ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ
РОССИЙСКОЙ
ФЕДЕРАЦИИ

ГОСТ (Р)
_____–_____.

Информационные технологии

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры.

(ISO/IEC WD 5259-1:202X(E))

Настоящий проект стандарта не подлежит применению до его принятия

Москва
Стандартинформ
20XX

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__ _____ 202_ г. № ____.

4 Настоящий стандарт идентичен разрабатываемому международному стандарту ISO/IEC WD 5259-1:202X(E) Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples, учитывающих специфику национальной стандартизации в области искусственного интеллекта и больших данных.

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 об-

щего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

© ISO, 20XX — Все права сохраняются
© Стандартинформ, оформление, 20XX

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

Содержание

Предисловие II
Введение V
1 Область применения 1
2 Нормативные ссылки.....1
3 Термины и определения.....1
4 Обозначения и сокращения 7
5 Концепции качества данных для аналитики и машинного обучения 8
5.1 Рекомендации по качеству данных для аналитики и машинного обучения..8
5.1.1 Общее 8
5.1.2 Машинное обучение и качество данных.....9
5.1.3 Большие данные и качество данных для аналитики и машинного обучения 10
5.1.4 Совместное использование данных, повторное использование данных и качество данных для аналитики и машинного обучения.....10
5.2 Концептуальная структура качества данных для аналитики и машинного обучения 11
5.2.1 Обзор.....11
5.2.2 Управление качеством данных 12
5.2.3 Стратегическое управление качеством данных.....16
5.2.4 Происхождение данных 17
5.3 Жизненный цикл данных для аналитики и машинного обучения 17
5.3.1 Обзор.....17
5.3.2 Модель жизненного цикла данных.....18
5.3.3 Многоэтапные процессы.....22
Приложение А 24
Библиография 26

Введение

Большие данные — это исходный материал для аналитики и машинного обучения, а их качество — критически важный аспект для связанных проектов и систем. Цель этого документа — помочь организациям достичь эффективного управления качеством данных с помощью стандартизированного процесса обеспечения качества данных.

Этот документ является частью серии ИСО/МЭК 5259. Цель серии ИСО/МЭК 5259 — предоставить инструменты и методы для оценки и улучшения качества данных, используемых для аналитики и машинного обучения.

Другие части серии ИСО/МЭК 5259 включают:

ИСО/МЭК 5259-2 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 2. Меры качества данных;

ИСО/МЭК 5259-2 предоставляет модель качества данных, показатели качества данных и рекомендации по отчетности о качестве данных в контексте аналитики и машинного обучения. ИСО/МЭК 5259-2 основан на серии стандартов ИСО 8000, ИСО/МЭК 25012 и ИСО/МЭК 25024.

Цель ИСО/МЭК 5259-2 состоит в том, чтобы помочь организациям достичь своих целей в отношении качества данных. Стандарт применим в организациях любого типа.

ИСО/МЭК 5259-3 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 3. Требования и рекомендации по управлению качеством данных;

ИСО/МЭК 5259-3 устанавливает требования и предоставляет рекомендации по установлению, внедрению, поддержанию и постоянному улучшению качества данных, используемых в областях аналитики и машинного обучения.

ИСО/МЭК 5259-3 не определяет подробный процесс, методы или показатели. Скорее, он определяет требования и руководство для процесса управления качеством, а также справочный процесс и методы, которые могут быть адаптированы для выполнения требований, изложенных в этом документе.

Требования и рекомендации, изложенные в ИСО/МЭК 5259-3, являются общими и предназначены для применения в любых организациях независимо от их типа, размера или характера.

ИСО/МЭК 5259-4 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 4. Структура процесса обеспечения качества данных.

ИСО/МЭК 5259-4 обеспечивает общие организационные подходы, независимо от типа, размера или характера организации-заявителя, для обеспечения качества данных для обучения и оценки в области аналитики и машинного обучения. Он включает в себя рекомендации по:

- Машинному обучению с учителем в отношении маркировки данных, используемых для обучения систем машинного обучения, включая общие организационные подходы к маркировке обучающих данных;
- Машинному обучению без учителя;
- Машинному обучению с частичным привлечением учителя;
- Машинному обучению с подкреплением.

ИСО/МЭК 5259-4 применим к данным обучения и оценки, которые поступают из разных источников, включая комплектование и композицию данных, предварительную обработку данных, маркировку данных, оценку и использование данных. ИСО/МЭК 5259-4 не определяет конкретные услуги, платформы или инструменты.

ГОСТ (Р) _____—_____.
(проект)

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Информационные технологии

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры

Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples

---

Дата введения – 2022-__-__

1 Область применения

Данный документ предоставляет средства для понимания и сопоставления отдельных документов серии ИСО/МЭК «Искусственный интеллект — качество данных для аналитики и машинного обучения» и является основой для концептуального понимания качества данных для аналитики и машинного обучения. В нем также обсуждаются взаимосвязанные технологии и примеры (например, варианты использования и сценарии использования).

2 Нормативные ссылки

В настоящем документе нет нормативных ссылок.

3 Термины и определения

Для целей настоящего документа применяются термины и определения, приведенные в ИСО/МЭК 20546, ИСО/МЭК 20547-3, ИСО/МЭК 22989 и ИСО/МЭК 23053, а также следующие.

ИСО и МЭК поддерживают терминологические базы данных для использования в стандартизации, расположенные по следующим адресам:

Платформа ИСО для онлайн-просмотра материалов по стандартам (Online Browsing Platform, OBP) доступна по адресу https://www.iso.org/obp

База данных МЭК «Электропедия» (IEC Electropedia) доступна по адресу http://www.electropedia.org/

3.1

Жизненный цикл данных

жизненный цикл данных

все этапы в процессе использования данных от концепции идеи до его прекращения

[ИСТОЧНИК: ISO 21710:2020, 3.12, изменено — жизненный цикл данных был добавлен в качестве предпочтительного термина, а «первоначальное создание» заменено на «концепцию идеи»]

3.2

Оператор (держатель) данных

сторона, имеющая юридический контроль над разрешением обработки данных другими сторонами

[ИСТОЧНИК: ИСО/МЭК 23751:2022, 3.4, изменено — исходное примечание к записи было удалено.]

