ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 71484.2— 2024 (ИСО/МЭК 5259-2:2024)

Искусственный интеллект

КАЧЕСТВО ДАННЫХ ДЛЯ АНАЛИТИКИ И МАШИННОГО ОБУЧЕНИЯ

Часть 2

Показатели качества данных

(ISO/IEC 5259-2:2024, MOD)

Издание официальное

Москва
Российский институт стандартизации
2024

Предисловие

1 ПОДГОТОВЛЕН Федеральным государственным бюджетным образовательным учреждением высшего образования «Московский государственный университет имени М.В.Ломоносова» (МГУ имени М.В. Ломоносова) в лице Научно-образовательного центра компетенций в области цифровой экономики МГУ и Обществом с ограниченной ответственностью «Институт развития информационного общества» (ИРИО) на основе собственного перевода на русский язык англоязычной версии стандарта, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 28 октября 2024 г. № 1551-ст

4 Настоящий стандарт является модифицированным по отношению к международному стандарту ИСО/МЭК 5259-2:2024 «Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 2. Показатели качества данных» (ISO/IEC 5259-2:2024 «Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures», MOD) путем изменения отдельных фраз (слов, значений, показателей, ссылок), которые выделены в тексте курсивом.

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте, приведены в дополнительном приложении ДА

5 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

Содержание

Введение

Ввиду того, что сегодня решения все активнее принимаются на основе данных, встают новые задачи по управлению качеством данных в области аналитики и искусственного интеллекта на основе машинного обучения. Проблемы с качеством данных, такие как неполные, ложные или устаревшие данные, могут отрицательно повлиять на процессы и результаты аналитики и машинного обучения. Данные из различных источников, включая структурированные данные (например, содержащиеся в реляционных базах данных) и неструктурированные данные (например, документы, изображения, аудио), могут быть напрямую использованы в жизненном цикле данных для аналитики и разработки моделей машинного обучения. Данные преобразуются на каждом этапе жизненного цикла данных аналитики и машинного обучения. Чтобы анализ данных и модели машинного обучения были безопасными, надежными и совместимыми, необходим целостный стандартизированный подход к контролю, производству и поставке достаточного количества высококачественных данных. Для разработки надежного управления качеством данных для аналитики и машинного обучения можно рассмотреть внутренние международные стандарты качества данных, включая концепции и варианты использования, характеристики и измерения, требования к управлению и структуру процессов.

Настоящий стандарт является частью серии ИСО/МЭК 5259 и основан на серии стандартов ИСО 8000, ИСО/МЭК 25012 и ИСО/МЭК 25024. Целью настоящего стандарта является описание модели качества данных посредством определения характеристик качества данных и показателей качества данных на основе ИСО/МЭК 25012 и ИСО/МЭК 25024. Модели качества данных могут быть расширены или изменены в соответствии с настоящим стандартом.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Искусственный интеллект

КАЧЕСТВО ДАННЫХ ДЛЯ АНАЛИТИКИ И МАШИННОГО ОБУЧЕНИЯ

Часть 2

Показатели качества данных

Artificial intelligence. Data quality for analytics and machine learning.
Part 2. Data quality measures

Дата введения — 2025—01—01

1 Область применения

В настоящем стандарте представлена модель качества данных, показатели качества данных и рекомендации по составлению отчетов о качестве данных для аналитики и машинного обучения.

Документ применим для всех типов организаций, которые хотят достичь своих целей в области качества данных.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ Р 70889—2023 (ИСО/МЭК 8183:2023) Информационные технологии. Искусственный интеллект. Структура жизненного цикла данных

ГОСТ Р 71476 (ИСО/МЭК 22989:2022) Искусственный интеллект. Концепции и терминология искусственного интеллекта

ГОСТ Р 71484.1—2024 (ИСО/МЭК 5259-1:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры

ГОСТ Р 71484.3 (ИСО/МЭК 5259-3:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 3. Требования и рекомендации по управлению качеством данных

ГОСТ Р 71484.4 (ИСО/МЭК 5259-4:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 4. Структура процесса управления качеством данных

ГОСТ Р ИСО/МЭК 25000 Требования и оценка качества систем и программных средств (SQuaRE). Руководство

ГОСТ Р ИСО/МЭК 25010 Информационные технологии. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Модели качества систем и программных продуктов

ГОСТ Р ИСО/МЭК 25020—2023 Системная и программная инженерия. Требования и оценка качества систем и программной продукции (SQuaRE). Основные принципы измерения качества

ГОСТ Р ИСО/МЭК 29100 Информационная технология. Методы и средства обеспечения безопасности

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указа-

Издание официальное

телю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающая положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

3 Термины и определения

В настоящем стандарте применены термины по *ГОСТ Р 71476*, а также следующие термины с соответствующими определениями:

3.1

данные (data): Представление информации в формальном виде, пригодном для передачи, интерпретации или обработки.

Примечание — Данные могут быть обработаны автоматически или вручную.

[*ГОСТ Р ИСО/МЭК 20546—2021, пункт 3.1.5*]

3.2 **фрейм данных** (data frame): Множество записей данных с общей структурой элементов данных, связанных с определенной предметной областью или предназначением.

Примечание — Фрейм данных является двумерным, как таблица со строками и столбцами. Этот термин специально используется в аналитике и машинном обучении, например, в языке R, в то время как в других языках термин «набор данных» означает то же самое. В настоящем стандарте термин «набор данных» имеет более общее значение.

3.3 **тип данных** (data type): Категоризация абстрактного набора возможных значений, характеристик и набора операций для атрибута.

Примечания

1 Примерами типов данных являются строки символов, тексты, даты, числа, изображения, звуки и т. д.

2 *См. [1], пункт 4.16.*

3.4 **значение данных** (data value): Содержимое элемента данных.

Примечания

1 В *[2],* пункт 5.1.1 указано, что качество данных является внутренне присущей характеристикой самих данных, такой как допустимые значения данных и возможные ограничения.

2 Номер или категория, присвоенные атрибуту сущности путем проведения измерения.

3 *См. [1], пункт 4.17.*

3.5 **пустой элемент данных** (empty data item): Элемент данных, содержимое которого (3.4) имеет пустое значение, т. е. Null или None.

Примечание — Это определение в целом означает отсутствие значения данных (т. е. значение NULL или None). Элемент данных со строковым типом данных может быть пустым элементом данных, использующим либо пустую строку, либо значение Null. Однако есть исключение для некоторых приложений: строка может быть пустой (например, «»), но не нулевой и, следовательно, не подразумевает пустого элемента данных.

3.6 **сущность** (entity): Конкретная или абстрактная вещь в рассматриваемой предметной области.

3.7 **необработанные данные** (raw data): Данные в первоначально полученной, прямой форме из источника перед последующей обработкой.

Примечание — *См. [3], пункт 3.1.10.04.*

3.8 **целевые данные** (target data): Данные, используемые в задаче аналитики или машинного обучения, качество которых измеряется.

3.9 **целевая аудитория** (target population): Генеральная совокупность, в отношении которой необходимо сделать выводы в проекте аналитики данных или машинного обучения.

3.10 **предмет качества данных** (data quality subject): Сущность, на которую влияет качество данных.

3.11

элемент показателя качества (quality measure element): Показатель, определенный в терминах свойства и метода измерения для количественного определения этого свойства, включая выборочно преобразования с помощью математической функции.

[ГОСТ Р ИСО/МЭК 25021—2014, пункт 4.14]

3.12 **количество** (quantity): Свойство явления, тела или вещества, когда свойство имеет величину, которая может быть выражена количественно в виде числа с указанием отличительного признака как основы для сравнения.

Примечание — См. [4], пункт 1.1.

3.13 **значение количества** (quantity value): Число с указанием основы для сравнения, выражающее размер величины количества.

Примечание — См. [4], пункт 1.1.

3.14

функция измерения (measurement function): Алгоритм или вычисление, выполняемое для комбинации не менее чем двух элементов показателя качества.

[ГОСТ Р ИСО/МЭК 25023—2021, пункт 4.6]

3.15 **результат измерения** (measurement result, result of measurement): Набор значений количества, приписываемых измеряемой величине вместе с любой другой доступной и существенной информацией.

Примечание — См. [4], пункт 2.9.

3.16

показатель (measure): Переменная, которой присваивается какое-то значение как конкретный результат измерения.

Примечание — Форма множественного числа «показатели» используется для ссылки на основные показатели, производные показатели и индикаторы.

[ГОСТ Р 58606—2019, пункт 3.15]

3.17

измерять (measure): Производить измерение.

[ГОСТ Р ИСО/МЭК 25000—2021, пункт 4.19]

3.18 **ограничивающий прямоугольник** (bounding box): Прямоугольная область, охватывающая аннотируемый объект.

Примечания

1 Большая и малая оси прямоугольника параллельны краям изображения, в другом случае следует использовать ограничивающий многоугольник.

2 См. [5], пункт 3.3.

3.19

кластер (cluster): Автоматически группируемая категория элементов, являющихся частью набора данных и имеющих общие атрибуты.

Примечание — Наличие имен для кластеров не обязательно.

[[6], пункт 3.3.2]

3.20 **алгоритм кластеризации:** Алгоритм, группирующий кластеры (3.19) по входным данным.

Примечание — Примеры алгоритмов кластеризации включают кластеризацию на основе центроидов, кластеризацию на основе плотности, кластеризацию на основе распределения, иерархическую кластеризацию и кластеризацию на основе графов.

3.21

переобучение (overfitting): <машинное обучение> Создание модели, слишком точно соответствующей обучающим данным и не способной к обобщению при использовании новых наборов данных.

Примечания

1 Переобученность может возникнуть, если в обученной модели избыточно учтены несущественные признаки обучающих данных (т. е. признаков, обобщение которых не приводит к полезным результатам), если обучающие данные содержат много шума (например, имеют чрезмерное количество выбросов), или же модель слишком сложна для конкретного набора обучающих данных.

2 Признаком переобученности модели является значительная разница между ошибками, измеренными на обучающих данных и на отдельных тестовых и валидационных данных. На производительность переобученных моделей особенно влияет значительная разница между обучающими и эксплуатационными данными.

[[6], пункт 3.1.4]

3.22 **верность** (fidelity): Степень, в которой модель или симуляция воспроизводит состояние и поведение объекта реального мира или восприятие объекта реального мира, функции, состояния или выбранного стандарта измеримым или воспринимаемым образом.

Примечание — См. [7], пункт 3.1.4.

3.23 **сопровождаемость** (maintainability): Способность функционального блока при данных условиях использования сохраняться или восстанавливаться до состояния, в котором он может выполнять требуемую функцию, когда обслуживание выполняется в данных условиях и с использованием установленных процедур и ресурсов.

Примечание — См. [8], пункт 2123027.

3.24 **надежность** (reliability): Величина, которой измеряется оценка.

Пример — Оценка будет иметь низкую надежность, если две формы оценки имеют неодинаковую сложность или охват, или если есть ошибки в процедурах начисления баллов или в отчетности о баллах.

3.25 **достоверность** (validity): Степень, в которой оценка соответствует поставленной цели путем измерения того, что она должна измерять, и получения результатов, которые можно использовать по их прямому назначению.

Примечание — Оценка имеет низкую достоверность, если на результаты оказывают чрезмерное влияние навыки, не имеющие отношения к заявленным целям оценки.

4 Сокращения

В настоящем стандарте применены следующие сокращения:

ИИ — искусственный интеллект (artificial intelligence);
МО — машинное обучение (machine learning);
ПДн — персональные данные (personally identifiable information);
CSV — значения, разделенные запятыми (comma separated values);
HDF — иерархический формат данных (hierarchical data format);
JSON — текстовый формат описания объектов JavaScript (javascript object notation);
IP — интернет-протокол (internet protocol);
UML — унифицированный язык моделирования (Unified Modeling Language).

