9

Companies developing compliance with

Guide 9. Accuracy

European
Artificial Intelligence Act

Financiado por la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de Recuperación, Transformación y Resiliencia

España | digital 2026

This guide has been developed within the framework of the development of the Spanish pilot for the regulatory AI Sandbox, through collaboration among participants, technical assistance providers, potential competent national authorities, and the sandbox's expert advisory group.

The aim of the guide is to serve as an introductory support to the European Regulation on Artificial Intelligence and its applicable obligations. Although **it is not legally binding and does not replace or develop the applicable legislation, it provides practical recommendations** aligned with regulatory requirements, pending the approval of the harmonised implementing standards for all Member States.

This document **is subject to an ongoing process of evaluation and review, with periodic updates** in line with the development of standards and the various guidelines published by the European Commission, and it will be updated once the Digital Omnibus amending the Artificial Intelligence Act is approved.

Among the currently applicable relevant technical references under development is **prEN 18229-2 "AI Trustworthiness Framework – Part 2: Accuracy and Robustness"** stands out, as it will serve as the basis for the evaluation of the accuracy and robustness of AI systems once adopted as a harmonized standard within the context of compliance with the European Regulation on Artificial Intelligence.

Revision date: 10, December 2025

General content

Detailed Index

1. Preamble

1.1 Purpose of the document

High-risk AI systems are named that way mainly, because of the potential health, safety and fundamental rights risks that their use represents. This is indicated in several points of **the European Regulation on Artificial (*AI Act*),** as well as throughout the guides that accompany this sandbox. A keyway to be able to mitigate these risks as much as possible is specifically with the **accuracy** of this AI system; through the accuracy of the system, we obtain a quantitative measure of the relationship between the **intended purpose** of the system and its performance from design to operation after the implementation of the AI system.

A series of organizational and technical measures are developed throughout the guide, aimed first at selecting and evaluating accuracy metrics for the AI system. Then, proceed to apply the model's quality controls that help verify and validate the reasons that lead to using such metrics. We will also deal with complementary aspects that are key to being able to implement the system, as they go beyond the immediate result of the model for a reference framework, and may have implications of discrimination, bias or imprecision.

It is important to note that the application of accuracy concepts in an AI system to comply with the requirements of the European Regulation on Artificial Intelligence requires, both on the part of providers, a knowledge of the state of the art of these accuracy-related techniques and how they can be applied to their AI system, according to their intended purpose. Therefore, accompanied by the mechanisms and information provided in this guide, providers must keep an eye not only on the evolution of regulations, but also on the ever-rapid evolution of the state of the art.

1.2 How to read this guide?

The process of achieving the appropriate accuracy of the model goes through a series of steps that we have understood to help cover the requirements established in Article 15, accuracy, robustness and cybersecurity of the European Regulation on Artificial Intelligence, see section on the content of the article and correspondence of the article with the sections of the guide.

1.3 Who is it for?

To accommodate all the issues detailed in this guide it is the responsibility of the provider of the high-risk artificial intelligence system by taking the appropriate measures proposed here, both organizational and technical. All this with the aim of ensuring that the accuracy requirements of the system are met.

Within its scope of application, the user of the system also has responsibilities that will materialize in concrete measures, again organizational and technical. The guide will indicate, in each case, which ones are applicable and the scope of these.

1.4 Use cases and examples throughout the guide

Throughout the guide, two use cases will be used as examples of how to develop the technical documentation. Examples will focus only on the provider who is responsible for generating and maintaining documentation. A detailed description of the use cases used can be found in the Cross-Cutting Information Guide.

Note: Whenever an example is given, it will be done in an illustrative way. Provider and user must consider the application of all the measures indicated in this guide.

The use cases have been selected based on their ability to explain the information and procedures detailed in this guide.

The cases selected in this case for the preparation of the guide are:

- Detection of false reports.
- Employee promotion system.

2. Introduction

2.1 What is accuracy for AI?

In the context of the European Regulation on Artificial Intelligence, recital (66) states that accuracy is among the key requirements for mitigating the risks associated with the AI system:

AI Act

(66)

Requirements should apply to high-risk AI systems as regards risk management, the quality and relevance of data sets used, technical documentation and record-keeping, transparency and the provision of information to deployers, human oversight, and robustness, accuracy and cybersecurity. Those requirements are necessary to effectively mitigate the risks for health, safety and fundamental rights. As no other less trade restrictive measures are reasonably available those requirements are not unjustified restrictions to trade.

We can consider that accuracy , understood as we have indicated in the previous note, allows us to secure, delimit and know the behaviour of the AI system according to its intended purpose, the datasets with which it will work, and the relationship of accuracy with the **rest** of the requirements, among others: cybersecurity, robustness, transparency, data governance, supervision. In addition, accuracy is a fundamental metric within the quality management system that surrounds the high-risk AI system.

The AI Act also indicates, in its recital (74), the need for the level of accuracy to be up to date with the advances and state of the art in the field and for this level to be maintained throughout the life cycle of the AI system, from its placing on the market to its withdrawal.

AI Act

(74)

High-risk AI systems should perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity, in light of their intended purpose and in accordance with the generally acknowledged state of the art. The Commission and relevant organisations and stakeholders are encouraged to take due consideration of the mitigation of risks and the negative impacts of the AI system. The expected level of performance metrics should be declared in the accompanying instructions of use. Providers are urged to communicate that information to deployers in a clear and easily understandable way, free of misunderstandings or misleading statements. Union law on legal metrology, including Directives 2014/31/EU and 2014/32/EU of the European Parliament and of the Council, aims to ensure the accuracy of measurements and to help the transparency and fairness of commercial transactions. In that context, in cooperation with relevant stakeholders and organisation, such as metrology

and benchmarking authorities, the Commission should encourage, as appropriate, the development of benchmarks and measurement methodologies for AI systems. In doing so, the Commission should take note and collaborate with international partners working on metrology and relevant measurement indicators relating to AI.

As we can see in the previous recital, the European Regulation on Artificial Intelligence can be understood as a clear relationship between accuracy and transparency, by relating both to information to the user.

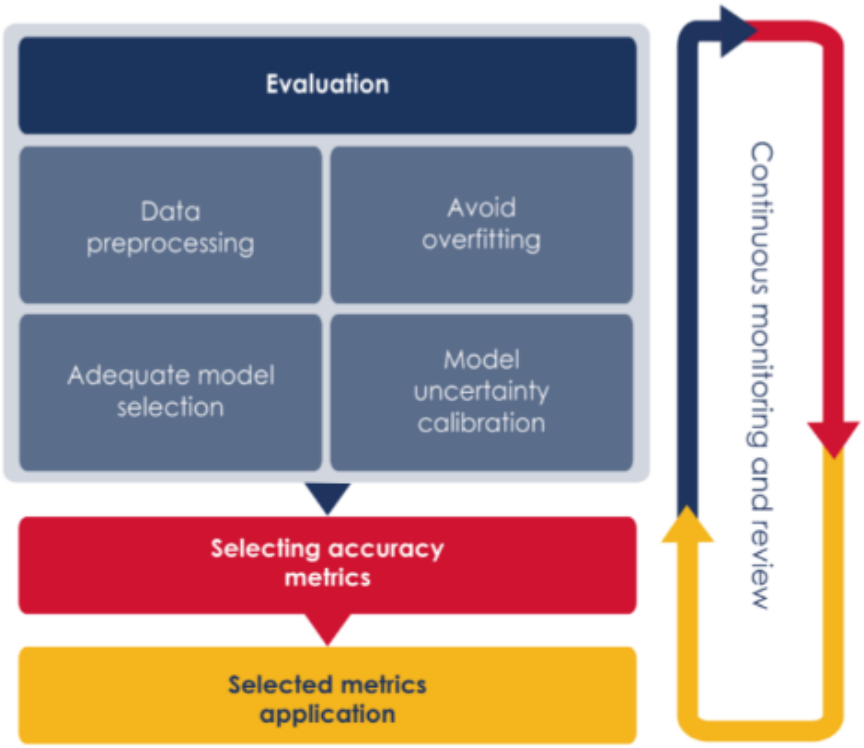
On the other hand, accuracy is very present throughout Annex IV of the AI Act, which explains how technical documentation should be approached. An example is the following extract from paragraph 3 of Annex IV, concerning the relationship between technical documentation and accuracy.

AI Act

Annex IV.3 – Technical documentation referred to in Article 11(1)

Detailed information about the monitoring, functioning and control of the AI system, in particular with regard to: its capabilities and limitations in performance, including the degrees of accuracy for specific persons or groups of persons on which the system is intended to be used and the overall expected level of accuracy in relation to its intended purpose; […].

The process to achieve accuracy can be visualized as:



3. European Regulation on Artificial Intelligence

The putting into service or use of high-risk AI systems should be subject to compliance with certain mandatory requirements, including accuracy. Those requirements aim to ensure that high-risk AI systems available in the Union or whose result outputs are used in the Union do not pose unacceptable risks to important public interests recognised and protected by Union law.

This section includes the articles referring to the generation of accuracy of Regulation 2024/1689 of the European Parliament and of the Council, of 13 June 2024 (European Regulation on Artificial Intelligence) and details in which sections of this guide the different elements of these articles are addressed.

3.1 Previous analysis and relationship of the articles

As we have indicated in the introduction, the accuracy in the European Regulation on Artificial Intelligence is clearly related and present with other areas of the AI system itself. It is within Article 15, accuracy, robustness and cybersecurity, of the AI Act, where the requirements that the AI system must meet are found, in aspects related to accuracy, are specifically addressed.

This article establishes the requirements that must be met in terms of three fundamental aspects *"Accuracy, robustness and cybersecurity"*. Cybersecurity and robustness are specifically addressed in their guides.

In this guide, we are going to emphasize the paragraphs of this article that are specifically oriented to accuracy in AI.

Therefore, we will indicate a series of measures aimed at ensuring that AI systems do not degrade their performance and accuracy specifications once they are commissioned, throughout their life cycle.

AI systems must not present problems of operation (compatibility with old libraries they use or data they process) or quality problems in terms of their accuracy, as they are used over time. To do this, they must respect minimum levels of accuracy and/or specific metrics associated with the task, pre-established, thus ensuring that it is consistent throughout the life cycle.

In addition, the European Regulation on Artificial Intelligence states (Art. 15, accuracy, robustness and cybersecurity Paragraph 2):

Instructions on how to obtain the proposed levels of accuracy, how to use and interpret them (transparency guide), their thresholds and associated metrics should be documented according to the technical documentation guide, see section 5.

Within the analysis of the articles defined in this guide, it can be summarised in the following points:

- Analyse and establish the relationship between the life cycle of the high-risk AI system and accuracy, at critical points of the life cycle. This aspect is reviewed in the section4.1 Accuracy and lifecycle. This section addresses how aspects of the AI system's life cycle can

influence the final accuracy of the system, and how these should be considered, with the aim of ensuring that accuracy is consistent throughout it.

- According to the model of our AI system, select metrics most suitable for measuring accuracy, in section 4.2, discusses how these metrics should be selected.
- These metrics will also include the selection of a target function that will be used to achieve the indicated accuracy during training.
- Once the metrics and their values have been selected according to the intended purpose, the provider must ensure accuracy throughout the life cycle in a *consistent* as established by the European Regulation on Artificial Intelligence. In the section 4.3, the measures that we understand help to cover that specific requirement are provided.
- Finally, the entire process must be accompanied by correct documentation in accordance with both the technical documentation (see the technical documentation guide itself in detail) and the measures proposed in this guide, which must be adequately documented. We address in section 5, how this action is proposed.

3.2 Content of the articles in the AI Act

It should be clarified that within Article 15 in which accuracy, robustness and cybersecurity are concerned, accuracy is specifically spoken of exclusively in points one, two and three.

AI Act

Art.15 – Accuracy, robustness and cybersecurity

1. High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle.

2. To address the technical aspects of how to measure the appropriate levels of accuracy and robustness set out in paragraph 1 and any other relevant performance metrics, the Commission shall, in cooperation with relevant stakeholders and organisations such as metrology and benchmarking authorities, encourage, as appropriate, the development of benchmarks and measurement methodologies.

3. The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use.

4. High-risk AI systems shall be as resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems. Technical and organisational measures shall be taken in this regard.

The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans.

High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures.

5. High-risk AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities.

The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks.

The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws.