3.3

Пользователь данных

сторона, которая уполномочена выполнять обработку данных под юридическим контролем держателя данных (3.2)

[ИСТОЧНИК: ИСО/МЭК 23751:2022, 3.5]

3.4

Качество данных

свойство данных соответствовать требованиям организации к данным в определённых условиях

3.5

Характеристика качества данных

категория атрибутов качества данных, имеющих отношение к качеству данных (3.4)

[ИСТОЧНИК: ИСО/МЭК 25012:2008, 4.4]

3.6

Модель качества данных

определенный набор характеристик, обеспечивающий основу для определения требований к качеству данных и оценки качества данных (3.4)

[ИСТОЧНИК: ИСО/МЭК 25012:2008, 4.6]

3.7

Мера качества данных

переменная, которой присваивается значение в результате измерения характеристики качества данных (3.5)

[ИСТОЧНИК: ИСО/МЭК 25012:2008, 4.5]

3.8

Измерение

совокупность операций, имеющих целью определение значения меры

[ИСТОЧНИК: ИСО/МЭК 25010:2011, 4.4.7]

3.9

Метод измерения

логическая организация операций, описанная в общем, используемая при измерении (3.8)

[ИСТОЧНИК: ИСО/МЭК 25021:2012, 4.8]

3.10

Требование по качеству

Требование к характеризующим качество свойствам или атрибутам продукта, данных или услуги ИКТ, которые удовлетворяют потребности, вытекающие из цели, для которой этот продукт, данные или услуга ИКТ должны использоваться

[ИСТОЧНИК: ИСО/МЭК 25030:2019, 3.15, изменено — исходное примечание к записи было удалено.]

3.11

Атрибут

Свойство или характеристика объекта, которое может быть количественно или качественно определено человеком или с помощью автоматизированных средств

[ИСТОЧНИК: ИСО/МЭК/ИИЭР 15939:2017, 3.2]

3.12

Шкала

упорядоченный набор значений, непрерывный или дискретный, или набор категорий, которым сопоставлен атрибут

Примечание 1: Тип шкалы зависит от характера взаимосвязи между значениями на шкале. Обычно выделяют четыре типа шкал:

- номинальный: значения измерений являются категориальными;
- порядковый: значения измерений являются ранжированием;
- интервал: значения измерений имеют равные расстояния, соответствующие равным количествам признака; а также
- отношение: значения измерений имеют равные расстояния, соответствующие равным количествам атрибута, где значение нуля не соответствует ни одному из атрибутов.

Это всего лишь примеры типов шкалы. Робертс [22] определяет больше типов шкал. Приложение А содержит примеры каждого типа шкалы.

Примечание 2: На основании определения «шкала (измерительного прибора)» в Международном словаре по метрологии – Основные и общие понятия и связанные термины, 2012 г.

[ИСТОЧНИК: ИСО/МЭК/ИИЭР 15939:2017, 3.34]

3.13

Аналитика

аналитика данных

составная концепция, охватывающая комплектование данных, сбор данных, проверку данных, обработку данных, включая количественную оценку данных, визуализацию данных, документирование данных и интерпретацию данных.

Примечание 1: Аналитика данных используется для понимания объектов, представленных данными, чтобы делать прогнозы для конкретной ситуации и рекомендовать шаги для достижения целей. Новые знания, полученные с помощью аналитики, используются для различных целей, таких как принятие решений, исследования, устойчивое развитие, проектирование, планирование и т. д.

[ИСТОЧНИК: ИСО/МЭК 20546:2019, 3.1.6, изменено — аналитика добавлена в качестве предпочтительного термина, добавлена документация данных]

3.14

Большие данные

Большие данные (big data): Большие массивы данных, отличающиеся главным образом такими характеристиками, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа.

Примечание – Термин «большие данные» широко применяется в различных значениях, например: в качестве наименования технологии масштабирования, используемой для обработки больших массивов данных.

[ИСТОЧНИК: ИСО/МЭК 20546:2019, 3.1.2]

3.15

Признак

<машинное обучение> измеримое свойство объекта или события по отношению к набору характеристик

Примечание 1: Признаки используются в обучении и прогнозировании.

Примечание 2: Признаки обеспечивают машиночитаемый способ описания соответствующих объектов. Поскольку алгоритм не будет возвращаться к самим объектам или событиям, представления признаков разработаны таким образом, чтобы содержать всю полезную информацию.

[ИСТОЧНИК: ИСО/МЭК 23053: 2022, 3.4.5]

3.16

Управление качеством данных

скоординированные действия по оперативному руководству и контролю организации в отношении качества данных (3.4)

[ИСТОЧНИК: ИСО 8000-2:2020, 3.8.2]

3.17

Стратегическое управление данными

Разработка и обеспечение исполнения политик, связанных с менеджментом данных. [ГОСТ Р ИСО/МЭК 24668 (проект) «Информационные технологии. Искусственный интеллект. Структура управления процессами аналитики больших данных», 3.4.]

Свойство или способность, которые необходимо координировать и реализовывать с помощью набора действий, направленных на разработку, реализацию и мониторинг стратегического плана управления активами данных. [ИСО/МЭК 5259-1]

Примечание 1: Стратегическое управление данными описано в ИСО/МЭК 38505-1.

Примечание 2: Актив данных понимается как набор элементов данных или объектов данных, которые имеют реальную или потенциальную выгоду для организации. Актив данных — это подмножество активов, определенных в ISO 55000 [21]. Выгода — это преимущество для организации на основе практических знаний, полученных из аналитической системы. Ее часто приписывают большим данным из-за понимания того, что данные имеют потенциальную пользу, которая ранее обычно не рассматривалась.

Примечание 3: Стратегический план управления активами данных — это документ, определяющий, как управление данными должно быть согласовано со стратегией организации. Этот термин имеет то же значение, что и план стратегического управления активами (SAMP), определенный в ISO 55000 с точки зрения данных.

[ИСТОЧНИК: ИСО/МЭК 20547-3:2020, 3.7]

3.18

Происхождение данных

происхождение

Сведения об организации или физическом лице, которое создало, накопило, сохранило и использовало данные

[ИСТОЧНИК: ISO 19160-1:2015, 4.16, изменено — в качестве предпочтительного термина добавлено происхождение данных, «использовало запись» заменена на «использовало данные», а исходное примечание к записи было удалено.]

3.19

Визуализация

научная визуализация

<компьютерная графика> использование компьютерной графики и обработки изображений для представления моделей или характеристик процессов или объектов для поддержки человеческого понимания

Примечание 1: Примеры: Отображаемое изображение, созданное путем объединения магнитно-резонансных сканирований опухоли; объемные виды озера сверху и сбоку, показывающие данные о температуре; двумерная модель ЭКГ.