5 Компоненты качества данных и модели качества данных для аналитики и машинного обучения

5.1 Компоненты качества данных в жизненном цикле данных

На рисунке 1 показаны компоненты качества данных, соответствующие модели жизненного цикла данных, представленной в *ГОСТ Р 71484.1—2024* (рисунок 3), которая может поддерживать процессы управления качеством данных. *ГОСТ Р 71484.1* определяет модель качества данных как определенный набор характеристик качества данных. Характеристика качества данных обеспечивает основу для требований к качеству данных и методов оценки. Показатели качества данных — это присвоенные переменные, значения которых являются результатами измерений характеристик качества данных. Показатели качества данных используются для оценки того, соответствуют ли данные требованиям к качеству данных. Показатели качества данных также можно использовать для мониторинга и составления отчетов о качестве данных.

Рисунок 1 — Компоненты качества данных в жизненном цикле данных для аналитики и машинного обучения

Целевые данные — данные, которые являются предметом измерения качества. Целевые данные могут быть необработанными данными либо данными, подвергшимися одному или нескольким преобразованиям. Целевыми данными для измерения качества могут быть обучающие, тестовые, валидационные или эксплуатационные данные, а также данные результатов использования аналитики или машинного обучения (см. *[6]*). Целевые данные могут быть сформированы как элементы данных или как наборы данных. Элемент данных состоит из имени элемента, значения данных и типа данных, представляющего область значений (например, строки символов, тексты, даты, числа, изображения, звуки). Набор данных можно разделить на три формы:

- совокупность элементов данных;
- совокупность записей данных;
- совокупность фреймов данных.

Целевые данные могут быть немаркированными или маркированными в зависимости от ассоциации с метками данных при использовании аналитики или машинного обучения.

Примечание — В настоящем стандарте не делается различия между структурами данных, такими как структурированные данные, полуструктурированные данные и неструктурированные данные, или способом их использования, например, в качестве основных (мастер-) данных, транзакционных данных или справочных данных.

Отчеты о качестве данных — это документы, в которых представлены требования к качеству данных, модель качества данных, показатели качества данных, результаты измерений качества данных и оценка того, соответствуют ли данные предъявляемым требованиям к качеству данных.

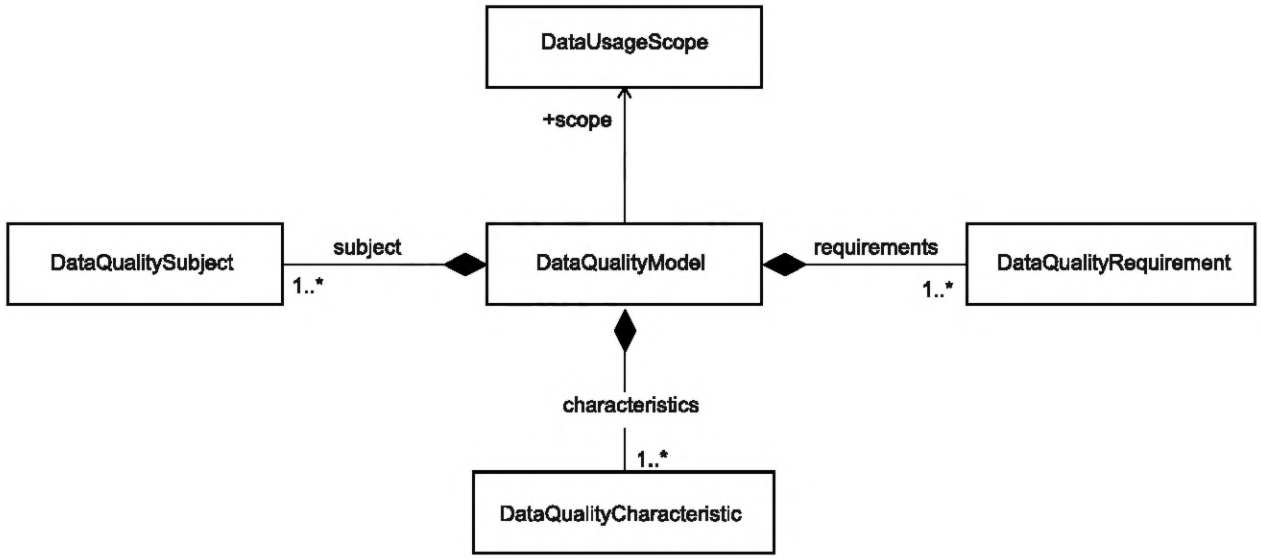
5.2 Модель качества данных

Модель качества данных задает структуру, позволяющую определить требования к качеству данных и проводить оценку их качества. На практике модель качества данных объединяет предметы качества данных, характеристики качества данных и требования к качеству данных в контексте их использования. Организация может уточнить модели качества данных, выбрав характеристики качества данных и показатели для достижения целевых требований к качеству целевых данных. На рисунке 2 представлена диаграмма отношений (в нотации UML) между компонентами модели качества данных.

Область применения данных описывает, как и где данные могут использоваться в задачах аналитики или машинного обучения и как они вписываются в систему ИИ.

Пример — Данные можно использовать для обучения модели машинного обучения глубокой нейронной сети с целью прогнозирования продаж продукта на основе особенностей маркетинговой стратегии. Модель можно обучить и развернуть с помощью облачных сервисов.

Предмет качества данных представляет собой сущность, на которую влияет качество данных. Характеристика качества данных — это категория атрибутов качества данных, которые влияют на качество данных (такие как аккуратность, наполненность, точность). Требование к качеству данных описывает свойства или атрибуты данных, а также критерии приемлемости для конкретной области применения данных. Критерии приемлемости могут быть количественными либо качественными.



DataUsageScope – область использования данных;
DataQualitySubject – предмет качества данных;
DataQualityModel – модель качества данных;
DataQualityRequirement – требование к качеству данных;
DataQualityCharacteristic – характеристика качества данных;

scope – область;
subjects – предметы;
requirements – требования;
characteristics – характеристики

Рисунок 2 — UML-диаграмма модели качества данных

Когда одна качественная характеристика влияет на другую, можно найти компромисс, оценивая каждое требование с точки зрения его важности и воздействия. Кроме того, крайне важно сбалансировать затраты на управление качеством данных с приоритетом требований к качеству данных при определении того, как характеристики и показатели качества данных включаются в модель качества данных. Организация может отобрать характеристики и показатели качества данных, которые соответствуют ее потребностям и требованиям. Качество данных следует оценивать путем сравнения результатов измерения выбранных показателей качества данных с целевыми показателями, установленными в требованиях к данным. В тех случаях, когда выполнить требования к качеству данных не удается, должны быть приняты корректирующие меры. *ГОСТ Р 71484.3* описывает требования и рекомендации системы управления качеством данных, которые должна применять организация.

Стандарты *[9]* и *[2]* описывают модели качества данных. В *[9]* определено три характеристики качества данных: синтаксическая (формат), семантическая (значение) и прагматическая (полезность) для поддержки промышленных данных в целом как продукта деловых и производственных процессов. В *[2]* определена общая модель качества данных, хранящихся в структурированном формате в компьютерной системе как часть программного продукта. В *[2]* учтены все типы данных (такие, как символы, строки, текст, даты, числа, изображения, звуки). Стандарт *[2]* предоставляет пятнадцать характеристик качества данных: аккуратность, наполненность, согласованность, достоверность, актуальность, доступность, соответствие, конфиденциальность, эффективность, точность, прослеживаемость, понятность, готовность, переносимость и восстанавливаемость.

В *[10]* рассмотрены различные аспекты качества данных, такие как управление данными, управление качеством данных (включая обработку) и оценку зрелости. В *ГОСТ Р ИСО/МЭК 25000* рассмотрены требования к качеству продукции (программного обеспечения, систем, данных, услуг) и ее оценку. В настоящем стандарте описывается, как характеристики качества данных *[2]* могут быть применены к модели качества данных для аналитики и машинного обучения. Кроме того, в нем определены дополнительные характеристики, которые могут способствовать более высокому качеству моделей и приложений МО, как показано на рисунке 3. Организации должны использовать характеристики качества данных и меры качества данных, описанные в этом документе, когда это возможно. Однако характеристики качества данных из настоящего стандарта не могут всесторонне охватывать аспекты для поддержки потребностей всех организаций в качестве данных. Организации могут разработать свою собственную модель качества данных, расширив характеристики качества данных и показатели качества данных в соответствии со своими требованиями к данным.

Примечания

1 См. приложение А для получения информации о разработке и документировании функций измерения.

2 См. приложение Е для сравнения характеристик качества данных в *[2]* и настоящем стандарте.

Рисунок 3 — Характеристики качества данных для аналитики и машинного обучения

6 Характеристики качества данных и показатели качества

6.1 Общие сведения

Характеристики и показатели качества данных используются при определении и проверке требований к качеству данных для определенных атрибутов целевых данных. Каждая характеристика качества данных связана с одним или несколькими показателями качества данных для количественной оценки. Показатель качества данных — это переменная, которой присваивается значение в результате применения функции измерения. Показатели качества данных в настоящем стандарте выбраны на основе контекста использования аналитики и машинного обучения.

Примечание — В приложении В показана структура предоставления общих словарей и взаимосвязей между компонентами показателей качества данных. В приложениях С и D показано, как показатели качества группируются с разных точек зрения.

В контексте аналитики и машинного обучения общее качество набора обучающих данных, набора валидационных данных или набора тестовых данных может быть столь же важным, как и качество отдельных значений данных в конкретном наборе данных. Даже если каждое значение данных в наборе данных является аккуратным, использование набора данных, который некорректно отражает основное распределение данных, может привести к некорректному результату анализа или к созданию модели МО, не соответствующей требованиям. Организация должна документировать целевые данные для каждого показателя качества данных.

Примечание — Характеристики статистических показателей (например, доступность для уполномоченных пользователей, аккуратность, согласованность, актуальность, понятность, релевантность, своевременность *[11]*), определенные такими учреждениями, как Статистический отдел Организации Объединенных Наций или Евростат, также могут использоваться для оценки того, соответствует ли качество набора данных установленным требованиям.

Там, где это возможно, следует использовать показатели качества данных и функции измерения, представленные в настоящем стандарте. В случае создания нового специального показателя качества данных и функции измерения качества данных следует использовать приложение А. При использовании любых модифицированных или вновь создаваемых показателей качества следует выбирать характеристики качества данных, определенные настоящим стандартом, и приводить обоснование изменений в соответствии с *[1]*, раздел 2.

6.2 Внутренне присущие характеристики качества данных

6.2.1 Аккуратность

6.2.1.1 Общие сведения

Аккуратность набора данных — это степень, в которой элементы данных в наборе данных имеют корректные значения данных или корректные метки данных. В *[2]* описана аккуратность как степень, в которой значения данных имеют атрибуты, корректно представляющие истинное значение предполагаемых атрибутов. В *[2]* описана аккуратность с точки зрения:

- синтаксической аккуратности, которая учитывает близость значений данных к набору синтаксически корректных значений данных для соответствующей области;
- семантической аккуратности, которая учитывает близость значений данных к набору семантически корректных значений данных для соответствующей области.

Элемент данных является синтаксически корректным, если его значение данных имеет тот же тип, что и его явный тип данных, и семантически корректным, если его значение данных имеет ожидаемое значение, соответствующее задаче МО. Модели машинного обучения представляют собой математические конструкции, а это означает, что низкая синтаксическая или семантическая аккуратность значений данных в наборах данных для обучения, валидации, тестирования или эксплуатации может привести к тому, что сама модель будет некорректной, либо выводы, сделанные моделью, будут некорректными.

Для системы обучения с учителем, предназначенной для классификации, корректность содержимого последовательности меток может повлиять на аккуратность вывода обученной модели. Факторы, которые следует учитывать при измерении аккуратности разметки, включают:

- корректность значений меток;
- корректность маркировки тегов;
- корректность содержания последовательности меток.

