International Standard

ISO/IEC 5259-2

First edition
2024-11

Artificial intelligence — Data quality for analytics and machine learning (ML) —

Part 2: Data quality measures

Intelligence artificielle — Qualité des données pour les analyses de données et l'apprentissage automatique —

Partie 2: Mesure de la qualité des données

Reference number
ISO/IEC 5259-2:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial Intelligence.*

A list of all parts in the ISO/IEC 5259 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

Data-supported decision-making brings new challenges to data quality management in data analytics and artificial intelligence (AI) based on machine learning (ML). Issues in data quality, such as incomplete, false or outdated data, can adversely affect analytics and ML processes and outcomes. Data from various sources, including structured data (e.g. relational databases) and unstructured data (e.g. documents, images, audios), can be directly consumed into the data life cycle for analytics and ML model development. Data are transformed in each stage of the data life cycle of analytics and ML. A holistic standardized approach to control, produce and deliver sufficient high-quality data is necessary for data analytics and ML models to be safe, reliable and interoperable. To develop credible data quality management for analytics and ML, intrinsic data quality International Standards, including concepts and use cases, characteristics and measurements, management requirements, and process framework, can be considered.

This document is a part of the ISO/IEC 5259 series. This document builds upon the ISO 8000 series, ISO/IEC 25012 and ISO/IEC 25024. The purpose of this document is to describe a data quality model through the definition of data quality characteristics and data quality measures based on ISO/IEC 25012 and ISO/IEC 25024. Data quality models can be extended or modified according to this document.

Artificial intelligence — Data quality for analytics and machine learning (ML) —

Part 2:
Data quality measures

1 Scope

This document specifies a data quality model, data quality measures and guidance on reporting data quality in the context of analytics and machine learning (ML).

This document is applicable to all types of organizations who want to achieve their data quality objectives.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 5259-1, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples*

ISO/IEC 25024, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality*

ISO/IEC 22989, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 5259-1, ISO/IEC 22989 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
data
re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing

Note 1 to entry: Data can be processed by humans or by automatic means.

[SOURCE: ISO/IEC 2382:2015, 2121272]

3.2
data frame
set of data records represented by a specific domain or purpose, with a shared structure of data items

Note 1 to entry: A data frame is two-dimensional, like a table with rows and columns. The term is specifically used in analytics and ML, e.g. in the R language, while other languages use “data set” to mean the same thing. In this document, “dataset” has a more generic meaning.

3.3
data type
categorization of an abstract set of possible values, characteristics, and set of operations for an attribute

Note 1 to entry: Examples of data types are character strings, texts, dates, numbers, images and sounds.

[SOURCE: ISO/IEC 25024:2015, 4.16]

3.4
data value
content of data item

Note 1 to entry: In ISO/IEC 25012:2008, 5.1.1, it is specified that from the inherent point of view, data quality refers to data itself such as data domain values and possible restrictions.

Note 2 to entry: Number or category assigned to an attribute of a target entity by making a measurement.

[SOURCE: ISO/IEC 25024:2015, 4.17]

3.5
empty data item
data item whose *data value* (3.4) has no value, i.e. Null or None

Note 1 to entry: This definition in general signifies non-existence of a data value (i.e. Null or None). A data item with string data type can be an empty data item by using either the empty string or Null. However, there is an exception for some application a string can be empty (e.g. “”) but not Null and hence not imply an empty data item.

3.6
entity
concrete or abstract thing in the domain under consideration

[SOURCE: ISO 8000-2:2022, 3.3.3]

3.7
raw data
data in its originally acquired, direct form from its source before subsequent processing

[SOURCE: ISO 5127:2017, 3.1.10.04]

3.8
target data
data (3.1) used in an analytics or ML task whose quality is measured

3.9
target population
population of interest in the analytics or ML project to which inferences are to be made

3.10
data quality subject
entity (3.6) affected by data quality

3.11
quality measure element
measure defined in terms of a property and the measurement method for quantifying it, including optionally the transformation by a mathematical function

[SOURCE: ISO/IEC 25024:2015, 4.32]

3.12
quantity
property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference

[SOURCE: ISO/IEC Guide 99:2007, 1.1, modified — Notes to entry deleted.]

3.13
quantity value
number and reference together expressing magnitude of a *quantity* (3.12)

[SOURCE: ISO/IEC Guide 99:2007, 1.9, modified — Examples deleted.]

3.14
measurement function
algorithm or calculation performed to combine one or more *quality measure elements* (3.11)

[SOURCE: ISO/IEC 25021:2012, 4.7, modified — Definition revised.]

3.15
measurement result
result of measurement
set of *quantity values* (3.13) being attributed to a measurand together with any other available relevant information

[SOURCE: ISO/IEC Guide 99:2007, 2.9, modified — Notes to entry deleted.]

3.16
measure
<noun> variable to which a value is assigned as the result of measurement

Note 1 to entry: The plural form "measures" is used to refer collectively to base measures, derived measures and indicators.

[SOURCE: ISO/IEC/IEEE 15939:2017, 3.15]

3.17
measure
<verb> make a measurement

[SOURCE: ISO/IEC 25000:2014, 4.19]

3.18
bounding box
rectangular region enclosing annotated object

Note 1 to entry: The major and minor axes of the rectangle are parallel to the edges of the images. For rotated boxes, the polygon annotation is to be used.

[SOURCE: ISO/IEC 30137-4:2021, 3.3]

3.19
cluster
automatically induced category of elements that are part of the dataset and that share common attributes

Note 1 to entry: Clusters do not necessarily have a name.

[SOURCE: ISO/IEC 23053:2022, 3.3.2]

3.20
clustering algorithm
algorithm which groups *clusters* (3.19) from input data

Note 1 to entry: Examples of clustering algorithms include centroid-based clustering, density-based clustering, distribution-based clustering, hierarchical clustering and graph-based clustering.

3.21
overfitting
<machine learning> creating a model which fits the training data too precisely and fails to generalize on new data

Note 1 to entry: Overfitting can occur because the trained model has learned from non-essential features in the training data (i.e. features that do not generalize to useful outputs), excessive noise in the training data (e.g. excessive number of outliers), a significant mismatch between training data and production data distributions or because the model is too complex for the training data.

Note 2 to entry: Overfitting can be identified when there is a significant difference between errors measured on training data and on separate test and validation data. The performance of overfitted models is especially impacted when there is a significant mismatch between training data and production data.

[SOURCE: ISO/IEC 23053:2022, 3.1.4]

3.22
fidelity
degree to which a model or simulation reproduces the state and behaviour of a real-world object or the perception of a real-world object, feature, condition, or chosen standard in a measurable or perceivable manner

[SOURCE: ISO 16781:2021, 3.1.4]

3.23
maintainability
ability of a functional unit, under given conditions of use, to be retained in, or restored to, a state in which it can perform a required function when maintenance is performed under given conditions and using stated procedures and resources

Note 1 to entry: The term used in IEV 191-02-07 is “maintainability performance” and the definition is the same.

Note 2 to entry: maintainability: term and definition standardized by ISO/IEC [ISO/IEC 2382-14:1997].

Note 3 to entry: 14.01.06 (2382)

[SOURCE: ISO/IEC 2382:2015, 2123027]

3.24
reliability
consistency with which an assessment measures

EXAMPLE An assessment will have low reliability if two assessment forms are of unequal difficulty or coverage or if there are errors in the scoring procedures or in the reporting of scores.

[SOURCE: ISO/IEC 23988:2007, 3.21]

3.25
validity
extent to which an assessment achieves its aim by measuring what it is supposed to measure and producing results which can be used for their intended purpose

Note 1 to entry: An assessment has low validity if the results are unduly influenced by skills which are irrelevant to the stated aims of the assessment.