Примечание 2: научная визуализация; визуализация: термины и определения, стандартизированные ИСО/МЭК [ИСО/МЭК 2382-13:1996].

Примечание 3: 13.01.07 (2382)

[ИСТОЧНИК: ИСО/МЭК 2382-13:1996. изменено — поменять местами предпочтительный термин с допустимым термином]

4 Обозначения и сокращения

ИИ — искусственный интеллект (artificial intelligence)

BDC — потребитель больших данных (big data consumer)

DL — глубокое обучение (deep learning)

DLC — жизненный цикл данных (data life cycle)

DQ — качество данных (data quality)

ML — машинное обучение (machine learning)

PII — персональные данные (personally identifiable information)

5 Концепции качества данных для аналитики и машинного обучения

5.1 Рекомендации по качеству данных для аналитики и машинного обучения

5.1.1 Общее

Существующие стандарты качества данных, такие как серия ISO 8000 [15], были разработаны с точки зрения производства данных и управления ими. Это связано с тем, что производители данных (или сборщики данных) были крупнейшими потребителями данных. Поскольку большая часть данных использовалась для заранее определенной цели, а соответствующие стандарты качества данных были сосредоточены только на характеристиках, необходимых для определенной цели, данные, полученные таким образом, могут потребовать дополнительной обработки для повторного использования.

В области анализа данных и машинного обучения пользователи данных (т. е. потребители больших данных) обычно не производят данные. Они ищут, собирают и обрабатывают данные, которые, по их мнению, необходимы и подходят для их аналитики и машинного обучения. В этом случае качество данных влияет на качество результатов анализа и производительность модели машинного обучения. Каким бы хорошим ни был анализ данных или алгоритм машинного обучения, результаты могут быть ненадежными при использовании данных низкого качества. Даже качественные данные для конкретного приложения могут быть оценены как некачественные для других приложений. Данные низкого качества для конкретной цели могут привести к неверным прогнозам, сбоям в работе или угрозам безопасности. Таким образом, для обеспечения качества данных, используемых для аналитики и машинного обучения, в серии стандартов ИСО/МЭК 5259 определены характеристики качества данных, меры качества данных, тре-

бования к управлению качеством данных и репрезентативный процесс управления качеством данных в течение жизненного цикла данных.

5.1.2 Машинное обучение и качество данных

ИСО/МЭК 22989 определяет машинное обучение как процесс оптимизации параметров модели с помощью вычислительных методов, чтобы поведение модели отражало данные или опыт. В ИСО/МЭК 23053 машинное обучение далее описывается как ветвь ИИ, в которой используются вычислительные методы, позволяющие системам учиться на основе данных или опыта. Машинное обучение можно использовать для выполнения самых разных задач на основе данных и алгоритмов машинного обучения. Данные, используемые в машинном обучении, классифицируются как данные обучения, данные проверки, данные тестирования и производственные данные. В машинном обучении с учителем модель машинного обучения создается путем обучения алгоритма машинного обучения с помощью обучающих данных. Затем данные проверки и тестирования используются, чтобы убедиться, что обученная модель машинного обучения работает в пределах допустимых параметров. Далее обученная модель машинного обучения используется для расчета выводов на основе производственных данных. Производительность обученной модели машинного обучения зависит от характеристик качества всех этих типов данных. ИСО/МЭК 23053 описывает несколько общих типов алгоритмов машинного обучения, и они могут иметь различную чувствительность к различным характеристикам качества данных.

ПРИМЕР 1

Репрезентативность — одна из важнейших характеристик качества данных для машинного обучения. Когда обучающие данные не представляют совокупность, включенную в производственные данные, обученная модель машинного обучения с большей вероятностью сделает неправильные выводы из производственных данных. При использовании для принятия решений о людях это может привести к дискриминационным или предвзятым действиям в отношении недостаточно представленных групп людей.

ПРИМЕР 2

Обучение алгоритма машинного обучения для создания обученной модели машинного обучения — это математический процесс, который перебирает набор обучающих данных, представляющих атрибуты вещи или явления. Точность каждой выборки в обучающих данных будет влиять на обученную модель машинного обучения. Если слишком много выборок в обучающих данных неточны, модель, скорее всего, сделает неверные выводы о производственных данных.

ПРИМЕЧАНИЕ См. ИСО/МЭК 5259-2 для получения подробной информации о том, как характеристики качества данных влияют на производительность моделей машинного обучения.

5.1.3 Большие данные и качество данных для аналитики и машинного обучения

Наборы данных, имеющие характеристики больших данных (то есть объем, разнообразие, скорость и изменчивость), часто используются в аналитике и машинном обучении. Разнообразие и изменчивость больших данных может повлиять на модель качества данных и связанные с ней меры качества данных. Объем и скорость больших данных могут потребовать использования автоматизированных инструментов для измерения качества данных и оценки того, соответствуют ли данные требованиям. Объем больших данных также может создавать проблемы для своевременного измерения и оценки качества данных.

5.1.4 Совместное использование данных, повторное использование данных и качество данных для аналитики и машинного обучения

Одни и те же данные можно использовать для разных задач аналитики или машинного обучения. Например, оператор (держатель) данных может предоставить к ним доступ ряду пользователей данных (как внутренних, так и внешних по отношению к организации, к которой принадлежит оператор данных). Точно так же пользо-

вателю данных может быть разрешено использовать данные для более чем одной задачи.

Различные задачи аналитики и машинного обучения могут иметь разные требования к качеству данных. Различные требования к качеству данных могут повлиять на выбор модели качества данных, соответствующих показателей качества данных и критериев оценки.

5.2 Концептуальная структура качества данных для аналитики и машинного обучения

5.2.1 Обзор

Рис. 1. Концептуальная структура качества данных (КД) для аналитики и машинного обучения

На рис. 1 представлена репрезентативная структура для определения, оценки и улучшения качества набора данных для использования в аналитике и машинном обучении. Цель структуры на рис. 1 — определить процессы, которые можно использовать для определения и обеспечения того, чтобы набор данных соответствовал нуждам и требованиям организации.

