

# ISO/IEC 5259-4:2024-07 (E)

## Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 4: Data quality process framework

---

### 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
<b>9</b>	<b>Roles of participants</b>	<b>21</b>
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
<b>10</b>	<b>Data quality process for semi-supervised ML</b>	<b>22</b>
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
<b>11</b>	<b>Data quality process for reinforcement learning</b>	<b>23</b>
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
<b>12</b>	<b>Data quality process for analytics</b>	<b>24</b>
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
	<b>Bibliography</b>	<b>28</b>