ISO/IEC 25000 and AI Product Quality Measurement Perspectives

Andrea Trenta [1]

[1] *UNINFO UNI TC 533 Technical Committee Artificial Intelligence, Turin, Italy*

Abstract

In previous papers [10, 8, 9] we discussed ISO/IEC 25000 application when new quality measures are defined. The definition of product quality measures for ML is challenging, because of the huge number of algorithms and their implementations, that implies a huge number of measures, too. In continuity with papers above, and consistently with ISO standards, we show through examples of measures of ML accuracy and explainability, how to define practical ISO/IEC 25000 compliant product quality measures for AI. Moreover, the paper can be considered for the works in AI standardization area.

Keywords

product quality, measures, accuracy, explainability, ISO, ISO/IEC 25059, ISO/IEC 5259-2, ISO/IEC 25000, metric, AI, ML, Machine Learning, Artificial Intelligence

1. Introduction

Policy makers, industries, and academia are facing the problem of building trust in AI; in the following we present a positive perspective, based on the actual scientific and standardization context, that can contribute to building trust in AI through a quantitative approach. Then, the paper focuses on open quality metric issues and proposes a solution.

2. Standardization context in AI

Policy makers have addressed the issue of AI trustworthiness mainly, but not only, to the international standardization body ISO/IEC SC42 and to the European standardization body CEN/CENELEC JTC21 that have in charge the drafting of technical standards in support of industry and of lawful rules. For the scope of this paper, we consider, among the others, the standards based on ISO 25000 series that define or contribute to define product quality for an AI product [21]. The assessment of product quality, possibly together with the assessment of process quality [22], will be performed in the near future on voluntary or mandatory basis, in the former case to promote trustworthiness in AI systems, in the latter case to get compliance to rules [23]. In the following, we focus on ML based AI systems [15].

3. Quality models for AI

In the following we sum up the status of AI product quality standardization and possible future direction to move on.

Product quality is faced by SQuaRE project that is described in ISO/IEC 25000 series and is based on quality model and its measures.

In a very brief manner, there are 4 quality models: (1) software, (2) data, (3) quality in use, (4) service. Over each model is defined a set of characteristics (for (1) reliability, defectivity, etc. for (2) currentness, accuracy, consistency, etc. for (3) usability, freedom from risk, etc. for (4) responsiveness, IT service maintainability, etc.), and in turn, over each of the characteristic are defined basic measures, (e.g. the measure 'number of duplicated items' for characteristic

Proceedings of 4th International Workshop on Experience with SQuaRE series and its Future Direction IWESQ@APSEC, December 6th, 2022, Tokyo (Virtual), Japan
EMAIL: andrea.trenta@dataqualitylab.it

'consistency' in (2) or the measure 'failure rate' for characteristic 'maturity' in (1)).

The ISO/IEC 25000 itself foresees the possibility to extend the model to specific technologies like AI, through the definition of new characteristics and new measures.

Exploiting this possibility, SQuaRE quality models for ML data and product were faced in [17] and in [16], where peculiar ML aspects, such as bias and trustworthiness are addressed through new specific characteristics like, but not only, 'similarity' for the former and 'accuracy' and 'transparency' for the latter.

Explainability and controllability, are also dealt in specific ISO/IEC standards under development.

4. Categorization

Firstly, it should be noted that generally is easier to design data quality measures that software quality ones, because most of former are influenced simply by data values [7][17][12][13][14][8][9] while the latter are influenced by many context variables [6][16] and so they are harder to measure; moreover, to get the software measures comparable, the software should be categorized. The need of software categorization emerged since the early stage of SQuaRE project and was addressed in [24], that, among the purposes, includes the quality support through the appropriate association and weight between type of software and quality characteristics (e.g. for an home-banking software it is important 'accessibility', for a defense or medical software it is important 'reliability'); this association and its weight (see also [26]), allows in turn the design of the relevant measures and helps homogeneous product quality evaluation by software category.

Applying the categorization approach to AI requires more careful analysis than non-AI software, as further considerations are needed: for example, the characteristic of 'reliability' should lead the evaluation of software for x-ray image processing, but the characteristic of 'transparency' or 'explainability' should lead the evaluation of software for an x-ray automated diagnosis instead.