К качеству данных в рамочной структуре имеют отношение следующие процессы: построение модели качества данных, выбор мер обеспечения качества данных, оценка качества данных, улучшение качества данных и подготовка отчетности по качеству (информирование о качестве) данных. Среди других важных процессов можно назвать процессы стратегического и оперативного управления, обеспечения безопасности, защиты персональных данных и отслеживания происхождения данных.

На рис. 1 показано, что меры качества данных, оценка качества данных и процессы улучшения качества данных могут быть итеративными, когда это необходимо для удовлетворения потребностей организаций и требований к набору данных.

Кроме того, для непрерывного машинного обучения (т. е. когда алгоритм машинного обучения непрерывно обучается на новых данных) эти процессы также могут применяться непрерывно в течение жизненного цикла системы.

5.2.2 Управление качеством данных

5.2.2.1 Модель качества данных

Для целей настоящего документа модель качества данных представляет собой определенный набор характеристик, который обеспечивает основу для определения требований к качеству данных и оценки качества данных. Пользователи данных могут создавать модели качества данных для аналитики и машинного обучения в соответствии со своими бизнес-целями. Модель качества данных может включать следующее:

- описание контекста использования аналитики и машинного обучения;
- требования к качеству данных;
- характеристики качества данных, связанные с жизненным циклом данных;
- функция измерения и шкала для каждой характеристики качества данных.

Пункты 5.4 и 5.5 стандарта ИСО/МЭК 5259-2:20XX [1] с точки зрения внутренне присущих данным характеристик, и с точки зрения системно-зависимых характе-

ристик определяют восемнадцать характеристик качества данных для аналитики и машинного обучения.

ИСО/МЭК 25012 описывает два типа характеристик качества данных как:

- Неотъемлемое качество данных относится к степени, в которой характеристики качества данных имеют внутренний потенциал для удовлетворения заявленных и подразумеваемых потребностей, когда данные используются в определенных условиях;
- Качество данных, зависящее от системы, относится к степени, в которой качество данных достигается и сохраняется в компьютерной системе, когда данные используются в определенных условиях.

Характеристики качества данных, описанные в ИСО/МЭК 5259-2:

- корректность (accuracy, правильность измерений);
- точность (precision, прецизионность измерений);
- полнота;
- согласованность;
- релевантность;
- масштабируемость данных;
- охват контекста;
- переносимость;
- своевременность;
- актуальность;
- идентифицируемость;
- проверяемость;
- достоверность;
- понятность;
- сбалансированность;
- эффективность;
- сходство;
- репрезентативность.

ПРИМЕЧАНИЕ См. ИСО/МЭК 5259-2 для описания этих характеристик качества данных, используемых для аналитики и машинного обучения.

ПРИМЕР

На рис. 2 показана взаимосвязь модели качества данных для аналитики и машинного обучения, жизненного цикла данных, требований к качеству данных и характеристик качества данных для аналитики и машинного обучения. Пользователи данных создают модели качества данных для аналитики и машинного обучения в соответствии со своими бизнес-целями. Модель качества данных может отражать измеримые характеристики качества данных для каждого этапа жизненного цикла данных.

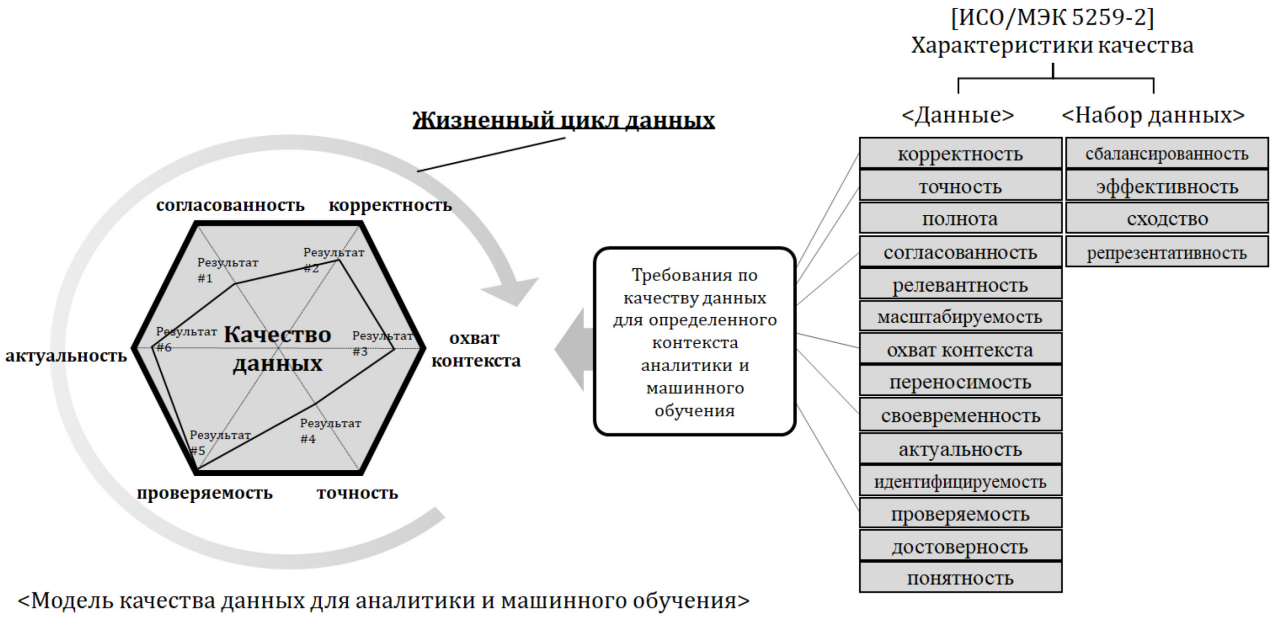


Рисунок 2. Пример применения характеристик из 5259-2

5.2.2.2 Меры качества данных

Как только модель качества данных определена в соответствии с контекстом использования аналитики и машинного обучения, пользователи данных могут выбрать соответствующие меры качества данных для оценки каждой из характеристик качества данных в модели качества данных.

ПРИМЕЧАНИЕ См. ИСО/МЭК 5259-2 для получения дополнительной информации о мерах качества данных, используемых в аналитике и машинном обучении.

5.2.2.3 Оценка качества данных

Организация может использовать результаты выбранных мер качества данных, чтобы оценить, соответствует ли набор данных ее нуждам и требованиям.

Если организация определяет, что набор данных соответствует ее потребностям и требованиям, этот набор данных можно использовать для обучения, проверки и тестирования алгоритма машинного обучения или для работы с обученной моделью или для использования в аналитических процессах.