3.3 Correspondence of the articles with the sections of the guide

The following table details the sections of this guide that address the different elements of this article:

Article	AI Act requirement	Guide section
15.1	Right level of accuracy	Section 4.2 and section 4.3
15.1	Consistent accuracy throughout the lifecycle	Section 4.1
15.2	Technical aspects of accuracy	Section 5
15.3	Relevant instructions and accuracy levels	Section 5

4. How to approach the requirements?

4.1 Accuracy and lifecycle

Establishing and measuring the accuracy of AI system is a process that covers the entire life cycle of the, as we have indicated. There are, however, points during the life cycle where establishing the accuracy of the system according to its intended purpose is critical. Below, to complete the work process described, we present these critical points, how they affect accuracy, and how to address them.

4.1.1 Data preprocessing

The accuracy of a model will depend on the quality of the training data and therefore its preprocessing. In addition to the measurements in the Data guide, relevant issues to consider in the preprocessing of the data to ensure the accuracy of the model are:

- The data used to test the *machine learning* model must be representative of the intended purpose of the system (see Data Guide).
- The model's training data must be free of sampling bias. In addition, training data for a particular task is not necessarily extensible to other different tasks. Special attention should be paid when dividing unbalanced databases to ensure that similar distributions are maintained between training data, validation and evaluation (see Data Guide).
- When data is pre-processed differently, with the aim of finding the accuracy appropriate to the intended purpose, the difference in accuracy cannot be attributed to the algorithm being evaluated (the model with the final objective or task, "d*ownstream algorithm*"). Therefore, identical ways of preprocessing data should be used to compare the accuracy between several models, and the same evaluation set (*test*) should be used to compare two models, so that the validation and evaluation sets will never contain samples also included in the training set. See ISO/IEC TS 4213:2022, *Information technology – Artificial intelligence – Assessment of machine learning classification performance* [116].
- The pre-training tasks of the model, if they differ from the tasks of the final model, shall be documented in an accurate and reproducible manner as well, and it shall be ensured that they coincide both in the training, validation and production phases.

More information on bias in AI systems can be found in ISO/IEC TR 24027 and in the Data Guide.

Example - False Reporting Detection

During the design phase of the system, it is established that the accuracy selected for the system must be guaranteed for a wide range of complaints, and that accuracy must not be affected by elements of categorization of the complaint itself as address or minority membership.

To this end, the decision is made, as indicated in this section, to carry out work on these datasets, with the aim of having a database that allows the chosen accuracy metric to be defined, established and documented throughout the life cycle. To do this:

- Complaints are collected from all over the national territory in Spanish. The collection is representative in equal proportion for each territory, to avoid population size biases.
- In addition to the previous point, to mitigate sampling bias, a random mixture of the complaints is carried out, which guarantees an even distribution throughout the training, testing and validation sets.
- The preprocessing required for the anonymization process is identical for all territorial sources and datasets. To this end, a specific piece of preprocessing software is developed that is unique for all territories and will be used to send the data. In this way, the design, development and implementation team already receive anonymized data.

This entire process is documented in each of the steps indicating the decisions made and the motivation.

4.1.2 Overfitting

Generative algorithms are trained by optimizing parameters of the learning model in such a way that the probability of success is maximized, on the sample of available training data and according to the established classification categories, while discriminative algorithms optimize their parameters to maximize the accuracy of the classification [116].

To avoid the common problem of overfitting, measures will be taken in accordance with the type of model and intended purpose. In general:

- All hyperparameters must be reported in the training/testing and validation processes of the model, as well as their values for each model. Hyperparameter selection bias must be taken into account when comparing models, as different models have different tuning capabilities, so the level of overlearning during training may differ between algorithms, especially in deep learning.
- No information from the testing data set will be used when adjusting hyperparameters, as this typically leads to overestimating accuracy metrics with optimism. When label information is needed for such tuning, it typically uses that of a separate dataset, the validation set, which is disjoint from the test set. This challenge can be approached, for example, with nested cross-validation, where an outer loop measures prediction accuracy while the inner loop adjusts the hyperparameters of individual models. In this way, methods can choose the optimal parameter settings for the predictive models to be chosen in the outer loop.

- Tools for automating hyperparameter search should ideally be used, e.g., *Hyperband, Hyperopt, or Optuna*[1] (commonly used in reinforcement learning).

Example - False Reporting Detection

The provider's risk management system for the AI system has established that there is a risk to rights and freedoms if the system is overfitting with respect to training sets. An *overfitting* can cause true complaints to be processed as false leading to a different processing procedure, so this *overfitting* can affect the final process. In this way, overfitting the training data, test, and validation set can provide an excellent measure of the chosen accuracy metric, but when the system is in production, cause results that are inadequate for its intended purpose.

To this end, and following the indications in this guide, the provider establishes that:

- The training hyperparameters will be stored associated with the code of the AI model, so that they are always accessible and identifiable, so that the development team can consult a history of them and study the possibility of *overfitting*.
- The training, testing, and validation sets are organized and structured so that they all have a random combination of origins, so that the origins of the already classified anonymous reports are evenly distributed in the sample. For each complaint, a unique identifier is generated (through the calculation of a *256 hash* of its text content). This system is used so that the training, testing, and validation sets are completely disjoint and do not contain any common samples.
- When training is being carried out, the indication proposed in this section of two programming loops is applied, which also keep track of the origin of the data, in order to detect data exchange between sets. This not only ensures hyperparameter tuning, but also adequately ensures that the sets are disjoint.
- Within the *framework* for the AI system, the development team has used, combined with the process described in the previous points, a specific library of the *framework* for the automated adjustment of hyperparameters.

4.1.3 Use of appropriate models

The accuracy of a model will have relevance with respect to how it is positioned with respect to the state of the art in models of its category and/or with base reference models. That is, the selected accuracy metrics must be compared and established in context, along with those base models with which it is related.

- For the reproducibility and sustainability of model accuracy, studies on the removal of components and elements (ablation) from the model should be noted or reported, to justify the composition and complexity of the components and elements of the model necessary to achieve the guaranteed accuracy.
- In them, appropriate baseline reference models (*Baselines*) will be used. *Trivial baselines* such as those that always predict the majority class are useful for gauging

[1] Words and descriptions **Outstanding** correspond to terms that are developed in the glossary (see 7.4).

the interpretation of the metric (see Transparency guide) but should not be the only point of comparison (ISOIEC TS 4213_2)[116].

Example - Employee Promotion System

Throughout the design of the AI system, and considering its intended purpose, the system provider analyses the baseline reference models within the state of the art. The objective of this analysis is to select one that serves both as a starting point for the design and implementation of the AI system, as well as to know how it is positioned within the context of systems in the same category.

For the analysis, the provider considers the intended purpose of establishing a mechanism for the promotion of employees that will be the final element of decision in promotion mechanisms.

As a base model selected for this task, a vector space model or VSM (*vector space mode*) is chosen, where the employees will be described through the variables (coordinates) established for their evaluation. This trivial base model will be used together with the rest of the techniques indicated in the guide, for the selection of the appropriate accuracy metric(s) and for the evaluation of the accuracy of the AI system.

The provider records the motivations and evaluations made on the base model candidates, as well as the indication of the final choice of the latter.

Important Note: The technical nature of this guide makes the examples specific to the use cases. This implies that the proposals are specific to the models considered as examples, and not a general solution for other types of models, or even models of the same typology. Examples should be considered demonstrative of the operation to obtain a technical conclusion, but never as demonstrative of that conclusion itself. Each provider must, in accordance with this guide, establish the specific technical measure for its type of AI system and its intended purpose.

4.1.4 Uncertainty and Accuracy

The accuracy of a model is associated with a level of certainty or confidence that does not always accompany the outputs of a model, if they are not calibrated or take into account the uncertainty of the model. The AI system can provide the result of its operation accompanied by an indication, for example, as a percentage (or in any case normalized) of the certainty it gives to that result. The uncertainty of a model, understood as confidence in its results, must therefore be documented to ensure accuracy and facilitate its monitoring and transparency, according to the robustness guide.

4.2 Assessing Accuracy

The applicable accuracy metrics guaranteed by the provider in the documentation should reflect and be a sign of the quality of the system. Otherwise, when these minimum established accuracy metrics cannot be guaranteed, the provider will provide mechanisms to notify the human that their supervision is required (according to human oversight guide).

At each step of the life cycle of a high-risk AI system, from its design, development and validation to its introduction to the market, a level of accuracy must be established in accordance with the intended purpose of the system, establishing appropriate measures.

Depending on the phase of the life cycle, the measurements must be carried out by provider (design, development) or user (during its production or operation), avoiding degradation during its life cycle.

Provider

The provider must take the following organizational steps to ensure that the high-risk AI system has an adequate level of accuracy. It is the responsibility of the provider to make the selection of the most appropriate accuracy metric (or metrics), from the conception and design of the AI system. This selection process must consider two important aspects:

- Metrics should be related to and motivated by the intended purpose.
- The metrics must be motivated and allow to mitigate the risks to the rights and freedoms of natural persons, health and damage to property/environment identified in the risks analysis.

Once the functional (typically specified by the functionality that the user will request in typical system use cases) and non-functional (specified by the provider technical staff as technical implementation characteristics or quantifiable metrics) of the system have been established, a battery of meaningful and representative tests of the application domain should reflect for compatible inputs, expected outputs within the possible ranges of data output contemplated, always in direct relation to the intended purpose and to the mitigation of risks.

The battery of tests may include, but is not limited to:

- Those unit and stress tests accumulated after the development of the model (for example, following methodologies such as *Test Driven Development*).
- Integration test for the encasing of functional and non-functional requirements with the intended purpose, and that have the established accuracy as a fundamental piece.
- Track the evolution of model accuracy when in production.
- The degradation of the model will be monitored with monitoring panels and accuracy visualization tools. In the guide we will indicate other possible associated model quality metrics such as target functions. This aspect is related to the information presented in the robustness guide.

As a complement to these organisational measures, the provider shall align the accuracy of the AI system with the following technical measures:

- Providers, designers and developers must provide detailed installation and commissioning documentation (see technical documentation guide) to use the system, indicating dependencies and other requirements or potential special cases that the user should take into account in order to deal with all types and formats of data supported by the system).
- The complete process necessary for the fine-tuning of data that requires pre-processing must be indicated beforehand to be used as input to the system provided (data guide).
- You should specify in sufficient detail any post-processing necessary to be able to interpret the accuracy of the model in a clear and concise manner, and thus be able to notify the provider when this is not the case.
- They must provide (according to the technical documentation guide) details on the type, format and amount of expected input data and output, and the requirements of these during their use in the training, validation and testing stages of the system

to obtain the documented level of accuracy, with examples of use of execution of the system for its understanding and transparency towards the user.
- For the reported accuracy to be actionable and useful, the provider must include the possible ranges of configurable parameters and both inputs data and outputs, as well as the expected latency measurements to obtain an expected accuracy (see robustness guide).

To reproduce the behaviour of the model in terms of its accuracy metrics, it is recommended to use documentation and formatting of the model in the ONNX https://github.com/onnx/onnx open standard for machine learning interoperabilityhttps://github.com/onnx/onnx.

To detail the process of selecting accuracy metrics and how they can ensure the specific requirements of Article 15, accuracy, robustness and cybersecurity, of the European Regulation on Artificial Intelligence both in its definition and throughout the lifecycle, as such, consistently:

- The sections 4.2.1 and 4.2.2 address the selection of the proposed metrics both of accuracy and those related to the objective function, which, as indicated, is very relevant to the concept of accuracy. Both sections are reflected in the annexes 7.1 Accuracy Metrics and 7.2 Objective respectively.
- Section 4.2.3 provides general and complementary aspects related to the metrics selected in the previous sections and which must be taken into account.
- Both aspects are directly related to the previous section 4.1, where the relationship of accuracy to critical milestones in the life cycle has been indicated.

The technical measures to be applied by the provider will be translated into controls based on model accuracy metrics, which will be related and established with the intended purpose of the model, and with the aim that accuracy is always aimed at guaranteeing it, but with the main focus on mitigating risks for the rights and freedoms of natural persons who have been located in the risks management system (see risks management guide).

4.2.1 Selecting Accuracy Metrics

The provider will decide on the relevant accuracy metrics or quality control KPIs (according to the quality management guide) of the system to be measured and evaluated, and will choose a mechanism to store and report a historical record or log (according to the records guide) with the values of these in the time of use of the system, in order to monitor (see human oversight guide) the accuracy and performance of the system. As we can see, the relationship of accuracy is totally transversal to the aspects requested by the European Regulation on Artificial Intelligence and its relationship with the rest of the sandbox guides.

To provide providers with a benchmark of accuracy metrics see the annex 7.1 Accuracy Metrics, where a list of applicable accuracy metrics is presented, for a typology of models. The annex provides a non-exhaustive classification, which may serve to assist in the selection of the metric.

For the process of selecting the accuracy metric appropriate to the AI system, we recommend that the provider, in addition to the strongly technical aspect of the selection, take into account two fundamental aspects:

- The intended purpose of the system, considering what the system is going to accomplish and its objective.