[SOURCE: ISO/IEC 23988:2007, 3.25]

4 Symbols and abbreviated terms

AI	artificial intelligence
CSV	comma separated values
HDF	hierarchical data format
JSON	JavaScript object notation
ML	machine learning
IP	internet protocol
PII	personally identifiable information
QM	quality measure
UML	unified modelling language

5 Data quality components and data quality models for analytics and machine learning

5.1 Data quality components in data life cycle

Figure 1 shows data quality components aligned with the data life cycle model shown in ISO/IEC 5259-1:2024, Figure 3, which can support data quality management processes. ISO/IEC 5259-1 defines a data quality model as a defined set of data quality characteristics. The data quality characteristic provides a framework for data quality requirements, implementation and evaluation methods. Data quality measures are variables assigned to which values are the results of measurements of data quality characteristics. Data quality measures are used to assess whether the data meet data quality requirements. Data quality measures can also be used to monitor and report data quality.

Target data are the data subject to data quality measurements. Target data can be raw data or data that has undergone one or more processes or transformations. Target data for measuring quality can be training, testing, validation, production and output data in the context of the use of analysis and ML (as described in ISO/IEC 23053).[1] Target data can be formed as either data items or datasets. A data item consists of an item name, data value and data type representing a domain of values (e.g. character strings, texts, dates, numbers, images, sounds). A dataset can be classified into three forms:

— a collection of data items;

— a collection of data records;

— a collection of data frames.

The target data can be unlabelled or labelled depending on the association with data labels in the use of analytics or ML task.

NOTE This document makes no distinction between data structures, such as structured data, semi-structured data and unstructured data, or data roles, such as master data, transaction data and reference data.

Data quality reports are documents that express data quality requirements, the data quality model of data quality characteristics, data quality measures, the results of data quality measurements and an assessment of whether the data meet data quality requirements.

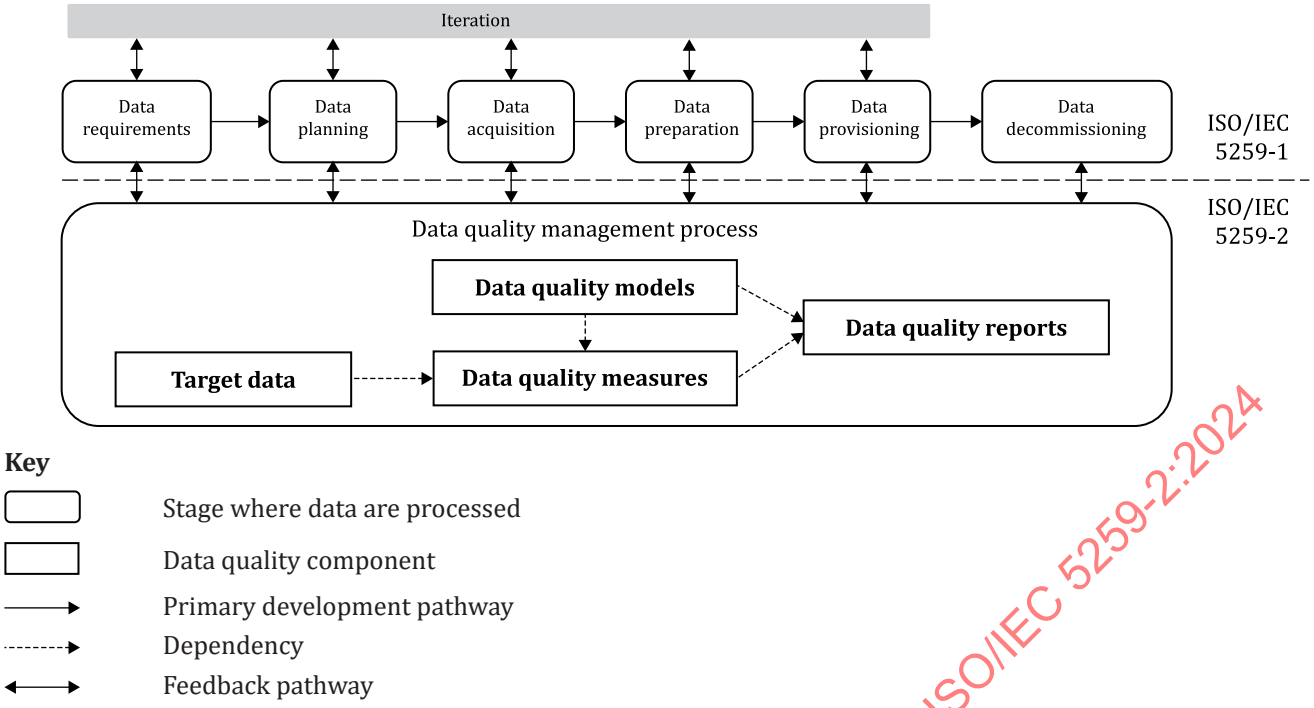

Figure 1 — Data quality components in data life cycle for analytics and ML

5.2 Data quality model

The data quality model provides a framework for specifying data quality requirements and evaluating data quality. In practice, a data quality model brings together data quality subjects, data quality characteristics and data quality requirements, for the context of the use of the data. The organization can specify data quality models by selecting data quality characteristics and measures to achieve target quality requirements for target data. Figure 2 provides a UML diagram of the relationships between the components of the data quality model.

A data usage scope describes how and where the data can be used in an analytics or ML task and how it fits into an AI system.

EXAMPLE The data can be used to train a deep neural network ML model to predict product sales based on the features of a marketing strategy. The model can be trained and deployed using cloud services.

A data quality subject represents an entity affected by data quality. A data quality characteristic is a category of data quality attributes that bear on data quality (e.g. accuracy, completeness, precision). A data quality requirement describes properties or attributes of the data along with acceptance criteria relative to the data usage scope. Acceptance criteria can be quantitative or qualitative.

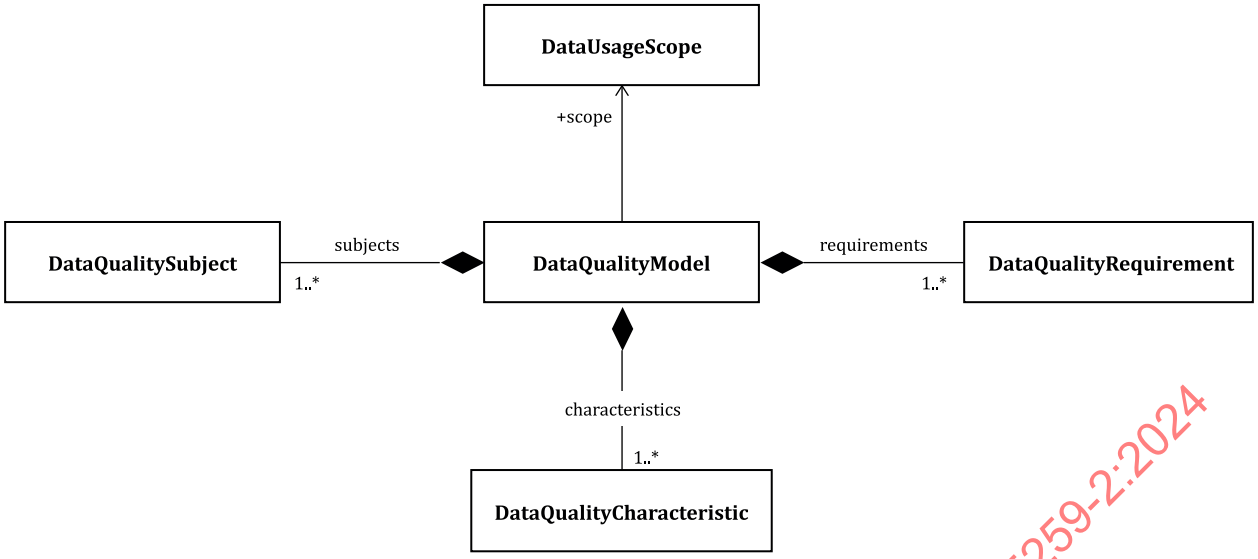

Figure 2 — Data quality model

When one quality characteristic affects another, trade-offs can be made by evaluating each requirement regarding importance and impact. In addition, it is crucial to balance the cost of data quality management with the priority of data quality requirements in determining how data quality characteristics and measures are incorporated into the data quality model. The organization can select the data quality characteristics and measures that correspond to their needs and requirements. Data quality should be assessed by comparing the results of selected data quality measures against established targets as established by data requirements. Any failures to achieve data quality requirements should be mitigated. ISO/IEC 5259-3[2] describes the requirements and recommendations of a data quality management system to be applied by the organization.