Если организация определяет, что набор данных не соответствует ее нуждам и требованиям, организация может:

- попытаться улучшить набор данных;
- прекратить использование набора данных;
- приобрести новый набор данных.

5.2.2.4 Улучшение качества данных

Преобразования данных могут быть применены к набору данных для улучшения его качества в той мере, в какой он соответствует нуждам и требованиям организации.

ПРИМЕЧАНИЕ См. ИСО/МЭК 5259-2 и ИСО/МЭК 5259-4 для получения дополнительной информации об улучшении качества данных.

5.2.2.5 Отчетность о качестве данных

Организация может разрабатывать и публиковать отчеты о качестве данных в соответствии со своей внутренней политикой. Эти отчеты могут быть полезны для определения первопричин низкой производительности моделей машинного обучения и других аналитических задач, а также могут способствовать прозрачности и объяснимости машинного обучения. Отчеты о качестве данных могут включать:

- предполагаемое использование набора данных;
- пороговые значения характеристик качества данных, имеющих отношение к нуждам и требованиям организации к набору данных;
- характеристики качества данных, выбранные для модели качества данных;
- объяснение причин, по которым некоторые характеристики качества данных не были включены в модель данных;

- меры качества данных, используемые для каждой характеристики качества данных;
- результаты измерений качества данных;
- действия, предпринятые для улучшения качества набора данных;
- оценку того, соответствует ли набор данных нуждам и требованиям организации;
- лиц, участвующих в модели качества данных, процессах измерения и улучшения.

Визуализация данных предоставляет инструменты для изучения данных, а также способы эффективного информирования о результатах процессов обеспечения качества данных. От комплектования данных до этапов вывода данных из эксплуатации визуализация данных может облегчить проверку состояния данных с помощью методов, описанных в ИСО/МЭК 20547-3, пункт 9.2.2.5. Особенно на этапе подготовки данных визуализация данных может помочь:

- в очистке данных путем нахождения неверных значений, отсутствующих значений, выбросов и повторяющихся значений;
- в создании и выборе атрибутов;
- в объединении категорий в рамках процесса сокращения данных.

5.2.3 Стратегическое управление качеством данных

Процессы обеспечения качества данных должны соответствовать политике стратегического управления данными организации. С точки зрения качества данных стратегическое управление данными может предоставить:

- набор передовых методов и средств контроля, внедренных организацией для активного управления и улучшения качества данных;
- процессно-ориентированную структуру для внедрения и выполнения действий, связанных с качеством данных, на каждом этапе модели жизненного цикла данных;
- организационные роли и обязанности для обеспечения качества данных посредством повторяющихся процессов.

ПРИМЕЧАНИЕ Для получения дополнительной информации об управлении данными см. ИСО/МЭК 38505 [24] и ИСО/МЭК 38507 [25]. Стратегическое управление данными для больших данных описано в пункте 8.4 и приложении А стандарта ИСО/МЭК 20547-3.

5.2.4 Происхождение данных

Процессы качества данных для аналитики и машинного обучения могут состоять из нескольких этапов. Документированные сведения о происхождении данных могут использоваться для сбора и хранения информации о происхождении данных, которая может служить основой для определения того, были ли данные несанкционированно обработаны или изменены. Документированные сведения о происхождении данных могут включать:

- источник или происхождение данных;
- все процессы, применяемые к данным;
- все изменения, примененные к данным (например, статистические преобразования, изменение значений данных);
- все организации или лица, которые хранили данные с момента их создания.

ПРИМЕЧАНИЕ Для получения дополнительной информации о записях о происхождении данных см. ИСО/МЭК 8000-2 [16].

5.3 Жизненный цикл данных для аналитики и машинного обучения

5.3.1 Обзор

Аналитика и машинное обучение применяются для получения прогнозов на основе текущих данных, и их результаты используются для принятия решений. Таким образом, аналитика и машинное обучение сильно зависят от характеристик данных (например, характеристик больших данных: объема, разнообразия, скорости и изменчивости) и характеристик качества данных с точки зрения контекста использования. В то же время на аналитику и машинное обучение оказывает влияние высо-

коуровневый жизненный цикл данных. Согласно пункту 5.2.2 требования к качеству данных зависят от цели анализа и машинного обучения, поэтому требования к качеству данных должны управляться в соответствии с жизненным циклом используемых данных.

Жизненный цикл данных для аналитики и машинного обучения предоставляет сквозное описание того, как данные используются и генерируются дополнительные данные в системе аналитики или машинного обучения.

Этот документ определяет жизненный цикл данных для аналитики и машинного обучения посредством:

- 6-этапной модели жизненного цикла данных (пункт 5.3.2);
- процессов на нескольких этапах модели жизненного цикла данных (пункт 5.3.3).

5.3.2 Модель жизненного цикла данных

5.3.2.1 Общая модель

Шестиэтапная модель жизненного цикла данных для аналитики и машинного обучения, показанная на рис. 3, определяет ряд событий (стадий), которые можно использовать для определения деталей того, как следует обрабатывать данные. На рис. 3 прямоугольники представляют этапы обработки данных, однонаправленные стрелки представляют основное продвижение по этапам, а двунаправленные стрелки представляют обратную связь с другими этапами модели жизненного цикла данных.