- The risks found detected in the risks management system in relation to the system's decisions and that must be mitigated.

As a technical measure, it is important to have a centralized repository where the information on metrics associated with the model can be managed at any point in its life cycle with associated changes in order to trace the causes of its loss of accuracy, along with identified actors responsible for it. The provider must provide the system with the capabilities that allow the user to access and monitor the different accuracy metrics and metrics associated with the model, as well as the protected variables, always in accordance with the data guide, in order to be able to observe and report possible changes in the system during its use, potential risks that may arise, or biases that may emerge, in the high-risk AI system.

Example - False Reporting Detection

The provider, considering that the intended purpose of the AI system is to classify between false and true reports, with a degree of probability, in a first approximation to select an accuracy metric, is considered to be performing a binary classification. A risk has also been identified that there is no discriminatory capacity and that a true complaint is not properly categorized, causing the complainant not to be treated properly.

With both criteria, it is established that one of the accuracy metrics selected for this AI system is the use of the area under the ROC curve (see Annex 7.1Accuracy Metrics). A value of 1 represents a perfect classification of the complaint and a value of 0.5 means that the assessment has zero discriminatory capacity. In this way, through the selection of this accuracy metric throughout the entire life cycle, it is possible to measure as the false complaint detection system behaves in the range 1 (perfect classification) and 0.5 (zero discrimination).

4.2.2 Selecting the Target Function

As we have mentioned in the introduction to this section, in order to optimize the accuracy of the model, the objective functions to be optimized (minimize or maximize) will be determined by the specific problems and function to be learned by the AI system and, therefore, directly related to the intended purpose. These are used to monitor the machine learning process of models.

The system provider must select an objective function that allows, in the same way as the accuracy metric, to achieve the intended purpose. In annex 7.1, a series of objective-classified functions are presented, which can be used as the basis for the AI system.

4.2.3 Dimensions complementary to accuracy

As complementary dimensions, which support accuracy, the following points are considered:

- Once the calculation of the accuracy metrics has been implemented, it must be monitored (according to human oversight guide), during the life cycle of operation of the system, that the notions relevant to the use case, and the accuracy of the global model do not vary or deteriorate significantly.

- Both bias metrics, impartiality and explainability metrics, will play a particularly relevant role in a phase of the AI system's life cycle (see guide to data, transparency and human oversigth).
- Both *fair judgment* and non-proprietary explainability tools are open source and preferable due to the principle of transparency towards responsible AI (see transparency guide). More tools to consider to achieve responsible AI systems can be found in [15].

The aspects relating to these aspects complementary to accuracy are presented in greater detail in the annex 7.3.

A relevant complementary aspect, and which is the responsibility of the provider, is related to the capabilities that the AI system is provided to the user. To facilitate the communication of the metrics and proper functioning of the system to the user, the user must have access to:

- An interface to obtain metrics of accuracy and performance of the system that allows notification of deficiencies of the system, potential operating errors. It is understood, therefore, that the availability of such an interface will be the responsibility of the provider, who will be obliged to provide the system with these capabilities, with the appropriate instructions (see technical documentation guide) and with the appropriate supervision and transparency capabilities (see supervision and transparency guides respectively).
- Inspect both input and output data.

Example - Employee Promotion System

Following the prompts set out in this guide, the employee promotion system provider has selected the metrics for its AI system. The selection is made based on providing a measure of accuracy appropriate to the intended purpose established for the AI system and by design and also on selecting an objective function also suitable for the intended purpose, thereby:

- For one, it selects *Discounted Cumulative Gain* as a metric to evaluate the accuracy of the system. This metric is considered to evaluate the relevance of the results obtained in relation to a query or *query*. Given the intended purpose, it states that your query is built based on the parameters configured for promotion.
- On the other hand, within the objective functions (See annex 7.2), rules out the use of *Pointwise* or *Pairwise* for oversimplifying its approach by approximating its AI system to a regression model or a binary classification respectively, which are not aligned with the intended purpose of the AI system and which could incur in excluding valid candidates. It is therefore opted for the use of an objective function of type *Listwise* which allows you to maximize the selected accuracy metric.

The entire process of analysis and selection, both of the accuracy metric and of the target function, are documented, explaining motivations and approaches, as well as all the results obtained from their application during the design and validation phases of the AI system. All this will be incorporated into the technical documentation as established in the corresponding guide.

Deployer

The deployer of the high-risk AI system must know and have trained personnel to understand the level of accuracy of the system during operation. Therefore, the accuracy of the model will not be complete without providing transparency of how accuracy metrics and related performance metrics are computed and computed.

Example - Employee Promotion System

One of the companies that has contracted the use of the AI system for the promotion of employees, following the indications in this section for the user, decides that those personnel who are going to operate with the AI system receive training that allows them to understand, at the level of operation, the concept of accuracy of the system and its relationship with the intended purpose (to have an indicator for the promotion of employees).

To do this, they consult with a company specialized in training courses, and following the instructions provided by the provider of the AI system on accuracy, they add to the training on the use of the system, a few hours dedicated to additionally providing the following information:

- The concept of high-level accuracy and its relationship with the results of the system within its operations.
- A high-level explanation of how the accuracy metric selected by the provider works and how it relates to the system's information panels.

This training is complemented with a periodic update plan, to ensure that the concept is taken into account in the operation of the system.

Transparency will therefore provide quality to the model, in terms of the degree of availability of information about the AI system and the way in which the information is communicated to relevant parties in accordance with its objectives and know-how (referring to ISO/IEC 25059, *Software engineering — Systems and software Quality Requirements and Evaluation (73) — Quality model for AI systems*). Given the context of the guide in which we find ourselves, one of the key aspects for this transparency information will be the information associated with the selected accuracy metrics (for more details on transparency, see the transparency guide).

The user should train members of their organization who interact with the High-Risk AI system to:

- Know how to use and interpret accuracy monitoring visualizations related to all metrics, according to the intended purpose and their explainability and associated uncertainties at any time in the system's life cycle.
- Know how to use the system's output, identifying requirements that might be necessary to use it as input in another AI system. (See ISO on Biases, SC42_N1011 ISOIEC_TS_4213_2 [116]).
- Be able to inspect both input data and output and have permission to correct and/or notify the provider of potential erroneous data or outputs, or the absence thereof.
- Have knowledge about the interfaces that provide transparency on the operation, output, and evaluation metrics of the AI model, and their interpretation according to the explainability of the model [35] to interpret it.

- Be responsible for understanding potential biases and impartiality that may compromise the accuracy of the model. For example, to know the concepts of algorithmic discrimination, bias, and impartiality, to detect possible problems or potential degradations of the output of the model. For example, the degradation of accuracy in a way that is not favourable to a minority class or (group of) protected variables.

In line with the example indicated above, on the AI system for the promotion of employees, the company using it must ensure that all the above points are properly transferred to the human resources staff.

4.3 Ensuring accuracy

The applicable accuracy metrics guaranteed by the provider in the technical documentation should reflect and be a sign of the quality of the system. Otherwise, when these minimum established accuracy metrics cannot be guaranteed, the provider will provide the user with mechanisms to notify the human required for monitoring (according to human oversight guide). In the same way, the user will be obliged to know this information and have internal processes to be able to carry them out.

In the previous section we have indicated how to select the accuracy metrics for the AI system and how these relate to other complementary actions.

In this section we are going to provide the necessary information so that the selected metrics are maintained consistently throughout the life cycle, as established by the AI Act.

Provider

The AI system provider is the primary guarantor that accuracy is consistently maintained throughout the AI lifecycle, with the implementation of measures to ensure this.

4.3.1 Technical measures

The **technical measures** that the provider must implement in relation to the inventories of accuracy metrics and target functions mentioned in the previous section (see section 4.2) are identified with the following actions:

- For the accuracy reported by the AI system to be actionable and useful, such documentation must include the possible ranges of configurable parameters, as well as the system's input data and output, and the latency measurements expected to achieve the desired accuracy (see Robustness Guide).
- The output of the system, ideally, should be accompanied by a measure of uncertainty associated with the accuracy of that output.
- The provider shall choose a mechanism to store all the accuracies communicated to the user in a historical record, in accordance with the Records Guide, in order to allow the traceability of the accuracy metrics over the time of use of the system and thus monitor its performance according to the Robustness Guide.

The user will be provided with a graphical interface that allows observing the metrics of accuracy over time and detecting inconsistencies or malfunctions of the system (according to the Human Oversight Guide) and monitoring them (according to the Robustness Guide).

The rationale for selecting one metric or another to assess model accuracy should be documented in accordance with the Technical Documentation Guide; See also section 5, which addresses the aspects related to the documentation of the actions indicated in this guide.

Example - Employee Promotion System

During the analysis phase of the AI system, it has been concluded that one of the measures established to maintain accuracy over time is to have a panel in the *operation* of the system where the accuracy metrics and their evolution over time are shown, in this way it is possible to keep the progress of the accuracy visible and therefore to be able to act on it in case it leaves the work ranges Defined. The system, in addition to displaying the data, makes a recording of the accuracy information presented according to what is established in the records.

In the same way, during the analysis of the system, it is found that it is important to accompany, as recommended in this guide, information on the uncertainty, the result of the output of the system, for this the use of a set of combined models (ensemble learning) is proposed. The use of this technique allows establishing a confidence value of the result, based on the responses of the different models (ensembles) combined in order to provide a measure of the uncertainty of the result.

4.3.2 Statistical significance assessments

During the process of selecting the accuracy metric (as well as the target functions), in order to give statistical validity to the results for the AI system, specific metrics should be used and accompanied by statistical evaluations. Assessments of statistical significance should consider the distribution of the data (parametric or non-parametric assessments), the dependence between them (independent and identically distributed, i.i.d, or not) and other specific assumptions of each assessment, in order to choose the relevant assessment appropriately (see data guide). Among them:

- The Wilcoxon signed *ranking* evaluation. Non-parametric, i.e. free of assumptions about distribution, and dependent data.
- Parametric or non-parametric evaluation.
- The Paired student T-Test, *analysis of variances,* Kruskal-Wallis assessment, Chi-Squared assessment, Fisher exact assessment, McNamar test, (including multiple comparison, Bonferroni correction and false discovery rate) or others may be used.

Many of these will require additional normality tests on the data, e.g., analysis of variances (ANOVA) to determine if the means of more than two groups are equal, or their differences in accuracy between 3 or more models are statistically significant, for which ANOVA assumes normal distributions and that the variance is homogeneous. Also take into account the applicability of this according to the *Central Limit Theorem*.

Example - Employee Promotion System

In this AI system, as explained in the previous section, the provider has made the selection of *Discounted Cumulative Gain* as a metric to evaluate the accuracy of the system.

On this AI system, a battery of statistical significance evaluations are carried out using the established metric. To this end, a series of testing data and validation sets are carried out grouped by professional sectors, to keep them in datasets related to each other and that allow the results to be properly consolidated. Having considered that the possible results of the different data sets do not have a normal distribution, *a ranking* with a Wilcoxon sign has been selected to perform this evaluation as the best tool to validate the process.

The evaluation of statistical significance is used during the training/testing process to verify that the model conforms to the specifications defined in the design phase, and according to the intended purpose (to have an indicator for the promotion of employees). It allows you to compare the evolution of the model with each other throughout its evolution over time, or to compare different candidate models with each other at a point in the process (*ensemble learning*), while considering the accuracy metric and specifications and guiding the training and testing process.

For more information, see ISO SC42 N1011. ISOIEC TS 4213_2 (Classification. Perf) [116]

4.3.3 *Database and Model* Benchmarks

In the complete information about the accuracy metric(s) selected for the AI system, information must be provided on the benchmarking *(Performance Tests)* of the data and the model. These *benchmarks* must take into consideration the intended purpose of the system, so that accuracy can be put into context not only internally within the AI system itself, but also in the external context in relation to data and models in relation to repeatable and measurable tests.

All appropriate metrics will be used and reported to measure the accuracy of the model and according to the protocols established by the benchmark *evaluations* of each machine learning task. The *benchmarks* most commonly used by the relevant scientific community in machine learning and AI and their fields of application should be used and documented.

Benchmarks *depend* on the task and the state of the art of the field; therefore, they will need to be updated as time goes by, and models and databases evolve. Examples of databases and corresponding models that act either as state-of-the-art or as baseline reference models in the image classification task are:

- In computer vision or image-processing models:
  - Benchmark/ImageNet Database: Base models AlexNet, VGG-19, ResNet-50, EfficientNet-B7.
  - MNEST database: RDML, MCDNN, LeNet base models.
  - CIFAR-10 database: EfficientNet-B7, ColorNet, DenseNet base models.
  - LFW database: FaceNet, DeepID3, DeepFace base models...
- In language models, LSTM, BERT, GPT1, 2, 3, [...], etc. models are considered *baselines* with respect to the state of the art. Also consider the metrics specific to the specific task

(machine translation, reading comprehension, summary, etc.). To do this, see table H.1 [4] for a set of significant tasks and metrics in language models.