Примеры

1 Если фраза «ленивая собака» (Lazy dog) введена как «lzy dg», система МО для понимания естественного языка может неверно интерпретировать эту фразу.

2 Если в обучающих данных число 100 введено как 1000, регрессионная модель может неверно рассчитать вес соответствующего признака, а если это было введено в эксплуатационных данных, выводы могут быть неверными.

6.2.1.2 Показатели качества для аккуратности

В таблице 1 представлены показатели качества данных для обеспечения аккуратности в конкретном контексте использования аналитики и машинного обучения.

Таблица 1 — Показатели качества для аккуратности

Идентификатор	Наименование показателя	Описание	Функция измерения
Acc-ML-1	Синтаксическая аккуратность данных	См. *[1]*, таблица 1	См. *[1]*, таблица 1
Acc-ML-2	Семантическая аккуратность данных	См. *[1]*, таблица 1	См. *[1]*, таблица 1
Acc-ML-3	Гарантия аккуратности данных	См. *[1]*, таблица 1	См. *[1]*, таблица 1
Acc-ML-4	Риск неаккуратности набора данных	См. *[1]*, таблица 1	См. *[1]*, таблица 1
Acc-ML-5	Аккуратность модели данных	См. *[1]*, таблица 1	См. *[1]*, таблица 1
Acc-ML-6	Диапазон аккуратности данных	См. *[1]*, таблица 1	См. *[1]*, таблица 1
Acc-ML-7	Аккуратность метки данных	Корректно ли присвоена метка данных каждому элементу в наборе данных?	$\frac{A}{B}$, где A — количество меток данных, которые предоставляют соответствующую необходимую информацию; B — количество меток данных, определенных в наборе данных

6.2.2 Наполненность

6.2.2.1 Общие сведения

В *[2]* описана наполненность с точки зрения наличия значений данных для всех предполагаемых атрибутов и экземпляров сущностей. В некоторых случаях алгоритмы МО могут давать сбой, когда они обнаруживают один или несколько пустых элементов данных в наборах данных для обучения, валидации или тестирования. Кроме того, обученные модели машинного обучения также могут давать сбой, если эксплуатационные данные содержат элементы с отсутствующим значением данных.

Показатели наполненности могут помочь специалистам по МО удовлетворить свои требования к данным и указать, следует ли предпринять дополнительные шаги по заполнению данных (imputation), как описано в *ГОСТ Р 71484.4*.

Характеристика наполненности размеченных данных в наборе данных является относительной. В разных сценариях представление о полноте может быть разным, и его следует учитывать в зависимости от конкретного варианта использования. Факторы, которые следует учитывать при измерении наполненности набора данных, включают следующие:

- проверка на наполненность набора данных, используемого для классификации изображений на основе МО, должна заключаться в проверке неразмеченных выборок в наборе данных, которые нельзя сразу использовать при машинном обучении с учителем;
- проверка на наполненность набора данных, используемого для обнаружения объектов на основе МО, должна заключаться в проверке на неполноту размеченных ограничивающих прямоугольников на объектах.

В частности, в реальной жизни часто бывает, что в выборке имеется несколько объектов разных категорий, поскольку сложно запечатлеть сцену с одним изолированным объектом, занимающим все пространство обзора. В этом случае в целях измерения наполненности набора данных для распознавания изображений на основе МО следует учитывать следующие факторы:

- в выборке существует какой-либо целевой объект;
- все целевые объекты категорированы;
- все обнаруженные целевые объекты размечены ограничивающими прямоугольниками или другими методами.

Примеры

1 Показатель наполненности набора данных указывает, что в наборе данных отсутствует более половины значений данных для признака, связанного с почтовым индексом. Исследователь данных решает, что признак почтового индекса не является необходимым предиктором для его задачи классификации, и решает удалить признак почтового индекса из наборов данных для обучения, валидации, тестирования и эксплуатации.

2 Показатель наполненности набора данных, используемого для задачи регрессии на основе МО, указывает на то, что один процент значений данных для признака, который является хорошим предиктором, пуст. Остальные данные имеют нормальное распределение. Исследователь данных решает заполнить пустые значения средним статистическим значением доступных значений данных.

3 Показатель наполненности набора данных, используемого для задачи кластеризации на основе МО, указывает на то, что небольшое количество записей содержит один или несколько пустых элементов данных. Исследователь данных решает удалить эти записи из обучающих данных.

4 Показателем наполненности вхождений значений в набор данных для задачи классификации на основе МО является отношение отсутствующих значений данных к целевому количеству элементов данных, ожидаемых для обеспечения надлежащей верности набора данных.

6.2.2.2 Показатели качества для наполненности

В таблице 2 представлены показатели качества данных для обеспечения наполненности в конкретном контексте использования аналитики и машинного обучения.

Таблица 2 — Показатели качества для наполненности

Идентификатор	Наименование показателя	Описание	Функция измерения
Com-ML-1	Наполненность значения	Отношение количества непустых элементов данных к общему количеству элементов данных в наборе данных, где существует хотя бы один элемент данных	$\frac{A}{B}$, где A — количество непустых элементов данных; B — общее количество элементов данных в наборе данных, где существует хотя бы один элемент данных
Com-ML-2	Наполненность появления значения	Отношение количества появлений заданного значения данных к ожидаемому количеству появлений значения данных, описанному в требованиях к качеству данных, в элементах данных из той же области, что и набор данных	$\frac{A}{B}$, где A — количество появлений заданного значения данных в элементах данных; B — ожидаемое количество появлений этого значения данных в элементах данных из той же области, что и набор данных
Com-ML-3	Наполненность признака	Отношение количества непустых элементов данных, связанных с признаком, к общему количеству элементов данных, связанных с этим признаком	$\frac{A}{B}$, где A — количество непустых элементов данных, связанных с заданным признаком; B — общее количество элементов данных, связанных с заданным признаком в наборе данных, где существует хотя бы один элемент данных

Окончание таблицы 2

Идентификатор	Наименование показателя	Описание	Функция измерения
Com-ML-4	Наполненность записи	Отношение количества непустых записей данных к общему количеству записей данных в наборе данных, где существует хотя бы одна запись данных	$\frac{A}{B}$, где A — количество непустых записей данных в наборе данных; B — общее количество записей данных в наборе данных, где существует хотя бы одна запись данных
Com-ML-5	Наполненность метки	Доля размеченных или неполностью размеченных выборок в наборе данных	$1-\frac{A}{B}$, где A — количество неразмеченных или не полностью размеченных выборок; B — количество всех выборок в наборе данных

6.2.3 Согласованность

6.2.3.1 Общие сведения

В *[2]* описана согласованность с точки зрения когерентности данных с другими данными и отсутствия противоречий. Когерентность — это ключевой аспект данных, используемых для машинного обучения, поскольку признаки, используемые в обучающих данных, вместе должны обеспечивать модель, позволяющую делать корректные выводы на основе производственных данных. Кроме того, машинное обучение может быть буквальным в интерпретации значений данных. Дублирующиеся записи могут привести к переоценке определенных признаков. Противоречия между признаками в обучающих данных могут привести к тому, что обученная модель не будет соответствовать требованиям. Качество обучающих данных зависит от согласованности меток, присваиваемых схожим элементам данных. Для улучшения производительности моделей МО метки данных во избежание несоответствий необходимо присваивать согласованно.

Пример — Веб-форма используется для сбора данных о предпочтениях избирателей в отношении политических кандидатов. Организованная группа людей наводняет веб-сайт записями о своем любимом кандидате. При использовании для обучения модели МО эти повторяющиеся данные могут привести к тому, что модель будет переоценивать конкретного кандидата при формировании выводов о людях, характеристики которых аналогичны характеристикам тех, кто заполнил веб-форму.

6.2.3.2 Показатели качества для согласованности

В таблице 3 представлены показатели качества данных для обеспечения согласованности в контексте использования аналитики и машинного обучения.

Таблица 3 — Показатели качества для согласованности

Идентификатор	Наименование показателя	Описание	Функция измерения
Con-ML-1	Согласованность записи данных	Доля повторяющихся записей в наборе данных	$1-\frac{A}{B}$, где A — количество повторяющихся записей в наборе данных; B — общее количество записей в наборе данных
Con-ML-2	Согласованность меток данных	Согласованность меток данных для схожих элементов данных	$\frac{A}{B}$, где A — количество пар схожих элементов, которым присвоены одинаковые метки; B — общее количество проведенных сравнений меток для схожих элементов

Окончание таблицы 3

Идентификатор	Наименование показателя	Описание	Функция измерения
Con-ML-3	Согласованность формата данных	См. *[1]*, таблица 3	См. *[1]*, таблица 3
Con-ML-4	Семантическая согласованность	См. *[1]*, таблица 3	См. *[1]*, таблица 3

6.2.4 Достоверность

6.2.4.1 Общие сведения

[*2*] определяет достоверность как степень, в которой данные обладают атрибутами, которым пользователи склонны доверять в конкретном контексте использования. Достоверность применима к отдельным элементам данных, к связанным элементам данных в записи данных и ко всему набору данных. Контекст, в котором используются данные, может повлиять на их воспринимаемую достоверность и правдоподобность. Данные могут быть изменены во время обработки (например, при перемещении, хранении, вычислении) уполномоченными и неавторизованными сторонами. Растущая озабоченность в МО связана с тем, что посторонние лица искажают данные для обучения, валидации, тестирования и эксплуатации, чтобы намеренно сделать обученные модели непригодными для использования либо манипулировать выводами, сделанными обученной моделью.

Процессы, используемые при подготовке данных, могут изменять данные, не меняя их значения (например, при нормализации, разделении или объединении признаков). В этих случаях данные сохраняют свою достоверность.

Примеры

1 Набор данных используется для обучения, валидации и тестирования модели МО, которая затем не достигает требуемой производительности на эксплуатационных данных. Аудит безопасности показывает, что неавторизованная сторона случайным образом изменила значения данных в наборе обучающих данных.

2 Набор обучающих данных содержит числовые признаки, которые имеют широко варьирующиеся диапазоны. Специалист по данным решает нормализовать значения данных для этих признаков, чтобы сделать их более сопоставимыми. Хотя значения данных могут измениться после нормализации, они по-прежнему заслуживают доверия, поскольку их содержание не изменилось в контексте машинного обучения.

3 Достоверность подготовки определенных типов наборов данных можно повысить при некоторых обстоятельствах, используя статистический метод случайности для составления выборок.

4 Доверие к подготовке набора данных может быть повышено путем указания происхождения данных в соответствии со структурой жизненного цикла данных (см. ГОСТ Р 70889—2023, пункт 6.6).

6.2.4.2 Показатели качества для достоверности

В таблице 4 представлены показатели качества данных для обеспечения их достоверности в контексте использования аналитики и машинного обучения.

Таблица 4 — Показатели качества для достоверности

Идентификатор	Наименование показателя	Описание	Функция измерения
Cre-ML-1	Достоверность значений	См. *[1]*, таблица 4	См. *[1]*, таблица 4
Cre-ML-2	Достоверность источника	См. *[1]*, таблица 4	См. *[1]*, таблица 4
Cre-ML-3	Достоверность словаря данных	См. *[1]*, таблица 4	См. *[1]*, таблица 4
Cre-ML-4	Достоверность модели данных	См. *[1]*, таблица 4	См. *[1]*, таблица 4

6.2.5 Актуальность

6.2.5.1 Общие сведения

В *[2]* описана актуальность данных с точки зрения надлежащего возраста относительно использования данных. Для машинного обучения актуальность может определяться возрастным диапазоном, подходящим для задачи МО. Например, данные о людях могут быть неполными для недостаточно представленных групп населения до изменений в законодательстве и социальных нормах. Модели МО, основанные на экономических данных, собранных за несколько десятилетий, могут быть неверными, если данные не скорректированы с учетом инфляции, обменных курсов и других факторов, которые меняются со временем. Проблемы, возникающие из-за различий в эксплуатационных данных по сравнению с данными, используемыми для обучения и тестирования модели, обычно известны как проблемы дрейфа данных, и их можно решить, обеспечив актуальность данных.