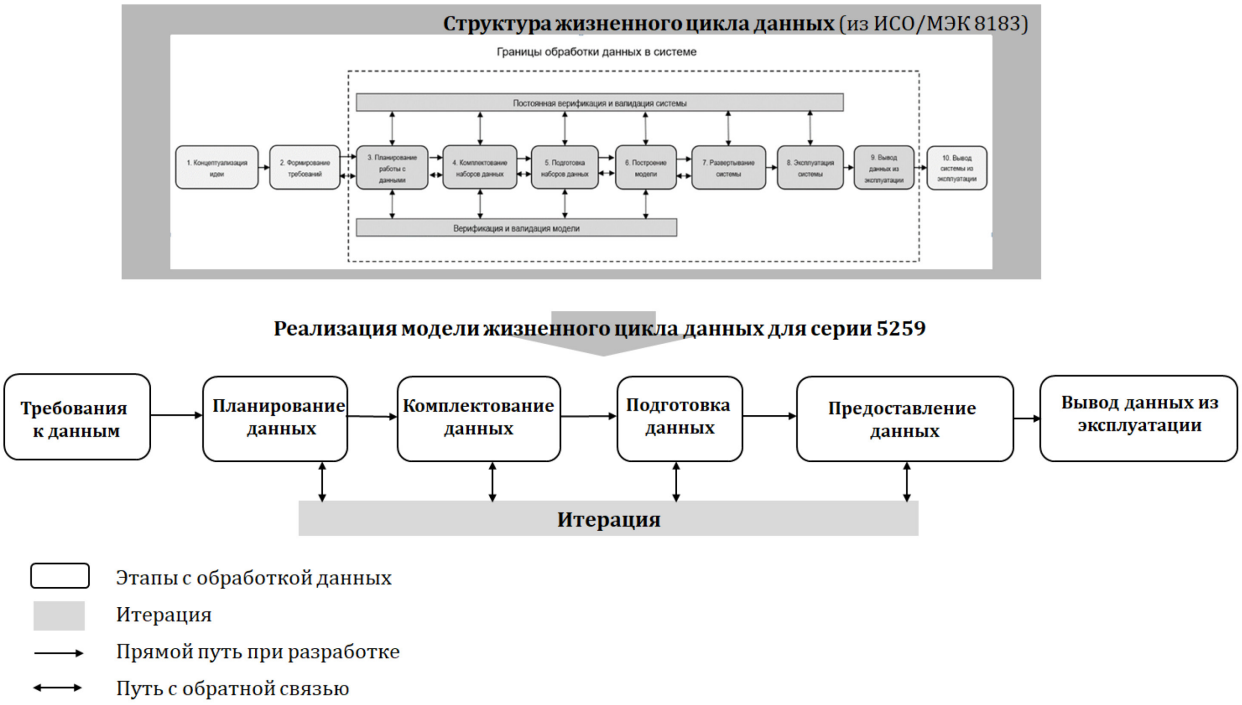


Рис. 3. Модель жизненного цикла данных для аналитики и машинного обучения.

5.3.2.2 Этап 1: Требования к данным

Этап формирования требований к данным включает в себя определение того, какие данные требуются для проекта аналитики или машинного обучения. Этот этап включает в себя:

- понимание потребностей в данных для аналитики или проекта машинного обучения;
- проверку наличия данных для проекта аналитики или машинного обучения;
- определение релевантных характеристик качества данных, связанных с потребностями в данных.

5.3.2.3 Этап 2: Планирование данных

Этап планирования данных гарантирует, что данные, которые будут использоваться, соответствуют требованиям этапа требований к данным, поддерживая при этом цели аналитики и машинного обучения, которые используют данные. Этот этап включает в себя:

- разработку архитектуры данных, которая определяет полный характер и объем необходимых данных, а также способы их использования;

- оценку усилий, необходимых для сбора и подготовки данных для поддержки проекта аналитики или машинного обучения.

Эта оценка должна включать любую необходимую реструктуризацию данных, время на передачу или сбор данных и построение модели качества данных для проекта аналитики или машинного обучения.

5.3.2.4 Этап 3: Комплектование данных

Этап комплектования данных включает в себя сбор данных, которые используются в проекте аналитики или машинного обучения. Оперативные и исторические данные собираются из источников, определенных на этапе планирования данных. В зависимости от вопросов, на которые требуется дать ответы, и требований проекта аналитики или машинного обучения эти данные могут поступать в виде потока или пакетами. Этот этап включает в себя:

- защиту персональных данных для собранных и сохраненных данных, которая соответствует законодательству о персональных данных и иному законодательству;
- улучшение или модификацию методов сбора данных с учетом тестового периода.

ПРИМЕЧАНИЕ. Для получения дополнительной информации об этапе комплектования данных см. ИСО/МЭК 8183.

5.3.2.5 Этап 4: Подготовка данных

На этапе подготовки данных собранные данные преобразуются в форму для проведения аналитики и машинного обучения. Этот этап играет важную роль в соблюдении требований к качеству данных и может быть повторен по результатам аналитики или производительности обученной модели машинного обучения. Этот этап включает в себя следующие необязательные процессы в зависимости от выявленных требований к качеству данных:

- преобразование данных: преобразование данных из одного формата в другой с минимальной потерей содержимого;

- проверка данных: обеспечение корректности данных на основе параметров проверки, таких как правильность, значимость, безопасность и защита персональных данных и т. д.;
- очистка данных: обнаружение неточных данных и исправление данных путем замены, изменения или удаления;
- агрегирование данных: объединение двух или более наборов данных в один набор данных в сводной форме;
- выборка данных: выбор подмножества набора данных;
- создание признаков: создание новых атрибутов, которые могут фиксировать основную информацию в данных более эффективно, чем исходные атрибуты;
- выбор признаков: уменьшение размерности данных за счет использования подмножества доступных признаков;
- обогащение: связывание различных источников данных и добавление дополнительного контекста к данным.
- маркировка данных: данные обучения, проверки и тестирования для машинного обучения с учителем требуют значений для одной или нескольких целевых переменных. Маркировка данных — это процесс присвоения значений целевым переменным, если они не включены в полученные данные. Маркировка данных обычно используется для задач распознавания изображений [7].

ПРИМЕЧАНИЕ Для получения дополнительной информации о подготовке данных см. ИСО/МЭК 8183 и ИСО/МЭК 5259-4.

5.3.2.6 Этап 5: Предоставление данных

На этапе предоставления данных подготовленные данные применяются к проекту аналитики и машинного обучения. На этом этапе оценивается производительность аналитики или обученной модели машинного обучения, чтобы определить, соответствуют ли они требованиям. если результаты анализа или производительность модели машинного обучения не соответствуют ожиданиям, можно выполнить следующие шаги:

- Для аналитики и машинного обучения определить, что является основной причиной: обучающие данные или алгоритм;

- улучшить качество данных, повторив этапы со 2 по 4;

  ПРИМЕЧАНИЕ. Иногда единственный способ добиться приемлемой производительности модели машинного обучения — использовать другие данные.

- повторить анализ или перестроить модель машинного обучения;

5.3.2.7 Этап 6: Вывод данных из эксплуатации

На этапе вывода данных из эксплуатации любые использованные данные сохраняются в хранилище или архивируются, чтобы быть доступными в будущем. При архивации данных следует также сохранить измеренные характеристики качества данных (а так же требования), а в случае системно-зависимых характеристик также следует сохранить сведения о контексте использования данных.