- In generative image models, photorealism or sample diversity (SSIM, MMD, IS, MS, FID, LPIPS) are measured. For example, to assess physical consistency: IoU (in image segmentation models) or FVPS [5].
- Databases for fairness include COMPAS recid, COMPAS viol. recid, Diabetes, OULAD, *Credit card clients and many others* (see Table 1 in [9]).

Further measures to ensure model accuracy and documentation can be seen in ISO SC42_N1011_ISOIEC_TS_4213_2 [116].

Deployer

Organizationally, the user must consider:

- The deployer is responsible for consulting the AI system instruction manual to know, apply, and keep an eye on the model (according to Data Guide and Human Oversight Guide).
- Both the deployer and all the parties involved (from the largest responsible manager or top manager, *product owner*, *project manager and data scientist*) must have access to the justification of the accuracy of the system and its associated metrics at any time during its useful life cycle. To this end, interfaces adapted to different audiences that may require an explanation of the model or its audit will be used, in order to facilitate the transparency of the reasoning process.

The deployer should become familiar with and understand at the appropriate level the interpretation of the notions and metrics of the model in terms of its accuracy and its relationship to the purpose.

5. Technical documentation

In the technical documentation guide provided within the framework of the AI sandbox, the structure and content of the technical documentation for high-risk AI systems are described in detail and in accordance with the European Regulation on Artificial Intelligence.

This section provides a detailed explanation of how to document the process of selecting and assuring accuracy, as outlined in this guide.

The AI Act establishes that not only the definition and selection of accuracy metrics are relevant, but that their scope should also be extended to the deployer of the system through the corresponding instructions. This is reflected in paragraphs 2 and 3 of Article 15 on accuracy, robustness, and cybersecurity:

AI Act

Art.15.2 and 3 – Accuracy, robustness and cybersecurity

2. To address the technical aspects of how to measure the appropriate levels of accuracy and robustness set out in paragraph 1 and any other relevant performance metrics, the Commission shall, in cooperation with relevant stakeholders and organisations such as metrology and benchmarking authorities, encourage, as appropriate, the development of benchmarks and measurement methodologies.

3. The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use.

Instructions on how to obtain accuracy, how to report it (transparency guide), and how to interpret and use its thresholds and associated metrics shall be documented in the specific sections detailed in the Technical Documentation Guide, in accordance with their selection and evaluation, as described in previous sections and in Annex 7.1.

In **Annex IV** of the minimum required technical documentation for high-risk AI systems, several points specifically refer to the accuracy of the system:

- **Point 2(g):** *"The validation and testing procedures used, including information on the validation and testing data employed and their main characteristics; the parameters used to measure accuracy, robustness, and compliance with other relevant requirements set out in Chapter III, Section 2, as well as potential discriminatory effects; the test logs and all test reports dated and signed by the responsible persons..."* This point constitutes the documentary core of accuracy documentation, as it requires describing how accuracy is measured, with which data, under what conditions, and using which metrics. The design of the tests, validation sets, formulas or accuracy indicators (see Annex 7.1), the results obtained, and the supporting evidence must be documented.
- **Point 3:** *"Detailed information on the monitoring, operation, and control of the AI system, in particular with regard to its capabilities and limitations, including the levels of*

accuracy for specific persons or groups of persons for whom the system is intended to be used, and the overall expected level of accuracy in relation to its intended purpose." This section requires declaring the achieved and expected accuracy levels of the system under real operating conditions, distinguishing between groups or contexts when relevant. The methodology for measuring operational accuracy, known limitations, error margins, and the justification that the performance is adequate for the intended purpose must be documented.

- **Point 4:** "A description of the suitability of the performance parameters for the specific AI system."
  Here it is required to justify that the metrics or indicators used to measure accuracy are appropriate and representative of the system's intended use. The reasons for selecting those parameters, their technical relevance, and how they correctly reflect the actual or expected model performance must be documented.

- **Point 9:** *"A detailed description of the system established to evaluate the performance of the AI system in the post-market phase, including the post-market monitoring plan..."*
  This point links accuracy with its ongoing monitoring after commercialization. It requires maintaining a documented procedure to monitor the evolution of performance (including accuracy) over time, record degradations or deviations, and establish corrective actions.

If one wishes to go beyond the minimum required by the Regulation and properly document accuracy—which depends on data quality—the provider should include in the documentation two relevant descriptive elements: **the Model Card and the Dataset Card.**

5.1 Model Card

A Model Card describing the model(s) used by the system, including their accuracy and robustness metrics, operational capabilities, limitations, and interrelations regarding changes in accuracy and robustness.

This section aligns with the Technical Documentation Guide, as it must be incorporated into the overall documentation.

Model Cards must include sections specifying, at a minimum:

- Model size, including the number of parameters in key configurations.
- Date of publication.
- Type of model, indicating whether it has been developed from scratch or if a base model has been used.
- Intended purpose.
- Identity terms, with reference to frequently vulnerable or affected groups, such as those focused on sexual orientation, gender, and race.
- Key articles and references, if any.
- Details of the model's use in training, validation, and evaluation, along with information on performance, efficiency, and limitations.
- Broader implications and ethical considerations, risks and recommendations [50].

In particular, model cards should include, with respect to performance or quality metrics, answers to the following questions in a specific section for the metrics, which measure and illustrate disproportionate model accuracy errors between subgroups

- Model Accuracy Measures: What metrics are reported and why were they selected? All metrics that reflect the actual potential impact of the model should be specified, including accuracy measures and other performance-related metrics.
- Decision thresholds: If used, why were those specific values selected? For digital model cards, we recommend that you include an interactive slider to display accuracy metrics based on different decision thresholds.
- Approaches to Uncertainty and Variability: How are the uncertainty measures and estimates of these metrics calculated? This could include standard deviation, variance, confidence intervals or KL divergence.
- Calculation methods and source of values: Details on how the values of these metrics are obtained should be included (e.g., average of five runs, cross-validation of 10 iterations or folds, etc.).
- Model interpretability: What methods or tools have been implemented to facilitate the understanding of the model and its predictions? (For example, SHAP, LIME, or feature importance visualizations.)
- Model adaptability: Information about the model's ability to adapt to new data or conditions, including details about fine-tuning or periodic update processes.
- Responsible use: Warnings or limitations of the model in terms of its application to avoid misunderstandings or misuse. This could include warnings about sensitive applications or inappropriate contexts, and how these could negatively impact the results.

The model cards are a transparency tool, also important in relation to robustness (see corresponding robustness guide) and technical documentation, favouring technical and non-technical auditing by analysts, as well as more inclusive user feedback mechanisms.

However, until standardization or formalization is carried out in a way that avoids misleading or confusing representations of results, the usefulness and accuracy of model cards are based on the integrity of the creator of the card itself [50].

Example - False Reporting Detection

To complete the form card, following the guides established in this guide, the AI system provider provides the following information on the card, among other things:

- It is indicated that the area under the ROC curve has been selected as a accuracy metric and this has been complemented with the information of the system's confusion matrix, as we have indicated, understanding the AI system as a binary classification.
- Since the AI system performs text processing of complaints, perplexity has also been considered as a metric of accuracy. The model sheet details how the perplexity results have been analysed during training and testing.
- To manage uncertainty and variability, add information to the model card, of the procedure followed. Random validation data have been divided, and by specific source sets, to study how accuracy metrics behave in different sets.

Note: The provider has completed the information on the model card by adding the fields that have been indicated in the guide, and in this example we have indicated those of greater relevance and in line with the treatment of this use case developed throughout the guide. Sandbox participants, in the case of model cards, must complete all the data indicated in this section.

5.2 Database Card

Datasheets for datasets (see [10]) are advisable to give a more holistic view of the accuracy metrics provided and the provenance of the data used for model training. If this changes with the use of the model, the cards in the model must also be updated.

Documentation methodologies associated with transparency such as those mentioned above, among others suitable for the potential future to come, should be studied to give a broader view to the accuracy metrics and visualize, if any, possible intersectional disparities, for example, in the accuracy of the task performed by the model.

An example of intersectional disparity for *accuracy* in commercial gender classification is available at [17], and an example of a model card can be viewed at https://github.com/openai/whisper/blob/main/model-card.md.

The provider should consider fairness and bias taxonomies to select complementary metrics that enrich the concept of accuracy in relation to the system's intended purpose.

The provider shall document the tests carried out for possible types of bias that the system may suffer affecting the accuracy for categories of data that may be discriminated against. In the same way, it should be documented how the same metrics provided, at the global level of models, do not present an unequal or disparate impact (*disparate impact*) for different groups belonging to sensitive variables. For these, the output of the model must be impartial and non-discriminatory (see data guide).

6. Self-assessment questionnaire

To carry out a self-assessment of compliance with the requirements of the European Regulation on Artificial Intelligence referred to in this guide, a global self-assessment questionnaire has been generated with a series of questions with the key points to be taken into account with respect to the obligations dictated by the articles of the AI Act mentioned in this guide.

It will be necessary to refer to this document in order to carry out the section of the self-assessment questionnaire corresponding to this guide.

7. Annexes

7.1 Accuracy Metrics

In this annex we present a series of accuracy metrics that can be used and the types of models with which they are related. This Annex is not intended to be an exhaustive list of existing metrics, but rather a presentation of those that can be applied to the types of models presented. The provider must take into account that a specific type of model or intended purpose (or the combination of both) may require a more specific analysis of the metric to be selected and that it may not be found in the list presented.

The following points describe the accuracy metrics proposed to measure accuracy, grouped by the typology of the associated model.

1. Error metrics for regression models: MSE, RMSE, MAE, MAPE. R^2 and R^2 adjusted.

2. In classification models, the computation of metrics shall include the use of basic elements of accuracy, hit rates and other performance metrics of the model, in accordance with the relevance of the intended purpose of the AI system system. In this regard, the following will be included:

- Confusion matrix.
- Accuracy or accuracy rate.
- Accuracy, *recall*, and specificity and sensitivity, the F1 value, F beta, and/or the Kullback-Lieber divergence.

2.i. Binary Classification. In particular, the models implemented to perform binary classification should report:

- Confusion matrix.
- Accuracy rate.
- F-value (F1 or F beta when you want to give greater importance to accuracy (than to completeness or, vice versa), and/or the Kullback-Lieber divergence.
- Accuracy and *recall curve*, area under the ROC curve (Receiver Operating Characteristic) and area under the AUROC curve.
- Cumulative response curve, and lift *curve*.

2.ii Evaluation metrics in multiclass classification: Coverage error, the average accuracy value of the label ranking, and *Ranking loss*, AUROC curve (useful when there is unbalanced data and the ranking between predictions is important). In addition, multiclass classification models shall report:

- Accuracy, macro-average, micro-average and weighted average.
- Distance metrics or differences in distributions. For example, in distillation, fidelity between distributions (the mean agreement metric or the mean KL divergence [33] between distributions) can be measured.

For example, fidelity (mean KL divergence) can be used to measure the degree of alignment between the model learned by the student model and the teacher model, when the student simplifies the teacher's model by applying distillation learning.

2.iii Multi-label classification evaluation metrics: These must report at least one of the following metrics:

- Hamming Loss Objective Function.
- Exact match ratio.
- Jaccard index, or IoU (intersection over junction).
- Distance metrics or distribution differences.

3. Clustering evaluation metrics*: these include the* adjusted mutual information value, Rand index, the Calinski and Harabaz value, the Davies-Bouldin value, contingency matrix, the completeness metric, the Fowlkes-Mallows index, the mean silhouette coefficient, etc.

4. In other more specific machine learning tasks, study the most convenient metric to quantify the accuracy of the model. For example:

4.i Evaluation metrics in machine learning problems with unbalanced data**:** Unary *or* one-class classifiers *and anomaly detection metrics are used*. In particular, a accuracy-recall curve (PRC) and the area under it (AUPRCs) are more suitable metrics than an ROC curve and the AUROC metric for showing accuracy with unbalanced data. F1 value with weighted-average F1 *is also a representative metric in these cases.*

4.ii Evaluation of language models: In speech recognition, the perplexity of the evaluation data, *Word Error Rate* (WER) is used. Other more generic metrics are Language Model Probability and Word Accuracy, BLEU, METEOR, NIST LRE and others.

4.iii To evaluate image models, use task-specific metrics; For example, in segmentation of objects in images: intersection over junction (IoU). In generative models, the Frechet Information Distance (FID) or Inception Score (IS) can be reported. In autoencoder models such as VAE, *sharpness*, *Inception score can be used*. FID *score*, among other metrics [3].

