frac{\text{size in bytes of record in a data file unnecessarily occupied due to the data format definition}}{\text{size in bytes of record in a data file due to data format definition}}$$

Data processing efficiency: Working time lost due to data item representation (data format)

$$1-\frac{time\ lost\ due\ to\ data\ item\ representation\ (data\ format)\ during\ a\ work}{time\ of\ processing}$$

Risk of wasted space: Wasted space in comparison with benchmarked average space.

$$Sum\ (size\ in\ bytes\ used\ for\ data\ in\ any\ physical\ data\ files\ of\ the\ database) \\ -\ size\ in\ bytes\ assumed\ as\ target\ (i.e., from\ a\ benchmark)\ for\ efficient\ data\ storage\ of\ the\ database$$

2.1.10 Precision

ISO/IEC 25012 defines precision as the exactness or ability of data to discriminate. ISO/IEC 25024 illustrates this through examples such as the number of decimal places in real numbers. In machine learning contexts, the precision of data—such as the decimal places in data values—can influence the significance of a feature in a trained ML model. For instance, a feature with multiple data items at 99.4 may carry more weight than a feature with values rounded to 99. Conversely, features with values rounded up may weigh more than those with finer precision. Data users need to consider how precision impacts the performance of the ML model when setting data precision requirements.

Precision of data values: Degree of data values precision according to the specification.

$$\frac{number\ of\ data\ values\ with\ the\ requested\ precision}{total\ number\ of\ data\ values\ with\ the\ precision\ requirement\ defined}$$

2.1.11 Traceability

Traceability measures indicate the extent to which data possesses attributes that enable an audit trail, documenting access and any modifications to the data within a particular usage context.

Traceability of data values: Degree to which the information of user access to the data value was traced.

$$\frac{number\ of\ data\ values\ for\ which\ required\ access\ traceability\ of\ vlaues\ exist}{number\ of\ data\ values\ for\ which\ access\ traceability\ is\ expected}$$

User Access traceability: Possibility to keep information about users access to data using system capabilities, for investigating who read/wrote data.

$$\frac{number\ of\ data\ items\ for\ which\ user\ access\ traceability\ is\ expected\ and\ realized}{number\ of\ data\ items\ for\ which\ user\ access\ traceability\ is\ expected}$$

Data values traceability: Possibility to trace the history of a data item value using system capabilities.

$$\frac{number\ of\ data\ items\ for\ which\ values\ are\ traceable\ using\ system\ capabilities}{number\ of\ data\ items\ for\ which\ values\ are\ expected\ to\ be\ traceable\ using\ system\ capabilities}$$

2.1.12 Understandability

Understandability refers to the ability of users to read and interpret data effectively. This includes the appropriate use of symbols, units, and languages. In the context of ML where models rely on numerical magnitudes, incorrect unit applications can lead to model failures. Similarly, for tasks involving natural language processing, the improper use of languages and symbols can obstruct successful language comprehension and generation. While data quality metrics are often quantitative, the qualitative assessments made by humans utilizing data for machine learning are also crucial. Correct application of symbols, units, and languages plays a key role in facilitating these qualitative judgments.

Symbols understandability: Degree to which comprehensible symbols are used

$$\frac{\textit{number of data values represented by known symbols}}{\textit{total number of data values for which symbols understandability is requested}}$$

Semantic understandability: Ratio of the common recognized vocabulary which is used in terms of definitions given in the data dictionary

$$\frac{\textit{number of data values defined in the data dictionary}}{\textit{total number of data values defined in the data dictionary}}$$

Data values understandability: Data values are understandable by intended users in the specific context of use.

$$\frac{\textit{number of data values easily understandable by intended users}}{\textit{total number of data values that users attempt to understand during an observation period}}$$

Data representation understandability: Degree to which data is represented in a comprehensible way to users by system and software.

$$\frac{\textit{number of data items considered understandable by intended users}}{\textit{total number of data items presented in a specific device}}$$

2.1.13 Availability

Availability measures assess the extent to which data attributes allow it to be accessed by authorized users and applications within a defined usage context.