ISO 8000-8[3] and ISO/IEC 25012[4] describe data quality models. ISO 8000-8 defines three data quality characteristics as being syntactic (format), semantic (meaning), and pragmatic (usefulness) to support industrial data generally as a product of business and manufacturing processes. ISO/IEC 25012 defines a general data quality model for data retained in a structured format within a computer system as a part of a software product. ISO/IEC 25012 takes into account all data types (e.g. characters, strings, texts, dates, numbers, images, sounds). ISO/IEC 25012 provides fifteen data quality characteristics: accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability and recoverability.

The ISO 8000 series[5] addresses various aspects of data quality such as data governance, data quality management (including processing) and maturity assessment. The ISO/IEC 25000 series[6] addresses product (software, systems, data, services) quality requirements and evaluation. This document describes how the data quality characteristics of ISO/IEC 25012 can be applied to a data quality model for analytics and ML. Furthermore, this document defines additional characteristics that can contribute to higher-quality ML models and applications, as shown in Figure 3. Organizations should use the data quality characteristics and data quality measures described in this document whenever possible. However, the data quality characteristics in this document cannot comprehensively cover aspects that support all organizations' needs regarding data quality. Organizations may design their own data quality model by extending the data quality characteristics and data quality measures to fit their data requirements.

NOTE 1 See Annex A for information on designing and documenting measurement functions.

NOTE 2 See Annex E for a comparison between the data quality characteristics in ISO/IEC 25012 and those in this document.

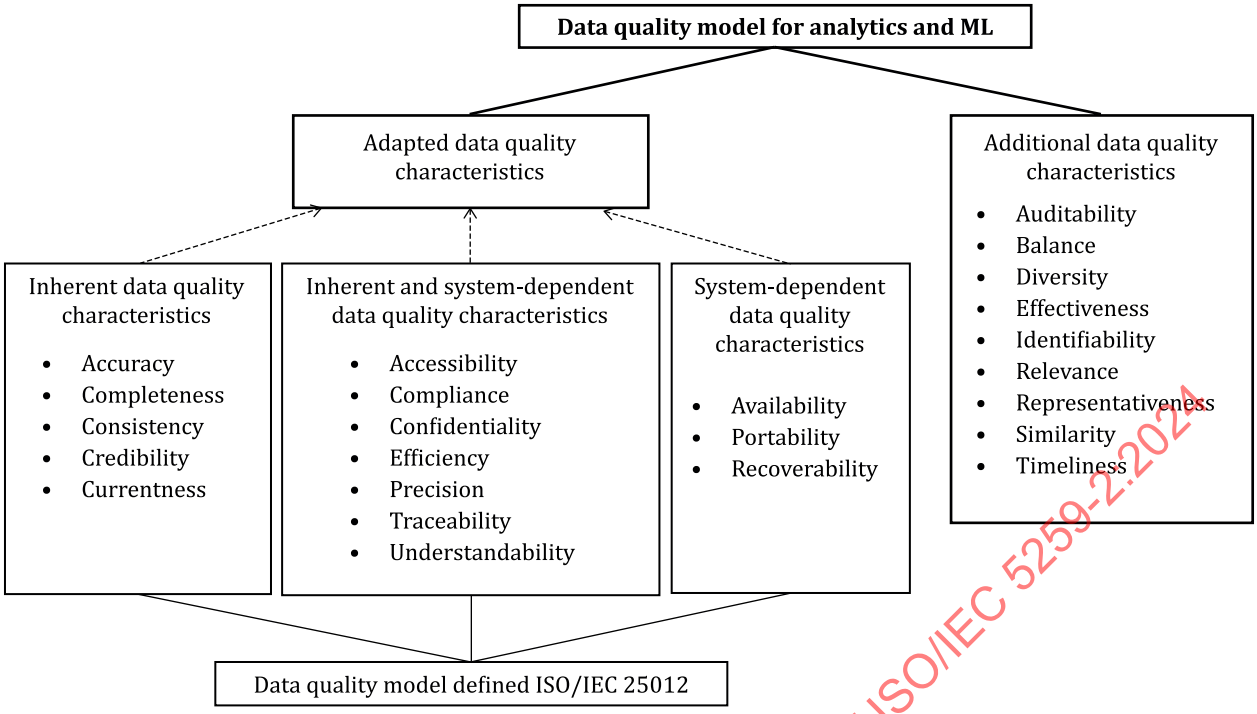

Figure 3 — Data quality characteristics for analytics and ML

6 Data quality characteristics and quality measures

6.1 General

Data quality characteristics and measures are used to specify and verify data quality requirements for identified attributes for target data. Each data quality characteristic is associated with one or more data quality measures for quantification. A data quality measure is a variable to which a value is assigned as the result of a measurement function. The data quality measures in this document are selected based on the context of use of analytics and ML.

NOTE 1 Annex B shows a framework for providing common vocabularies and relationships between the components of data quality measures.

NOTE 2 Annex C and Annex D show how quality measures are grouped from different perspectives.

In the context of analytics and ML, the overall quality of a training dataset, a validation dataset or a test dataset can be just as important as the quality of the individual data values in the dataset. Even though every data value in a dataset is accurate, a dataset that does not correctly reflect the underlying distribution of data can cause an incorrect analysis result or the creation of an ML model that does not meet requirements. The organization should document the target data for each data quality measure.

NOTE 3 Characteristics for statistical measures (e.g. accessibility by authorized users, accuracy, consistency, currentness, understandability, relevance, timeliness)[7] as defined by institutions such as the United Nations Statistics Division (UNSD) and European Statistics (EUROSTAT) can also be used to assess whether the quality of a dataset meets requirements.

The data quality measures and measurement functions in this document should be used when appropriate. Refer to Annex A in cases where the user of this document needs to create a new, bespoke data quality measure and data quality measurement function. Any quality measure, when using modified or newly defined, shall select data quality characteristics defined in this document and shall provide the rationale for changes in accordance with ISO/IEC 25024:2015, Clause 2.

6.2 Inherent data quality characteristics

6.2.1 Accuracy

6.2.1.1 General

Accuracy of a dataset is the degree to which data items in the data set have the correct data values or correct data labels. ISO/IEC 25012 describes accuracy as the degree to which data values have attributes that correctly represent the true value of the intended attributes. ISO/IEC 25012 further describes accuracy in terms of:

— syntactic accuracy which considers the closeness of the data values to a set of syntactically correct data values in a relevant domain;

— semantic accuracy which considers the closeness of the data values to a set of semantically correct data values in a relevant domain.

A data item is syntactically correct if its data value is the same type as its explicit data type and semantically correct if its data value has an expected value corresponding to the ML task. ML models are mathematical constructs, which means that low syntactic or semantic accuracy of the data values in training, validation, testing or production datasets can cause the model itself to be incorrect or the inferences made by the model to be incorrect.

For a supervised learning classification system, the correctness of the label sequence contents can affect the inference accuracy of a trained model. Factors that should be considered for measuring the accuracy of labelling include:

— correctness of label values;

— correctness of labelled tags;

— correctness of label sequence contents.

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.

6.2.1.2 QMs for accuracy

Table 1 provides data quality measures for accuracy in a specific context of use of analytics and ML.

Table 1 — Accuracy measures

ID	Name	Description	Measurement function
Acc-ML-1	Syntactic data accuracy	See ISO/IEC 25024:2015, Table 1	See ISO/IEC 25024:2015, Table 1
Acc-ML-2	Semantic data accuracy	See ISO/IEC 25024:2015, Table 1	See ISO/IEC 25024:2015, Table 1
Acc-ML-3	Data accuracy assurance	See ISO/IEC 25024:2015, Table 1	See ISO/IEC 25024:2015, Table 1
Acc-ML-4	Risk of dataset inaccuracy	See ISO/IEC 25024:2015, Table 1	See ISO/IEC 25024:2015, Table 1
Acc-ML-5	Data model accuracy	See ISO/IEC 25024:2015, Table 1	See ISO/IEC 25024:2015, Table 1
Acc-ML-6	Data accuracy range	See ISO/IEC 25024:2015, Table 1	See ISO/IEC 25024:2015, Table 1
Acc-ML-7	Data label accuracy	Does data label correctly assign to each element in the dataset?	$\frac{A}{B}$ where A is the number of data labels that provide the appropriate required information; B is the number of data labels defined in the dataset.