7.2 Functions Objective

In this annex we present a series of objective functions, grouped by the types of model or tasks with which they are related.

7.2.1 Objective functions regression, classification or *ranking*

Below, we list objective regression, classification, or ranking functions:

1. In regression, the following can be used: MAE (*Mean Average Error*), MSE (*Mean Squared Error* or L2 *loss*), MAPE (*Mean Absolute percentage Error*), *Mean Squared Logarithmic Error* (robust to anomalous values), cosine similarity, logarithm of the hyperbolic cosine (LogCosh), Huber *loss* (less sensitive to anomalous values).

2. In classification depending on the mechanism of the model used.

2.i In Binary classification the use of the target functions is recommended: Hinge *embedding loss* (also to learn non-linear representations or *embeddings* or semi-supervised tasks), binary cross-entropy (BCE), KL divergence, etc.

2.ii Multi-class classification objective functions such as negative log likelihood *(NLL)*, Crossentropy, Categorical cross-entropy, Sparse categorical cross-entropy, *Poisson's* Divergence KL (to ensure that the distribution of predictions is similar to that of the predictions) can be used. training data) and to approximate complex functions).

2.iii Multi-Tag Sorting: The Hamming Loss Target Function.

7.2.2 Target functions in other model types

For other high-risk AI system model types, a compendium of other possible applicable target functions is listed in this section, with the aim of establishing the accuracy of the models.

- In classifiers with unbalanced data, or object detection: Focal *loss.*
- In ranking problems, you can use target functions such as the *Margin Ranking Loss function* to predict the relative distance between entries, and metrics such as AUC among others.
- In ranking problems, you can also use *Pointwise Methods, Pairwise Methods, Listwise Methods.*
- In problems of learning representations or *embeddings* and learning relative similarity between inputs and problems based on content-based retrieval: *Triplet margin loss* objective.
- In generative models such as GANs, discriminator and generator target functions can be used as classic basic target functions of min-max loss *of GAN,* non-saturating GAN loss, and alternatives such as *Wasserstein GAN* loss, conditional GAN target function (CGAN), *Sharpness Loss,* Cook Distance, or other https://neptune.ai/blog/gan-loss-functions. In addition, *Log Hyperbolic Cosine Loss* **(**LogCosh**) can be used on VAE models**.
- In reinforcement learning, the objective functions will be given by reward functions adapted to the problem and environment to be solved. The specific choice in each case must be evidenced in the technical documentation. The motivation for the choice of the selected target function must be described in this documentation, and how this relates to the accuracy established for the intended purpose of the system.

7.3 Accuracy, bias, and impartiality

In this annex we will address the relationship of accuracy with two aspects of the high-risk AI system in particular: bias and impartiality. The aim is therefore to establish the boundaries between accuracy and the rest of the aspects discussed here so that the provider of the AI system has a complete view.

7.3.1 Bias and accuracy

The presence of biases can strongly affect the accuracy of the system, even without a correct analysis of biases, falsifying an accuracy metric, invalidating its usefulness, which is why it is very important, in the process of establishing the appropriate metrics for the accuracy of the AI system, in its development and validation. Perform bias verification and mitigation of bias in the data, and in the operation and output of the AI system.

Possible biases to be avoided in the design, and to be verified and mitigated during the life cycle of use of the AI model according to data guide must be verified; these are detailed in ISO in Bias ISO/IEC TR 24027:2021, *Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making*, the bias catalogue (https://catalogofbias.org/) and the exemplified bias catalogue [7].

Especially at the design stage, the BIASeD taxonomy of cognitive biases, or equivalent, has to be methodically taken into account to consider and avoid each bias and each effect listed in it [112].

- Verification of notions of impartiality. Accuracy metrics will be reliable if they are fair or fair and convey the information necessary to facilitate the transparency of the model to arrive at them (according to the transparency guide).
- Verification of the explainability of the model, in order to facilitate the two previous points (see data guide).

Several biases in the data should be avoided, considering the prospects for using the model in the context of ethical AI:

- Representation bias.
- Historical bias.
- Population bias, sampling bias, omitted variables.
- Masking *bias*, intersectional bias and algorithmic bias [18] among other applicable bias that are directly related to the intended purpose of the AI system (see data guide).

Other more generic biases could also be present depending on applicability: selection bias, collider *bias*, information bias, *recall*, ascertainment, *attrition*, immortal time, *confounding* bias, misclassification, or the Hawthorne effect, among many [8].

All these biases should be monitored from data collection to the end of the operating life cycle, especially after commissioning [62]. To complete the list of tools to facilitate the application of bias and impartiality monitoring, there is again a clear relationship between data guides and human oversight and the notion and metric of accuracy.

There are tools in the Fairness Ecosystem for AI systems that can be used by the provider, both commercial and largely open source. Notions of impartiality should be ensured, ideally and where it makes sense, in accordance with the risk scales specified in the Risks Guide.

For example, crime recidivism prediction models must exhibit *accuracy equity* and predictive parity [16]. For example, given the proximity to the indicated concept, and the intended purpose of the false complaint detection system, the provider adds as an additional accuracy criterion to the one already selected, complying with both concepts.

7.3.2 Fairness to mitigate bias

Assuming that the preprocessing of the data has been done in a way that is appropriate to the intended purpose (according to the data guide) and in a fair way for all variables (protected and not), when a bias in the data guide or notion of impartiality has been detected that is not satisfied, impartiality techniques can be applied.

Relevant fairness metrics aim to uncover potential bias in the data, model, or the model designer/developer itself, which may require mitigation actions. Generally, these are based on differences, ratios, and evaluations of statistical impartiality. The ontology of possible notions and metrics of impartiality to measure types of bias is divided into the following two categories. See Figure 2 of reference [18].

The calculation of the accuracy of the model will take into account the fairness metrics to minimize the possible discrimination of the model towards groups of values belonging to sensitive variables (e.g., minorities due to sexual orientation, race, gender or sex). According to applicability, the different notions of impartiality will be implemented through metrics of ratios, differences or statistical tests according to the applicability of each metric. For example:

Notions of group impartiality:

- Impartiality of class independence (statistical parity, conditional statistical parity).
- Impartiality of class separation: Equalized *Odds*, Equal Opportunity, Predictive Equality, Total Impartiality.
- Class sufficiency fairness (conditional usage accuracy equality, predictive parity, Well calibration).
- Total, impartiality.
- Equal treatment.
- Overall accuracy equality.

Notions of relaxed impartiality:

- Notions of impartiality with a threshold and based on statistical evaluations [18].

7.3.2.1 Taxonomies and metrics of impartiality.

A comprehensive guide should be evaluated prior to use. They can be seen in [9], the *Fairness Ontology* [18], and in the BIASeD taxonomy of cognitive biases [112]. There will be cases where not all the principles can be obtained [63], and the commitment of such a conflict should be documented according to the technical documentation guide and on the model card.

Techniques for evaluating and establishing impartiality include, for example, but are not limited to:

- Reweighting to mitigate bias in the training data collection phase.

- Techniques to evaluate the new transformed data, such as *balanced accuracy* or *average odds difference*, to avoid disparate impact_or results that do not preserve equality (*unequal outcomes*) between protected variables [62].

- Calculate the Area Under the Absolute ROC Curves *for protected and non-protected groups (*ABROCA) [113]).

- When in doubt as to which impartiality metrics should be assessed, the ontology of impartiality can be applied [18]. The privileged group will be considered as the one that was historically observed to have a systematic advantage, while the non-privileged group would represent the one with systematic disadvantage in history.

- Variations according to these groups and the notion of impartiality chosen can be graphically represented to visually describe biases in layered combinations of attributes protected with *plotly sunburst* **plots. Example**: See Fig. 6 in [50].

- For the verification of notions of impartiality, equivalences can be established: e.g., equality between different metrics of the confounding matrix equals equality of opportunity; equal rates of false negatives and false positives between groups equals satisfying *Equality of Odds* [50].

7.4 Glossary

Term	Definition
Absolute Between-ROC Area, ABROCA	ABROCA measures the absolute value of the area between the ROC curve of the reference group and the ROC curves of one or more comparison groups. In this way, ABROCA quantifies the divergence between the ROC curves of different groups through all possible thresholds, aggregating this divergence regardless of which subgroup performs better at a specific threshold. This allows to evaluate impartiality in the performance of the model for different subgroups. See development of the technique in https://homes.cs.washington.edu/~jpgard/papers/lak19_slicing.pdf
Accuracy	Accuracy, or accuracy, in the context of classification models is the proportion of correct predictions over the total predictions made. It's a performance measure that indicates how well the model correctly classifies instances in the dataset.
Accuracy equity	An AI system shows accuracy *equity* if it can equally discriminate between the possible ranking of its output space for different groups.
Macro-media, micro-media and weighted average-average accuracy	A macro-mean will calculate the metric independently for each class, and then take the mean (thus treating all classes equally), while a micro-mean will aggregate the contributions of all classes to calculate the average metric. In a multi-class classification, micro-media is preferable if it is suspected that there may be an imbalance between classes (i.e., if you have many more examples of one class than others). On the other hand, to calculate a weighted average, each number in the dataset is multiplied by a predetermined weight before the final calculation is made.
ANOVA	*Analysis of Variance* (ANOVA). A statistical formula used to compare the variances between the means (or averages) of different groups. It is used to determine if there is any difference between the means of different groups. This quotient shows the difference between the variance within the group and the variance between groups, which ultimately produces a figure that allows us to conclude that the null hypothesis is supported or rejected. If there is a significant difference between the groups, the null hypothesis is not supported.
Average odds difference	It is the average of the difference in false positive and true positive rates between the underprivileged and privileged groups. A value of 0 implies that both groups benefit equally.

Term	Definition
Balanced accuracy	This metric can be used to analyse the accuracy (performance) of a classification model. It is calculated by adding the ratio of true positives with the ratio of true negatives and dividing it by two. It is especially useful when classes are not balanced. The closer this value is to 1, the better the model is for making correct classifications.
Base Line model	They are simple models that work as a reference for the AI system. Its main function is to contextualize the result of the AI model's training. Generally, these models lack complexity and have little predictive power. They serve as a reference and for establishing comparisons with the AI system and allow a better understanding of the work data.
BIASeD	It is a categorization or taxonomy of known cognitive biases, from the perspective of AI systems. This catalogue and the research work carried out aims to align biases in AI with the biases of the human race itself. The full study can be found in https://arxiv.org/abs/2210.01122
BLEU	*BLEU is a metric used in machine translation to evaluate the quality of a generated translation compared to a reference translation. Its calculation is based on the coincidence of n-grams between the translated text and the reference text, providing a score that reflects the similarity and accuracy of the generated translation.*
Central Limit Theorem	The central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a "bell curve") as the sample size increases, assuming that all samples are identical in size and regardless of the actual shape of the population distribution. In other words, CLT is a statistical premise according to which, given a sufficiently large sample size of a population with a finite level of variance, the mean of all variables sampled from the same population will be approximately equal to the mean of the entire population.
Classification (binary, multiclass, multitag	Binary classification involves predicting between two possible categories or labels. Multiclass classification includes more than two possible tags, assigning a single tag to each instance. Multi-tag sorting allows multiple tags to be assigned to each instance, which is useful in situations where the same example may belong to more than one category simultaneously
Clustering	As a type of unsupervised learning, *clustering* is the process of ordering a group of objects in such a way that objects in the same group (which is called a cluster) are more similar to each other than to objects in any other group. There are different types of *clustering algorithms* (K-means, MeanShift, DBSCAN etc.).

Term	Definition
Average Silhouette Coefficient	This is a coefficient that measures the quality of clustering, in which higher values indicate *better* defined clusters, taking values between -1 and +1. The interpretation is as follows: - Values close to -1 indicate an incorrect grouping. - Values close to zero indicate *overlapping* clusters. - Values close to +1 indicate *highly dense clusters.* It is calculated considering: - a: Average of the distance between a sample and all other points of the same class (*cluster*) - b: Distance between the sample and all points of the next closest *cluster*. so s = (b-a)/max(a,b)
Binary Cross-Entropy	It is an objective function, which measures the performance of a classification model whose output is a probability value between 0 and 1. The loss of cross-entropy increases as the predicted probability diverges from the actual label. Thus, predicting a probability of 0.012 when the actual observation tag is 1 would be bad and would result in a high loss value. A perfect model would have a logarithmic loss of 0.
Lift curve	Measure the performance of a chosen classifier against a random classifier. The curve shows the relationship between the number of cases that were predicted positive and those that are actually positive, and thus measures the performance of a chosen classifier against a random classifier. The graph is constructed with the cumulative number of cases (in descending order of probability) on the abscissa axis and the cumulative number of true positives on the ordinate axis.
Cumulative Response Curve	This curve shows the ratio of gains in total positives to the proportion of records in the test set. Thus, we can analyse what proportion of the set of tests is necessary to obtain a certain percentage of gain in the total number of positives.
DCG (Discounted Cumulative Gain)	It is used for AI systems where ranking or ordering is established, i.e. when the true score of a sample d is a discrete value on a scale that measures relevance with respect to a query q. For a given query q and the samples D = {d_1, ..., dn}, the k-th sample is considered. The G k gain measures the usefulness of this sample, while the discount Dk = 1/log(k+1) penalizes retrieved documents with a lower range. The sum of the discounted profit terms GkDk for k = 1...n is the discounted cumulative profit (DCG)
Absurd impact	It is the probability ratio of favorable results between disadvantaged and privileged groups, allowing the detection of biases or errors in accuracy between both groups, privileged and disadvantaged.

