ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

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

(ISO/IEC 5259-1:2024, 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. № 1537-ст

4 Настоящий стандарт является модифицированным по отношению к международному стандарту ИСО/МЭК 5259-1:2024 «Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры» (ISO/IEC 5259-1:2024 «Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples», MOD) путем изменения отдельных фраз (слов, значений, показателей, ссылок), которые выделены в тексте курсивом. В результате модификации отдельные термины из международного стандарта были приведены в соответствие с терминологией, принятой в российском техническом регулировании.

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

Данные являются исходным материалом для аналитики и машинного обучения, а их качество — это критически важный аспект для соответствующей аналитики, а также проектов и систем машинного обучения. Цель серии стандартов ГОСТ Р 71484 — предоставить инструменты и методы для оценки и повышения качества данных, используемых для аналитики и машинного обучения.

В состав серии стандартов ГОСТ Р 71484 входят:

- ГОСТ Р 71484.2 описывает модель качества данных, показатели качества данных и рекомендации по предоставлению сведений о качестве данных в контексте аналитики и машинного обучения. ГОСТ Р 71484.2 опирается на стандарты [1]—[3]. Цель ГОСТ Р 71484.2 заключается в том, чтобы способствовать организациям в достижении ими своих целей в отношении качества данных. Стандарт применим в организациях любого типа;

- ГОСТ Р 71484.3 устанавливает требования и предоставляет рекомендации по установлению, внедрению, поддержанию и постоянному повышению качества данных, используемых в областях аналитики и машинного обучения. ГОСТ Р 71484.3 не дает детального описания процессов, методов и показателей. Данный стандарт определяет требования и дает рекомендации в отношении процесса управления качеством, а также описывает эталонный процесс и методы, которые могут быть адаптированы для выполнения требований, изложенных в этом документе. Требования и рекомендации, изложенные в ГОСТ Р 71484.3, являются типовыми и предназначены для применения в любых организациях независимо от их типа, размера или характера;

- ГОСТ Р 71484.4 описывает общие типовые организационные подходы, не зависящие от типа, размера или характера применяющей их организации и используемые для обеспечения качества данных для обучения и оценки в области аналитики и машинного обучения. Стандарт включает в себя рекомендации:

- по обучению с учителем в отношении разметки данных, используемых для обучения систем МО, включая распространенные организационные подходы к разметке обучающих данных;

- обучению без учителя;

- обучению с частичным привлечением учителя;

- обучению с подкреплением;

- аналитике.

ГОСТ Р 71484.4 применим к данным обучения и оценки, которые поступают из разных источников, включая комплектование и композицию данных, предварительную обработку данных, разметку данных, оценку и использование данных. ГОСТ Р 71484.4 не определяет конкретные услуги, платформы или инструменты;

- *[1]* описывает структуру стратегического управления качеством данных для аналитики и машинного обучения, дающую возможность органам стратегического управления организации направлять и контролировать внедрение и функционирование показателей обеспечения качества данных, оперативного управления качеством данных и связанных с ними процессов посредством использования адекватных мер контроля и управления в рамках описанной в настоящем стандарте модели жизненного цикла данных;

- *[2]* описывает структуру визуализации качества данных в аналитике и машинном обучении. Его цель заключается в том, чтобы способствовать использованию заинтересованными сторонами методов визуализации для оценки результатов измерения показателей качества данных. Эта структура визуализации поддерживает достижение целей по обеспечению качества данных.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Искусственный интеллект

КАЧЕСТВО ДАННЫХ ДЛЯ АНАЛИТИКИ И МАШИННОГО ОБУЧЕНИЯ

Часть 1

Обзор, терминология и примеры

Artificial intelligence. Data quality for analytics and machine learning. Part 1. Overview, terminology and examples

Дата введения — 2025—01—01

1 Область применения

Настоящий стандарт служит основой для концептуального понимания качества данных для аналитики и машинного обучения. В нем также приводятся взаимосвязанные технологии и примеры (например, варианты использования и сценарии применения).