6.2.2 Completeness

6.2.2.1 General

ISO/IEC 25012 describes completeness in terms of data having values for all expected attributes and entity instances. In some cases, ML algorithms can fail when they encounter one or more empty data items in training, validation or testing datasets. Additionally, trained ML models can also fail when production data contains null data values.

Measures for completeness can help ML practitioners meet their data requirements and can indicate whether additional imputation steps should be taken as described in ISO/IEC 5259-4.[8]

The completeness characteristic of the labelled data in a dataset is relative. In different scenarios, the meaning of completeness can be different and should be considered with a specific usage scope. Factors that should be considered for measuring the completeness of a dataset include:

— The completeness of a dataset being used for an ML-based image classification should check the unlabelled samples in a dataset, which cannot be directly used in supervised ML.

— The completeness of a dataset being used for an ML-based object detection should check the incompleteness of labelled bounding boxes on objects.

In particular, it is common in real life that a sample has multiple objects in various categories since it is difficult to capture a scene with a single isolated object taking the entire view space. In this case, to measure the completeness of the dataset for an ML-based image recognition, the following factors should be considered:

— there exists any target object in a sample;

— all target objects are categorized;

— all target objects detected are labelled with bounding boxes or other methods.

EXAMPLE 1

A completeness measure for a dataset indicates that the dataset is missing more than half of the data values for the zip code feature. The data scientist decides the zip code feature is not a necessary predictor for their classification task and elects to remove the zip code feature from the training, validation, testing and production datasets.

EXAMPLE 2

A completeness measure for a dataset being used for an ML regression task indicates that one percent of the data values for a feature that is a good predictor are empty. The rest of the data has a normal distribution. The data scientist chooses to fill the null data values with the statistical mean of the available data values.

EXAMPLE 3

A completeness measure for a dataset being used for an ML clustering task indicates that a small number of records have one or more empty data items. The data scientist chooses to delete those records from the training data.

EXAMPLE 4

A completeness measure for value occurrences in a dataset for an ML classification task is the ratio of missing data values to the target number of data items expected for the proper fidelity of the dataset.

6.2.2.2 QMs for completeness

Table 2 provides data quality measures for completeness in a specific context of use of analytics and ML.

Table 2 — Completeness measures

ID	Name	Description	Measurement function
Com-ML-1	Value completeness	Ratio of data items with no null data values to the total number of data items in the dataset where at least one data item exists	$\frac{A}{B}$ where A is the number of data items whose value is not null; B is the total number of data items in the dataset where at least one data item exists.
Com-ML-2	Value occurrence completeness	Ratio of the number of occurrences of a given data value to the expected number of data value occurrences, described in the data quality requirement, in data items with the same domain in a dataset	$\frac{A}{B}$ where A is the number of occurrences of the data value in the data items; B is the expected number of occurrences of that data value in data items with the same domain in the dataset.
Com-ML-3	Feature completeness	Ratio of data items, associated with a feature, with no null data values to the total number of data items associated with the feature	$\frac{A}{B}$ where A is the number of data items, associated with the given feature, with no null data values; B is the total number of data items, associated with the given feature in the dataset where at least one data item exists.
Com-ML-4	Record completeness	Ratio of non-empty data records to the total number of data records in a dataset where at least one data record exists	$\frac{A}{B}$ where A is the number of non-empty data records in the dataset; B is the total number of data records in the dataset where at least one data record exists.
Com-ML-5	Label completeness	Ratio of unlabelled or incompletely labelled samples in a dataset	$1-\frac{A}{B}$ where A is the number of unlabelled or incompletely labelled samples; B is the number of all samples in the dataset.

6.2.3 Consistency

6.2.3.1 General

ISO/IEC 25012 describes consistency in terms of the data being coherent with other data and free of contradictions. Consistency is a key aspect of data used for ML as the features used in training data should together provide a model that permits correct inferences on production data. Additionally, ML can be literal in its interpretation of data values. Duplicate records can cause over-weighting of certain features. Contradictions between features in training data can cause a trained model to perform below requirements.

The data quality of a training data depends on the consistency of the labels assigned to similar data items. To improve the performance of machine learning models, data labels is needed to be assigned consistently to avoid inconsistencies.

EXAMPLE

A web form is used to collect voter preferences for political candidates. An organized group of people floods the site with entries for their favourite candidate. When used to train an ML model, these duplicate data can cause the model to over-weight a particular candidate when making inferences for people who have characteristics similar to those who flooded the web form.

6.2.3.2 QMs for consistency

Table 3 provides data quality measures for consistency in a specific context of use of analytics and ML.

Table 3 — Consistency measures

ID	Name	Description	Measurement function
Con-ML-1	Data record consistency	The ratio of duplicate records in the dataset	$1-\frac{A}{B}$ where A is the number of duplicate records in the dataset; B is the total number of records in the dataset.
Con-ML-2	Data label consistency	Consistency of data labels of similar data items	$\frac{A}{B}$ where A is the number of pairs of similar items that have been given the same label; B is the total number of comparisons made between labels of similar items.
Con-ML-3	Data format consistency	See ISO/IEC 25024:2015, Table 3	See ISO/IEC 25024:2015, Table 3
Con-ML-4	Semantic consistency	See ISO/IEC 25024:2015, Table 3	See ISO/IEC 25024:2015, Table 3

6.2.4 Credibility

6.2.4.1 General

ISO/IEC 25012 defines credibility in terms of the degree to which data has attributes that are regarded as believable by users in a specific context of use. Credibility is applicable for single data items, for related data items in a data record and for the entire dataset. The context in which the data are used can impact its perceived trueness and believability. Data can be perturbed during processing (e.g. transit, storage, computation) by authorized and unauthorized parties. An emerging concern for ML is unauthorized parties perturbing training, validation, testing and production data to deliberately render trained models as unusable or to manipulate the inferences made by a trained model.

Processes used in data preparation can change the data without changing its meaning (e.g. normalization, splitting or combining features). In these cases the data maintain its credibility.

EXAMPLE 1

A dataset is used to train, validate and test an ML model which then fails to achieve the required performance on production data. A security audit indicates that an unauthorized party has randomly changed data values in the training dataset.

EXAMPLE 2

A training dataset contains numerical features which have widely varying ranges. A data scientist elects to normalize the data values for these features to make them more comparable. Although the data values can be changed after being normalized, they are still credible as the meaning has not changed within the context of ML.

EXAMPLE 3

The credibility of the preparation of certain types of datasets can be improved in some circumstances using a randomness statistical method to compose the sample cases.

EXAMPLE 4

The credibility of the preparation of a dataset can be improved by declaring the data provenance of the data in the data life cycle framework (see ISO/IEC 8183:2023, 6.6).[9]

6.2.4.2 QMs for credibility

Table 4 provides data quality measures for credibility in a specific context of use of analytics and ML.

Table 4 — Credibility measures

ID	Name	Description	Measurement function
Cre-ML-1	Values credibility	See ISO/IEC 25024:2015, Table 4	See ISO/IEC 25024:2015, Table 4
Cre-ML-2	Source credibility	See ISO/IEC 25024:2015, Table 4	See ISO/IEC 25024:2015, Table 4
Cre-ML-3	Data dictionary credibility	See ISO/IEC 25024:2015, Table 4	See ISO/IEC 25024:2015, Table 4
Cre-ML-4	Data model credibility	See ISO/IEC 25024:2015, Table 4	See ISO/IEC 25024:2015, Table 4

6.2.5 Currentness

6.2.5.1 General

ISO/IEC 25012 describes data currentness in terms of being the right age relative to the use of the data. For ML, currentness can be in terms of an age range that is appropriate for the ML task. For example, data about people can be incomplete for under-represented populations prior to shifts in regulations and social norms. ML models based on economic data collected over several decades can be incorrect if the data are not corrected for inflation, exchange rates and other factors that vary with time. The issues arising due to the variations in the production data, compared to the data used to train and test the model, are generally known as data-drift issues and can be addressed by ensuring the currentness of the data.