5.3.3 Многоэтапные процессы

5.3.3.1 Общее

На рис. 4 показаны процессы, которые должны применяться на нескольких этапах модели жизненного цикла данных для аналитики и машинного обучения. Управление качеством данных, стратегическое управление качеством данных и происхождение данных описаны в 5.2.

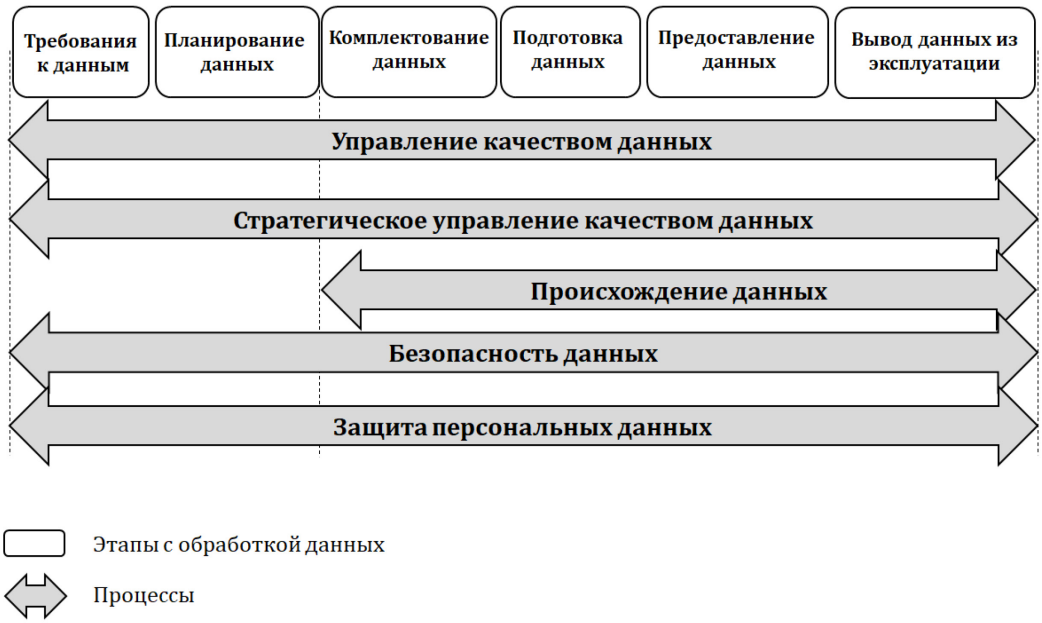


Рисунок 4. Процессы на нескольких этапах

5.3.3.2 Безопасность данных

Набор данных должен храниться защищенным образом во время процессов обеспечения качества данных, чтобы обеспечивать его доступность авторизованным лицам и средствам и процессам обработки, а также отсутствие внесения в данные несанкционированных изменений. Несанкционированные изменения в наборе данных сами по себе могут привести к неправильным результатам моделей машинного обучения и других аналитических задач.

ПРИМЕЧАНИЕ Для получения дополнительной информации о безопасности данных см. серию стандартов ИСО/МЭК 27000 [26].

5.3.3.3 Защита персональных данных

Наборы данных, используемые для машинного обучения и аналитики, могут содержать персональные данные, которые должны быть защищены во время процессов обеспечения качества данных. Методы деидентификации могут использоваться для удаления персональных данных, но производственные данные, используемые для прогнозирования в отношении отдельных людей, могут по-прежнему содержать персональные данные или быть связаны с ними.

ПРИМЕЧАНИЕ Для получения дополнительной информации о конфиденциальности данных см. ИСО/МЭК 27701 [27].

Приложение А

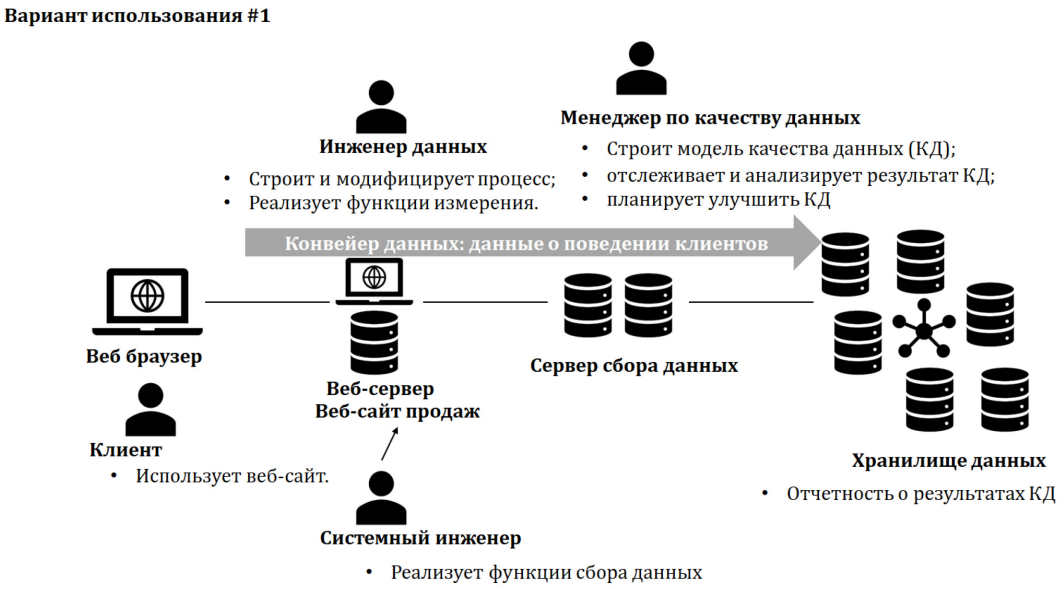
Примеры и сценарии

В этом приложении представлены либо примеры, либо сценарии, либо и то и другое в отношении качества данных для аналитики и машинного обучения (например, традиционные статистические методы, машинное обучение или глубокое обучение).

В таблице А.1 показан сценарий сбора и хранения данных с использованием конвейера данных в хранилище данных.