Data availability ratio: Ratio of data items available when required (e.g., during backup/restore procedures).

$$\frac{\textit{number of data items available in a specific period of time}}{\textit{number of data items requested in the same period of time}}$$

Portability of data available: Probability of successful requests trying to use data items during requested duration.

$$\frac{\textit{number of times that data items are available for the requested duration}}{\textit{number of times that data items are requested for the requested duration}}$$

Architecture elements availability: Degree to which architecture elements are available.

$$\frac{\textit{number of elements of the architecture available for the intended users}}{\textit{number of elements of the architecture}}$$

2.1.14 Portability

ISO/IEC 25012 defines the data quality characteristic of portability as the ability to transfer data from one system to another within a specific context, while maintaining its quality. In the realm of analytics and machine learning, data might be processed across various systems—for instance, data might be collected on one system, undergo quality processing on a second system, and then transferred to a third system for training an ML model. If the data does not retain its quality during these transfers, the effectiveness of the trained ML model could be compromised. It's crucial that data portability requirements are clearly established to ensure data maintains its integrity throughout its movement across systems.

Data portability ratio: Data quality does not decrease after porting (or migration).

$$\frac{\textit{number of data items that preserve existing quality after porting}}{\textit{total number of data items ported}}$$

Perspective data portability: Degree to which portability of data item conforms to requirements

$$\frac{\textit{number of data items that can be moved to a target system}}{\textit{total number of data items for which portability is expected}}$$

2.1.15 Recoverability

Recoverability measures assess the extent to which data attributes support the maintenance and preservation of a specified level of operations and quality, even during failures, within a particular usage context.

Data recoverability: Degree to which data stored in a device are successfully and correctly recovered

$$\frac{\textit{number of data items successfully and correctly recovered by the system}}{\textit{total number of data items that are required to be recovered}}$$

Periodical backup: Data is backed up periodically as stated in requirements

$$\frac{\textit{number of data items (or data file) successfully backed up periodically}}{\textit{total number of data items (or data file) to be backed up}}$$

Architecture recoverability: Degree to which architecture elements are recoverable

$$\frac{\text{number of elements of the architecture successfully recovered}}{\text{total number of elements of architecture that shall be managed by backup or restore procedures}}$$

2.2 Additional data quality characteristics (ISO/IEC 5259 series, 25024)

2.2.1 Auditability

Auditability is defined as the feature of a dataset where either the entire dataset or sections of it have been audited, or where the data are accessible to relevant stakeholders for audit purposes. Conducting audits on datasets used in analytics and machine learning enhances the credibility of the data and may be necessary to meet compliance requirements.

Audited records: *Ratio of the records in the dataset that have been audited*

$$\frac{\text{number of records in the dataset that have been audited}}{\text{total number of records in the dataset}}$$

Auditable records: *Ratio of the records in the dataset that are available for audit.*

$$\frac{\text{number of records in the dataset available for audit}}{\text{total number of records in the dataset}}$$

2.2.2 Identifiability

ISO/IEC 29100 defines identifiability as the ability to recognize an individual either directly or indirectly through specific personally identifiable information (PII) within a dataset. It is crucial to determine if any PII within a dataset can identify an individual, as legal constraints in various regions may regulate or prohibit such activities. To mitigate risks of identifiability, de-identification processes can be implemented across training, validation, testing, and production data sets.

For instance, an ML model designed for targeted advertising might use data from search engine queries, including users' IP addresses, which are recognized as PII under certain legal frameworks. To ensure compliance and enhance privacy, anonymization techniques are employed to remove the IP addresses before dividing the dataset into training, validation, and testing sets.

Identifiability ratio: *Ratio of data records in the dataset that can be used for identifiability.*

$$\frac{\begin{array}{c}\text{number of data records that contain data items that can be used for identifiability,}\\ \text{either on their own or in conjuction with other data items}\end{array}}{\text{total number of data records in the dataset}}$$

2.2.3 Effectiveness

Effectiveness of a dataset is defined as its ability to meet the requirements for a specific machine learning (ML) task. Here are examples illustrating how dataset effectiveness is assessed in different ML applications: 1) *Computer Vision System:* In an ML-based computer vision system, the effectiveness of the dataset could be measured by the lowest acceptable ratio of images with brightness or resolution below a certain threshold relative to all images or videos in the dataset. This metric ensures that the quality of the visual data is sufficient to support accurate processing and analysis by the system, 2) *Image Classification System:* For an ML-based image classification system, dataset effectiveness might be determined by the minimum acceptable proportion of images that belong to a specific category compared to the total number of images in the dataset. This measure helps evaluate whether there is enough representational data for each category to train the model effectively, 3) *Object Detection System:* In an ML-based object detection system, the effectiveness of the dataset could be evaluated by the lowest acceptable ratio of images that meet specific criteria necessary for the object detection task, such as clarity, object placement, or other relevant attributes. This ensures the dataset is capable of supporting the training of a model that can accurately identify and locate objects within images. Each of these examples underscores the importance of having a dataset that is not only large but also qualitatively aligned with the demands of the specific ML tasks it supports.