Актуальность набора данных можно описать с точки зрения общего периода времени, который охватывает конкретный набор данных (например, изображения или высказывания, собранные с 2010 по 2021 год), периода времени между последней датой элемента данных и текущей датой (например, 8 месяцев) и цикла обновления (например, каждые 6 месяцев). Актуальность следует рассматривать как составной показатель, основанный на этих трех аспектах.

Примеры

1 Модель МО, используемая для прогнозирования будущих продаж, постоянно не учитывает фактическую сумму продаж. В ходе расследования выяснилось, что при создании модели использовались обучающие данные за 10 лет продаж, но значения данных не были скорректированы с учетом инфляции.

2 Модель МО используется для прогнозирования того, за какого кандидата может проголосовать конкретный человек. Модель была обучена на протоколах голосования за предыдущие 20 лет. Некоторые недостаточно представленные группы начали регулярно голосовать лишь в последние 10 лет. Поэтому модель МО регулярно допускает ошибки в прогнозе выбора, сделанного избирателями, входящими в такие группы.

6.2.5.2 Показатели качества для актуальности

В таблице 5 представлены показатели качества данных для обеспечения их актуальности в контексте использования аналитики и машинного обучения.

Таблица 5 — Показатели качества для актуальности

Идентификатор	Наименование показателя	Описание	Функция измерения
Cur-ML-1	Актуальность признака	Доля элементов данных с некоторым признаком в наборе данных, попадающих в допустимый возрастной диапазон, как указано в требованиях организации к качеству данных	$\frac{A}{B}$, где A — количество элементов данных с некоторым признаком, попадающих в требуемый возрастной диапазон; B — общее количество элементов данных для данного признака
Cur-ML-2	Актуальность записи	Доля записей данных в наборе данных, где все элементы данных из записи попадают в требуемый возрастной диапазон	$\frac{A}{B}$, где A — количество записей данных, попадающих в требуемый возрастной диапазон; B — общее количество записей данных в наборе данных

6.3 Внутренне присущие и системно-зависимые характеристики качества данных

6.3.1 Доступность

6.3.1.1 Общие сведения

Согласно *[1]* доступность означает степень, в которой данные могут быть доступны в конкретном контексте использования, особенно для лиц, которым необходимы вспомогательные технологии или специальная конфигурация из-за ограниченных возможностей здоровья. Кроме того, для аналитики и

машинного обучения должны быть обеспечены беспрепятственный доступ к наборам данных и простое развертывание наборов данных с помощью соответствующих инструментов.

6.3.1.2 Показатели качества для доступности

В таблице 6 представлены показатели качества данных для обеспечения доступности в контексте использования аналитики и машинного обучения.

Таблица 6 — Показатели качества для доступности

Идентификатор	Наименование показателя	Описание	Функция измерения
Acs-ML-1	Доступность для пользователей	См. *[1]*, таблица 6.1	См. *[1]*, таблица 6.1
Acs-ML-2	Доступность формата данных	См. *[1]*, таблица 6.1	См. *[1]*, таблица 6.1
Acs-ML-3	Доступность данных	Доля доступных записей в наборе данных	$\frac{A}{B}$, где A — количество доступных записей в наборе данных; B — общее количество записей данных в наборе данных

6.3.2 Соответствие

6.3.2.1 Общие сведения

В *[2]* описано соответствие с точки зрения степени, в которой данные удовлетворяют требованиям нормативного регулирования, стандартов, соглашений или других правил. Например, персональные данные, используемые для аналитики или МО, могут подпадать под действие законодательных и нормативных требований. Аналогично пользователи данных могут иметь собственные требования соответствия, а схемы сертификации также могут иметь требования соответствия.

6.3.2.2 Показатели качества для соответствия

В таблице 7 представлены показатели качества данных для обеспечения соответствия в контексте использования аналитики и машинного обучения.

Таблица 7 — Показатели качества для соответствия

Идентификатор	Наименование показателя	Описание	Функция измерения
Cmp-ML-1	Соответствие элементов данных	Степень, в которой элементы данных соответствуют требованиям	$\frac{A}{B}$, где A — количество элементов данных, соответствующих требованиям; B — общее количество элементов данных в наборе данных

6.3.3 Эффективность

6.3.3.1 Общие сведения

Эффективность означает степень, в которой данные обладают атрибутами, которые можно обрабатывать и обеспечивать ожидаемый уровень производительности при использовании соответствующих объемов и типов ресурсов в конкретном контексте. Например, эффективность формата данных важна при совместном использовании наборов обучающих данных, особенно когда размер данных велик, например CSV, HDF, JSON. Кроме того, оптимальный размер набора данных в памяти может снизить стоимость обучения.

6.3.3.2 Показатели качества для эффективности

В таблице 8 представлены показатели качества данных для обеспечения эффективности в контексте использования аналитики и машинного обучения.

Таблица 8 — Показатели качества для эффективности

Идентификатор	Наименование показателя	Описание	Функция измерения
Eff-ML-1	Эффективность формата данных	См. *[1]*, таблица 9.2	См. *[1]*, таблица 9.2
Eff-ML-2	Эффективность обработки данных	См. *[1]*, таблица 9.2	См. *[1]*, таблица 9.2
Eff-ML-3	Риск потери пространства в памяти	См. *[1]*, таблица 9.2	См. *[1]*, таблица 9.2

6.3.4 Точность

6.3.4.1 Общие сведения

Согласно *[1]* точность означает степень, с которой можно утверждать, что данные являются точными или их можно различить. Например, можно утверждать, что данные совпадают с точностью до нескольких десятичных знаков после запятой для действительного числа. В контексте машинного обучения точность, выраженная количеством десятичных знаков для значения элемента данных, может влиять на вес данного признака в обученной модели машинного обучения. Например, признак со многими значениями элементов данных, равными 99,4, может иметь больший вес, чем другой признак, у которого то же значение округлено в меньшую сторону до 99. Аналогично признак, значения которого были округлены в большую сторону, может иметь больший вес, чем признак с большей точностью. При указании требований к точности данных пользователи данных должны учитывать общее влияние на обученную модель МО.

6.3.4.2 Показатели качества для точности

В таблице 9 представлены показатели качества данных для обеспечения точности в контексте использования аналитики и машинного обучения.

Таблица 9 — Показатели качества для точности

Идентификатор	Наименование показателя	Описание	Функция измерения
Pre-ML-1	Точность значений данных	См. *[1]*, таблица 10.1	См. *[1]*, таблица 10.1

6.3.5 Прослеживаемость

6.3.5.1 Общие сведения

Прослеживаемость показывает, в какой степени данные обладают атрибутами, позволяющими отслеживать доступ и любые внесенные в них изменения в конкретном контексте использования.

6.3.5.2 Показатели качества для прослеживаемости

В таблице 10 представлены показатели качества данных для обеспечения прослеживаемости в контексте использования аналитики и машинного обучения.

Таблица 10 — Показатели качества для прослеживаемости

Идентификатор	Наименование показателя	Описание	Функция измерения
Tra-ML-1	Прослеживаемость доступа пользователей к значениям данных	См. *[1]*, таблица 11.1	См. *[1]*, таблица 11.1
Tra-ML-2	Прослеживаемость доступа пользователей к значениям данных системными средствами	См. *[1]*, таблица 11.1	См. *[1]*, таблица 11.1
Tra-ML-3	Прослеживаемость изменений значений данных системными средствами	См. *[1]*, таблица 11.1	См. *[1]*, таблица 11.1

6.3.6 Понятность

6.3.6.1 Общие сведения

В *[2]* описана понятность с точки зрения возможности пользователей читать и интерпретировать данные. Кроме того, понятность подразумевает использование соответствующих символов, единиц измерения и языков. Модели МО могут не соответствовать требованиям, если единицы измерения признаков используются ненадлежащим образом. Понятность — важная характеристика, обеспечивающая объяснимость системы ИИ и помогающая заинтересованным сторонам взаимодействовать с эксплуатационными данными системы ИИ (меткой вывода или значениями весов модели МО).

В задачах обработки естественного языка ненадлежащее использование человеческих языков и символов может привести к сбою при решении таких задач, как понимание и генерация языка.

Хотя показатели качества данных из настоящего стандарта являются количественными, люди, использующие данные для МО, также проводят качественную оценку данных. Соответствующее использование символов, единиц измерения и языков может помочь в вынесении качественных суждений.

6.3.6.2 Показатели качества для понятности

В таблице 11 представлены показатели качества данных для понятности в контексте использования аналитики и машинного обучения.

Таблица 11 — Показатели качества для понятности

Идентификатор	Наименование показателя	Описание	Функция измерения
Und-ML-1	Понятность символов	См. *[1]*, таблица 12.1	См. *[1]*, таблица 12.1
Und-ML-2	Семантическая понятность	См. *[1]*, таблица 12.1	См. *[1]*, таблица 12.1
Und-ML-3	Понятность значений данных	См. *[1]*, таблица 12.1	См. *[1]*, таблица 12.1
Und-ML-4	Понятность представления данных	См. *[1]*, таблица 12.2	См. *[1]*, таблица 12.2

6.4 Системно-зависимые характеристики качества данных

6.4.1 Готовность

6.4.1.1 Общие сведения

Готовность означает степень, в которой авторизованные пользователи или приложения могут извлекать наборы данных для конкретной задачи аналитики или МО на стадии комплектования данных жизненного цикла данных.

6.4.1.2 Показатели качества для готовности

В таблице 12 представлены показатели качества данных для готовности в контексте использования аналитики и машинного обучения.

Таблица 12 — Показатели качества для готовности

Идентификатор	Наименование показателя	Описание	Функция измерения
Ava-ML-1	Коэффициент готовности данных	См. *[1]*, таблица 13	См. *[1]*, таблица 13

6.4.2 Переносимость

6.4.2.1 Общие сведения

В *[2]* описана характеристика переносимости для качества данных с точки зрения возможности в пределах заданного контекста переносить данные из одной системы в другую, сохраняя при этом их качество.

Данные, используемые в аналитике и МО, могут обрабатываться в нескольких системах. Например, данные для машинного обучения можно собирать в одной системе, процессы оценки качества используемых данных можно выполнять во второй системе, а затем данные передавать в третью систему для обучения модели МО.

Если качество данных (т. е. соответствие требованиям) не поддерживается при передаче данных из одной системы в другую, то сама обученная модель МО может не соответствовать требованиям.

Примечание — Требования к переносимости данных устанавливаются на стадии формирования требований к данным жизненного цикла данных и зависят как от системы, так и от окружения.

6.4.2.2 Показатели качества для переносимости

В таблице 13 представлены показатели качества данных для переносимости в контексте использования аналитики и машинного обучения.

Таблица 13 — Показатели качества для переносимости

Идентификатор	Наименование показателя	Описание	Функция измерения
Por-ML-1	Коэффициент переносимости данных	См. *[1]*, таблица 14	См. *[1]*, таблица 14
Por-ML-2	Потенциальная переносимость данных	См. *[1]*, таблица 14	См. *[1]*, таблица 14

6.4.3 Восстанавливаемость

6.4.3.1 Общие сведения

Восстанавливаемость означает степень, в которой наборы данных могут поддерживаться и сохраняться на определенном уровне операций и качества (даже в случае сбоя) для конкретной задачи анализа и МО на стадиях подготовки и предоставления данных жизненного цикла данных, особенно с большим объемом наборов данных.

6.4.3.2 Показатели качества для восстанавливаемости

В таблице 14 представлены показатели качества данных для восстанавливаемости в контексте использования аналитики и машинного обучения.

