International Standard

ISO/IEC 5259-4

First edition
2024-07

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

Part 4: Data quality process framework

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

Partie 4: Cadre pour le processus de qualité des données

Reference number
ISO/IEC 5259-4:2024(en)

© ISO/IEC 2024



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2024

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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 ..... **v**
Introduction ..... **vi**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Symbols and abbreviated terms **3**
5 Data quality process principles **3**
6 Data quality process framework **3**
6.1 General 3
6.2 Data quality planning 5
6.3 Data quality evaluation 6
6.4 Data quality improvement 6
6.5 Data quality process validation 6
6.6 Using the DQPF 7
7 Data quality process for ML **7**
7.1 General 7
7.2 Data requirements 8
7.3 Data planning 9
7.4 Data acquisition 9
7.5 Data preparation 10
7.5.1 General 10
7.5.2 Supervised ML 10
7.5.3 Unsupervised ML 10
7.5.4 Semi-supervised ML 10
7.5.5 Dataset composition 11
7.5.6 Data labelling 11
7.5.7 Data annotation 11
7.5.8 Data quality assessment 12
7.5.9 Data quality improvement 13
7.5.10 Data de-identification 15
7.5.11 Data encoding. 16
7.6 Data provisioning 16
7.6.1 General 16
7.6.2 Supervised ML 16
7.6.3 Unsupervised ML 16
7.6.4 Semi-supervised ML 16
7.7 Data decommissioning 16
8 Data labelling methods and process **17**
8.1 General 17
8.2 Data labelling principles 17
8.3 Data labelling methods 17
8.4 Data labelling process 18
8.4.1 General 18
8.4.2 Labelling specifications 18
8.4.3 Labelling participant roles 18
8.4.4 Labelling tools or platforms 19
8.4.5 Labelling task establishment 19
8.4.6 Labelling task assignment 19
8.4.7 Labelling process control 20
8.4.8 Labelling result quality checking 20
8.4.9 Labelling result revision 20

© ISO/IEC 2024 – All rights reserved

9 Roles of participants **21**
9.1 General 21
9.2 Data planner 21
9.3 Data originator 21
9.4 Data collector 21
9.5 Data engineer 21
9.6 Data holder 21
9.7 Data user 21

10 Data quality process for semi-supervised ML **22**
10.1 General 22
10.2 Data requirements 22
10.3 Data planning 22
10.4 Data acquisition 22
10.5 Data preparation 22
10.6 Data provisioning 22
10.7 Data decommissioning 23

11 Data quality process for reinforcement learning **23**
11.1 General 23
11.2 Data requirements 23
11.3 Data planning 23
11.4 Data acquisition 23
11.5 Data preparation 23
11.5.1 General process 23
11.5.2 Data recording 24
11.6 Data provisioning 24
11.7 Data decommissioning 24

12 Data quality process for analytics **24**
12.1 General 24
12.2 Data requirements 24
12.3 Data planning 24
12.4 Data acquisition 25
12.4.1 General 25
12.4.2 Data loading 25
12.4.3 Data storage 25
12.5 Data preparation 25
12.5.1 General 25
12.5.2 Data cleaning 25
12.5.3 Data transformation 25
12.5.4 Data aggregation 26
12.5.5 Data quality assessment 26
12.5.6 Data quality improvement 26
12.6 Data provisioning 27
12.7 Data decommissioning 27

Bibliography **28**

© ISO/IEC 2024 – All rights reserved

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.

© ISO/IEC 2024 – All rights reserved

Introduction

Artificial intelligence (AI)-related products, systems or solutions have developed quickly in recent years. One of the common characteristics of an AI system, especially for systems using supervised machine learning (ML), is whether the AI system can be trained on a dataset before deployment or trained dynamically as the system is used.

Data have been recognized as one of the most important aspects of ML-based AI systems. For all supervised, semi-supervised, unsupervised and reinforcement learning approaches, data quality can be a primary concern in creating and using data for training and evaluating ML systems. It has been shown that with more accurate and richer data, the results of analytics and ML can be more useful and reliable. In addition, for the development of supervised learning-based AI systems, a large number of task-specific labelled training data is needed. This makes accurately labelled data one of the most important resources in the AI industry. Nowadays, there is a verified market of industrial services and tools for training data labelling. This market is now reaching a level of maturity that justifies the development of International Standards for the benefit of providers and users of these services and tools to ensure high-quality labelled data.

This document describes the implementation of a standardized common procedure of data processing with regard to data quality for analytics and ML. Clause 5 describes principles about data quality process and Clause 6 describes a data quality process framework. Clause 7 describes the data quality process for ML approaches, Clause 8 describes data labelling methods and process, Clause 9 provides roles of participants in data quality processes, Clauses 10 and 11 then describe the additional considerations that apply to semi-supervised learning and reinforcement learning. Clause 12 describes how the data quality process framework applies to analytics.

This document provides the process framework on a detailed level which can be used to fulfil the requirements specified in ISO/IEC 5259-3. It also links processes that are mapped on the data life cycle model in ISO/IEC 5259-1.

© ISO/IEC 2024 – All rights reserved

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

Part 4:
Data quality process framework

1 Scope

This document establishes general common organizational approaches, regardless of the type, size or nature of the applying organization, to ensure data quality for training and evaluation in analytics and machine learning (ML). It includes guidance on the data quality process for:

— supervised ML with regard to the labelling of data used for training ML systems, including common organizational approaches for training data labelling;

— unsupervised ML;

— semi-supervised ML;

— reinforcement learning;

— analytics.

This document is applicable to training and evaluation data that come from different sources, including data acquisition and data composition, data preparation, data labelling, evaluation and data use. This document does not define specific services, platforms or tools.

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 leaning (ML) — Part 1: Overview, terminology and examples*

ISO/IEC 22989:2022, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053:2022, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 5259-1, ISO/IEC 22989 and ISO/IEC 23053 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/

© ISO/IEC 2024 – All rights reserved