Term	Definition
Cook Distance	Cook distance is a measure used to identify outliers in regression models. Calculates the change in model predictions by deleting a specific observation. A high value in the Cook distance indicates that this observation has a great influence on the model, which may indicate that it is an outlier
KL Divergence or Kullback-Lieber Divergence	The main characteristic of KL divergence is to measure how one probability distribution differs from another. By its nature, it is a non-symmetrical metric, so it is not in itself a metric or measure of distance. Considered as an optimization approach, it can be considered to have two forms: Forward and Reverse KL.
Specificity and sensitivity	Sensitivity is the metric that evaluates a model's ability to predict true positives for each category to be predicted by our model. Specificity is the metric that assesses a model's ability to predict the true negatives of each available category. These metrics apply to any categorical model.
F-beta	The F-beta score is a metric used in the field of artificial intelligence to evaluate the performance of a machine learning model on a classification problem, taking into account both precision (accuracy) and *recall* (also known as sensitivity or true positive *rate*). This metric is a combined measure that provides a balance between accuracy and *recall*.
F1 with Weighted Average	The *Weighted Averages of F1 Score* is a metric that is used when classes are unbalanced, and you want to take into account the relative importance of each class in the overall evaluation of the model. The F1 score is a metric that combines both accuracy and *recall*. It represents the balance between the model's ability to correctly identify positive cases (*recall*) and the ability to correctly classify positive cases among all positive predictions (accuracy).
Fisher's Exact Test	Fisher's exact test is a statistical test used to determine if there is a significant association between two categorical variables in a 2x2 contingency table. It is especially useful when cell counts are low, as it does not require the assumptions of the Chi-square test.

Term	Definition
Focal loss	A focal loss function addresses class imbalance during training in tasks such as object detection. Focal loss applies a modulator term to cross-entropy loss to focus learning on misclassified examples. It's a dynamically escalated loss of cross-entropy, where the scaling factor drops to zero as confidence in the right class increases. Intuitively, this scaling factor can automatically reduce the contribution of easy examples during training and quickly focus the model on difficult examples.
Frechet Information Distance (FID)	The *Frechet Inception Distance*, or FID for short, is a metric for evaluating the quality of generated images and developed specifically to evaluate the performance of generative adversarial networks. It was proposed as an improvement over *the Inception Score* (IS).
Conditional GAN (CGAN)	Conditional GANs (GANs) are a variant of Generative Adversarial Networks (GANs) that incorporate an additional condition variable to allow for the controlled and conditional generation of synthetic data. This enables targeted data generation based on additional information supplied, extending the capabilities of GANs in various data generation applications.
Hamming Loss	It is a metric that measures the accuracy of a multi-label classification model when calculating the average rate of incorrectly classified labels. It provides an overall measure of model performance in multi-label classification, regardless of label imbalance.
Hinge embedding loss	It is a loss function used in binary classification algorithms, such as support vector machines (SVMs). Their goal is to maximize the margin between classes and penalize incorrect classifications and samples close to the margin.
Hubber loss	It's a loss function used in machine learning. Unlike other loss functions, such as Mean Squared Error (MSE), Huber Loss is less sensitive to outliers in the data. Huber Loss combines the advantages of Mean Square Error and Mean Absolute Error, adapting its behaviour according to a tolerance parameter.
Hyperband (tool)	Hyperband is a hyperparameter optimization algorithm that selectively allocates resources to different hyperparameter configurations. It employs an early elimination approach, quickly discarding low-performing configurations and allocating more resources to promising configurations, enabling faster and more efficient hyperparameter searching.
Hyperopt	It is a Python library that is used for automatic hyperparameter optimization. It provides tools and algorithms that enable efficient hyperparameter search in machine learning.

Term	Definition
Inception Score (IS)	It is a metric used to evaluate the quality and diversity of images generated by imaging models, such as GANs. It is based on image classification using a pre-trained Inception network and combines diversity and quality measures to provide a score that indicates the overall quality of the images generated.
Fowlkes-Mallows Index	It is a metric used to evaluate the quality of the groupings obtained in data analysis. It is based on the comparison of real tags and tags assigned by a clustering algorithm and calculates a value that indicates the similarity between the clusters. A value closer to 1 indicates greater similarity, while a value of 0 indicates complete differences between the groupings.
Jaccard Index (also IoU)	It is a metric that measures the similarity between two sets or regions. It is used to calculate the overlap between sets by dividing the size of the intersection by the size of the joint. It is widely used in the field of computer vision and other areas where similarity or match between sets needs to be measured.
Rand Index	It is a metric used to assess the similarity between two groupings. Calculates the proportion of data pairs that are mapped in the same way in both pools. A higher Rand Index indicates greater similarity between clusters, while a lower Rand Index indicates lower similarity or random assignment of clusters.
Adjusted Mutual Information	Adjusted mutual information is a metric used in clustering analysis to measure the similarity between two sets of labels. This metric adjusts mutual information to correct for chance, providing an estimate of the dependency between labels without inflating when the number of clusters is high.
Kruskal-Wallis	It is a nonparametric test used to determine if there are significant differences between the medians of two or more independent groups. It is based on comparing the ranges of the data and uses a test statistic to evaluate the null hypothesis of equality of medians.
Language Model Probability	Language model probability is a metric that assesses how likely a language model is to generate a specific sequence of words. This metric is based on the conditional probabilities of each word given its previous context and measures both the consistency and naturalness of the sequences generated.
LogCosh (hyperbolic cosine logarithm)	It is a soft loss function used in optimization problems in machine learning. It is less sensitive to outliers and seeks to minimize the difference between predictions and true values in a stable manner.

Term	Definition
Mean Average Error (MAE)	It is an evaluation metric that measures the average of the absolute differences between predictions and actual values. It is a commonly used measure in regression problems to evaluate the accuracy of a machine learning model.
Mean Absolute Percentage Error (MAPE)	It is one of the most widely used KPIs to measure forecast accuracy. ASM is the sum of individual absolute errors divided by demand (each period separately). It is the average of the percentage errors.
Margin Ranking Loss	It is a loss function used in classification and peer learning problems. Its goal is to maximize the difference between the scores of similar and non-similar items to train models that can effectively classify items or compare pairs.
Confusion matrix	The confounding matrix is a table that shows the predictions of a classification model compared to the actual values. It has four cells: true positives, false positives, false negatives, and true negatives. This matrix allows you to evaluate the performance of the model and calculate metrics such as accuracy, completeness, and F1 score. It is a useful tool for understanding how the model classifies different classes and improving its accuracy.
Contingency matrix	A contingency matrix is a table that shows the joint distribution of two or more categorical variables. It allows you to visualize the relationship and frequencies of each combination of categories. It is useful for analysing associations between variables and revealing patterns in the data
McNemar	It is a non-parametric test for paired nominal data. It is used when you are interested in finding a change in the proportion of the paired data. For example, you could use this test to look at retrospective case-control studies, in which each treatment is paired with a control. It could also be used to look at an experiment in which two treatments are given to paired pairs. This test is sometimes called McNemar's Chi-square test because the test statistic has a Chi-square distribution.
Mean Squared Error (MSE or L2 loss)	It is an evaluation metric that measures the average of the squared differences between predictions and actual values in a regression problem. It is widely used and penalizes large errors more but can be sensitive to the scale of the data.
Mean Squared Logarithmic Error	It is an evaluation metric used in regression problems. Calculates the squared difference between the logarithms of predictions and the logarithms of actual values. It is useful for penalizing errors more at the extremes and when you want to focus on relative accuracy instead of absolute accuracy .

Term	Definition
METEOR	*Metric for Evaluation of Translation with Explicit Ordering*, is a translation evaluation metric that takes into account both the fluency of the text and the semantic correspondence between the translation and the reference.
Completeness Metric	Also known as *recall*, it measures a model's ability to identify all the positive elements in a dataset. It is calculated as the ratio of true positives to the sum of true positives and false negatives. High completeness indicates a good ability to identify positive elements but should be considered in conjunction with other metrics to evaluate the overall performance of the model.
Min-max Loss	It is a loss function used in sorting problems. Penalizes incorrect predictions by maximizing the difference between the probability assigned to the correct class and the maximum probability between the incorrect classes. Its goal is to improve the model's ability to distinguish between classes and reduce classification errors.
Vector Space Model (VSM)	It is an approach used in natural language processing to represent and analyse text documents. Documents are represented as vectors in a multidimensional space, where each dimension corresponds to a term or a characteristic of the text. This allows you to calculate similarities or distances between documents and perform various text analysis tasks.
NIST LRE	NIST *Language Recognition Evaluation*, the goal of this metric is to establish a baseline measure for spoken language recognition performance and ability over the phone.
Non saturating GAN loss	It is used in GANs to train the imager and overcome saturation and gradient fading issues. It provides an effective alternative to classic GAN loss and has proven useful in training more stable, high-quality generators.
Optuna (tool)	It is a hyperparameter optimization library that is used to automate and facilitate the search for the optimal configuration of the hyperparameters of a machine learning model.
Paired student T-Test	It is a statistical test used to compare the means of two sets of related data. It is used to determine whether there is a significant difference between the two datasets and is based on the assumption of normal distribution of paired data.
Plotly sunburst plot (tool)	It is an interactive visual representation that shows the hierarchical structure of the data in the form of concentric rings. Each ring represents a category or subcategory, and the size of the portions in the chart reflects a specific metric. It is a useful tool for visualizing hierarchical data and exploring its distribution intuitively.

Term	Definition
Pointwise Methods, Pairwise Methods, Listwise Methods	All *Learning to Rank* models use a base model of learning to calculate s = f (x). The choice of loss function is the distinguishing element for *Learning to Rank models*. In general, we have 3 approaches, depending on how the loss is calculated. *Pointwise Methods*: Total loss is calculated as the sum of the loss terms defined in each di document (therefore punctually) as the distance between the predicted score si and the fundamental truth yi, for i = 1... By doing this, we transform our task into a regression problem, where we train a model to predict and. *Pairwise Methods*: Total loss is calculated as the sum of the loss terms defined in each pair of documents di, dj (therefore in pairs), for i, j = 1... n. The objective on which the model is trained is to predict whether yi > yj or not, i.e. which of the two documents is more relevant. By doing this, we transform our task into a binary classification problem. *Listwise Methods*: The loss is calculated directly on the entire document list (hence *listwise*) with the corresponding predicted ranges. This way, ranking metrics can be incorporated more directly into the loss.
Precision Recall (curve) PRC and Area Under PRC (AUPRC)	Precision-recall (PRC) is a metric that evaluates the performance of a classification model in terms of accuracy and *recall* as the classification threshold is varied. The PRC curve plots these values, and the area under the curve (AUPRC) summarizes the overall performance of the model.
Predictive parity	An equity metric that checks whether, for a given classifier, the accuracy indices are equivalent for the subgroups considered. For example, a model that predicts college acceptance would satisfy predictive parity for nationality if its accuracy index is the same regardless of nationality of origin.
R2	R^2 is a statistical metric that measures the proportion of variation in the dependent variable that is explained by the regression model. It indicates how well the data fits the model, with values from 0 to 1, where a value closer to 1 suggests a better fit.
R2 Adjusted	It is a corrected measure of the goodness of fit of a linear model. It is used to evaluate the accuracy of the model and determine the percentage of variance in the dependent variable that is explained by the independent variables. R2 tends to overestimate the fit of the linear regression model and always increases as more variables are included in the model. Adjusted R2 attempts to correct for this overestimation and may decrease if the inclusion of a variable does not improve the model.