Таблица 14 — Показатели качества для восстанавливаемости

Идентификатор	Наименование показателя	Описание	Функция измерения
Rec-ML-1	Коэффициент восстанавливаемости данных	См. *[1]*, таблица 15	См. *[1]*, таблица 15
Rec-ML-2	Коэффициент восстанавливаемости признаков	Доля передаваемых поэтапно признаков набора данных, которые могут быть восстановлены	$\frac{A}{B}$, где A — количество успешно восстановленных признаков набора данных; B — признаки набора данных, которыми можно управлять с помощью процедур резервного копирования и восстановления

6.5 Дополнительные характеристики качества данных

6.5.1 Проверяемость

6.5.1.1 Общие сведения

Для целей настоящего стандарта проверяемость относится к характеристике набора данных, заключающейся в том, что весь набор данных или его часть подверглись аудиту или что данные доступны соответствующим заинтересованным сторонам с целью проведения аудита. Аудит наборов данных, используемых для аналитики и МО, может способствовать повышению достоверности данных и может потребоваться для обеспечения соответствия требованиям.

Пример — Набор данных с изображениями используется для распознавания изображений и был размечен сторонним исполнителем. Чтобы убедиться, что изображения размечены корректно, организация привлекает третью сторону для проверки подмножества размеченных изображений.

6.5.1.2 Показатели качества для проверяемости

В таблице 15 представлены показатели качества данных для проверяемости в контексте использования аналитики и машинного обучения.

Таблица 15 — Показатели качества для проверяемости

Идентификатор	Наименование показателя	Описание	Функция измерения
Aud-ML-1	Проверенные записи	Доля записей в наборе данных, прошедших аудит	$\frac{A}{B}$, где A — количество записей в наборе данных, которые прошли аудит; B — общее количество записей в наборе данных
Aud-ML-2	Проверяемые записи	Доля записей в наборе данных, доступных для аудита	$\frac{A}{B}$, где A — количество записей в наборе данных, доступных для аудита; B — общее количество записей в наборе данных

6.5.2 Сбалансированность

6.5.2.1 Общие сведения

Под сбалансированностью набора данных понимается распределение выборок по всем признакам конкретного набора данных. Например, если набор данных представляет X категорий элементов данных, количество выборок на категорию должно быть распределено равномерно, чтобы набор данных был сбалансирован. Для набора данных с изображениями такие признаки могут включать метки, значимые для логики деловых процессов; разрешение; яркость; соотношение ширины и высоты размеченных ограничивающих прямоугольников; размер размеченных ограничивающих прямоугольников и любые другие, которые потенциально влияют на производительность модели МО.

Сбалансированность набора данных может частично повлиять на общую производительность модели МО. Для системы компьютерного зрения на основе машинного обучения также следует учитывать сбалансированность набора данных.

Примеры

1 Когда существуют значительные различия в яркости или разрешении между выборками из набора обучающих данных и реальными данными, модели МО могут потерпеть неудачу из-за зашумленных данных, вызванных тусклостью или нечеткостью.

2 В классификационной системе, основанной на машинном обучении, наличие несбалансированной категории выборочной совокупности может привести к невозможности обнаружения и классификации редких экземпляров. Такие экземпляры могут быть неверно классифицированы или даже идентифицированы как зашумленные данные.

3 В системе обнаружения объектов на основе машинного обучения значительные различия в соотношении ширины и высоты или в размере ограничивающих прямоугольников могут привести к несоответствию размеров обнаруживаемых объектов при фиксированном размере рецептивного поля. Следовательно, это также может привести к потере возможности обобщения, если не применяются дополнительные подходы к проверке или корректировке объектов разного размера.

Хотя представленные здесь примеры относятся к изображениям, концепцию сбалансированности можно применить и к другим типам данных.

6.5.2.2 Показатели качества для сбалансированности

В таблице 16 представлены показатели качества данных для сбалансированности в контексте использования аналитики и машинного обучения.

Таблица 16 — Показатели качества для сбалансированности

Идентификатор	Наименование показателя	Описание	Функция измерения
Bal-ML-1	Сбалансированность яркости	Обратное значение максимального отношения разницы яркости выборки изображений к средней яркости выборок в наборе данных	$\frac{A}{B}$, где A — среднее значение яркости выборки; B — максимальное значение абсолютной разницы между значением яркости каждого изображения в выборке и A
Bal-ML-2	Сбалансированность разрешения	Обратное значение максимального отношения разницы разрешения выборки изображений к усредненному разрешению выборок в наборе данных	$\frac{A}{B}$, где A — среднее значение разрешения выборки; B — максимальное значение абсолютной разницы между значением разрешения каждого изображения в выборке и A
Bal-ML-3	Сбалансированность изображений между категориями	Обратная величина максимального отношения разницы размера категории (количества содержащихся выборок) к усредненному размеру категории набора данных	$\frac{A}{B}$, где A — средний размер категории набора данных; B — максимальное значение абсолютной разницы между размером каждой категории в наборе данных и A
Bal-ML-4	Сбалансированность соотношения высоты и ширины ограничивающего прямоугольника	Обратная величина максимального соотношения высоты и ширины ограничивающего прямоугольника) к усредненному соотношению высоты и ширины ограничивающего прямоугольника для конкретной выборки в наборе данных	$\frac{A}{B}$, где A — усредненный ограничивающий прямоугольник с соотношением высоты и ширины по всем образцам в наборе данных; B — максимальное значение абсолютных различий между ограничивающей рамкой с соотношением высоты и ширины каждого образца в наборе данных и A
Bal-ML-5	Сбалансированность площади ограничивающего прямоугольника категории	Обратное значение максимального отношения усредненной площади ограничивающего прямоугольника категории к усредненной площади ограничивающего прямоугольника всех выборок в наборе данных	$\frac{A}{B}$, где A — усредненная площадь ограничивающего прямоугольника по всем выборкам в наборе данных; B — максимальное значение абсолютных различий между усредненной площадью ограничивающего прямоугольника каждой категории в наборе данных и A
Bal-ML-6	Сбалансированность площади ограничивающего прямоугольника выборки	Обратное значение максимального отношения усредненной площади ограничивающего прямоугольника выборки к усредненной площади ограничивающего прямоугольника по всем выборкам в наборе данных	$\frac{A}{B}$, где A — усредненная площадь ограничивающего прямоугольника по всем выборкам в наборе данных; B — максимальное значение абсолютной разницы между усредненной площадью ограничивающего прямоугольника для каждой выборки из набора данных и A

Окончание таблицы 16

Идентификатор	Наименование показателя	Описание	Функция измерения
Bal-ML-7	Сбалансированность пропорций метки	Разница в пропорциях элементов данных из двух разных категорий данных, имеющих определенное значение метки	$A - B$, где A — доля элементов данных в категории C_A имеющих значение метки L в наборе данных, т. е. $n_A^{(L)}/n_A$; B — доля элементов данных в категории C_B имеющих значение метки L в наборе данных, т. е. $n_B^{(L)}/n_B$; n_A — количество элементов данных, принадлежащих категории C_A; n_B — количество элементов данных, принадлежащих категории C_B (не C_A); $n_A^{(L)}$ — количество элементов данных, имеющих значение метки L в категории C_A; $n_B^{(L)}$ — количество элементов данных, имеющих значение метки L в категории C_B
Bal-ML-8	Сбалансированность распределения меток	Расхождение между распределением меток и равномерным распределением меток	$f(A,B)$, где A — распределение меток элементов данных с разными значениями меток в оцениваемом наборе данных, т. е. $[n_{L1}/n, n_{L2}/n, ..., n_{LN}/n]$; B — равномерное распределение меток элементов данных, т. е. $[n/N, n/N, ..., n/N]$; f — функция, измеряющая расхождение между двумя распределениями, например, расхождение Кульбака—Лейблера, расхождение Дженсена—Шеннона, L_p— норма, общее расстояние вариации и статистика критерия Колмогорова—Смирнова; N и n — количество различных значений меток и общее количество элементов данных в наборе соответственно; n_{Li} — количество элементов данных, имеющих i-е значение метки $\{L_1, L_2, ... L_N\}$ в наборе данных

6.5.3 Разнообразность

6.5.3.1 Общие сведения

Разнообразность набора данных означает различия между выборками с точки зрения целевых данных. В наборе данных, используемом для модели МО, важно адекватное различие между выборками. Если все записи или большинство записей данных в наборе данных одинаковы, модель МО, обученная на основе этого набора данных, может иметь риск переобученности и, следовательно, быть менее подходящей для обобщения. Разнообразность набора данных представляет собой степень, в которой набор данных содержит различные диапазоны для различных признаков, значений, меток, кластеров или источников среди отдельных данных. Улучшение данных с помощью генеративных моделей МО может улучшить разнообразность данных, но эти подходы могут потерпеть неудачу, если разнообразность исходного набора данных ограничено. Разнообразность тесно связана с репрезентативностью и сбалансированностью. Это характеристика качества данных, которую также можно использовать для оценки верности набора данных.

Измерение разнообразности может выполняться в контексте конкретных целевых данных, как это определено требованиями задачи МО.

6.5.3.2 Показатели качества для разнообразности

В таблице 17 представлены показатели качества данных для разнообразности в контексте использования аналитики и машинного обучения.

Таблица 17 — Показатели качества для разнообразности

Идентификатор	Наименование показателя	Описание	Функция измерения
Div-ML-1	Разнообразность меток	Доля различных меток в наборе данных	$\frac{A}{B}$, где A — количество различных меток в наборе данных; B — количество элементов данных в наборе данных
Div-ML-2	Относительная разнообразность меток	Доля данных (например, элементов данных, записей данных, фреймов данных), имеющих одну и ту же метку в наборе данных	$\frac{A}{B}$, где A — количество данных (например, элементов данных, записей данных, фреймов данных), в которых есть целевые метки; B — количество данных (например, элементов данных, записей данных, фреймов данных) в наборе данных
Div-ML-3	Разнообразность размеров категорий	Доля категорий, в которых количество классифицированных элементов данных ниже порогового значения, определенного требованиями к качеству	$\frac{A}{B}$, где A — количество категорий, в которых количество классифицированных элементов данных ниже порогового значения требования к качеству; B — общее количество категорий

6.5.4 Результативность

6.5.4.1 Общие сведения

Результативность набора данных показывает, соответствует ли набор данных требованиям для использования в конкретной задаче МО.

Примеры

1 Для системы компьютерного зрения на основе машинного обучения результативность набора данных может быть наименьшим приемлемым соотношением, при котором количество изображений с яркостью или разрешением ниже требуемого порога делится на общее количество изображений или видео в наборе данных.

2 Для системы классификации изображений на основе машинного обучения результативность набора данных может относиться к наименьшему приемлемому соотношению количества изображений в категории, деленному на общее количество изображений в наборе данных.

3 Для системы обнаружения объектов на основе машинного обучения результативность набора данных может относиться к наименьшему приемлемому соотношению количества изображений, значения площади которых в ограничивающих прямоугольниках ниже требуемого порога, деленному на общее количество изображений или видео в наборе данных.

6.5.4.2 Показатели качества для результативности

В таблице 18 представлены показатели качества данных для результативности в контексте использования аналитики и машинного обучения.

Таблица 18 — Показатели качества для результативности

Идентификатор	Наименование показателя	Описание	Функция измерения
Eft-ML-1	Результативность признака	Доля выборок с приемлемым признаком в наборе данных	$\frac{A}{B}$, где A — количество выборок с приемлемым признаком; B — количество всех выборок в наборе данных
Eft-ML-2	Результативность размера категории	Доля категорий, в которых количество классифицированных выборок ниже порогового значения	$\frac{A}{B}$, где A — количество категорий, в которых количество классифицированных выборок ниже порогового значения; B — общее количество категорий
Eft-ML-3	Результативность метки	Доля выборок с приемлемой меткой в наборе данных	$\frac{A}{B}$, где A — количество выборок с приемлемой меткой; B — количество всех выборок в данных

6.5.5 Идентифицируемость

6.5.5.1 Общие сведения

В *ГОСТ Р ИСО/МЭК 29100* описана идентифицируемость как возможность идентифицировать субъект персональных данных (ПДн) прямо или косвенно на основе заданного набора ПДн. Важно понимать, могут ли какие-либо персональные данные в наборе данных использоваться для идентификации субъекта ПДн, поскольку требования законодательства в некоторых юрисдикциях могут ограничивать такую деятельность. Чтобы снизить возможность идентификации, к обучающим, валидационным, тестовым и эксплуатационным данным могут применяться процессы обезличивания.