Dataset currentness can be described in terms of what is the overall time period that the dataset covers (e.g. images or sentences collected from 2010 until 2021), time period between the last date of a data item and current date (e.g. 8 months) and what is the update cycle (e.g. every 6 months). Currentness should be considered a composite metric based on these three aspects.

EXAMPLE 1

An ML model used to predict future sales consistently misses the actual sales amount. Upon investigation, it is determined that the training data from 10 years of sales transactions was used to create the model but the data values were not adjusted for inflation.

EXAMPLE 2

An ML model is used to classify which candidate a particular person can vote for. The model was trained on voting records from the prior 20 years. Certain under-represented groups didn't regularly vote until the last 10 years. The ML model consistently fails to correctly predict the choices made by under-represented voters.

6.2.5.2 QMs for currentness

Table 5 provides data quality measures for currentness in a specific context of use of analytics and ML.

Table 5 — Currentness measures

ID	Name	Description	Measurement function
Cur-ML-1	Feature currentness	Ratio of data items for a feature in the dataset that falls within the acceptable range of dates as specified in the organization's data quality requirements	$\frac{A}{B}$ where A is the number of data items for a feature that falls within the required age range; B is the total number of data items for the feature.
Cur-ML-2	Record currentness	Ratio of data records in the dataset where all data items in the record fall within the required age range	$\frac{A}{B}$ where A is the number of data records that fall within the required age range; B is the total number of data records in the dataset.

6.3 Inherent and system-dependent data quality characteristics

6.3.1 Accessibility

6.3.1.1 General

ISO/IEC 25024 describes that accessibility refers to the degree to which data can be accessed in a specific context of use, particularly by people who need assistive technology or special configuration because of some disability. In addition, seamless access to datasets and easy deployment of datasets through appropriate tools should be provided for analytics and ML.

6.3.1.2 QMs for accessibility

Table 6 provides data quality measures for accessibility in a specific context of use of analytics and ML.

Table 6 — Accessibility measures

ID	Name	Description	Measurement function
Acs-ML-1	User accessibility	See ISO/IEC 25024:2015, Table 6.1	See ISO/IEC 25024:2015, Table 6.1
Acs-ML-2	Data format accessibility	See ISO/IEC 25024:2015, Table 6.2	See ISO/IEC 25024:2015, Table 6.2
Acs-ML-3	Data accessibility	Ratio of accessible records in the dataset	$\frac{A}{B}$ where A is the number of accessible records in the dataset; B is the total number of data records in the dataset.

6.3.2 Compliance

6.3.2.1 General

ISO/IEC 25012 describes compliance in terms of the data meeting regulations, standards, conventions or other rules. For example, personal data used for analytics or ML can be subject to legal and regulatory requirements. Likewise, data users can have their own compliance requirements and certification schemes can have compliance requirements.

6.3.2.2 QMs for compliance

Table 7 provides data quality measures for compliance in the context of use of analytics and ML.

Table 7 — Compliance measures

ID	Name	Description	Measurement function
Cmp-ML-1	Data item compliance	Degree to which data items meet compliance requirements	$\frac{A}{B}$ where A is the number of data items that meet compliance requirements; B is the total number of data items in the dataset.

6.3.3 Efficiency

6.3.3.1 General

Efficiency refers to the degree to which data have attributes that can be processed and provide the expected performance levels by using the appropriate amounts and types of resources in a specific context of use. For example, the efficiency of the data format is important when sharing training datasets, especially when the data size is large (e.g. CSV, HDF, JSON). Also, an optimal size of the dataset in memory can reduce the training cost.

6.3.3.2 QMs for efficiency

Table 8 provides data quality measures for efficiency in a specific context of use of analytics and ML.

Table 8 — Efficiency measures

ID	Name	Description	Measurement function
Eff-ML-1	Data format efficiency	See ISO/IEC 25024:2015, Table 9.2	See ISO/IEC 25024:2015, Table 9.2
Eff-ML-2	Data processing efficiency	See ISO/IEC 25024:2015, Table 9.2	See ISO/IEC 25024:2015, Table 9.2
Eff-ML-3	Risk of wasted space	See ISO/IEC 25024:2015, Table 9.2	See ISO/IEC 25024:2015, Table 9.2

6.3.4 Precision

6.3.4.1 General

ISO/IEC 25012 describes precision as the data being exact or providing discrimination. ISO/IEC 25024 provides an example of precision in terms of the number of decimal places for a real number. In the context of ML precision, as expressed by the number of decimal places for the value of a data item, can affect the weight of a given feature in a trained ML model. For example, a feature with many data item values of 99.4 can achieve a greater weight than another feature where the same value has been rounded down to 99. Likewise, a feature whose values were rounded up can have a greater weight than a feature with more precision. When specifying data requirements for precision, data users should consider the overall effect on the trained ML model.

6.3.4.2 QMs for precision

Table 9 provides data quality measures for precision in a specific context of use of analytics and ML.

Table 9 — Precision measures

ID	Name	Description	Measurement function
Pre-ML-1	Precision of data values	See ISO/IEC 25024:2015, Table 10.1	See ISO/IEC 25024:2015, Table 10.1

6.3.5 Traceability

6.3.5.1 General

Traceability refers to the degree to which data has attributes that provide an audit trail of access to the data and any changes made to the data in a specific context of use.

6.3.5.2 QMs for traceability

Table 10 provides data quality measures for traceability in a specific context of use of analytics and ML.

Table 10 — Traceability measures

ID	Name	Description	Measurement function
Tra-ML-1	Traceability of data values	See ISO/IEC 25024:2015, Table 11.1	See ISO/IEC 25024:2015, Table 11.1
Tra-ML-2	User access traceability	See ISO/IEC 25024:2015, Table 11.2	See ISO/IEC 25024:2015, Table 11.2
Tra-ML-3	Data values traceability	See ISO/IEC 25024:2015, Table 11.2	See ISO/IEC 25024:2015, Table 11.2

6.3.6 Understandability

6.3.6.1 General

ISO/IEC 25012 describes understandability in terms of users being able to read and interpret the data. Additionally, understandability includes the use of appropriate symbols, units and languages. ML models can fail to perform to requirements if units for features are used inappropriately. Understandability is an important characteristic for supporting the explainability of the AI system and helping the stakeholders interact with the AI system's production data (inference label or weight data of the ML).

For natural language processing tasks, inappropriate use of human languages and symbols can cause tasks such as language understanding and generation to fail.

While the data quality measures described in this document are quantitative, the humans using data for ML also make qualitative assessments about the data. Appropriate use of symbols, units and languages can assist in making qualitative judgments.

6.3.6.2 QMs for understandability

Table 11 provides data quality measures for understandability in a specific context of use of analytics and ML.

Table 11 — Understandability measures

ID	Name	Description	Measurement function
Und-ML-1	Symbols understandability	See ISO/IEC 25024:2015, Table 12.1	See ISO/IEC 25024:2015, Table 12.1
Und-ML-2	Semantic understandability	See ISO/IEC 25024:2015, Table 12.1	See ISO/IEC 25024:2015, Table 12.1
Und-ML-3	Data values understandability	See ISO/IEC 25024:2015, Table 12.1	See ISO/IEC 25024:2015, Table 12.1
Und-ML-4	Data representation understandability	See ISO/IEC 25024:2015, Table 12.2	See ISO/IEC 25024:2015, Table 12.2

6.4 System-dependent data quality characteristics

6.4.1 Availability

6.4.1.1 General

Availability refers to the degree to which datasets can be retrieved by authorized users or applications for a specific task of analytics or ML in the data acquisition stage of the data life cycle.

6.4.1.2 QMs for availability

Table 12 provides data quality measures for availability in the context of use of analytics and ML.