5. Quality measures for ML

In this perspective, an overall quality score Qs could be a sum of j-measurements M_{ij} for each of the W_i weighted i-characteristics selected for the evaluation, and should be comparable with the relevant benchmarks B_{ij}:

$$Qs = \sum_{i=1}^{n} W_i \cdot \sum_{j=1}^{m} \frac{M_{ij}}{B_{ij}}$$

The main issue for designing and applying measures in a ML context seems the manifold of implementations, more than 82000 according to [27].

We can define the implementation I as a function of

1) I = I(method, algorithm(library, parameters), training(dataset, process))

where:

'method' is the high-level categorization, about 40 in [18] like decision tree, k-means clustering, neural networks, and others

'algorithm' is the type of method[2] (es. ResNet for method=NN)

'library' contains the code to be invoked for evaluation (see machine learning process in [52])

'parameters' are the configuration data of the algorithm.

'training' includes dataset (ImageNet, MNIST,…) and process (initialization, retraining,…).

Then, we can define

2) $M_{ij}=M_{ij}(I)$

and taking into account 1)

3) $M_{ij}=M_{ij}$(method, algorithm(library, parameters), training(dataset, process))

With those definitions, benchmark B_{ij} is the best value M_{ij} for the time being (e.g., for a full year)

[2] for 'algorithm' it is intended the categorization of the code that perform the task, e.g., for the classification task, the 'algorithm' can be either a neural network, or a decision tree, or a support vector machine, or other.

for the i-characteristic and the j-measure[3] among all the K implementations of I_k

4) $B_{ij}=\max_k M_{ij}(I_k)$ k=1,..,K

In the function I, the argument 'library' specifies the code or library that represents or simulates the code to be measured. This approach follows the global research community attempt to describe performance of most of the papers through their code, that is often available and public (see e.g. GitHub that hosts Linux Foundation projects in the category of Trusted and Responsible AI e.g. [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], and others [38][43] that are relevant for explainability metrics). It should be noted that some of the biggest metric projects are led or supported by big companies like Meta Research [27], IBM [42], Microsoft Research [43]. In addition to these resources, there are others like Scikit-learn, and computing tools, like Matlab or Wolfram, that have developed their own ML libraries, mainly on the most consolidated algorithms, for free or commercial use.

6. Use case 1: accuracy

In figure 1 below an example where the i-characteristic is accuracy, and the j-measure is based on multi-class classification metrics [25] and calculated for the test dataset ImageNet for various image classification algorithms (from ZFNet to NFNet of the neural network method); figure 1 shows the progress B_{ij} of different implementations since 2014 (points in grey are non-top performing implementations for the date).
See also [25] for comparison with ISO benchmark definition.

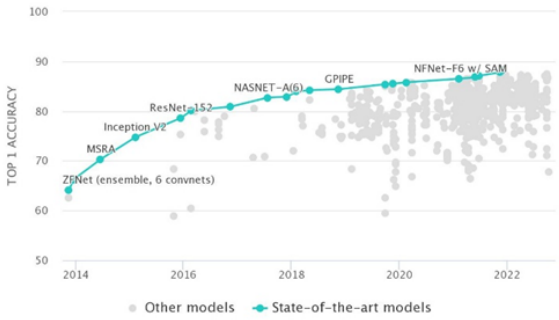

Figure 1: Benchmarks for Image Classification through neural network implementations [27]

As an example, the grey points of the figure 1 can be calculated repeating along the time the measurement M_{ij}, where the i-characteristic= accuracy, and the j-measure Hamming loss (j=1) where I_k is defined in (1).

To get measurements as homogeneous as possible while grouping commonalities, it is advisable that in (1) some variables are not varying, for figure 1 they are:

'method' = Neural Network,

'training dataset' = Imagenet,

'process' = one-step training, and

'library' = library_url,

then 1) becomes:

1a) I = I(algorithm(parameters))

With such assumptions, we can define a k-family measure, as a group of measures where any measure belonging to a family differs from any other of the same family only for the value of a subset of variables of the relevant implementation. An example of a k-family measure is in table 1, where, for the measurement function 'Hamming loss', the measures of the same k-family share the same method, library, and training; moreover, each family differs one from the other for the algorithm and its parameters.

[3] in (4) the j-measure is supposed as scalar; if the j-measure is a vector or a matrix, the expression (4) should be adapted.