Feature effectiveness: *Ratio of samples with acceptable feature in a dataset*

$$\frac{number\ of\ the\ samples\ with\ acceptable\ feature}{total\ number\ of\ samples\ in\ the\ dataset}$$

Category size effectiveness: *Ratio of categories where the number of categorized samples is lower than a threshold*

$$\frac{number\ of\ categories\ where\ the\ number\ of\ categorized\ samples\ is\ lower\ than\ a\ threshold}{total\ number\ of\ categories}$$

Label effectiveness: *Ratio of samples with acceptable label in a dataset*

$$\frac{number\ of\ the\ samples\ with\ acceptable\ label}{total\ number\ of\ samples\ in\ the\ data}$$

2.2.4 Balance

Balance refers to the equitable distribution of samples across all aspects relevant to the dataset. For instance, in a dataset with multiple categories, balance would mean having a roughly equal number of samples in each category. In image datasets, important factors might include label relevance to business logic, image resolution, brightness, and attributes like the width-to-height ratio and size of labeled bounding boxes. These factors are critical because they can significantly impact the performance of a machine learning (ML) model.

The balance of a dataset is crucial for ensuring reliable performance in ML applications. For example, in an ML-based computer vision system, ensuring a balanced dataset is vital for accurate system functioning.

Example 1:

In scenarios where there are considerable differences in brightness or resolution between the training dataset samples and the real-world data, ML models may perform poorly. Issues such as faintness or blurriness introduce noisy data, which can degrade model accuracy.

Example 2:

In ML-based classification systems, an imbalance in the sample population across categories can hinder the discovery and correct classification of rare instances. These instances might be incorrectly labeled as noise or misclassified, due to the model's overfitting to more frequently represented categories.

Example 3:

For ML-based object detection systems, significant variations in the width-to-height ratios or the sizes of bounding boxes can result in inconsistencies in the detected object sizes, particularly if the receptive field size of the model is fixed. This discrepancy can affect the model's ability to accurately detect and categorize objects.

Overall, maintaining dataset balance is essential for optimizing ML model performance and ensuring that the system functions effectively in varied real-world conditions.

Brightness balance: Reciprocal of the maximal ratio of the brightness difference of an image sample over the averaged brightness of samples in a dataset.

$$\frac{\text{average value of brightness of the samples}}{\text{maximum value of absoluted differences between the brightness value of each image in the sample and } A}$$

Resolution balance: Reciprocal of the maximal ratio of the resolution difference of an image sample over the averaged resolution of samples in a dataset.

$$\frac{\text{average value of resolution of the samples}}{\text{maximum value of absolute differences between the resolution value of each image in the sample and } A}$$

Balance of images between categories: Reciprocal of the maximal ratio of the category size (number of contained samples) difference over the averaged category size of a dataset.

$$\frac{\text{average category size of the dataset}}{\text{maximum value of absolute differences between the size of each category in the dataset and } A}$$

Bounding box height to width ratio balance: Reciprocal of the maximal ratio of the bounding box height to width ratio difference over the averaged bounding box height to width ratio of the samples in a dataset.

$$\frac{\text{average bounding box with height to width ratio over all the samples in the dataset}}{\begin{array}{c}\text{maximum value of absolute differences between bounding box with height to width ratio of each sample}\\ \text{in the dataset and } A\end{array}}$$

Category of bounding box area balance: Reciprocal of the maximal ratio of the averaged bounding box area of a category over the averaged bounding box area of all the samples in a dataset.

$$\frac{\text{average bounding box area over all the samples in the dataset}}{\begin{gathered}\text{maximum value of absolute differences between averaged bounding box area of each category}\\ \text{in the dataset and } A\end{gathered}}$$