Table 12 — Availability measures

ID	Name	Description	Measurement function
Ava-ML-1	Data availability ratio	See ISO/IEC 25024:2015, Table 13	See ISO/IEC 25024:2015, Table 13

6.4.2 Portability

6.4.2.1 General

ISO/IEC 25012 describes the portability data quality characteristic in terms of the ability to move data from one system to another, within a specified context, while preserving its quality.

Data used in analytics and ML can undergo processing on multiple systems. For example, data can be acquired for ML on one system, data quality processes can be performed on the data using a second system and the data can then be transferred to a third system to train an ML model.

If the quality of the data (i.e. that it meets requirements) is not maintained when the data are transferred from one system to another then the trained ML model itself can fail to meet requirements.

NOTE Requirements for data portability are established as part of the data requirements portion of the data life cycle and are system and environment dependent.

6.4.2.2 QMs for portability

Table 13 provides data quality measures for portability in the context of use of analytics and ML.

Table 13 — Portability measures

ID	Name	Description	Measurement function
Por-ML-1	Data portability ratio	See ISO/IEC 25024:2015, Table 14	See ISO/IEC 25024:2015, Table 14
Por-ML-2	Prospective data portability	See ISO/IEC 25024:2015, Table 14	See ISO/IEC 25024:2015, Table 14

6.4.3 Recoverability

6.4.3.1 General

Recoverability refers to the degree to which datasets can be maintained and preserved at a specific level of operations and quality, even in the event of failure, for a specific task of analysis and ML in data preparation and provisioning stages of the data life cycle, especially with a large volume of datasets.

6.4.3.2 QMs for recoverability

Table 14 provides data quality measures for recoverability in the context of use of analytics and ML.

Table 14 — Recoverability measures

ID	Name	Description	Measurement function
Rec-ML-1	Data recoverability ratio	See ISO/IEC 25024:2015, Table 15	See ISO/IEC 25024:2015, Table 15
Rec-ML-2	Feature recoverability ratio	Degree to which features of a dataset transferred in stages are recoverable	$\frac{A}{B}$ where A is the number of features of the dataset successfully recovered; B is the features of the dataset can be managed by backup and restore procedures.

6.5 Additional data quality characteristics

6.5.1 Auditability

6.5.1.1 General

For the purposes of this document, auditability refers to the characteristic of a dataset that all or part of the dataset has undergone an audit or that the data are available to relevant stakeholders for the purposes of conducting audits. Auditing of datasets used for analytics and ML can contribute to the credibility of the data and can be required for compliance with requirements.

EXAMPLE

A dataset of images is used for image recognition and is labelled by a third-party contractor. To ensure that the images are labelled properly, the organization uses another third party to audit a subset of the labelled images.

6.5.1.2 QMs for auditability

Table 15 provides data quality measures for auditability in a specific context of use of analytics and ML.

Table 15 — Auditability measures

ID	Name	Description	Measurement function
Aud-ML-1	Audited records	Ratio of the records in the dataset that were audited	$\frac{A}{B}$ where A is the number of records in the dataset that were audited; B is the total number of records in the dataset.
Aud-ML-2	Auditable records	Ratio of the records in the dataset that are available for audit.	$\frac{A}{B}$ where A is the number of records in the dataset available for audit; B is the total number of records in the dataset.

6.5.2 Balance

6.5.2.1 General

For a dataset, the balance refers to the distribution of the samples for all features of the dataset. For example, if the dataset represents X number of categories of data items, the number of samples per category should be evenly distributed for the dataset to be balanced. For an image dataset, such features can include labels meaningful to business logic, resolution, brightness, the width-to-height ratio of labelled bounding boxes, the size of labelled bounding boxes and any others that potentially influence the ML model performance.

The balance of a dataset can affect a part of the overall performance of an ML model. For an ML-based computer vision system, the balance of the dataset should be considered.

EXAMPLE 1

When considerable differences of the brightness or the resolution exist between the samples of a training dataset and the real-world data, ML models can fail due to noisy data introduced by faintness or vagueness.

EXAMPLE 2

In an ML-based classification system, the presence of an imbalanced category of sample population can result in the failure of rare instance discovery and classification. Such instances can even be wrongly categorized or identified as noisy data.

EXAMPLE 3

In an ML-based object detection system, significant differences in the width-to-height ratios or the size of bounding boxes can lead to inconsistency in the size of detected objects, given a fixed size of the receptive field. Consequently, this can also cause a loss of generalizability if extra multi-size object checks or adjustment approaches are not applied.

Although the examples presented here relate to images, the concept of balance can also be applied to other types of data.

6.5.2.2 QMs for balance

Table 16 provides data quality measures for balance in a specific context of use of analytics and ML.

Table 16 — Balance measures

ID	Name	Description	Measurement function
Bal-ML-1	Brightness balance	Reciprocal of the maximal ratio of the brightness difference of an image sample over the averaged brightness of samples in a dataset	$\frac{A}{B}$ where A is the average value of brightness of the samples; B is the maximum value of absolute differences between the brightness value of each image in the sample and A.
Bal-ML-2	Resolution balance	Reciprocal of the maximal ratio of the resolution difference of an image sample over the averaged resolution of samples in a dataset	$\frac{A}{B}$ where A is the average value of resolution of the samples; B is the maximum value of absolute differences between the resolution value of each image in the sample and A.
Bal-ML-3	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	$\frac{A}{B}$ where A is the average category size of the dataset; B is the maximum value of absolute differences between the size of each category in the dataset and A.
Bal-ML-4	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	$\frac{A}{B}$ where A is the averaged bounding box with height to width ratio over all the samples in the dataset; B is the maximum value of absolute differences between bounding box with height to width ratio of each sample in the dataset and A.
Bal-ML-5	Category 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	$\frac{A}{B}$ where A is the averaged bounding box area over all the samples in the dataset; B is the maximum value of absolute differences between averaged bounding box area of each category in the dataset and A.

Table 16 *(continued)*

ID	Name	Description	Measurement function
Bal-ML-6	Sample 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	$\frac{A}{B}$ where A is the averaged bounding box area over all the samples in the dataset; B is the maximum value of absolute differences between averaged bounding box area of each sample in the dataset and A.
Bal-ML-7	Label proportion balance	Difference in proportion of data items from different two data categories having a certain label value	$A - B$ where A is the proportion of data items in category C_A having label value L in the dataset, i.e. $n_A^{(L)} / n_A$; B is the proportion of data items in category C_B having label value L in the dataset, i.e. $n_B^{(L)} / n_B$; n_A is the number of data items belonging to category C_A; n_B is the number of data items belonging to category C_B (not C_A); $n_A^{(L)}$ is the number of data items having label value L in category C_A; $n_B^{(L)}$ is the number of data items having label value L in category C_B.
Bal-ML-8	Label distribution balance	Divergence between the label distribution and the uniform label distribution	$f(A,B)$ where A is the label distribution of data items with distinct values of labels in the dataset under assessment, i.e. $[n_{L1}/n, n_{L2}/n, ..., n_{LN}/n]$; B is the uniform label distribution of data items, i.e. $[n/N, n/N, ..., n/N]$; f is a function measuring divergence between two distributions such as Kullback-Leibler Divergence, Jensen-Shannon Divergence, L_p-norm, Total Variation Distance, and Kolmogorov-Smirnov test statistic; N and n is the number of distinct label values and the total number of data items, in the dataset, respectively; n_{Li} is the number of data items having the i-th label value of $\{L_1, L_2, ..., L_N\}$ in the dataset.

6.5.3 Diversity

6.5.3.1 General

Diversity of a dataset refers to the difference between samples in terms of the target data. In a dataset used for an ML model, an adequate difference between samples is important. If all or most data records in a dataset are alike, an ML model trained from that dataset can have the risk of overfitting and consequently being less generalizable. The diversity of a dataset represents the degree to which the dataset contains various ranges of different features, values, labels, clusters or sources among individual data. Data enhancement by generative ML models can improve the data diversity, but these approaches can fail if the diversity of the original dataset is limited. Diversity is closely related to representativeness and balance. It is a data quality characteristic that can be used to evaluate a dataset's fidelity.

Measurement of diversity can be done in the context of the specific target data, as determined by the ML task requirements.