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ Р ИСО/МЭК 27001 Информационная технология. Методы и средства обеспечения безопасности. Системы менеджмента информационной безопасности. Требования

ГОСТ Р 54911 (ИСО/TR 8000-120:2009) Качество данных. Часть 120. Основные данные. Обмен данными характеристик. Происхождение

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

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

ГОСТ Р 71484.2 (ИСО/МЭК 5259-2:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 2. Показатели качества данных

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

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 указателю «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, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется

Издание официальное

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* и *[3]*, а также следующие термины с соответствующими определениями:

3.1 **жизненный цикл данных** (data life cycle, life cycle of data): Все стадии в процессе использования данных от замысла до вывода из эксплуатации.

3.2 **создатель данных** (data originator): Сторона, которая создала данные и может обладать правами на них.

Примечания

1 Создателем данных может быть физическое лицо.

2 Создатель данных может отличаться от физического или юридического лица, которое упомянуто в данных, описано ими либо явно или неявно связано с ними. Например, создателем данных могут быть собраны персональные данные, идентифицирующие других физических лиц. Эти субъекты персональных данных также могут обладать правами в отношении такого набора данных.

3 Права могут включать право на публичное использование, право на отображение имени, право на идентичность, право запрещать использование данных оскорбительным образом.

4 См. [4], пункт 3.2.

3.3 **распорядитель данными** (data holder): Сторона, имеющая законную возможность авторизовать обработку данных другими сторонами.

Примечания

1 Распорядителем данными может быть создатель данных (3.2).

2 См. [4], пункт 3.4.

3.4 **пользователь данных** (data user): Сторона, которая авторизована распорядителем данными выполнять обработку данных.

Примечание —См. [4], пункт 3.5.

3.5 **качество данных** (data quality): Свойство данных соответствовать требованиям организации к данным в конкретных условиях.

3.6 **характеристика качества данных** (data quality characteristic): Атрибут данных, имеющий отношение к качеству данных.

Примечание — См. [5], пункт 4.4.

3.7 **модель качества данных** (data quality model): Заданный набор характеристик, который обеспечивает основу для определения требований к качеству и оценки качества данных.

Примечание — См. [5], пункт 4.6.

3.8 **показатель качества данных** (data quality measure): Переменная, которой присваивается значение в результате измерения характеристики качества данных.

Примечание — См. [5], пункт 4.5.

3.9 **требование к качеству** (quality requirement): Требование к характеризующим качество свойствам или атрибутам продукта, данных или услуги информационно-коммуникационных технологий, которые удовлетворяют потребности, вытекающие из цели, для которой этот продукт, данные или услуга информационно-коммуникационных технологий должны использоваться.

Примечание — См. [6], пункт 3.15.

3.10 **измерение** (measurement): Совокупность операций, имеющих целью определение значения показателя.

Примечание — См. [7], пункт 4.27.

3.11 **шкала измерений** (measurement scale, quantity-value scale): Упорядоченная совокупность количественных значений величин определенного вида, используемая при ранжировании по значению величин этого вида.

Примеры:
1 Температурная шкала Цельсия.
2 Шкала времени.
3 Шкала твердости С Роквелла.

Примечание — См. [8], пункт 1.28.

3.12

аналитика, аналитика данных (analytics, data analytics): Составная концепция, охватывающая комплектование данных, сбор данных, проверку данных, обработку данных, включая количественную оценку данных, визуализацию данных, документирование данных и интерпретацию данных.

Примечание — Аналитика используется для понимания объектов или событий, 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 и планирование.

[ГОСТ Р ИСО/МЭК 20546—2021, пункт 3.1.6]

3.13

атрибут (attribute): Свойство или характеристика сущности, которая может быть количественно или качественно различима человеком или автоматизированным средством.
[ГОСТ Р 58606—2019/ISO/IEC/IEEE 15939:2017, пункт 3.2]

3.14

признак (feature) (в машинном обучении): Измеримое свойство объекта или события, связанное с заданным набором характеристик.