Пример — Модель МО обучается на запросах поисковых систем с целью таргетированной рекламы. Набор данных включает IP-адрес пользователя, который в некоторых юрисдикциях относится к ПДн. К набору данных применяется процедура обезличивания с целью удаления IP-адреса перед разделением набора данных на обучающие, валидационные и тестовые наборы данных. При необходимости процедура обезличивания может быть применена и к эксплуатационным данным, передаваемым в модель МО.

6.5.5.2 Показатели качества для идентифицируемости

В таблице 19 представлены показатели качества данных для идентифицируемости в контексте использования аналитики и машинного обучения.

Таблица 19 — Показатели качества для идентифицируемости

Идентификатор	Наименование показателя	Описание	Функция измерения
Idn-ML-1	Коэффициент идентифицируемости	Доля записей данных в наборе данных, которые можно использовать для идентификации	$\frac{A}{B}$, где A — количество записей данных, содержащих элементы данных, которые можно использовать для идентифицируемости либо отдельно, либо в сочетании с другими элементами данных; B — количество записей данных в наборе данных

6.5.6 Релевантность

6.5.6.1 Общие сведения

Для целей настоящего стандарта релевантность означает степень, в которой набор данных подходит для данного контекста (при условии, что он аккуратен, наполнен, согласован, актуален и т. д.).

Для машинного обучения релевантность может означать, что отобранные признаки в обучающих данных и их значениях данных являются хорошими предикторами для целевой переменной.

Пример — Модель МО используется для определения кредитоспособности людей. Обучающие данные являются репрезентативными для выборочной совокупности, которая, как ожидается, появится в эксплуатационных данных. Обучающие данные включают соответствующие признаки, такие как предыдущая кредитная история, доход, стаж работы и собственный капитал, которые являются хорошими показателями кредитоспособности. В то же время обучающие данные включают рост и вес каждого человека. Статистические тесты не показывают корреляции между ростом и весом с предыдущей кредитной историей и считаются плохими предикторами будущих показателей кредитоспособности. Чтобы повысить общую релевантность набора данных, признаки роста и веса были удалены.

6.5.6.2 Показатели качества для релевантности

В таблице 20 представлены показатели качества данных для релевантности в конкретном контексте использования аналитики и машинного обучения.

Таблица 20 — Показатели качества для релевантности

Идентификатор	Наименование показателя	Описание	Функция измерения
Rel-ML-1	Релевантность признака	Доля признаков в наборе данных, соответствующих данному контексту	$\frac{A}{B}$, где A — количество признаков в наборе данных, которые считаются релевантными в данном контексте использования данных; B — общее количество признаков в наборе данных
Rel-ML-2	Релевантность записи	Доля записей в наборе данных, соответствующих данному контексту	$\frac{A}{B}$, где A — количество записей в наборе данных, которые считаются релевантными в данном контексте использования данных; B — общее количество записей в наборе данных

6.5.7 Репрезентативность

6.5.7.1 Общие сведения

В *[12]* определена репрезентативность как степень, в которой набор данных отражает изучаемую целевую совокупность. Для машинного обучения с учителем набор обучающих данных можно рассматривать как подмножество более крупной совокупности, а эксплуатационные данные — как целевую совокупность, на основе которой можно делать выводы. Если обучающие данные недостаточно представляют эксплуатационные данные, обученная модель МО может не работать должным образом. В *[13]* описана смещенность в данных, возникающая в результате процесса отбора данных или процесса разметки данных.

Репрезентативность как характеристика качества данных связана с релевантностью в том смысле, что набор данных, который не представляет изучаемую целевую совокупность, вряд ли обеспечит хорошие предикторы для целевой переменной.

Примеры

1 Система распознавания лиц, обученная только на изображениях людей со светлым оттенком кожи, может некорректно идентифицировать людей при применении к изображениям людей с темным оттенком кожи.

2 Система прогнозного технического обслуживания, обученная только на данных электродвигателей, может не суметь корректно спрогнозировать необходимое техническое обслуживание применительно к двигателям внутреннего сгорания.

6.5.7.2 Показатели качества для репрезентативности

В таблице 21 представлены показатели качества данных для репрезентативности в контексте использования аналитики и машинного обучения.

Таблица 21 — Показатели качества для репрезентативности

Идентификатор	Наименование показателя	Описание	Функция измерения
Rep-ML-1	Коэффициент репрезентативности	Отношение соответствующих атрибутов, обнаруженных у субъектов целевой совокупности, к атрибутам, обнаруженным в наборе данных	$\frac{A}{B}$, где A — количество целевых атрибутов в наборе данных; B — количество релевантных атрибутов в конкретном контексте

6.5.8 Сходство

6.5.8.1 Общие сведения

Сходство набора данных означает сходство между выборками с точки зрения интересующих признаков. Это актуально для задач классификации (см. *[6]*, пункт 6.2.3), которые обычно реализуются с использованием машинного обучения с учителем (см. *[6]*, пункт 7.2). Это также актуально для задач кластеризации (см. *[6]*, пункт 6.2.4), которые обычно реализуются с использованием обучения без учителя (см. *[6]*, пункт 7.3). Для успешного выполнения задач классификации и кластеризации требуется адекватный уровень различий между выборками (см. *[13]*, раздел 5.2).

Модель МО, обученная на наборе данных, содержащем достаточно похожие изображения (например, созданные путем небольшого сдвига пикселей из нескольких исходных изображений), может оказаться переобученной и, следовательно, обладать меньшей возможностью для обобщения. В этом случае можно применить подходы, основанные на манипулировании данными, такие как вращение и сдвиг, которые могут улучшить возможности обобщения у модели МО. Эти подходы не будут работать, если количество исходных изображений ограничено. В этом случае следует проверить долю одинаковых образцов. Иной подход заключается в рассмотрении алгоритмов кластеризации с использованием концептуальных методов уменьшения дрейфа *[14]*.

Другие показатели выявляют сходство данных с помощью геометрического подхода: т. е. набор данных, состоящий из N записей данных и M признаков, может быть представлен как совокупность N векторов в M-мерном пространстве, поэтому его можно анализировать и сравнивать с использованием геометрических инструментов; в частности, сходство может быть связано с взаимным расположением векторов в пространстве.

6.5.8.2 Показатели качества для сходства

В таблице 22 представлены показатели качества данных для сходства в контексте использования аналитики и машинного обучения.

Таблица 22 — Показатели качества для сходства

Идентификатор	Наименование показателя	Описание	Функция измерения
Sim-ML-1	Сходство выборок	Доля схожих выборок в наборе данных	$1 - \frac{A}{B}$, где A — количество всех выборок в наборе данных; B — количество кластеров, полученное в результате применения алгоритма кластеризации ко всем элементам набора данных (примечание 7)
Sim-ML-2	Плотность выборок	Плотность нормализованного набора данных	$A - B$, где A — максимальное собственное значение G^*; B — минимальное собственное значение G^*

Окончание таблицы 22

Идентификатор	Наименование показателя	Описание	Функция измерения
Sim-ML-3	Независимость выборок	Соотношение анализа главных компонентов (PCA) и размерности набора данных	$1 - \frac{A}{B}$, где A — число главных компонентов при методе PCA с 95 % охватом суммы собственных значений (примечание 3); B — общее количество измерений набора данных

* G представляет собой матрицу с M строками и M столбцами и равна Φ_{norm} (см. ниже примечание 1).

Примечания

1 Φ_{norm} — это нормализованный набор данных, рассчитанный на основе $\Phi_{N\times M}$ (примечание 2) после вычитания из каждого столбца его среднего значения и нормализации до 1. Визуально нормализованные данные представлены на поверхности гиперсферы с радиусом, равным единице, и с центром в начале координат ($M \le N$) [*15*].

2 $\Phi_{N\times M}$ — это матрица размера N на M с N записями данных (векторами) и M признаками (размерами).

3 Число главных компонентов $K \le M$ — это наименьшее число собственных значений $C_{M\times M}$ (примечание 4), начиная с самого большого, выбранного так, чтобы представлять 95 % их суммы [16].

4 $C_{M\times M}$ представляет собой матрицу размера M на M с M строками и M столбцами и равна $\Phi_{mean}{}^{T}\,\Phi_{mean}$ (примечание 5).

5 Φ_{mean} рассчитывается на основе $\Phi_{N\times M}$ после вычитания из каждого столбца его среднего значения. Визуально нормализованные данные Φ_{mean} соответствуют (гипер)эллипсоиду с собственными векторами в качестве оси и центром в начале координат.

6 Основные компоненты могут быть выбраны по разным критериям или в процентном отношении. В приложении А дан пример модификации показателя.

7 Нулевое значение означает наименьшее сходство. Показатель сходства дает ноль, когда количество выборок равно количеству кластеров, что указывает на то, что ни одна выборка не похожа на другую.

6.5.9 Своевременность

6.5.9.1 Общие сведения

Своевременность означает задержку (т. е. ΔT_1) между моментом возникновения явления и моментом, когда данные, связанные с этим явлением, станут доступны для использования. Своевременность отличается от актуальности тем, что актуальность представляет собой разницу ΔT_2 между временем записи выборки данных и временем ее использования. Своевременность может быть важным компонентом: если разница ΔT_1 между явлением и доступностью соответствующей выборки данных слишком велика, она больше не может быть хорошим предиктором в контексте машинного обучения. Например, задачи МО для потоковой передачи данных (например, анализ транзакций с ценными бумагами, обучение с подкреплением, поисковые запросы) могут использовать непрерывное обучение и делать выводы практически в режиме реального времени.

6.5.9.2 Показатели качества для своевременности

В таблице 23 представлены показатели качества данных для своевременности в контексте использования аналитики и машинного обучения.

Таблица 23 — Показатели качества для своевременности

Идентификатор	Наименование показателя	Описание	Функция измерения
Tml-ML-1	Своевременность элементов данных	Доля элементов данных, отвечающих требованиям своевременности	$\frac{A}{B}$, где A — количество элементов данных в наборе данных, отвечающих требованиям своевременности; B — количество элементов данных в наборе данных

7 Реализация модели качества данных и показателей качества данных для задач аналитики или машинного обучения

Общий процесс реализации модели качества данных и связанных с ней показателей качества данных для задач аналитики или машинного обучения может включать в себя:

- отбор характеристик качества данных из настоящего стандарта, подходящих для задачи аналитики или машинного обучения;
- выбор для каждой характеристики качества данных соответствующих показателей качества данных из настоящего стандарта и их доработка для конкретного приложения;
- если для оценки качества данных необходимы дополнительные показатели качества данных, разработку показателей и соответствующих функций измерения качества данных можно проводить с использованием приложения А;
- пересмотр требований к качеству данных с критериями приемлемости для каждого показателя качества данных (такими как минимальное или максимальное пороговое значение, диапазон значений);
- применение функции измерения к целевым данным на соответствующих стадиях модели жизненного цикла качества данных для аналитики и машинного обучения (см. *ГОСТ Р 71484.1*);
- оценку соответствия каждого результата измерения качества данных заданным требованиям;
- оценку соответствия всего набора данных заданным требованиям;
- применение (при необходимости) процессов улучшения качества данных (см. *ГОСТ Р 71484.4*).
- постоянное отслеживание и улучшение качества данных, а также валидация процесса управления качеством данных на протяжении жизненного цикла задачи аналитики или машинного обучения (например, применяя функции измерения качества данных всякий раз, когда данные или условия задачи изменяются).