6.5.3.2 QMs for diversity

Table 17 provides data quality measures for diversity in a specific context of use of analytics and ML.

Table 17 — Diversity measures

ID	Name	Description	Measurement function
Div-ML-1	Label richness	Ratio of distinct labels in a dataset.	$\frac{A}{B}$ where A is the number of distinct labels in the dataset; B is the number of data items in the dataset.
Div-ML-2	Relative label abundance	Portion of the number of individual data (i.e. data item, data record, data frame) having the same label in a dataset	$\frac{A}{B}$ where A is the number of individual data in which have target labels; B is the number of individual data in the dataset.
Div-ML-3	Category size diversity	Ratio of categories where the number of categorized data items is lower than a threshold defined by quality requirements.	$\frac{A}{B}$ where A is the number of categories where the number of categorized data items is lower than the threshold of the quality requirement; B is the number of categories in total.

6.5.4 Effectiveness

6.5.4.1 General

Effectiveness of a dataset indicates whether the dataset meets requirements for use in a specific ML task.

EXAMPLE 1

For an ML-based computer vision system, the dataset effectiveness can be the lowest acceptable ratio at which the number of images with brightness or resolution lower than a required threshold divided by the total number of images or videos in the dataset.

EXAMPLE 2

For an ML-based image classification system, the dataset effectiveness can refer to the lowest acceptable ratio of the number of images in a category divided by the total number of images in the dataset.

EXAMPLE 3

For an ML-based object detection system, the dataset effectiveness can refer to the lowest acceptable ratio of the number of images whose area values in the bounding boxes are below a required threshold divided by the total number of images or videos in the dataset.

6.5.4.2 QMs for effectiveness

Table 18 provides data quality measures for effectiveness in a specific context of use of analytics and ML.

Table 18 — Effectiveness measures

ID	Name	Description	Measurement function
Eft-ML-1	Feature effectiveness	Ratio of samples with acceptable feature in a dataset	$\frac{A}{B}$ where A is the number of the samples with acceptable feature; B is the number of all samples in the dataset.
Eft-ML-2	Category size effectiveness	Ratio of categories where the number of categorized samples is lower than a threshold	$\frac{A}{B}$ where A is the number of categories where the number of categorized samples is lower than a threshold; B is the number of categories in total.
Eft-ML-3	Label effectiveness	Ratio of samples with acceptable label in a dataset	$\frac{A}{B}$ where A is the number of the samples with acceptable label; B is the number of all samples in the data.

6.5.5 Identifiability

6.5.5.1 General

ISO/IEC 29100[10] describes identifiability as the capability to identify a personally identifiable information (PII) principal directly or indirectly on the basis of a given set of PII. It is important to understand whether any PII in a dataset can be used to identify a PII principal as legal requirements in some jurisdictions can restrict such activity. De-identification processes can be applied to training, validation, testing and production data to reduce the possibility of identifiability.

EXAMPLE

An ML model is trained on search engine queries for the purpose of targeted advertising. The dataset includes the user's IP address which is considered to be PII in some jurisdictions. Anonymization is applied to the dataset to remove the IP address before the dataset is split into training, validation and testing datasets. Where appropriate the anonymization process can be applied to production data passed to the model.

6.5.5.2 QMs for identifiability

Table 19 provides data quality measures for identifiability in a specific context of use of analytics and ML.

Table 19 — Identifiability measures

ID	Name	Description	Measurement function
Idn-ML-1	Identifiability ratio	Ratio of data records in the dataset that can be used for identifiability	$\frac{A}{B}$ where A is the number of data records that contain data items that can be used for identifiability, either on their own or in conjunction with other data items; B is the number of data records in the dataset.

6.5.6 Relevance

6.5.6.1 General

For the purposes of this document, relevance refers to the degree to which a dataset (assuming that it is accurate, complete, consistent, current, etc.) is suitable for a given context.

For ML, relevance can mean that the selected features in training data and their data values are good predictors for the target variable.

EXAMPLE

An ML model is used to determine the creditworthiness of people. The training data are representative of the sample of the population expected to appear in the production data. The training data include relevant features such as prior credit history, income, job tenure and net worth which are good predictors of creditworthiness. The training data also include the height and weight of each person. Statistical tests show no correlation of height and weight to prior credit history and are deemed to be poor predictors of future credit performance. To improve the overall relevance of the dataset, the height and weight features are dropped.

6.5.6.2 QMs for relevance

Table 20 provides data quality measures for relevance in a specific context of use of analytics and ML.

Table 20 — Relevance measures

ID	Name	Description	Measurement function
Rel-ML-1	Feature relevance	Ratio of features in the dataset that are relevant to the given context	$\frac{A}{B}$ where A is the number of features in the dataset deemed to be relevant in the context of the use of the data; B is the total number of features in the dataset.
Rel-ML-2	Record relevance	Ratio of records in the dataset that are relevant to the given context	$\frac{A}{B}$ where A is the number of record in the dataset deemed to relevant in the context of the use of the data; B is the total number of records in the dataset.

6.5.7 Representativeness

6.5.7.1 General

ISO 20252[11] defines representativeness as the degree to which a dataset reflects the target population being studied. For supervised ML, a training dataset can be considered as a subset of a larger population and the production data as a target population to which inferences can be made. When the training data do not sufficiently represent the production data, the trained ML model can fail to perform as required. ISO/IEC TR 24027[12] describes data biases derived from the data selection process or data labelling process.

The representativeness data quality characteristic is related to the relevance data quality characteristic in that a dataset that does not represent the target population under study is unlikely to provide good predictors for the target variable.

EXAMPLE 1

A facial recognition system trained only on images of humans with light skin tones can fail to correctly identify individuals when applied to images of humans with dark skin tones.

EXAMPLE 2

A predictive maintenance system trained only on data from electric motors can fail to correctly predict needed maintenance when applied to internal combustion engines.

6.5.7.2 QMs for representativeness

Table 21 provides data quality measures for representativeness in a specific context of use of analytics and ML.

Table 21 — Representativeness measure

ID	Name	Description	Measurement function
Rep-ML-1	Representa-tive-ness ratio	Ratio of relevant attributes found in the subjects of a target population to the attributes found in the dataset	$\frac{A}{B}$ where A is the number of target attributes in the dataset; B is the number of relevant attributes in a specific context.

6.5.8 Similarity

6.5.8.1 General

The similarity of a dataset refers to the similarity between samples in terms of interesting features. This is relevant for classification tasks (see ISO/IEC 23053:2022, 6.2.3) which are typically implemented using supervised learning (see ISO/IEC 23053:2022, 7.2). This is also relevant for clustering tasks (see ISO/IEC 23053:2022, 6.2.4) which are typically implemented using unsupervised learning (see ISO/IEC 23053:2022, 7.3). Both classification and clustering tasks require an adequate level of difference among samples to perform successfully (see ISO/IEC TR 24027:2021, 5.2)

An ML model trained on a dataset containing quite similar images (e.g. that are generated by a slight shift of pixels based on a few seed images) can have the risk of overfitting and consequently less generalizability. In this case, it is possible to consider the application of data manipulation approaches, such as rotation and shift that can improve the generalizability of the ML model. These approaches cannot work if the number of seed images is limited. In this case, the proportion of the similar samples should be checked. Another approach is to consider clustering algorithms with concept drift mitigation methods.[13]

Further measures identify data similarity through a geometrical approach: i.e. a dataset of N data records and M features can be represented such as N vectors in an M-dimensional space, so it can be analysed and compared using the tools of geometry; in particular, similarity can be associated to the mutual position of vectors in the space.

6.5.8.2 QMs for similarity

Table 22 provides data quality measures for similarity in a specific context of use of analytics and ML.

Table 22 — Similarity measures

ID	Name	Description	Measurement function
Sim-ML-1	Sample similarity	Ratio of similar samples in a dataset	$1-\frac{A}{B}$ where A is the number of all samples in the dataset; B is the number of the clusters, resulting from a clustering algorithm, on all samples of the dataset (NOTE 7).
Sim-ML-2	Samples tightness	Tightness of normalized dataset	$A-B$ where A is the max eigenvalue of G [a]; B is the min eigenvalue of G [a].
Sim-ML-3	Samples independency	Ratio of Principal Component Analysis (PCA) and dataset dimension	$1-\frac{A}{B}$ where A is the number of principal components with PCA method with 95 % coverage of eigenvalues sum (NOTE 3); B is the total number of dataset dimensions.