Table 1
Accuracy k-family measure – Hamming loss

ID	Accu-ML-1-k
Name	Accuracy of Neural Networks for classification task
Description	Hamming loss for classification Neural Networks
Measurement function	X= L (I_k, O, Q) L is the Hamming loss [27] I_k is the k-implementation NOTE 1 O is the set of observations Q is the set of predictions
NOTE 1 I_k = I_k (method, algorithm (library, parameters), training (dataset, process)) where: method = {Neural Network} algorithm = {type of NN_k} library = {library_url} parameters = {$parameters_k$} training = {Imagenet, one-step training}	

In the example above, the ID is in the format

(5) CCCC-ML-F-k,

where:

CCCC is the acronym of the characteristic relevant for the measure

ML identifies the Machine Learning application

F is the number assigned to the measurement function family

k is the number assigned to the measure of the G family

For example, the family of measures of accuracy trough F1 score of NNs trained with dataset MNIST over a certain library can be identified by ID = Accu-ML-5-k.

For the correct identification in case of more detailed measurement function, the ID can be further extended in the format [4][10]:

(6) CCCC-ML-F-k-AA-v

7. Use case 2: explainability

In the following it is provided a second use case of an ML measure relevant to the sub-characteristic[4] 'explainability'.

As an example, but not the only one, in the field of medicine, the questions to answer are: can an AI automated x-ray diagnosis compete with a professional diagnosis? Do the patient trust in the AI outcome? Some researchers are discouraged from answering such questions [19] and conclude that "unless there are substantial advances in explainable AI, we must treat these systems (AI automated diagnosis system) as black boxes". Being out of the scope of this paper to define any requirement towards clinical procedures or protocols, but keeping in mind health professionals' concerns, in the following we propose a possible design of an explainability ISO 25000 compliant measure.

Explainability [39][40] plays an important role as it can help evaluation and risk assessment [41]. As a rough definition, explainability helps humans to understand the work done by the ML system and can be measured considering the more salient features that influence the ML decision or forecast; usually the metric for explainability is based on the higher-scored features, measured through numerical values or heatmap pixels. In other words, it is attempted to explain decisions (e.g. a classification outcome like 'this is a dog') splitting the whole input data in smaller portions (features) and permuting the input example or altering it; those altered input data usually produce an altered decision and allow to identify which input alterations were most likely to change the output decision. If the input data is an image, for example by occluding one by one of the n featured parts of the image, the explanation will produce an heatmap that indicates the m<n image parts that contributed the most to the decision.

Explainability measurements can be used in conjunction with accuracy ones for assessing purposes; for example, in [47] the measurement function implements the F1 score for accuracy measure, and the CAM heatmap for explainability measure.

[4] According to [16], explainability is not a tier-1 characteristic; in this paper, 'explainability' is relevant to 'transparency' that in turn is a sub characteristic of 'usability'. For the sake of simplicity, in the following it will be referred as a tier-1 characteristic.

Figure 2a: Explainability through Grad-Class activation mapping (Grad-CAM) [20]- X-ray chest

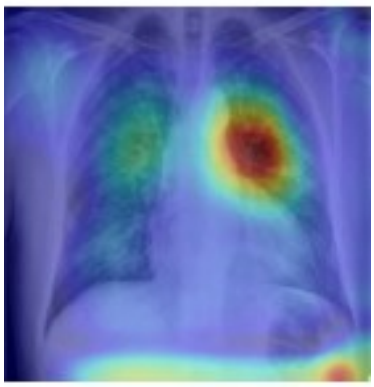

Figure 2b: Explainability through Grad-Class activation mapping (Grad-CAM) [20]- Diagnosis heatmap

The area[5] in red in figure 2b is the area (2D feature) of the chest that mainly driven the ML system to the diagnosis of pneumonia from the analysis of the x-ray image of figure 2a.

The same approach can be used for example to explain why a credit request on behalf of an enterprise is rejected by a ML decision support system used by a bank. In this case it is useful to understand which are the single items that more contributed to the credit rejection decision; in this example they are the enterprise financial health indexes (features) like Book Value of Total Debt (MVE_BVTD), Sales\Total assets (S_TA), industry, and other features. The ML was trained with the dataset of historical credit rating, that contained the decisions made on past lending requests based on financial indexes of requesting enterprises. A typical way to measure such contributions is calculation of Shapley values [48]. Following the previous use case and definition (1), we can define an example of the measurement M_{ij}, where the i-characteristic= explainability, and the j-measure is Shapley values (j=1) and method and algorithm in (1) are not referred to the original ML model but to the simulated one, the so-called 'explanation model':