Samples bounding box area balance: Reciprocal of the maximal ratio of the bounding box area of a sample over the averaged bounding box area of all the samples in a dataset.

$$\frac{\text{average bounding box area over all the samples in the dataset}}{\begin{gathered}\text{maximum value of absolute differences between averaged bounding box area of e⍰⍰ch sample}\\ \text{in the dataset and } A\end{gathered}}$$

2.2.5 Diversity

Diversity in a dataset signifies the variety among samples regarding the target data. For machine learning models, it is crucial that samples differ sufficiently. Homogeneous datasets can lead to overfitting, reducing the model's ability to generalize. Diversity in a dataset indicates a range of different value domains, labels, clusters, and distributions among the data entries. Using generative ML models to enhance data diversity can be beneficial, but these methods may not be effective if the original dataset lacks sufficient diversity. Diversity, closely linked with representativeness and balance, is a data quality attribute that helps assess the fidelity of a dataset. The measurement of diversity should be tailored to the specific requirements of the ML task concerning the target data.

Label richness: Number of different labels in a dataset.

$$\text{number of different labels in the dataset}$$

Relative label abundance: Portion of the number of individual data (i.e., item, record, frame) having the same label in a dataset.

$$\frac{\text{number of individual data in which target labels}}{\text{total number of individual data in the dataset}}$$

Component richness: Count of time series components in the dataset as a number between 1 and 4, divided by 4.

Time series components can be:
- *Trend* – data established by model that describes a stable, long-term tendency of data
- *Seasonal variations* – data established by a model that describes fundamental periodic information on timescales of hours, days, months, and/or quarters
- *Cyclical variations* – data established by a model that describes fundamental periodic information on timescales of more than a year, often related to a business cycle
- *Irregular variations* – data that is not provided by the other components– often considered as random. The difference in the value predicted by trend, seasonal, and cyclical variations are established in the irregular variations component.

$$\begin{gathered}\text{count of time series components in}\\ \text{the dataset as a number between } 1 \text{ and } 4,\\ \text{divided by } 4\end{gathered}$$

Category size diversity: Ratio of categories where the number of categorized samples is lower than a threshold.

$$\frac{\textit{number of categories where the number of categorized samples is lower than a threshold}}{\textit{total number of categories in total}}$$

2.2.6 Relevancy

Relevance is defined as how suitable a dataset is for a specific context, assuming it meets other quality criteria like accuracy, completeness, consistency, and currentness. For machine learning, relevance means that the features selected in the training data, and their values, effectively predict the target variable.

For example, consider an ML model designed to assess creditworthiness. The training dataset is representative of the population expected in the production data and includes pertinent features such as credit history, income, job tenure, and net worth—all of which are strong predictors of creditworthiness. However, it also includes data on individuals' height and weight. Statistical analysis reveals no significant correlation between these dimensions and credit history, indicating that they are ineffective predictors of credit performance. Therefore, to enhance the dataset's relevance, features like height and weight are omitted.

Feature relevance: *Ratio of features in the dataset that are relevant to the given context.*

$$\frac{\textit{number of features in the dataset deemed to be relevant in the context of the use of the data}}{\textit{total number of features in the dataset}}$$

Record relevance: *Ratio of records in the dataset that are relevant to the given context.*

$$\frac{\textit{number of record in the dataset deemed to be relevant in the context of the use of the data}}{\textit{total number of records in the dataset}}$$

2.2.7 Representativeness

ISO 20252 defines representativeness as the extent to which a sample accurately mirrors the target population under investigation. In supervised machine learning, the training dataset serves as the sample, while the production data represents the broader population. If the training data inadequately reflects the production data, the resulting ML model may not perform as intended. Representativeness is closely tied to the relevance data quality characteristic, as a dataset that does not faithfully represent the population being studied is unlikely to yield reliable predictions for the target variable.

For instance, a facial recognition system trained solely on images of individuals with light skin tones may struggle to accurately identify individuals with darker skin tones.