Таблица А.1 Сбор и хранение данных

Название	Сбор и хранение данных с использованием конвейера данных в хранилище данных
Описание	Компания, занимающаяся онлайн-продажами, хочет собирать данные о закономерностях поведения клиентов своего веб-сайта и использовать их в маркетинговых целях. С этой целью они создают конвейер данных для сбора данных о событиях, связанных с действиями клиентов в хранилище данных и используют их для различного анализа больших данных. Для этого: • Шаг 1: (Системный инженер) реализует функции сбора данных о пользовательских событиях; • Шаг 2: (Менеджер по качеству данных) создает универсальную модель качества данных для собранного набора данных; • Шаг 3: (инженер данных) создает конвейер для предварительной обработки данных, включая очистку данных, ETL (от англ. Extract, Transform, Load — дословно «извлечение, преобразование, загрузка») и сегментацию по единицам времени, а также их хранение в хранилище данных; • Шаг 4: (менеджер по качеству данных, инженер по данным) внедряет функции измерения для каждой характеристики качества данных на основе модели качества данных, созданной на шаге 2, и добавляет функции в конвейер;

	• Шаг 5: (Хранилище данных, система управления качеством данных) сообщает о результатах оценки качества набора данных; • Шаг 6: Если оцененное качество данных не достигает запланированного уровня, (менеджер по качеству данных) анализирует причину и предлагает план по улучшению качества данных; • Шаг 7: (инженер данных) изменяет конвейер данных, чтобы отразить план, созданный на шаге 6; При необходимости повторяйте шаги 1–7, пока качество данных не достигнет желаемого уровня. • Шаг 8: Когда качество данных достигает желаемого уровня, (инженер данных) завершает конвейер данных.
Рисунок	

Проблема качества данных и технические аспекты	• Построение модели качества данных для данных и наборов данных общего назначения;
	• Реализация измерительных функций;
	• Отчетность о качестве данных во время сбора и хранения данных для поддержки:
	  ○ Анализ причин низкого качества;
	  ○ Предложение путей улучшения качества данных.
	• Хранение информации о качестве данных в метаданных и каталоге данных.

Библиография

[1] ISO/IEC 5259-2:20XX, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures*

[2] ISO/IEC 5259-3:20XX, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines*

[3] ISO/IEC 5259-4:20XX, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework*

[4] ISO/IEC 20546:2019, *Information technology — Big data — Overview and vocabulary*

[5] *ISO/IEC 20547-3:2020, Information technology — Big data reference architecture — Part 3: Reference architecture*

[6] *ISO/IEC 22989:2022, Information technology — Artificial intelligence — Concepts and terminology*

[7] *ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

[8] ISO/IEC 25010:2011, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*

[9] ISO/IEC 25012:2008, *Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*

[10] ISO/IEC 25021:2012, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality measure elements*

[11] ISO/IEC 25024:2015, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality*

[12] ISO/IEC 25030:2019, *Systems and software engineering — Systems and software quality requirements and evaluation (SQuaRE) — Quality requirements framework*

[13] ISO/IEC/IEEE 15939:2017, *Systems and software engineering — Measurement process*

[14] ISO 21710:2020, *Biotechnology — Specification on data management and publication in microbial resource centers*

[15] ISO 8000 (all parts), *Data quality*

[16] ISO 8000-2:2020, *Data quality — Part 2: Vocabulary*

[17] ISO 8000-61:2016, *Data quality — Part 61: Data quality management: Process reference model*

[18] ISO 8000-120:2016, *Data quality — Part 120: Master data: Exchange of characteristic data: Provenance*

[19] ISO 19160-1:2015, *Addressing — Part 1: Conceptual mode*

[20] ISO/IEC 2382:2015, *Information technology — Vocabulary*

[21] ISO 55000:2014, *Asset management — Overview, principles and terminology*

[22] *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms ISO (VIM 3rd edition) JCGM 200:2012*

[23] *Systems Engineering Leading Indicators, version 2, LAI MIT, SEA RI, INCOSE, and PSM, January 29, 2010*

[24] ISO/IEC 38505 (all parts), *Information technology — Governance of IT — Governance of data*

[25] ISO/IEC FDIS 38507:2022, *Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations*

[26] ISO/IEC 27000 (all parts), *Information technology — Security techniques — Information security management systems*

[27] ISO/IEC 27701:2019, *Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines*

[28] ISO/IEC 23751:2022, *Information technology — Cloud computing and distributed platforms — Data sharing agreement (DSA) framework*

Приложение А
(справочное)
Сведения о соответствии ссылочных международных стандартов национальным стандартам

Таблица А

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	NEQ*	ПНСТ 553–2021** Информационные технологии. Искусственный интеллект. Концепции искусственного интеллекта и терминология (ISO/IEC DIS 22989, NEQ)
ISO/IEC 20546:2019, Information technology — Big data — Overview and vocabulary	EQ	ГОСТ Р ИСО/МЭК 20546-2021 Информационные технологии. Большие данные. Обзор и словарь.
ISO/IEC 25010:2011, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*	EQ	ГОСТ Р ИСО/МЭК 25010-2015 Информационные технологии. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Модели качества систем и программных продуктов.
ISO/IEC 25021:2012, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality measure elements	EQ	ГОСТ Р ИСО/МЭК 25021-2014 Информационные технологии. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Элементы показателя качества.
ISO/IEC/IEEE 15939:2017,	EQ	ГОСТ Р 58606-2019/ISO/IEC/IEEE

Systems and software engineering — Measurement process		15939:2017 Системная и программная инженерия. Процесс измерения.
ISO 8000-2:2020, Data quality — Part 2: Vocabulary	EQ	ГОСТ Р ИСО 8000-2-2019 Качество данных. Часть 2. Словарь.
ISO/IEC 2382:2015, Information technology — Vocabulary	EQ	ГОСТ 33707-2016 Информационные технологии. Словарь.
ISO 55000:2014, Asset management — Overview, principles and terminology	EQ	ГОСТ Р 55.0.01-2014 Управление активами. Общее представление, принципы и терминология.
*NEQ — неэквивалентный стандарт; EQ — эквивалентный стандарт; ** В России принят предварительный национальный стандарт, основанный на ранней редакции международного стандарта		

ГОСТ (Р) _____—_____.

(проект)

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: информационные технологии, искусственный интеллект, качество данных, машинное обучение, большие данные, аналитика больших данных, характеристики качества данных, стратегическое управление качеством данных, жизненный цикл данных, происхождение данных.

Председатель совета директоров ООО «Институт развития информационного общества», руководитель разработки стандарта		Ю.Е. Хохлов
Руководитель департамента развития распределенных вычислительных технологий ООО «Институт развития информационного общества», доцент, Санкт-Петербургский государственный университет, разработчик стандарта		В.В. Корхов