Примечание — *ГОСТ Р 71484.4* описывает структуру процесса управления качеством данных.

8 Отчетность о качестве данных

8.1 Структура отчетности о качестве данных

Отчеты о качестве данных могут предоставить соответствующим заинтересованным сторонам сведения по использованию качества данных, например модели качества данных, показатели качества данных и результаты их измерений, а также о том, соответствуют ли целевые данные требованиям к качеству данных. Данные и качество данных могут меняться со временем. Может возникнуть необходимость пересматривать отчеты о качестве данных по заданному графику в соответствии с рисками качества данных для задачи аналитики или машинного обучения.

Отчеты о качестве данных должны включать:

- назначение отчета (например, проинформировать соответствующие заинтересованные стороны, облегчить принятие решений, предоставить доказательства соответствия);
- предмет отчета (например, первоначальный отчет или редакция, охватываемый период времени, решаемая задача аналитики или МО);
- график пересмотра отчета;
- критерии завершения процесса отчетности;
- местонахождение и способ хранения отчетов (например, для последующего использования или аудита);
- элементы, описанные в 8.3.

8.2 Информация о показателях качества данных

На рисунке 4 показана модификация модели деятельности (см. *[9]*, приложение D), для аналитики и машинного обучения.

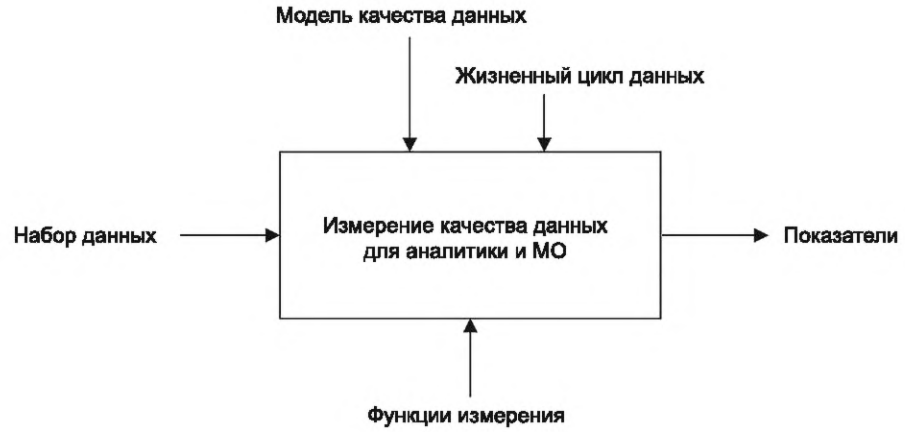


Рисунок 4 — Информация о показателях качества данных для отчетов о качестве

8.3 Руководство для организаций

При подготовке отчетов о качестве данных для задач аналитики или МО организация должна:

a) определить лиц, ответственных за подготовку, рассмотрение и утверждение отчетов о качестве данных;

b) определить соответствующие заинтересованные стороны, которые должны получать копии отчетов о качестве данных;

c) определить, в каких контрольных точках на стадиях жизненного цикла качества данных следует инициировать или пересматривать отчеты о качестве данных;

d) определить соответствующий интервал между пересмотрами отчета о качестве данных;

e) обеспечить, чтобы отчеты о качестве данных учитывались при планировании работы с данными;

f) определить охват целевых данных, включаемых в отчеты о качестве данных;

g) собрать требования к качеству данных;

h) документировать модель качества данных, включая характеристики качества данных, составляющие модель качества данных;

i) документировать отобранные показатели качества данных и их целевые значения;

j) документировать любые показатели качества данных, разработанные в соответствии с приложением А;

k) документировать результаты измерений всех выбранных показателей качества данных;

l) документировать все преобразования, произведенные с целевыми данными;

m) документировать оценку соответствия или несоответствия целевых данных требованиям к качеству данных;

n) документировать план улучшения качества данных, если целевые данные не соответствуют требованиям к качеству данных.

Приложение А
(справочное)

Проектирование и документирование функции измерения

Некоторые из показателей, перечисленных в серии стандартов *ГОСТ Р ИСО/МЭК 25000* и в настоящем стандарте, описаны с общей точки зрения. Для практического применения может возникнуть необходимость в детальном проектировании функции измерения и ее контекстной информации, как это описано в *ГОСТ Р ИСО/МЭК 25020*.

Ниже на примере показано как спроектировать и задокументировать соответствующую настоящему стандарту функцию измерения синтаксической аккуратности с помощью показателя Acc-I-1 (*см. [1]*).

В некоторых случаях (например, когда сравнение необходимо для оценки) показателя Acc-I-1 недостаточно, и необходимо получить представление о «синтаксически аккуратных значениях» для *A* и «количестве элементов данных, для которых требуется синтаксическая аккуратность» для *B*, то есть выбрать метод измерения расстояния между строками для *A* и определить область для *B*. Существует несколько возможностей для проектирования как *A*, так и *B*, например, задать:

A — как отношение количества строк, в которых расстояние равно нулю, к общему количеству строк в области *B*;

B — как количество допустимых строк.

Используя функцию измерения с *A* и *B*, пересмотренный показатель Acc-I-1-IT-2 приведен в таблице А.1.

Таблица А.1 — Синтаксическая аккуратность Acc-I-1-IT-2

Идентификатор	Наименование показателя	Описание	Функция измерения	Жизненный цикл данных. Целевые сущности. Свойства
Acc-I-1-IT-2	Синтаксическая аккуратность	Коэффициент близости значений данных к набору значений для заданной области	$1 - \frac{A}{B}$, где *A* — количество значений данных, для которых расстояние от заданной области равно нулю; *B* — количество значений для заданной области	Все стадии жизненного цикла данных, кроме стадии «Проектирование данных». Файл данных. Элемент данных, значение данных

Примечания

1 Наилучшее сходство между сравниваемыми строками и заданной областью достигается при меньшем расстоянии, поэтому чем меньше значение, тем лучше.

2 Идентификатор включает дополнение «IT-2», см. *ГОСТ Р ИСО/МЭК 25020—2023*, приложение С.

В таблице А.2 показано применение меры Acc-I-1-IT-2 для сравнения аккуратности двух баз данных, каждая из которых содержит по три имени. Сравнение производится с синтаксисом, состоящим из имен длиной 4. Имена длиной 4 представляют собой подмножество всех возможных строк длины 4.

По результатам применения функции измерения значения в базе данных *R*, выраженные меньшим значением показателя аккуратности, являются более аккуратными, чем значения в базе данных *W*, поскольку имя Marj не принадлежит к синтаксису имен.

Таблица А.2 — Примеры показателей для синтаксической аккуратности Acc-I-1-IT

Функция измерения	База данных R	База данных W
	John	John
	Jack	Jack
	Mary	Marj
$1 - \frac{A}{B}$,	$1 - \frac{3}{26^4 - Not\_adm}$	$1 - \frac{2}{26^4 - Not\_adm}$
Примечание — *Not_adm* — набор значений, которые не принадлежат синтаксису; 26^4 — *Not_adm* — набор значений, принадлежащих синтаксису.		

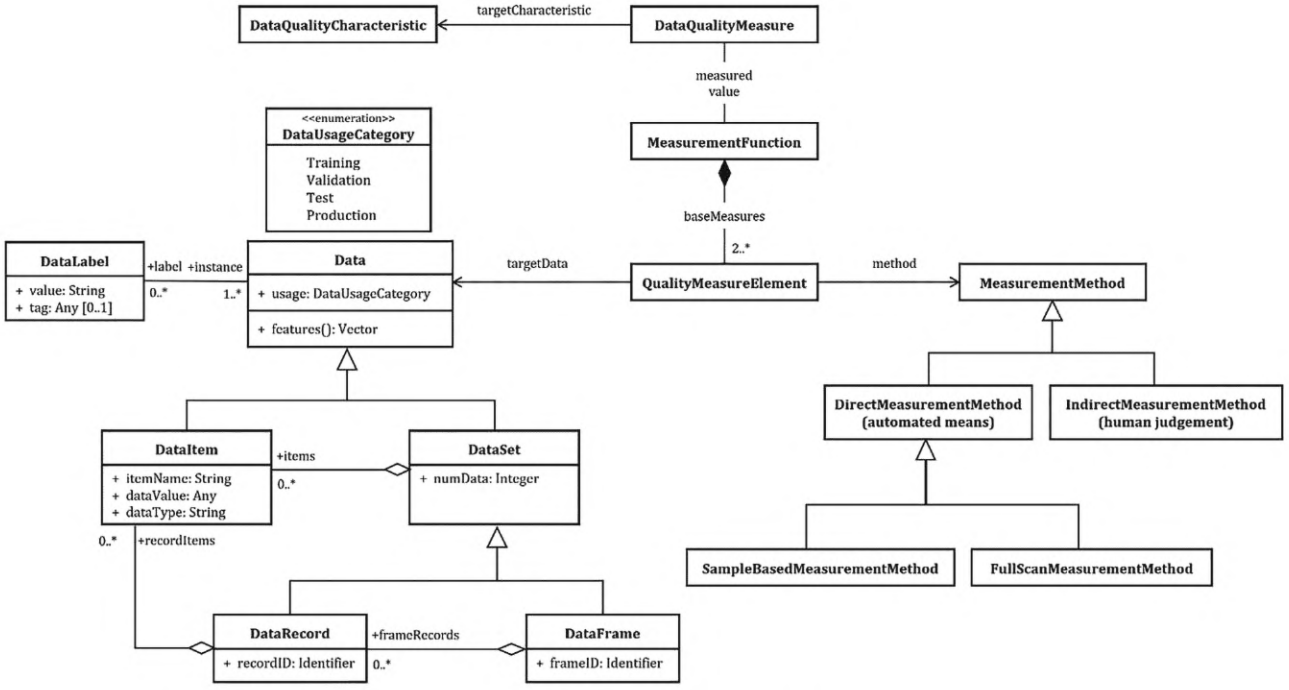
Примечание — Здесь 26 — количество возможных символов английского алфавита.

Приложение В
(справочное)

Модель структуры показателя качества данных (в нотации UML)

На рисунке В.1 показана диаграмма классов UML для структуры показателя качества данных, обеспечивающая общий словарь и взаимосвязи между компонентами показателей качества данных в настоящем стандарте. Подробную информацию о нотации UML см. *[17]*.

На рисунке В.1 описана структура измерения характеристик качества данных на основе эталонной модели измерения качества из *ГОСТ Р ИСО/МЭК 25020*. Показатель качества данных связан с целевой характеристикой качества данных. Каждый показатель качества данных определяется как результат измерения, возвращаемый функцией измерения двух или более элементов показателя качества данных. Элемент показателя качества данных — это базовая мера, связанная с методом измерения и с целевыми данными. Метод измерения — это общее описание логической структуры операций, используемых для количественной оценки базовой меры с помощью автоматизированных средств или человеческого суждения.



DataQualityCharacteristic – характеристика качества данных;
DataUsageCategory – категория использования данных;
Data – данные;
DataLabel – метка данных;
DataItem – элемент данных;
DataSet – набор данных
DataRecord – запись данных;
DataFrame – фрейм данных;
targetData – целевые данные;
label – метка;
instance – экземпляр;
items – элементы;
recorditems – элементы записи;
frameRecords – элементы фрейма;

DataQualityMeasure – показатель качества данных;
MeasurementFunction – функция измерения;
QualityMeasureElement – элемент показателя качества;
MeasurementMethod – метод измерения;
DirectMeasurementMethod (automated means) – прямой метод измерения (средства автоматизации);
IndirectMeasurementMethod (human judgement) – непрямой метод измерения (суждение человека);
SampleBasedMeasurementMethod – метод измерения на основе выборки;
FullScanMeasurementMethod – метод измерения полным сканированием;
targetCharacteristic – целевая характеристика;
enumeration – перечисление;
measured value – измеренное значение;
baseMeasures – базовые показатели;
method – метод

Рисунок В.1 — Структура измерения качества данных

Приложение С
(справочное)

Обзор характеристик качества данных

Чтобы представить взаимосвязь между элементами данных и общими характеристиками качества набора данных, полезно их объединить в четыре группы: сопровождаемость, достоверность, надежность и верность *[18]*.