Table 2
Explainability k-family measure – Shapley values

ID	Expl-ML-1-k
Name	Shapley values
Description	Explainability through Shapley values
Measurement function	X= S (I_k, G, O, Q) S is Shapley values vector (1xF) NOTE 1 I_k is the k-implementation NOTE 2 G is the training matrix (NxF) O is the training ouput vector (1xN) Q is the query point (1xF) NOTE 3
NOTE 1 F is the number of features NOTE 2 $I_k = I_k$ (method, algorithm (library, parameters), training (dataset, process)) where: method = {Regression model} algorithm = {Shapley $value_k$} library = {library_url} parameters = {$parameters_k$} training = {Credit Rating Historical, one-step training} NOTE 3 the query point belongs to a 'local' domain D	

Again, the expression (1a) holds, as we have chosen in table 2 to group by k the algorithms and parameters as there are more than one 'Shapley values' possible ways of calculation.

The influence of training data G and O on the regression parameters of the machine, suggests that even the description of the measure Expl-ML-1-k shall be intended as a general measure that is not suitable for practical purposes; therefore, the measures in table 2 should be further detailed through:

i. a new ID in the format (6),
ii. the specification of the domain D allowed for query points (as the explanation

[5] The explanation measurement in figure 2b is an heatmap matrix and not a scalar numeric value. At the same manner, the measurement function X in table 2 is a vector (1xF).

model works fine only for local query points).

Under the same code of the model and the same training dataset and training process, the specific measure can be applied in practice for all the query points in the local defined domain. Such a further specification is needed, for example, to check if the ML is 'explainable' for an enterprise that belongs to an industry that was not (outlier) in the training data (as the enterprise financial indexes are the query point).

8. Proposal

To establish a framework of meaningful and comparable measures for AI applications requires a new approach: the manifold issue compels us to define as many measures and specific benchmarks as the thousands of tasks, algorithms, dataset combinations are. This is why general measures would be hard to practice and would be not comparable.

Then, consistently with the structure of libraries available from the AI research community, we derived the definition (2) $M_{ij}=M_{ij}(I)$, that is applied both in the example in Table 1 for accuracy and in Table 2 for explainability. Both examples represent the proposal: it consists in a detailed product quality measure design and documentation that includes algorithm, training dataset, library code and parameters; moreover, it was considered the chance to group by family homogeneous measures. Last but not the least, the proposal is conceived to be compliant to ISO/IEC 25000.

9. Conclusion

The spread of AI applications in fields like finance, healthcare, transportation, urges to build trustworthiness in users; policy makers are facing this issue and so developing several norms, e.g. [49, 50, 51].

The contribution of ISO 25000 models and measures to the construction of a trusted AI environment is already recognized [16], but it will be maximally effective only if the measurements will be appropriately assessed, and benchmarks will be available. A way to do this, is through the approach proposed that in turn is strictly based on the actual categorization and organization of AI libraries developed by the research community [27-38, 42, 43] and inspired by successful similar experiences in creating listed measures [46].

The present proposal, as well as the measure naming and process described in [10], could be considered by the ISO SC42 and SC7 relevant working groups for the standardization work in progress.

10. References

[1] International Organization for Standardization, ISO/IEC 25000:2014 Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Guide to SQuaRE. URL: https://standards.iso.org/ittf/PubliclyAvailableStandards/index.html

[2] International Organization for Standardization, ISO/IEC 25010:2011 Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models. URL: https://www.iso.org/standard/35733.html

[3] International Organization for Standardization, ISO/IEC 25012:2008 Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Data quality model. URL: https://www.iso.org/standard/35736.html

[4] International Organization for Standardization, ISO/IEC 25020:2019, Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality measurement framework. URL: https://www.iso.org/standard/72117.html

[5] International Organization for Standardization, ISO/IEC 25022:2016, Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of quality in use. URL: https://www.iso.org/standard/35746.html

[6] International Organization for Standardization, ISO/IEC 25023:2016, Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of

system and software product quality. URL: https://www.iso.org/standard/35747.html

[7] International Organization for Standardization, ISO/IEC 25024:2015, Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality. URL: https://www.iso.org/standard/35749.html

[8] A. Trenta, Data bias measurement: a geometrical approach through frames, Proceedings of IWESQ@APSEC 2021. URL: http://ceur-ws.org/Vol-3114/

[9] A. Trenta: ISO/IEC 25000 quality measures for A.I.: a geometrical approach, Proceedings of IWESQ@APSEC 2020. URL: http://ceur-ws.org/Vol-2800/