Примечания
1 Признаки играют роль в обучении и прогнозировании.
2 Признаки предоставляют 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.
3 См. [3], пункт 3.3.3.

3.15

управление качеством данных (data quality management): Согласованная деятельность по контролю и управлению организацией, имеющей непосредственное отношение к качеству данных.
[ГОСТ Р ИСО 8000-2—2019, пункт 3.8.2]

3.16 **стратегическое управление данными** (data governance, governance of data): Система, посредством которой осуществляется стратегическое управление текущим и будущим использованием данных.

3.17 **происхождение данных** (data provenance): Сведения о месте и времени создания, получения или генерации набора данных, доказательства аутентичности набора данных и/или документированные сведения о прошлых и текущих распорядителях набора данных.

Примечание — См. [9], пункт 3.11.

3.18 **визуализация** (visualization, scientific visualization): <компьютерная графика> Использование компьютерной графики и обработки изображений для представления моделей или характеристик процессов или объектов для поддержки человеческого понимания.

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

Примечание — См. [10], пункт 2125942.

3.19 **проект машинного обучения** (machine learning project, ML project): Проект, использующий аналитику и машинное обучение и отвечающий за соответствующие данные на протяжении всего их жизненного цикла.

3.20 **архитектура данных** (data architecture): Описание структуры и взаимодействия основных типов и источников данных, логических активов данных, физических активов данных и ресурсов управления данными предприятия.

Примечания

1 Логические сущности данных могут быть привязаны к приложениям, хранилищам и сервисам, и могут быть структурированы в соответствии с соображениями практической реализации.

2 Понятие «данные» в *[11]* намеренно не определяется, поскольку оно является частью определения понятия «архитектура данных» для каждого сценария применения, и соответствует конкретным требованиям этого сценария.

3 См. [11], пункт 3.2.6.

3.21 **элемент данных** (data item): Наименьшая идентифицируемая единица данных в определенном контексте, для которой определение, идентификация, допустимые значения и иная информация заданы посредством набора свойств.

Примечания

1 Термин «поле» (field) рассматривается как синоним термина «элемент данных».

2 Элемент данных представляет собой физический объект — «контейнер», содержащий значения данных.

3 См. [7], пункт 4.9.

3.22 **запись данных** (data record): Набор взаимосвязанных элементов данных, обрабатываемый как единое целое.

Примечание — См. [7], пункт 4.15.

3.23 **метаданные** (metadata): Данные, которые определяют и описывают другие данные.

Примечания

1 В контексте аналитики и машинного обучения метаданные предоставляют сведения об элементах данных или записях данных, такие как свойства, структура, тип, контекст, целевое использование, владельцы, доступ и волатильность.

2 См. [12], пункт 3.2.26.

4 Сокращения

ИИ — искусственный интеллект (artificial intelligence);

МО — машинное обучение (machine learning).

5 Концепции качества данных для аналитики и машинного обучения

5.1 Рекомендации по качеству данных для аналитики и машинного обучения

5.1.1 Общие положения

Существующие стандарты качества данных, такие как *[13]*, были разработаны с точки зрения производства данных и управления ими. Это связано с тем, что производители и/или сборщики данных традиционно 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 этой определенной цели, данные, полученные таким образом, могут потребовать дополнительной обработки для использования в иных условиях.

В области анализа данных и машинного обучения пользователи данных обычно сами не производят данные. Они ищут, собирают и обрабатывают данные, которые, по их мнению, необходимы и подходят для их проектов в области аналитики или машинного обучения. В этом случае качество данных влияет на качество результатов анализа и производительность модели машинного обучения. Каким бы тщательным ни был анализ данных или алгоритм машинного обучения, результаты могут быть ненадежными при использовании не соответствующих требованиям данных. Даже если данные удовлетворяют требованиям для определенного приложения или контекста, это не значит, что они обязательно будут удовлетворять требованиям для других приложений или контекстов. Использование данных, которые не удовлетворяют требованиям для конкретного назначения, может привести к тому, что модели машинного обучения окажутся неточными и предрасположенными к ошибкам. В этой связи, чтобы помочь организациям обеспечить соответствие требованиям к данным, используемым для ана-

литики и машинного обучения, в стандартах серии *ГОСТ Р 71484*, *[1], [2]* определены характеристики качества данных, показатели, качества данных, требования к управлению качеством данных и репрезентативный процесс управления качеством данных на протяжении жизненного цикла данных, вместе с понятиями «запись данных» и «элемент данных» для применения в управлении качеством данных, а также со структурой стратегического управления для направления и контроля над реализацией и функционированием всего вышеперечисленного.