Representativeness ratio: Ratio of relevant attributes found in the subjects of a population to the attributes found in the sample.

$$\frac{\textit{number of target attributes in the samples}}{\textit{total number of attributes in the population}}$$

2.2.8 Similarity

The dataset's similarity pertains to how closely related samples are based on specific features of interest. This is crucial for tasks like classification, typically conducted through supervised learning, and clustering, commonly implemented via unsupervised learning. Both tasks require sufficient diversity among samples for effective performance. For instance, an ML model trained on a dataset with highly similar images risks overfitting and reduced generalizability, especially if generated from a limited set of seed images. Techniques such as data augmentation (e.g., rotation, shifting) can mitigate this issue if applied judiciously. Additionally, clustering algorithms with methods for handling topic drift can help manage datasets with varying levels of similarity. Geometric approaches can further analyze and compare datasets by representing data records as vectors in multi-dimensional space, where similarity is determined by their spatial relationships.

Sample similarity: Ratio of similar samples in a dataset; the lower, the better.

$$1 - \frac{number\ of\ all\ samples\ in\ the\ dataset}{total\ number\ of\ the\ clusters, resulting\ from\ a\ clustering\ algorithm, on\ all\ samples\ of\ the\ dataset}$$

Samples tightness: *Tightness of normalized dataset.*

$$Max\ eigenvalues\ of\ G^{a} - Min\ eigenvalue\ of\ G^{a}$$

Where G is a matrix with M rows and M columns and is equal to $\Phi_{norm}\ \mathrm{T}\ \Phi_{norm}$

NOTE 1 Φ_{norm} is the normalized dataset, calculated from Φ_{NXM} (NOTE 2) after subtracting from each column its mean, and normalization to 1. Visually, normalized data rely over the surface of a hypersphere of radius=1 and centered in the origin $(M \leq N)$.
NOTE 2 Φ_{NXM} is an N-by-M matrix, with N data records (vectors) and M features (dimensions).
NOTE 3 The number of principal components $K \leq M$ is the smallest number of eigenvalues of C_{MXM} (NOTE4), starting from the biggest, chosen in order to represent 95% of their sum.
NOTE 4 C_{MXM} is an M-by-M matrix, with M rows and M columns and is equal to $\Phi_{mean}\ \mathrm{T}\ \Phi_{mean}$ (NOTE 5).
NOTE 5 $\Phi_{mean}$is calculated from Φ_{NXM} after subtracting from each column of its mean. Visually, normalized data Φ_{mean} fit an (hyper)ellipsoid with eigenvectors as axis and centered in the origin.
NOTE 6 Principal components can be selected with criteria or percentage different; see Annex A in ISO/IEC DIS 5259-2 for measure modification.
NOTE 7 A measurement of zero means the least similarity. The similarity measure will yield zero when the number of samples is equal to the number of clusters indicating that no sample is similar to another.

Samples independency: *Ratio of Principal Component Analysis (PCA) and dataset dimension*

$$1 - \frac{number\ of\ PCs\ with\ PCA\ method}{total\ number\ of\ dataset\ dimensions}$$

2.2.9 Timeliness

Timeliness, also called the latency, is the ΔT between the time when a phenomenon occurs and the time when the recorded data for that phenomenon are available for use, which makes it different from currentness (i.e., ΔT between the time a data sample is recorded and the time it is used.

If the ΔT between a phenomenon and the availability of its corresponding data sample is too great, it can no longer be a good predictor in the context of ML. AI application tasks related to streaming data (e.g. analysis of securities transactions, reinforcement learning, search queries) can make use of continuous learning and inferencing in near real-time.

$$\frac{\textit{number of data items in the dataset that meet timeliness requirements}}{\textit{total number of data items in the dataset}}$$

2.3 Data type specific quality testing

Data quality testing can be performed by the general properties outlined previously, yet it may also necessitate specialized tests tailored to specific data types and use cases. Typically, the bulk of machine learning data falls into categories such as tabular, textual, and image data.

2.3.1 Tabular data

Tabular data is the most traditional form of data and is structured in rows and columns, similar to what you would find in a spreadsheet. Each row typically represents an instance or record, and each column represents a feature or attribute of the instance. This type of data is commonly used in predictive modelling, such as classification and regression tasks. Financial records, customer databases, and sales figures are classic examples of tabular data.

Tabular data is well-suited for statistical analysis and traditional machine learning algorithms such as decision trees, ensemble methods, and linear regression. Some relevant standards:

- **ISO/IEC 11179**[10] - Metadata Registries (MDR) series: This series of standards focuses on the management of metadata for data interchange between organizations and interoperability among disparate systems. It is particularly useful for managing the quality of structured (tabular).
- **ISO/IEC 38500**[11] - Governance of IT: Although not specific to data quality, this standard provides a framework for effective governance of IT to support data management operations and ensure data quality across the organization.