Term	Definition
Wilcoxon Signed Ranking	It is a nonparametric statistical test used to compare two paired samples and determine if there is a significant difference between them. It is a useful alternative when the data do not meet the assumptions of normality or when working with ordinal scales.
Regression	Regression is a statistical and supervised learning technique that seeks to model the relationship between a continuous dependent variable and one or more independent variables. Regression is used to make predictions and analyse how independent variables influence the dependent variable.
Reweighting	It is a technique that minimizes bias, adjusts the weights or probabilities of observations in a dataset to address imbalances or improve the representativeness of certain classes or instances.
RMSE	It is a metric that quantifies the average error between the predictions of a regression model and the actual values of the dataset. It is used to evaluate and compare the accuracy of different models, with the lowest possible RMSE value being desirable.
ROC (curve) AND Area under ROC (AUROC or AUC)	The ROC curve is a graphical representation that shows the relationship between the true positive rate and the false positive rate as the classification threshold is varied. The AUROC is a metric that summarizes the quality of the ROC curve and provides a quantitative measure of the performance of the binary classification model. The higher the AUROC value, the better the performance of the model.
Cosine similarity	It is a measure that evaluates the similarity between two vectors based on the angle between them. It is a measure commonly used in various applications to compare the similarity between texts, documents, user profiles, and other types of data represented in vector form.
Triplet Margin Loss	Also known as *Triplet Loss*, it is a loss function used in machine learning to learn representations of data where similar instances are closer to each other and different instances are further apart. It uses data triplets and is based on the comparison of distances in the representation space.
Unequal outcomes	This refers to situations in which *machine learning* models produce uneven or biased results for different groups or individuals.
Calinsk and Harabaz Valor	This index is a measure used in the field of statistics and machine learning to assess the quality of a grouping or *clustering* of data. It is based on the idea that good *clustering* should have high *intra-cluster* cohesion (the points within each group are similar) and low *inter-cluster separation* (groups are well differentiated from each other). The higher the index value, the better the quality of the *clustering*.

Term	Definition
F1 Value	The F1 score is a statistical measure that combines the accuracy and recall of a classification model. It is useful when you have unbalanced classes in your data. It provides a single metric that represents accuracy and recall in a balanced manner, and its value ranges from 0 to 1, with 1 being the best possible result. A high F1 score indicates a balance between accuracy and model accuracy in the class rankings.
Wasserstein GAN loss	This is one of the most powerful alternatives to the original GAN loss. It addresses the problem of mode collapse and gradient disappearance. In this implementation, the activation of the discriminator's output layer is changed from sigmoid to linear. This simple change influences the discriminator to output a score rather than a probability associated with the distribution of the data, so the output doesn't have to be in the range of 0 to 1.
Word Accuracy	Word Accuracy measures the fraction of words that a word recognition or processing system correctly classifies or identifies, compared to a reference. It is a metric that evaluates overall performance in accurately identifying words without considering accuracy or recall.
Word Error Rate	It is a measure that quantifies the accuracy of automatic speech recognition by comparing the transcription generated by the system to a reference transcript, providing a way to evaluate and compare the quality of different speech recognition systems.

8. References, Standards & Norms

8.1 References

8.1.1 General references

[1] Hullermeier et al. 2019 Randomic and Epistemic Uncertainty in Machine Learning: A Tutorial Introduction.

[2] Calibration of Machine Learning Models. Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. Department of Computer Systems and Computing, Polytechnic University of Valencia 2010

[3] Metrics for Deep Generative Models N Chen · 2018 ·

[4] Language Models are Few-Shot Learners T, Brown et al., 2020

[5] Physically Consistent Generative Adversarial Networks for Coastal Flood Visualization Bjorn Lutjens et al. 2021

[6] IBM Uncertainty Quantification 360 Toolkit https://uq360.mybluemix.net

[7] Questioning causality on sex, gender, and COVID-19, and identifying bias in large-scale data-driven analyses: the Bias Priority Recommendations and Bias Catalog for Pandemics. Díaz-Rodríguez et al. 2021 https://arxiv.org/abs/2104.14492

[8] https://catalogofbias.org

[9] A survey on datasets for fairness-aware machine learning Tai Le Quy∗1, Arjun Roy†12, Vasileios Iosifidis‡1, Wenbin Zhang§3, and Eirini Ntoutsi 2022

[10] Datasheets for Datasets. Timnit Gebru et al. 2021

[11] "Searching for a Search Method: Benchmarking Search Algorithms for Generating NLP Adversarial Examples",

[12] EMNLP 2020 Blackbox NLP Workshop track proceedings. https://github.com/QData/TextAttack

[13] TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP, J Morris et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020 [In Python]

[14] PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries K Kaczmarek-Majer, G Casalino, G Castellano... - Information ..., 2022 - Elsevier

[15] Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? Paul B. de Laat Philosophy & Technology volume 34, pages 1135–1193 (2021)

Note: SHAP is not from Amazon nor proprietary in Table 3 (SHAP is Scott Lundberg's creation with a MIT open source license, in Github, and was developed with Microsoft Research teams. AzureML also implements XAI algorithms and Amazon has as ML tool, Amazon SageMaker).

[16] COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity 2016

[17] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learn Research), Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New Ing

[18] An Ontology for Fairness Metrics. Franklin et al. https://dl.acm.org/doi/pdf/10.1145/3514094.3534137

[19] Human-centred artificial intelligence https://scilog.fwf.ac.at/en/environment-and-technology/15317/human-centred-artificial-intelligence

[20] A Holzinger et al. Digital Transformation in Smart Farm and Forest Operations Needs Human-Centered AI: Challenges and Future Directions

[21] Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics, 3, (2), 119-131, doi:10.1007/s40708-016-0042-6.

[22] Holzinger, A., Malle, B., Kieseberg, P., Roth, P.M., Müller, H., Reihs, R. & Zatloukal, K. 2017. Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. arXiv:1712.06657.

[23] Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R. & Zatloukal, K. 2017. Machine Learning and Knowledge Extraction in Digital Pathology needs an integrative approach. Springer Lecture Notes in Artificial Intelligence Volume LNAI 10344. Cham: Springer International, pp. 13-50. doi: 10.1007/978-3-319-69775-8_2

[24] Human-Centered Artificial Intelligence for Designing Accessible Cultural Heritage G Pisoni, N Díaz-Rodríguez, H Gijlers, L Tonolli Applied Sciences 11 (2), 870

[25] Continual Learning for Robotics: Definition, Framework, Learning Strategies, Opportunities and Challenges T Lesort, V Lomonaco, A Stoian, D Maltoni, D Filliat, N Díaz-Rodríguez Information Fusion 220

[26] 2020 A survey on ontologies for human behavior recognition ND Rodríguez, M.P. Cuéllar, J., Lilius, M.D. Calvo-Flores ACM Computing Surveys (CSUR) 46(4), 1-33 219 2014 A fuzzy ontology for semantic modelling and recognition of human behaviour ND Rodríguez, MP Cuéllar, J Lilius, MD Calvo-Flores Knowledge-Based Systems 66, 46-60 133 201

[27] Explainability in Deep Reinforcement Learning A Heuillet, F Couthouis, N Díaz-Rodríguez Knowledge-Based Systems 214, 106685 119 2021

[28] Don't forget, there is more than forgetting: new metrics for Continual Learning N Díaz-Rodríguez, V Lomonaco, D Filliat, D Maltoni NeurIPS workshop on Continual Learning 2018

[29] Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence A Holzinger, M Dehmer, F Emmert-Streib, R Cucchiara, I Augenstein, ... Information Fusion 79, 263-278 2022

[30] Personas for Artificial Intelligence (AI) An Open-Source Toolbox A Holzinger, M Kargl, B Kipperer, P Regitnig, M Plass, H Müller IEEE Access 10, 23732-23747 2022

[31] Measuring the quality of explanations: the system causability scale (SCS) A Holzinger, A Carrington, H Müller KI-Künstliche Intelligenz 34 (2), 193-19

[32] Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities R Kusters, D Misevic, H Berry, A Cully, and Le Cunff, L Dandoy, ... Frontiers in Big Data 3, 45 2020

[33] S. Stanton, P. Izmailov, P. Kirichenko, A. A. Alemi, A. G. Wilson, Does knowledge distillation really work? (2021). doi:10.48550/ARXIV.2106.05945. URL https://arxiv.org/abs/2106.05945

[34] A Neural-Symbolic learning framework to produce interpretable predictions for image classification, PhD Thesis 2022

[35] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. AB Arrieta, N Díaz-Rodríguez, J Del Ser, A Bennetot, S Tabik, A Barbado, ... Information Fusion 58, 82-115

[36] Wilcoxon, Frank. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945, 1 (6), 1095 pages 80-83.

[37] Dietterich et al. Approximate Statistical Tests for Comparing Supervised Classification 1092 Learning Algorithms. Neural Computation, Volume 10, Issue 7. 1998, 10 (7), pages 1895-1923. 1093 https://doi.org/10.1162/089976698300017197

[38] Akenine-Möller, Tomas, and Johnsson, Björn. Performance per what? Journal of Computer Graphics 1073 Techniques. 2012, 1, pages 37-41. http://jcgt.org/published/0001/01/03/paper.pdf

[39] Blouw, Peter and Xuan Choo and Hunsberger, Eric and Eliasmith, Chris. Benchmarking keyword 1075 spotting efficiency on neuromorphic hardware. Proceedings of the 7th Annual Neuro-inspired 1076 Computational Elements Workshop. 2019, pages 1-8. https://arxiv.org/pdf/1812.01739.pdf

[40] Suffering-focused AI safety: In favor of "fail-safe" measures Lukas Gloor Center on Long-Term Risk Report

[41] Superintelligence as a Cause or Cure for Risks of Astronomical Suffering

[42] Kaj Sotala and Lukas Gloor Foundational Research Institute, Berlin, Germany Superintelligence as a Cause or Cure ... Informatica 41 (2017) 389–400

[43] Safe Deep RL in 3D environments using human feedback. 2022.

[44] Safeguard By Design Lessons Learned from DOE Experience Integrating Safety in Design

[45] Sustainability at scale: towards bridging the intention-behavior gap with sustainable recommendations S Tomkins, S Isley, B London, L Getoor - Proceedings of the 12th ACM conference on ..., 2018

[46] Green AI. Schwartz et al.

[47] Distilling the Knowledge in a Neural Network

[48] EVALUATION METRICS FOR LANGUAGE MODELS Stanley Chen, Douglas Beeferman, Ronald Rosenfeld.

[49] Chip Huyen, "Evaluation Metrics for Language Modeling," The Gradient, 2019. https://thegradient.pub/understandSing-evaluation-metrics-for-language-models/

[50] Model cards for model reporting. M Mitchell, S Wu, A Zaldivar, P Barnes, L Vasserman... - Proceedings of the ..., 2019

[51] Frank McSherry Materialize: a platform for building scalable event based systems

[52] Frank McSherry, Kunal Talwar Mechanism design via Differential Privacy, 2008

[53] Nobel Prize Report, Mechanism Design. 2007 Scientific background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2007 Mechanism Design Theory, based in:

[54] L. Hurwicz & S. Reiter (2006) Designing Economic Mechanisms, p. 30

[55] https://divedeep.ai/2022/03/17/data-drift-vs-concept-drift/

[56] University of Oxford researchers have created a tool called capAI, a procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act. CapAI provides organizations with practical guidance on how to translate high-level ethics principles into verifiable criteria that help shape the design, development, deployment and use of ethical AI. This tool can be used to demonstrate that the development and operation of an AI system are trustworthy. The tool is being validated with firms at the moment and the most up-to-date version can be found here

[57] A survey on concept drift adaptation ACM computing surveys (CSUR), 46(4):1–37, 2014, Gama et al.

[58] Analysis of representations for domain adaptation Neurips 2007, Ben-David et al

[59] Dataset Shift in Machine Learning Quiñonero-Candela et al 2022

[60] Generalized out-of-distribution detection: A survey. Yang et al 2022

[61] Understanding Continual Learning Settings with Data Distribution Drift Analysis" Lesort et al. 2022 video: https://www.youtube.com/watch?v=WFhozvAgnsU

[62] Sagemaker Clarify: Amazon AI Fairness and Explainability Whitepaper https://pages.awscloud.com/rs/112-TZM-766/images/Amazon.AI.Fairness.and.Explainability.Whitepaper.pdf https://aws.amazon.com/blogs/machine-learning/learn-how-amazon-sagemaker-clarify-helps-detect-bias/

[63] Data Privacy and Trustworthy Machine Learning. Strobel et al. 2022

[64] Gradual (In)Compatibility of Fairness Criteria Corinna Hertweck, Tim Räz 2022 https://arxiv.org/abs/2109.04399

[65] Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat ICLR 2019 Workshop on Structure & Priors in Reinforcement Learning (SPIRL) 2019

[66] Continual reinforcement learning deployed in real-life using policy distillation and sim2real transfer RT Kalifou, H Caselles-Dupré, T Lesort, T Sun, N Diaz-Rodriguez, D Filliat ICML Workshop on Multi-Task and Lifelong Learning 2019

[67] S-RL Toolbox: Environments, Datasets and Evaluation Metrics for State Representation Learning

[68] A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat NeurIPS workshop on Deep Reinforcement Learning 2018

[69] Deep Unsupervised state representation learning with robotic priors: a robustness analysis

[70] T Lesort, M Seurin, X Li, N Díaz-Rodríguez, D Filliat. 2019 International Joint Conference on Neural Networks (IJCNN) 2017

[71] Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence A Holzinger, M Dehmer, F Emmert-Streib, R Cucchiara, I Augenstein, et al. Information Fusion.