5.1.2 Машинное обучение и качество данных

ГОСТ Р 71476 определяет машинное обучение как процесс оптимизации параметров модели с помощью вычислительных методов таким образом, чтобы поведение модели отражало данные и/или опыт. В *[3]* машинное обучение далее описывается как направление ИИ, использующее вычислительные методы для того, чтобы дать системам возможность учиться на основе данных и опыта. Машинное обучение можно использовать для выполнения разнообразных задач с использованием данных и алгоритмов машинного обучения. Данные, используемые в машинном обучении, подразделяются на обучающие данные, валидационные данные, тестовые данные и эксплуатационные данные. В машинном обучении с учителем модель создается путем обучения алгоритма на обучающих данных. Затем валидационные и тестовые данные применяются для того, чтобы обеспечить функционирование обученной модели МО в соответствии с установленными организацией требованиями. Далее обученная модель машинного обучения используется для получения логических выводов на основе эксплуатационных данных. Производительность обученной модели МО зависит от характеристик качества всех этих типов данных. *[3]* описывает несколько общих типов алгоритмов машинного обучения, которые могут иметь различную чувствительность к различным характеристикам качества данных.

Примеры

1 Репрезентативность — одна из важнейших характеристик качества данных для машинного обучения. Если обучающие данные не отражают характерные особенности совокупности, представленные в эксплуатационных данных, то обученная модель машинного обучения с бóльшей вероятностью сделает неправильные логические выводы на основе эксплуатационных данных. При принятии решений о людях это может привести к дискриминационным или предвзятым действиям в отношении недостаточно представленных групп людей.

2 Создание обученной модели МО посредством обучения алгоритма представляет собой математический процесс, в ходе которого обрабатывается набор обучающих данных, представляющих атрибуты объекта или события. Качество обучающих данных будет влиять на обученную модель машинного обучения. Если слишком большая доля обучающих данных неточна, то модель, скорее всего, сделает неверные логические выводы на основе эксплуатационных данных.

Примечание — См. *ГОСТ Р 71484.2* для получения подробной информации о том, как характеристики качества данных влияют на производительность моделей машинного обучения.

5.1.3 Характеристики данных, с которыми связаны проблемы качества при выполнении аналитики и машинного обучения

Наличие массивов данных, отличающихся существенным разнообразием и вариативностью, может повлиять на модель качества данных и связанные с ней показатели (меры) качества данных. Большие объемы и высокая скорость создания или изменения данных могут потребовать использования автоматизированных инструментов для измерения качества данных и оценки того, соответствуют ли данные требованиям. Большие объемы данных также могут 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