[a] G is a matrix with M rows and M columns and is equal to $\Phi_{norm}{}^{T}\,\Phi_{norm}$ (NOTE 1).

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 fit a hypersphere with a radius of one and centred in the origin ($M \leq N$).[14]

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} (NOTE 4), starting from the biggest, chosen in order to represent 95 % of their sum.[15]

NOTE 4 C_{MxM} is an M-by-M matrix, with M rows and M columns and is equal to $\Phi_{mean}{}^{T}\,\Phi_{mean}$ (NOTE 5).

NOTE 5 Φ_{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 centred in the origin.

NOTE 6 Principal components can be selected with criteria or percentage different. Annex A shows an example of measure modification.

NOTE 7 A measurement of zero means the least similarity. The similarity measure yields zero when the number of samples is equal to the number of clusters indicating that no sample is similar to another.

6.5.9 Timeliness

6.5.9.1 General

Timeliness refers to the latency (i.e. ΔT_1) between the time when a phenomenon occurs and the time when the data recorded for that phenomenon are available for use. Timeliness differs from currentness in that currentness is the ΔT_2 between the time a data sample is recorded and the time it is used. Timeliness can be a component of relevance in that if the ΔT_1 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. For example, ML tasks on streaming data (e.g. analysis of securities transactions, reinforcement learning, search queries) can make use of continuous learning and inferencing in near real-time.

6.5.9.2 QMs for Timeliness

Table 23 provides data quality measures for timeliness in a specific context of use of analytics and ML.

Table 23 — Timeliness measures

ID	Name	Description	Measurement function
Tml-ML-1	Timeliness of data items	The ratio of data items which meet timeliness requirements	$\frac{A}{B}$ where A is the number of data items in the dataset that meet timeliness requirements; B is the number of data items in the dataset.

7 Implementing a data quality model and data quality measures for an analytics or ML task

A general process for implementing a data quality model and related data quality measures for an analytics or ML task can include:

— Select data quality characteristics from this document that are appropriate for the analytics or ML task.

— For each data quality characteristic, select appropriate data quality measures from this document and rework them until they are useful for the application.

— If other, or additional, data quality measures are needed to assess the quality of the data, develop one or more new data quality measures and data quality measurement functions such as the example in Annex A.

— Revise the data quality requirements with acceptance criteria for each data quality measure (e.g. minimum or maximum threshold value, range of values).

— Apply measurement functions to the target data at appropriate stages in the data quality life cycle model for analytics and ML (see ISO/IEC 5259-1).

— Assess whether each data quality measurement result meets requirements.

— Assess whether the overall dataset meets requirements.

— If needed, apply data quality improvement processes (see ISO/IEC 5259-4).

— Continuously monitor and improve data quality and validate the data quality process over the life cycle of the analytics or ML task (e.g. apply data quality measurement functions whenever the data or the details of the task have changed).

NOTE ISO/IEC 5259-4 describes a data quality process framework.

8 Data quality reporting

8.1 Data quality reporting framework

Data quality reporting can provide documentation to appropriate stakeholders on data quality uses such as data quality models, data quality measures and their results, and whether the target data meet data quality requirements. Data and data quality can change over time. It can be necessary to revise data quality reports on an appropriate schedule according to the data quality risks for the analytics or ML task.

Data quality reports should include:

— the purpose of the report (e.g. to inform appropriate stakeholders, to facilitate decision-making, to provide evidence of compliance);

— the scope of the report (e.g. initial report or revision, time period covered, analytics or ML task covered);

— a schedule of revisions;

— criteria for concluding the reporting process;

— location and retention of the reports (e.g. for later reference or audit);

— the detailed items covered in 8.3.

8.2 Data quality measure information

Figure 4 shows a modification of the activity model described in ISO 8000-8:2015, Annex D for analytics and ML.

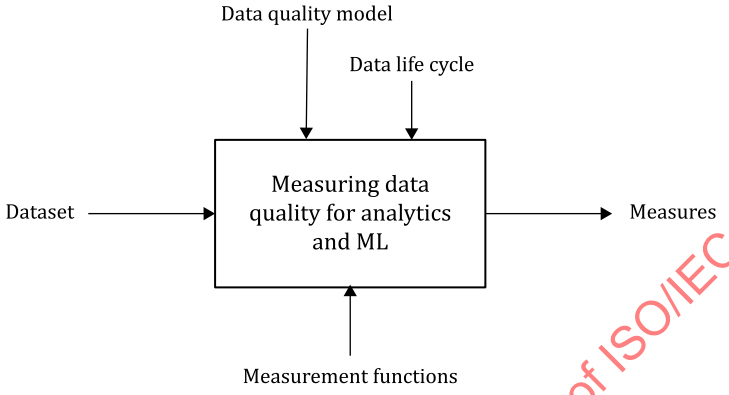

Figure 4 — Data quality measures information for quality reports

8.3 Guidance to organizations

When preparing data quality reports for an ML or analytics task, the organization should:

a) identify the persons responsible for preparing, reviewing and approving data quality reports;

b) identify the appropriate stakeholders that should receive copies of the data quality reports;

c) determine at which points in the data quality life cycle data quality reports should be initiated or revised;

d) determine the appropriate interval between revisions of the data quality report;

e) ensure that data quality reports are covered in data planning;

f) determine the scope of the target data covered by the data quality reports;

g) gather the data quality requirements;

h) document the data quality model including the data quality characteristics that constitute the data quality model;

i) document the selected data quality measures and their target values;

j) document any data quality measures developed according to Annex A;

k) document the results of all selected data quality measures;

l) document all transforms made to the target data;

m) document an assessment of whether or not the target data meet data quality requirements;

n) document a plan for improving the data quality, if the target data does not meet data quality requirements.

Annex A
(informative)

Design and document of a measurement function

Some of the measures listed in the ISO/IEC 25000 series and in this document are described from a general perspective. For a practical application, a detailed design of a measurement function and its contextual information can be needed and it is handled as described in accordance with ISO/IEC 25020.[16]

This example shows how to design and document a measurement function with the measure Acc-I-1 defined in ISO/IEC 25024 for syntactic accuracy conforming to this document.

For certain purposes (e.g. when a comparison is needed for assessment), Acc-I-1 from ISO/IEC 25024 is not sufficient and it is necessary to get insights for "values syntactically accurate" for *A* and "number of data items for which syntactic accuracy is required" for *B*, that is to choose the method for measuring string distance for *A*, and to determine a domain for *B*. There are several possibilities for designing both *A* and *B*, and among them are the following:

— *A* as the number of occurrences of the condition distance equals zero from a string against all strings in the domain of *B*;

— *B* as the number of admissible strings.

Using the measurement function with *A* and *B*, a revised measure Acc-I-1-IT-2 is shown in Table A.1.

Table A.1 — Syntactic accuracy Acc-I-1-IT-2

ID	Name	Description	Measurement function	Data life cycle(DLC) Target entities Properties
Acc-I-1-IT-2	Syntactic accuracy	Ratio of closeness of the data values to a set of values defined in a domain	$1-\frac{A}{B}$ where *A* is the number of data values for which the distance from the domain is null; *B* is the number of domain values.	All DLC except data design Data file Data item, data value
NOTE 1 The best similarity among the strings to be compared and their domain is when the distance is lower, so a lower value is better. NOTE 2 ID includes additional part "IT-2", see ISO/IEC 25020:2019, Annex C.				

Table A.2 shows the application of measure Acc-I-1-IT-2 for comparison of accuracy of two databases, each containing 3 names. The comparison is made against a syntax composed by names of length 4. Names of length 4 are a subset of all possible length 4 strings.

From the results of measurement function, the values in database *R* which is expressed by a lower value of the accuracy metric are more accurate than values in database *W*, as the name Marj doesn't belong to the syntax of names.