2.3.2 Textual data

Text data comprises strings of characters and is one of the most common unstructured data types. It is pervasive, found in emails, social media posts, news articles, documents, books, and more. Natural Language Processing (NLP) is the subset of machine learning that deals with understanding and processing text data. Text data requires specialized preprocessing techniques like tokenization, stemming, lemmatization, and the removal of stop words. Some relevant standards:

[10] ISO/IEC 11179-1:2023 Information technology — Metadata registries (MDR). Available at https://www.iso.org/standard/78914.html

[11] ISO/IEC 38500:2024 Information technology — Governance of IT for the organization. Available at https://www.iso.org/standard/81684.html

- **ISO/IEC 11179** - Metadata Registries (MDR) series: As discussed earlier, this series of standards focuses on the management of metadata for data interchange. Besides, structured (tabular), it is also relevant for managing the quality of semi-structured (some textual) data through comprehensive metadata registration.

2.3.3 Image data

Image data consists of arrays of pixel values, where each pixel can represent levels of brightness or color values across different channels, such as red, green, and blue (RGB). It is a prevalent form of unstructured data found in applications ranging from medical imaging and satellite photos to everyday photography and video content. Computer Vision is the subset of machine learning that focuses on interpreting and processing image data. Each image is typically represented as a matrix of pixels for grayscale images or a three-dimensional array for color images, where the third-dimension accounts for color channels. This format allows sophisticated operations and transformations that can extract features, detect objects, recognize patterns, and perform image classification and analysis.

Image data requires specific preprocessing steps to ensure that the data is in a suitable form for analysis by machine learning models. Common preprocessing tasks include resizing images to a uniform dimension, normalizing pixel values (scaling pixel values to a range, typically 0 to 1), and augmenting the dataset through techniques such as rotation, scaling, and cropping to improve the robustness of the model. Advanced computer vision applications often utilize convolutional neural networks (CNNs), which are specifically designed to process pixel data and recognize spatial hierarchies in images, making them effective for tasks such as image recognition, object detection, and semantic segmentation. Some relevant image data specific standards:

- **ISO/IEC 15948**[12] - Portable Network Graphics (PNG): Specifies a data format for lossless, portable, compressed graphics for raster images. It includes provisions that enhance the quality and integrity of the image data.
- **ISO/IEC 19794**[13] - Biometric Data Interchange Formats: Particular parts of this standard deal with image data, such as face, iris, and fingerprint images. It sets quality and formatting specifications for biometric data interchange.

2.4 Specialized tools for data quality testing

Tools utilized in data quality testing can range from software packages that automate the testing process to custom scripts designed to identify specific types of data anomalies. Some data quality tools are mentioned in data quality standards and some state-of-the-art tools can fulfil the required characteristics mentioned in the data quality standards. Below are discussed two of them:

[12] ISO/IEC 15948:2004 Information technology — Computer graphics and image processing — Portable Network Graphics (PNG): Functional specification. Available at https://www.iso.org/standard/29581.html

[13] ISO/IEC 19794-5:2011 Information technology — Biometric data interchange formats. Available at https://www.iso.org/standard/50867.html

Apache Griffin[14] **-** is an open source data quality solution for distributed data systems at any scale. It supports both batch and streaming modes and provides a way to measure data quality in multiple dimensions, from accuracy and completeness to timeliness and profiling. It includes a data quality service framework, a process engine to schedule and run jobs, and a metric model to define data quality measurements.

AI Fairness 360/AIF360[15] **(ISO/IEC TR 24027**[16]**) -** The AI Fairness 360 toolkit is an extensible open-source library containing techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle. AI Fairness 360 package is available in both Python and R.

RISE Research Institutes of Sweden AB
Box 857, 501 15 BORÅS, SWEDEN
Telephone: +46 10-516 50 00
E-mail: info@ri.se, Internet: www.ri.se

INDUSTRIAL SYSTEMS
RISE Report :
ISBN:

[14] Apache Griffin. Available at https://griffin.apache.org/
[15] AI Fairness 360. Available at https://aif360.res.ibm.com/
[16] ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making Available at https://www.iso.org/standard/77607.html