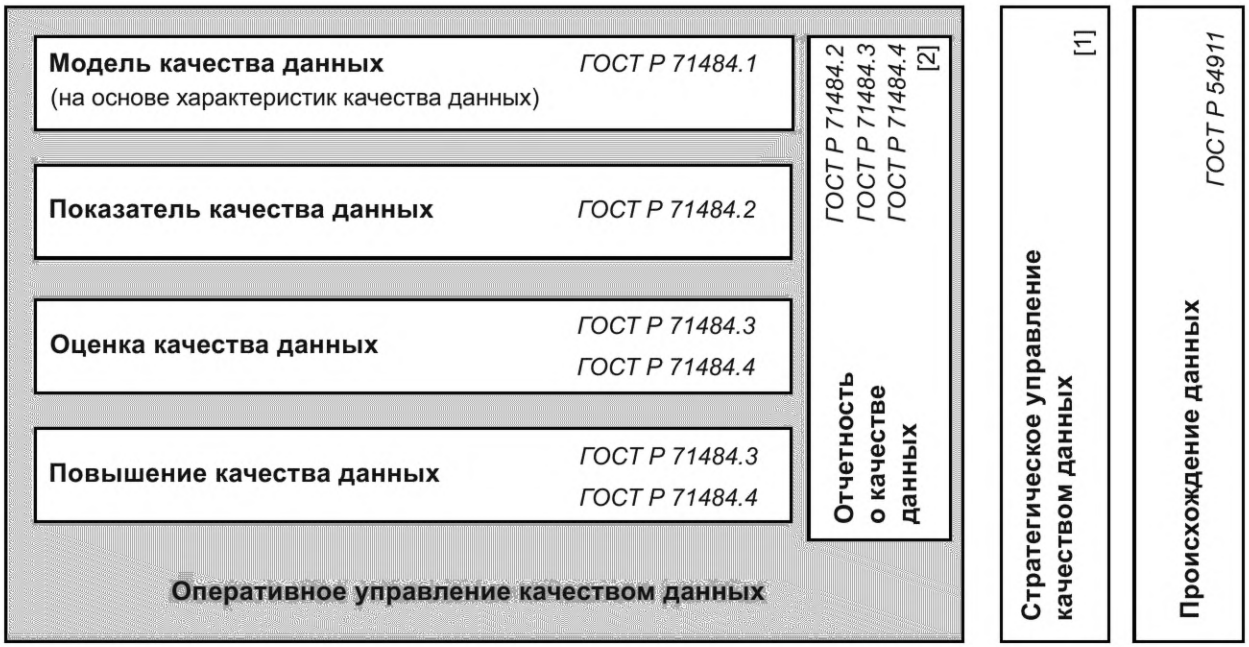
Одни и те же данные можно использовать для разных проектов аналитики и/или машинного обучения. Например, распорядитель данными может предоставить доступ к данным для многочисленных пользователей (как внутренних, так и внешних по отношению к организации распорядителя данными); или же пользователю данных может быть разрешено использовать данные для выполнения нескольких различных задач.

Разные проекты аналитики и машинного обучения могут предъявлять различающиеся требования к качеству данных, что может повлиять на выбор модели качества данных, соответствующих показателей качества данных и критериев оценки.

5.2 Структура обеспечения качества данных для аналитики и машинного обучения

5.2.1 Обзор

На рисунке 1 представлена структура и взаимосвязи стандартов серии *ГОСТ Р 71484*, *[1]*, *[2]* и *ГОСТ Р 54911*, предназначенная для определения, оценки и повышения качества набора данных при использовании в проектах аналитики или машинного обучения.



Обозначение:

Итерация

Рисунок 1 — Концептуальная структура обеспечения качества данных для аналитики и машинного обучения

Цель данной структуры заключается в идентификации процессов, которые могут быть использованы для определения и обеспечения того, чтобы набор данных соответствовал нуждам и требованиям организации.

В число специфических для качества данных элементов концептуальной структуры входят модель, показатели, оценка, повышение качества данных и подготовка отчетности по качеству данных. Среди других важных процессов можно назвать процессы стратегического и оперативного управления, а также отслеживание происхождения данных.

Процессы выбора показателей, оценки и повышения качества данных могут быть итеративными, когда это необходимо для удовлетворения потребностей организации и ее требований к набору данных.