[10] D. Natale, A. Trenta, Examples of practical use of ISO/IEC 25000, Proceedings of IWESQ@APSEC 2019. URL: http://ceur-ws.org/Vol-2545/

[11] International Organization for Standardization, ISO/IEC TR 24027 Information technology - Artificial Intelligence (AI) – Bias in AI systems and AI-aided decision making. URL: https://www.iso.org/standard/35749.html

[12] M. Mecati, F. E. Cannavò, A. Vetrò, M. Torchiano, Identifying Risks in Datasets for Automated Decision–Making, in: International Conference on Electronic Government, 2020, pp. 332-344, Springer,Cham. URL: https://link.springer.com/chapter/10.1007%2F978-3-030-57599-1_25

[13] E. Beretta, A. Vetrò, B. Lepri, J.C. De Martin, Detecting discriminatory risk through data annotation based on Bayesian inferences, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 794-804. URL: https://doi.org/10.1145/3442188.3445940

[14] E. Beretta, A. Vetrò, B. Lepri, J.C. De Martin, Ethical and socially-aware data labels, in Annual International Symposium on Information Management and Big Data, 2018, pp. 320-327, Springer, Cham. URL: https://link.springer.com/chapter/10.1007%2F978-3-030-11680-4_30

[15] International Organization for Standardization, ISO/IEC 22989:2022 Information technology — Artificial intelligence —Artificial intelligence concepts and terminology. URL: https://www.iso.org/standard/74296.html

[16] International Organization for Standardization, ISO/IEC DIS 25059 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality Model for AI-based systems. URL: https://www.iso.org/standard/80655.html

[17] International Organization for Standardization, ISO/IEC CD 5259-2 (under development) Artificial intelligence — Data quality for analytics and ML — Part 2: Data quality measures. URL: https://www.iso.org/standard/81860.html

[18] International Organization for Standardization, ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). URL: https://www.iso.org/standard/74438.html

[19] M. Ghassemi, L. Oakden-Rayner, A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care Lancet Digit Health, 2021, 3:e745–50. URL: https://www.thelancet.com/pdfs/journals/landig/PIIS2589-7500(21)00208-9.pdf

[20] M. Chetoui, M.A. Akhloufi, B. Yousefi, E.M. Bouattane, Explainable COVID-19 Detection on Chest X-rays Using an End-to-End Deep Convolutional Neural Network Architecture, Big Data and Cognitive Computing, 2021, 5,73. URL: https://doi.org/10.3390/bdcc5040073

[21] D. Natale, Extensions of ISO/IEC 25000 quality models to the context of Artificial Intelligence, Proceedings of IWESQ@APSEC 2022. To appear.

[22] International Organization for Standardization, ISO/IEC 42001 (draft) Information technology — Artificial intelligence — Management system. URL: https://www.iso.org/standard/81230.html

[23] European Commission, COM/2021/206 'Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts', 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[24] International Organization for Standardization, ISO/IEC TR 12182:2015 Systems and software engineering — Framework for categorization of IT systems

and software, and guide for applying it. URL: https://www.iso.org/standard/63611.html

[25] International Organization for Standardization, ISO/IEC TS 4213:2022 Information Technology — Artificial Intelligence — Assessment of machine learning classification performance. URL: https://www.iso.org/standard/79799.html

[26] International Organization for Standardization, ISO/IEC CD 25040 (under development) Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality evaluation framework. URL: https://www.iso.org/standard/83467.html

[27] Meta Research, Papers with Code resource. URL: https://paperswithcode.com/sota

[28] Trusted-AI, Contrastive explanation methods resource. URL: https://github.com/IBM/AIX360/tree/master/aix360/algorithms/contrastive/CEM.py for A. Dhurandhar et al., Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives, Advances in Neural Information Processing Systems (NeurIPS), 2018. https://arxiv.org/abs/1802.07623

[29] S. Lundberg, SHAP resource. URL: https://github.com/slundberg/shap for Lundberg, et al. A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems (NeurIPS), 2017. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpretingmodel-predictions

[30] Trusted-AI, Contrastive Explanations Method with Monotonic Attribute Functions resource. URL: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/contrastive/CEM_MAF.py for R. Luss et al., Leveraging Latent Features for Local Explanations, 2019, URL: https://arxiv.org/abs/1905.12698

[31] Trusted-AI, Explainability resource. URL: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/ted/TED_Cartesian.py for M. Hind et al., Teaching AI to Explain its Decisions, Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, URL: https://doi.org/10.1145/3306618.3314273