[72] Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study Zech et al 2018 (extended from Confounding variables can degrade generalization performance of radiological deep learning models. Zech et al. 2018.)

[73] ISO/IEC 25000, Systems and software engineering – Systems and software Quality 937 Requirements and Evaluation (SQuaRE) and ISO/IEC WD 25059:2021, Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) Quality Model for AI systems

[74] STRIDE-AI: An Approach to Identifying Vulnerabilities of Machine Learning Assets Lara Mauri et al. IEEE CSR 2021

[75] Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks Papernot 2016

[76] Steps Toward Robust Artificial Intelligence Thomas G. Dietterich 2017

[77] Improving the Robustness of Deep Neural Networks via Stability Training

[78] SECURING MACHINE LEARNING ALGORITHMS. December 2021 ANNEX D: REFERENCES by input data type and lifecycle stages.

[79] Towards Resilient Artificial Intelligence: Survey and Research Issues

[80] The Robustness of Counterfactual Explanations Over Time, A Ferrario et al.

[81] Research priorities for robust and beneficial artificial intelligence.

[82] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. 2016.

[83] Evaluating Robustness of Counterfactual Explanations Artelt et al 2021

[84] Exploring the Trade-off between Plausibility, Change Intensity and Adversarial Power in Counterfactual Explanations using Multi-objective Optimization. Javier Del Ser et al. 2020

[85] Towards Human-Compatible XAI: Explaining Data Di erentials with Concept Induction over Background Knowledge. Widmer et al 2022

[86] A survey on bias in visual datasets 2022. Fabrizzi et al.

[87] Omitted variable bias: A threat to estimating causal relationships. Wilms et al.

[88] Google. Machine Learning Glossary: Fairness. 2021 [cited 29 November, 2021]; available from: https://developers.google.com/machine-learning/glossary/fairness.

[89] David Lopez-Paz, Krikamol Muandet, Bernhard Scholkopf, and Ilya O. Tolstikhin. Towards a learning theory of cause-e ect inference. In Francis R. Bach and David M. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille,

France, 6-11 July 2015, volume 37 of JMLR Workshop and Conference Proceedings, pages 1452{1461. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/ lopez-paz15.html.

[90] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schlkopf, and Leon Bottou. Discovering causal signals in images. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 58{ 66, 2017. doi: 10.1109/CVPR.2017.14.

[91] ICO, Guidance on the AI auditing framework: draft guidance for consultation. Information [92] Commissioner's Office, 2020.

[93] PwC, PwC Ethical AI Framework. 2020.

[94] Deloitte, Deloitte introduces trustworthy AI framework to guide organizations in ethical application of technology. August 26, 2020. New York.

[95] Orcaa, It's the age of the algorithm and we have arrived unprepared. 2020.

[96] Epstein, Z., et al., Turingbox: an experimental platform for the evaluation of AI systems. IJCAI International Joint Conference on Artificial Intelligence, 2018. 2018-July: p. 5826-5828. #Discontinued.

[97]. Shi et al. Robustness Verification for Transformers. International Conference on Learning Representations. 2020. arXiv:2002.06622

[98] Incremental Bounded Model Checking of Artificial Neural Networks in CUDA Luiz H. Sena et al.

[99] A Raffin, A Hill, R Traoré, T Lesort, N Díaz-Rodríguez, D Filliat

[100] NeurIPS workshop on Deep Reinforcement Learning 2018

[101] Stable-Baselines3 Reliable Reinforcement Learning Implementations https://stable-baselines3.readthedocs.io/en/master/

[102] T Lesort, M Seurin, X Li, N Díaz-Rodríguez, D Filliat 2019 International Joint Conference on Neural Networks (IJCNN) 2017

[103] Error Analysis tool, part of the Responsible AI Dashboard in Azure: https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard

[104] Evolved from "Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure besmira Nushi Ece Kamar Eric Horvitz HCOM 2018.

[105] Deep Reinforcement Learning that Matters - P Henderson · 2017

[106] L. Hurwicz & S. Reiter (2006) Designing Economic Mechanisms,

[107] Facial Recognition: Analyzing Gender and Intersectionality in Machine Learning. Report. http://genderedinnovations.stanford.edu/case-studies/facial.html#tabs-2

[108] Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights. 2018

[109] A.S. Ross, M.C. Hughes, F. Doshi-Velez. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. IJCAI'17.

[110] K Burns, L. A. Hendricks, K Saenko, T Darrell, A Rohrbach. Women also Snowboard: Overcoming Bias in Captioning Models. ECCV'18, 771-787.

[111] A Rohrbach, L.A. Hendricks, K Burns, T Darrell, K Saenko. Object Hallucination in Image Captioning. EMNLP'18.

[112] A Gulati, MA Lozano, B Lepri, N Oliver. BIASeD: Bringing Irrationality into Automated System Design

[113] J Gardner, C Brooks, R Baker. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis

8.1.2 Glossary References

For the glossary, the following references have been used, which can be consulted to expand on the aspects described therein.

- https://homes.cs.washington.edu/~jpgard/papers/lak19_slicing.pdf,
- https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf
- https://www.tibco.com/reference-center/what-is-analysis-of-variance-anova
- https://medium.com/sfu-cspmp/model-transparency-fairness-552a747b444
- https://www.statology.org/balanced-accuracy/
- https://towardsdatascience.com/baseline-models-your-guide-for-model-building-1ec3aa244b8d
- https://arxiv.org/abs/2210.01122
- https://www.investopedia.com/terms/c/central_limit_theorem.asp
- https://machinelearningmastery.com/types-of-classification-in-machine-learning/
- https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms
- https://jdvelasq.github.io/courses/notebooks/sklearn_unsupervised_03_clustering/1-03_metodo_de_la_silueta.html
- https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html
- https://orange3.readthedocs.io/en/3.5.0/widgets/evaluation/liftcurve.html
- https://rstudio-pubs-static.s3.amazonaws.com/577248_f94b111668f546e896649e408011969d.html
- https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4
- https://medium.com/sfu-cspmp/model-transparency-fairness-552a747b444
- https://keepcoding.io/blog/que-es-la-distancia-de-cook/
- https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-use-kullback-leibler-divergence-kl-divergence-with-keras.md
- https://towardsdatascience.com/evaluating-categorical-models-ii-sensitivity-and-specificity-e181e573cff8#:~:text=Sensitivity%20is%20the%20metric%20that,negatives%20of%20each%20available%20category.
- https://machinelearningmastery.com/fbeta-measure-for-machine-learning/
- https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f
- https://www.statology.org/fishers-exact-test/

- https://paperswithcode.com/method/focal-loss#:~:text=Focal%20loss%20applies%20a%20modulating,in%20the%20correct%20class%20increases.
- https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/#:~:text=for%20Real%20Images-,What%20Is%20the%20Frechet%20Inception%20Distance%3F,performance%20of%20generative%20adversarial%20networks.
- https://machinelearningmastery.com/how-to-develop-a-conditional-generative-adversarial-network-from-scratch/
- https://www.linkedin.com/pulse/hamming-score-multi-label-classification-chandra-sharat/
- https://medium.com/udacity-pytorch-challengers/a-brief-overview-of-loss-functions-in-pytorch-c0ddb78068f7
- https://www.cantorsparadise.com/huber-loss-why-is-it-like-how-it-is-dcbe47936473
- https://arxiv.org/abs/1603.06560
- https://github.com/hyperopt/hyperopt
- https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a
- http://sedici.unlp.edu.ar/bitstream/handle/10915/76139/Documento_completo.pdf?sequence=1
- https://deepai.org/machine-learning-glossary-and-terms/jaccard-index
- https://towardsdatascience.com/7-evaluation-metrics-for-clustering-algorithms-bdc537ff54d2
- https://cloud.google.com/vertex-ai/docs/tabular-data/tabular-workflows/feature-engineering?hl=es-419#:~:text=combina%20los%20resultados.-,Informaci%C3%B3n%20mutua%20ajustada%20(AMI),si%20se%20share%20m%C3%A1s%20informaci%C3%B3n.
- https://deepai.org/machine-learning-glossary-and-terms/kruskal-wallis-test
- https://medium.com/ingeniouslysimple/language-models-15e45dce0805
- https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-use-logcosh-with-keras.md
- https://medium.com/@20__80__/mean-absolute-error-mae-machine-learning-ml-b9b4afc63077
- https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d
- https://pykeen.readthedocs.io/en/stable/api/pykeen.losses.MarginRankingLoss.html
- https://keepcoding.io/blog/medidas-de-calidad-en-matrices-de-confusion/
- https://keepcoding.io/blog/tipos-tests-estadisticos-para-big-data/
- https://statologos.com/prueba-de-mcnemar/
- https://www.britannica.com/science/mean-squared-error
- https://insideaiml.com/blog/MeanSquared-Logarithmic-Error-Loss-1035
- https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall
- https://neptune.ai/blog/gan-loss-functions
- https://en.wikipedia.org/wiki/Vector_space_model
- https://arxiv.org/abs/2010.08029
- https://optuna.org/
- https://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf
- https://www.geeksforgeeks.org/sunburst-plot-using-plotly-in-python/

- https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4
- Precision-Recall Curves. Sometimes a curve is worth a thousand... | by Doug Steen | Medium
- https://developers.google.com/machine-learning/glossary/fairness#predictive-parity
- https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=terms-r2
- https://www.ibm.com/docs/es/cognos-analytics/11.2.0?topic=terms-adjusted-r-squared
- https://www.cienciadedatos.net/documentos/18_prueba_de_los_rangos_con_signo_de_wilcoxon
- https://economipedia.com/definiciones/analisis-de-regresion.html
- https://towardsdatascience.com/fairmodels-lets-fight-with-biased-machine-learning-models-f7d66a2287fc
- https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d
- https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
- https://www.researchgate.net/publication/360499809_SHARP-GAN_SHARPNESS_LOSS_REGULARIZED_GAN_FOR_HISTOPATHOLOGY_IMAGE_SYNTHESIS
- https://ieeexplore.ieee.org/document/9761534 https://medium.com/beyondminds/advances-in-generative-adversarial-networks-7bad57028032
- https://keepcoding.io/blog/similitud-entre-vectores-o-cosine-similarity/
- https://towardsdatascience.com/triplet-loss-advanced-intro-49a07b7d8905
- https://cristhianhenry57.medium.com/clustering-6adbfae73ded
- https://towardsdatascience.com/the-f1-score-bec2bbc38aa6
- https://neptune.ai/blog/gan-loss-functions
- https://www.wonderflow.ai/blog/what-is-accuracy-in-text-analysis
- https://huggingface.co/spaces/evaluate-metric/wer

8.2 Standards

Concepts and terms

- [114] ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology, https://www.iso.org/standard/74296.html
- [115] ISO/IEC 23053:2022, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), https://www.iso.org/standard/74438.html

Precision/ Accuracy. Some of these standards have already been published, while others are currently under development or revision.

- [116] ISO/IEC TS 4213:2022, Information technology — Artificial intelligence — Assessment of machine learning classification performance, https://www.iso.org/standard/79799.html
- prEN 18229-2 AI trustworthiness framework – Part 2: Accuracy and robustness
- prEN Evaluation methods for accurate computer vision systems
- prEN ISO/IEC 23282 Evaluation methods for accurate natural language processing systems

Design and Development

- [117] ISO/IEC DIS 5338, Information technology – Artificial intelligence – AI system life cycle processes, https://www.iso.org/standard/81118.html
- [118] ISO/IEC DIS 5339, Information technology – Artificial intelligence – Guidance for AI applications, https://www.iso.org/standard/81120.html
- [119] ISO/IEC DIS 5392, Information technology – Artificial intelligence – Reference architecture of knowledge engineering, https://www.iso.org/standard/81228.html
- [120] ISO/IEC TR 24372:2021, Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems, https://www.iso.org/standard/78508.html
- [121] ISO/IEC CD TS 12791, Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks, https://www.iso.org/standard/84110.html

Financiado por
la Unión Europea
NextGenerationEU