Кроме того, в случае 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

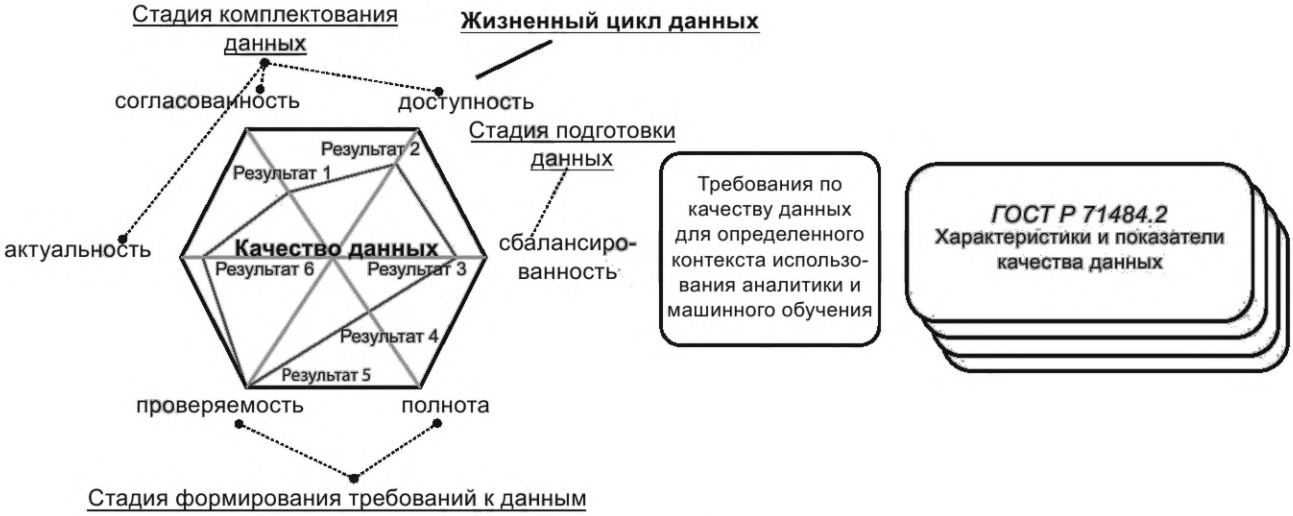
В настоящем стандарте модель качества данных представляет собой заданный набор характеристик, используемый в качестве основы для установления требований к качеству данных и при оценке качества данных. Пользователи данных могут создавать модели качества данных для аналитики и машинного обучения в соответствии со своими деловыми целями.

В разделе 6 стандарта *ГОСТ Р 71484.2* описываются как внутренние, так и системно-зависимые характеристики качества данных согласно *[5]*, наряду с дополнительными характеристиками качества данных для аналитики и машинного обучения.

В *[5]* следующим образом описаны две категории характеристик качества данных:

- под внутренне присущим качеством данных понимается степень, в которой характеристики качества самих данных потенциально способны удовлетворять явно сформулированные и подразумеваемые потребности в случае использования данных в заданных условиях;
- под системно-зависимым качеством данных понимается степень, в которой качество данных достигается и сохраняется в компьютерной системе, когда данные используются в заданных условиях.

На рисунке 2 показана взаимосвязь модели качества данных для аналитики и машинного обучения, жизненного цикла данных (см. 5.3), 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 деловыми целями. Модель качества данных представляет собой заданный набор измеримых характеристик качества данных. Измерения характеристик качества данных могут проводиться на соответствующих стадиях жизненного цикла данных.



<Модель качества данных для аналитики и машинного обучения>

Обозначение:

Жизненный цикл данных для аналитики и машинного обучения (см. 5.3)

Рисунок 2 — Пример применения характеристик качества данных из *ГОСТ Р 71484.2*

5.2.2.2 Показатели качества данных

После определения модели качества данных в соответствии с контекстом использования аналитики и машинного обучения пользователи данных могут выбрать соответствующие показатели качества данных для оценки каждой из характеристик качества данных в составе модели.

Примечание — Дополнительную информацию о показателях качества данных, используемых в аналитике и машинном обучении, см. *ГОСТ Р 71484.2*.

5.2.2.3 Оценка качества данных

Организация может использовать результаты измерения выбранных показателей качества данных для оценки соответствия набора данных ее нуждам и требованиям.

Если организация установит, что набор данных соответствует ее нуждам и требованиям, то этот набор данных в дальнейшем можно будет использовать для обучения, валидации и тестирования алгоритма машинного обучения, для работы с обученной моделью МО и/или для использования в процессах аналитики. Процесс определения соответствия набора данных предъявляемым к нему требованиям может происходить итеративно, как описано в 5.3.