[32] Trusted-AI, Faithfulness resource. URL: https://github.com/Trusted-AI/AIX360/blob/master/aix360/metrics/local_metrics.py, for D. Alvarez-Melis, T. Jaakkola, Towards Robust Interpretability with Self-Explaining Neural Networks, Advances in Neural Information Processing Systems 31 (NeurIPS), 2018, URL: https://papers.nips.cc/paper/8003-towards-robust-interpretabilitywith-self-explaining-neural-networks

[33] Trusted-AI, Monotonicity resource. URL: https://github.com/Trusted-AI/AIX360/blob/master/aix360/metrics/local_metrics.py for R. Luss et al., Leveraging Latent Features for Local Explanations, 2019, URL: https://arxiv.org/abs/1905.12698

[34] Trusted-AI, ProfWeight resource. URL: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/profwt/profwt.py for A. Dhurandhar et al., Improving Simple Models with Confidence Profiles, Advances in Neural Information Processing Systems 31 (NeurIPS), 2018, URL: https://papers.nips.cc/paper/8231-improving-simple-models-with-confidenceprofiles

[35] Trusted-AI, ProtoDash resource. URL: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/protodash/PDASH.py for S. Gurumoorthy et al., Efficient Data Representation by Selecting Prototypes with Importance Weights, International Conference on Data Mining (ICDM), 2019, URL: https://arxiv.org/abs/1707.01212

[36] Trusted-AI, Boolean Decision Rules via Column Generation resource. URL: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/rbm/BRCG.py (Light Edition) for S. Dash et al., Boolean Decision Rules via Column Generation, Advances in Neural Information Processing Systems 31 (NeurIPS), 2018, URL: https://papers.nips.cc/paper/7716-boolean-decision-rules-via-column-generation

[37] Trusted-AI, Generalized Linear Rule Models resource. URL: https://github.com/Trusted-AI/AIX360/blob/master/aix360/algorithms/rbm/GLRM.py for D. Wei et al., Generalized Linear Rule Models, Proceedings of the 36th International Conference on Machine Learning, PMLR 97:6687-6696, 2019, URL: http://proceedings.mlr.press/v97/wei19a.html

[38] H. X. Vinh, Quantitative Input Influence resource. URL: https://github.com/hovinh/QII, for A. Datta S.Sen, Y.Zick, Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems, IEEE Symposium on Security and Privacy (SP), 2016. URL: https://ieeexplore.ieee.org/document/7546525

[39] International Organization for Standardization, ISO/IEC AWI TS 6254 (under development) Information technology - Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems. URL: https://www.iso.org/standard/82148.html

[40] International Organization for Standardization, ISO/IEC TR 24028:2020 Information technology —Artificial intelligence (AI) — Overview of trustworthiness in artificial intelligence. URL: https://www.iso.org/standard/77608.html

[41] International Organization for Standardization, ISO/IEC CD TR 5469 (under development) Artificial intelligence — Functional safety and AI systems. URL: https://www.iso.org/standard/81283.html

[42] AI Explainability 360 – IBM Research Trusted AI URL: https://aix360.mybluemix.net/

[43] Microsoft Research, InterpretML toolkit resource. URL: https://interpret.ml

[44] M. Sameki, S. Bird, K. Walker, et al. InterpretML: A toolkit for understanding machine learning models, 2020. URL: https://www.microsoft.com/en-us/research/uploads/prod/2020/05/InterpretML-Whitepaper.pdf

[45] Scikit-learn resource. URL: https://scikit-learn.org/stable/index.html#

[46] International Organization for Standardization, ISO/IEC 5055:2021 Information technology - Software measurement - Software quality measurement - Automated source code quality measures. URL: https://www.iso.org/standard/80623.html

[47] P. Rajpurkar, J. Irvin, K. Zhu, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning, 2017. URL: http://arxiv.org/abs/1711.05225

[48] Mathworks resource. URL: https://it.mathworks.com/help/stats/shapley-values-for-machine-learning-model.html

[49] Federal Reserve Board, Equal Credit Opportunity Act, URL: https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_reg_b.pdf

[50] Consumer Financial Protection Bureau, Fair Credit Reporting Act, URL: https://www.consumerfinance.gov/rules-policy/regulations/1022/

[51] European Commission, Proposal for a 'Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts' COM/2021/206 final URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[52] International Organization for Standardization, ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). URL: https://www.iso.org/standard/74438.html