Примечание — Сопровождаемость, достоверность, надежность и верность не являются характеристиками сами по себе и не связаны с теми же или разными терминами, используемыми в *ГОСТ Р ИСО/МЭК 25010*. В частности, сопровождаемость не связана с возможностью поддерживать данные (как это предусмотрено в *ГОСТ Р ИСО/МЭК 25010*), а связана с прозрачностью данных (соответствие, подотчетность, документирование, идентификация происхождения).

- **Сопровождаемость.** Сопровождаемость данных относится к свойству набора данных (на основе лучших практик жизненного цикла разработки программного обеспечения), которое указывает на уровень соответствия практикам разработки набора данных и уровень прозрачности для поддержки принятия решений и подотчетности с набором данных. Это относится к существованию и актуальности документации, определяющей происхождение версий набора данных и его обладателей, а также к механизму контроля версий. Это группа качественных характеристик для доступности, проверяемости, переносимости и понятности целевых данных как элементов данных и наборов данных, для идентификации элементов данных и эффективности наборов данных.
- **Достоверность.** Достоверность набора данных относится к эксплуатационной достоверности набора данных, т. е. насколько хорошо он описывает (в целом или на уровне элемента) явления, для которых он предназначен, и какова его актуальность. Например, учитывает ли набор данных всю сложность и спектр субъективности явлений, с которыми он связан, позволяет ли он делать обобщения в рамках явлений и каковы справочные материалы для актуальности и доступности. Достоверность — это всеобъемлющая группа характеризующая актуальность как для элемента данных, так и для набора данных.
- **Надежность.** Надежность набора данных относится к внутренней достоверности набора данных, т. е. к его достоверности, вытекающей из его последовательности, тиражируемости и воспроизводимости в целевых данных элементов данных и наборов данных. Надежность тесно связана с сопровождаемостью, поскольку наборы данных, которые не поддерживаются должным образом, часто не воспроизводимы и не поддаются проверке на предмет согласованности. Кроме того, показатели аккуратности и аккуратности можно использовать для определения аспектов надежности набора данных.
- **Верность.** Верность набора данных относится к характеристикам наполненности, согласованности, сбалансированности и репрезентативности набора данных по отношению к явлению, с которым он связан. Как правило, способы выборки данных из более крупных источников или выбор самих источников могут повлиять на верность получаемого набора данных с точки зрения его разнообразности и появления возможных систематических ошибок, таких как временные, географические, культурные и другие социальные предубеждения. Верность также связана с надежностью с точки зрения таких свойств, как релевантность и своевременность.

В таблице С.1 представлена группа характеристик качества данных. Качество набора данных можно измерить двумя способами: качество отдельных элементов данных и интегральный показатель качества всего набора данных. Отсюда вытекает разделение характеристик качества данных (в таблице С.1) для элементов данных и для набора данных. Каждая характеристика качества данных может быть применена к целевым данным либо для элемента данных, либо для набора данных.

Таблица С.1 — Группа характеристик качества данных

Группы характеристик	Показатели качества данных	Целевые формы данных	
		Элемент данных	Набор данных
Сопровождаемость	Доступность	✓	✓
	Проверяемость	✓	✓
	Эффективность	—	✓
	Переносимость	✓	✓
	Идентифицируемость	✓	—
	Восстанавливаемость	✓	✓
	Понятность	✓	✓

Окончание таблицы С.1

Группы характеристик	Показатели качества данных	Целевые формы данных	
		Элемент данных	Набор данных
Обоснованность	Готовность	✓	✓
	Актуальность	✓	✓
	Результативность	—	✓
Надежность	Аккуратность	✓	✓
	Соответствие	✓	✓
	Достоверность	✓	✓
	Точность	✓	✓
	Прослеживаемость	✓	✓
Верность	Наполненность	✓	✓
	Сбалансированность	—	✓
	Согласованность	✓	✓
	Разнообразность	—	✓
	Релевантность	—	✓
	Репрезентативность	—	✓
	Сходство	—	✓
	Своевременность	—	✓

Приложение D
(справочное)

Альтернативные группы характеристик качества данных

В таблице D.1 показан пример сопоставления характеристик качества данных со следующими альтернативными группами:

- техничность;
- законность;
- ориентация на пользователя;
- ориентация на реальность.

Таблица D.1 — Группа характеристик качества данных

Группы характеристик	Характеристики качества данных	Формы целевых данных	
		Элемент данных	Набор данных
Технические особенности	Переносимость	✓	✓
	Эффективность	—	✓
	Проверяемость	✓	✓
	Прослеживаемость	✓	✓
	Восстанавливаемость	✓	✓
Законность	Соответствие	✓	✓
	Идентифицируемость	✓	—
Ориентированные на пользователя	Доступность	✓	✓
	Готовность	✓	✓
	Понятность	✓	✓
Ориентированные на реальность	Аккуратность	✓	✓
	Наполненность	✓	✓
	Согласованность	✓	✓
	Достоверность	✓	✓
	Актуальность	✓	✓
	Точность	✓	✓
	Результативность	—	✓
	Релевантность	—	✓
	Своевременность	—	✓
	Репрезентативность	—	✓
	Сбалансированность	—	✓
	Сходство	—	✓
	Разнообразность	—	✓

Приложение Е
(справочное)

Сравнение характеристик качества данных ИСО/МЭК 25012 с настоящим стандартом

В таблице Е.1 приведены характеристики качества данных ИСО/МЭК 25012 и дополнительные характеристики, представленные в настоящем стандарте.

Таблица Е.1 — Применение характеристик качества данных к элементам данных и наборам данных

ИСО/МЭК 25012		Настоящий стандарт		
Точка зрения	Характеристики качества данных	Дополнительные характеристики качества данных	Элемент данных	Набор данных
Внутренне присущие	Аккуратность		☐	☐
	Наполненность		☐	☐
	Согласованность		☐	☐
	Достоверность		☐	☐
	Актуальность*	Своевременность	☐	☐
		Проверяемость	☐	☐
		Результативность	—	☐
		Релевантность	—	☐
		Репрезентативность	—	☐
		Сбалансированность	—	☐
		Сходство	—	☐
Внутренне присущие и системно зависимые	Доступность		☐	☐
	Соответствие		☐	☐
	Конфиденциальность**	Идентифицируемость	☐	—
	Эффективность		☐	☐
	Точность		☐	☐
	Прослеживаемость		☐	☐
	Понятность		☐	☐
Системно зависимые	Готовность		☐	☐
	Переносимость		☐	☐
	Восстанавливаемость		☐	☐

* Характеристика, семантически связанная с понятием времени в характеристике актуальности, см. *[1]*, пункт 8.6.

** Характеристика, семантически связанная с понятием защиты персональных данных в конфиденциальности, см. *[1]*, пункт 8.9.

Примечание — Доступность (для предполагаемых пользователей) и эффективность (формат и пространство) могут применяться к элементу данных.

Приложение ДА
(справочное)

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в применяемом международном стандарте

Таблица ДА.1

Обозначение национального стандарта	Степень соответствия	Обозначение и наименование ссылочного международного стандарта
ГОСТ Р 71476—2024 (ИСО/МЭК 22989:2022)	MOD	ISO/IEC 22989:2022 «Искусственный интеллект. Концепции и терминология искусственного интеллекта»
ГОСТ Р 71484.1—2024 (ИСО/МЭК 5259-1:2024)	MOD	ISO/IEC 5259-1:2024 «Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, термины и примеры»
Примечание — В настоящей таблице использовано следующее условное обозначение степени соответствия стандартов: - MOD — модифицированные стандарты.		

Библиография

[1] *ИСО/МЭК 25024:2015 Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Определение качества данных [Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality]*

[2] ИСО/МЭК 25012:2008 Программная инженерия. Требования и оценка качества программного продукта (SQuaRE). Модель качества данных [Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model]

[3] ИСО 5127:2017 Информация и документация. Основные положения и словарь (Information and documentation — Foundation and vocabulary)

[4] ИСО/МЭК Guide 99:2007 Международный словарь по метрологии. Основные и общие понятия и соответствующие термины (VIM) [International vocabulary of metrology — Basic and general concepts and associated terms (VIM)]

[5] ИСО/МЭК 30137-4:2001 Информационные технологии. Применение биометрии в системах видеонаблюдения. Часть 4. Процедура видеоаннотации (Information technology — Use of biometrics in video surveillance systems — Part 4: Ground truth and video annotation procedure)

[6] *ПНСТ 838—2023/ИСО/МЭК 23053:2022 Искусственный интеллект. Структура описания систем искусственного интеллекта, использующих машинное обучение [Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)]*

[7] ИСО 16781:2021 Космические системы. Требования к моделированию системы управления (Space systems — Simulation requirements for control system)

[8] ИСО/МЭК 2382:2015 Информационная технология. Словарь (Information technology — Vocabulary)

[9] ИСО 8000-8:2015 Качество данных. Часть 8. Качество информации и данных. Понятия и измерение (Data quality — Part 8: Information and data quality: Concepts and measuring)

[10] ИСО 8000 (все части) Качество данных (Data quality)

[11] Natale D., Paoletti M.C., Simonetta A. Data Quality and Statistical Information. Journal of accidents and occupational diseases, 2012, 281-296 (in Italian) https://www.inail.it/cs/internet/docs/alg-ass-stat-la-qualita-dei-dati-e-l-informazione-statistica.pdf

[12] ИСО 20252:2019 Исследование рынка, общественного мнения и социальных проблем. Словарь и сервисные требования (Market, opinion and social research, including insights and data analytics — Vocabulary and service requirements)

[13] *ПНСТ 839—2023 (ISO/IEC TR 24027:2021) Искусственный интеллект. Смещенность в системах искусственного интеллекта и при принятии решений с помощью искусственного интеллекта [ISO/IEC TR 24027:2022, Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI-aided decision making]*

[14] Webb G.I., Lee L.K., Goethals B. et al., Analyzing concept drift and shift from sample data. Data Mining and Knowledge Discovery. 2018, 32, 1179−1199. doi:10.1007/s10618-018-0554-1

[15] Trenta A. Data bias measurement: a geometrical approach through frames. in Proceedings APSEC IWESQ 2021 [online]. 2021, 11-16. http://ceur-ws.org/Vol-3114/paper-03.pdf

[16] Trenta A. ISO/IEC 25000 quality measures for A.I.: a geometrical approach. in Proceedings APSEC IWESQ 2020 [online]. 2020, 20−21. http://ceur-ws.org/Vol-2800/paper-05.pdf

[17] ISO/IEC 19505-1:2012, Information technology — Object Management Group Unified Modeling Language (OMG UML) — Part 1: Infrastructure

[18] Aroyo, L., Lease, M., Paritosh, P. and Schaekermann, M. Data Excellence for AI: Why should you care? ACM Interactions. XXIX.2, 66. https://interactions.acm.org/archive/view/march-april-2022/data-excellence-for-ai

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: информационные технологии, искусственный интеллект, качество данных, машинное обучение, большие данные, аналитика больших данных, показатель качества данных, характеристика качества данных

Редактор *Н.А. Таланова*
Технический редактор *И.Е. Черепкова*
Корректор *Р.А. Ментова*
Компьютерная верстка *Е.А. Кондрашовой*

Сдано в набор 25.11.2024. Подписано в печать 04.12.2024. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 4,65. Уч.-изд. л. 3,95.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru