

# ISO/IEC TS 4213:2022-10 (E)

## Information technology - Artificial intelligence - Assessment of machine learning classification performance

---

<b>Contents</b>		<b>Page</b>
Foreword .....		v
Introduction .....		vi
<b>1</b>	<b>Scope .....</b>	<b>1</b>
<b>2</b>	<b>Normative references .....</b>	<b>1</b>
<b>3</b>	<b>Terms and definitions .....</b>	<b>1</b>
<b>3.1</b>	<b>Classification and related terms .....</b>	<b>1</b>
<b>3.2</b>	<b>Metrics and related terms .....</b>	<b>1</b>
<b>4</b>	<b>Abbreviated terms .....</b>	<b>3</b>
<b>5</b>	<b>General principles .....</b>	<b>4</b>
<b>5.1</b>	<b>Generalized process for machine learning classification performance assessment .....</b>	<b>4</b>
<b>5.2</b>	<b>Purpose of machine learning classification performance assessment .....</b>	<b>4</b>
<b>5.3</b>	<b>Control criteria in machine learning classification performance assessment .....</b>	<b>5</b>
<b>5.3.1</b>	<b>General .....</b>	<b>5</b>
<b>5.3.2</b>	<b>Data representativeness and bias .....</b>	<b>5</b>
<b>5.3.3</b>	<b>Preprocessing .....</b>	<b>5</b>
<b>5.3.4</b>	<b>Training data .....</b>	<b>5</b>
<b>5.3.5</b>	<b>Test and validation data .....</b>	<b>6</b>
<b>5.3.6</b>	<b>Cross-validation .....</b>	<b>6</b>
<b>5.3.7</b>	<b>Limiting information leakage .....</b>	<b>6</b>
<b>5.3.8</b>	<b>Limiting channel effects .....</b>	<b>6</b>
<b>5.3.9</b>	<b>Ground truth .....</b>	<b>7</b>
<b>5.3.10</b>	<b>Machine learning algorithms, hyperparameters and parameters .....</b>	<b>7</b>
<b>5.3.11</b>	<b>Evaluation environment .....</b>	<b>8</b>
<b>5.3.12</b>	<b>Acceleration .....</b>	<b>8</b>
<b>5.3.13</b>	<b>Appropriate baselines .....</b>	<b>8</b>
<b>5.3.14</b>	<b>Machine learning classification performance context .....</b>	<b>8</b>
<b>6</b>	<b>Statistical measures of performance .....</b>	<b>8</b>
<b>6.1</b>	<b>General .....</b>	<b>8</b>
<b>6.2</b>	<b>Base elements for metric computation .....</b>	<b>9</b>
<b>6.2.1</b>	<b>General .....</b>	<b>9</b>
<b>6.2.2</b>	<b>Confusion matrix .....</b>	<b>9</b>
<b>6.2.3</b>	<b>Accuracy .....</b>	<b>9</b>
<b>6.2.4</b>	<b>Precision, recall and specificity .....</b>	<b>9</b>
<b>6.2.5</b>	<b>F1 score .....</b>	<b>9</b>
<b>6.2.6</b>	<b>F .....</b>	<b>9</b>
<b>6.2.7</b>	<b>Kullback-Leibler divergence .....</b>	<b>10</b>
<b>6.3</b>	<b>Binary classification .....</b>	<b>10</b>
<b>6.3.1</b>	<b>General .....</b>	<b>10</b>
<b>6.3.2</b>	<b>Confusion matrix for binary classification .....</b>	<b>11</b>
<b>6.3.3</b>	<b>Accuracy for binary classification .....</b>	<b>11</b>
<b>6.3.4</b>	<b>Precision, recall, specificity, F1 score and F for binary classification .....</b>	<b>11</b>
<b>6.3.5</b>	<b>Kullback-Leibler divergence for binary classification .....</b>	<b>11</b>
<b>6.3.6</b>	<b>Receiver operating characteristic curve and area under the receiver operating characteristic curve .....</b>	<b>11</b>

6.3.7	Precision recall curve and area under the precision recall curve .....	11
6.3.8	Cumulative response curve .....	12
6.3.9	Lift curve .....	12
6.4	Multi-class classification .....	12
6.4.1	General .....	12
6.4.2	Accuracy for multi-class classification .....	12
6.4.3	Macro-average, weighted-average and micro-average .....	12
6.4.4	Distribution difference or distance metrics .....	13
6.5	Multi-label classification .....	14
6.5.1	General .....	14
6.5.2	Hamming loss .....	14
6.5.3	Exact match ratio .....	15
6.5.4	Jaccard index .....	15
6.5.5	Distribution difference or distance metrics .....	15
6.6	Computational complexity .....	16
6.6.1	General .....	16
6.6.2	Classification latency .....	16
6.6.3	Classification throughput .....	17
6.6.4	Classification efficiency .....	17
6.6.5	Energy consumption .....	17
7	Statistical tests of significance .....	18
7.1	General .....	18
7.2	Paired Student's t-test .....	18
7.3	Analysis of variance .....	19
7.4	Kruskal-Wallis test .....	19
7.5	Chi-squared test .....	19
7.6	Wilcoxon signed-ranks test .....	19
7.7	Fisher's exact test .....	19
7.8	Central limit theorem .....	20
7.9	McNemar test .....	20
7.10	Accommodating multiple comparisons .....	20
7.10.1	General .....	20
7.10.2	Bonferroni correction .....	20
7.10.3	False discovery rate .....	21
8	Reporting .....	21
	Annex A (informative) Multi-class classification performance illustration .....	22
	Annex B (informative) Illustration of ROC curve derived from classification results .....	24
	Annex C (informative) Summary information on machine learning classification benchmark tests ..	29
	Annex D (informative) Chance-corrected cause-specific mortality fraction .....	31
	Bibliography .....	32