Если организация установит, что набор данных не соответствует ее нуждам и требованиям, то она может:

- попытаться улучшить набор данных;
- прекратить использование набора данных;
- сформировать (приобрести или создать самостоятельно) новый набор данных.

5.2.2.4 Повышение качества данных

С целью повышения качества набора данных до уровня, соответствующего нуждам и требованиям организации, к нему могут применяться преобразования данных. Вопросы обеспечения качества данных должны решаться на как можно более ранних стадиях (например, создателями и распорядителями данных), чтобы уменьшить нагрузку на пользователей данных и обеспечить большую согласованность в последующих версиях набора данных.

Повышение качества данных следует рассматривать в контексте деятельности организации, требований к качеству данных и производительности модели машинного обучения. Чтобы соответствовать требованиям, не всегда имеет смысл тратить время и средства на повышение всех без исключения характеристик качества данных до идеального состояния.

Примечание — Дополнительную информацию о повышении качества данных см. в *ГОСТ Р 71484.4*.

5.2.2.5 Отчетность о качестве данных

Организация может готовить и публиковать отчеты о качестве данных в соответствии со своими внутренними политиками. Эти отчеты могут быть полезны для определения первопричин низкой эффективности моделей машинного обучения и выполнения других аналитических задач, а также могут способствовать прозрачности и объяснимости машинного обучения. Отчеты о качестве данных могут включать:

- предполагаемое (целевое) использование набора данных;
- пороговые значения характеристик качества данных, имеющих отношение к нуждам и требованиям организации к набору данных;
- характеристики качества данных, выбранные для включения в модель качества данных;
- объяснение причин, по которым некоторые характеристики качества данных не были включены в модель качества данных;
- показатели качества данных, используемые для каждой из характеристик качества данных;
- результаты измерения качества данных;
- тенденции изменения качества данных (например, наблюдается ли повышение или снижение качества данных);
- действия, предпринятые для повышения качества набора данных;
- оценку соответствия набора данных нуждам и требованиям организации;
- сведения о лицах, принимавших участие в процессах разработки модели качества данных, измерения и повышения качества данных.

Визуализация данных предоставляет инструменты для изучения данных, а также способы эффективного информирования о результатах процессов обеспечения качества данных. От стадии комплектования данных и до стадии вывода данных из эксплуатации, визуализация данных может облегчить проверку состояния данных с помощью методов, подобных тем, что описаны в *[14]*, 9.2.2.5. Визуализация данных, особенно на стадии подготовки данных, может помочь:

- в очистке данных посредством выявления неверных, отсутствующих и повторяющихся значений;
- в создании и отборе атрибутов или признаков;
- в объединении атрибутов или признаков в процессе сокращения объема данных;
- в объяснении наблюдающихся в данных тенденций, закономерностей, распределений и выбросов;
- в демонстрации взаимосвязи между показателями качества данных и установленными пороговыми значениями (например, с использованием красных, желтых и зеленых индикаторов).

Примечание — Дополнительную информацию о визуализации качества данных см. в *[2]*.

5.2.3 Стратегическое управление качеством данных

Процессы обеспечения качества данных должны соответствовать политике стратегического управления данными организации. Ключевую роль в обеспечении качества данных в организации играет культура подотчетности. С точки зрения качества данных стратегическое управление данными может предоставить:

- набор руководящих принципов, устанавливаемых организацией с целью активного управления и повышения качества данных;
- структуры принятия решений и подотчетности, благодаря которым лица, на которых возложена ответственность за качество данных, привлекаются к ответственности;

- роли и обязанности в организации, используемые для обеспечения качества данных посредством выполнения повторяющихся процессов.

Примечание — Для получения дополнительной информации о структуре стратегического управления качеством данных см. *[1].* Стратегическое управление рассматривается в *[15]* и *[16].* Стратегическое управление большими данными описано в подразделе 8.4 и приложении А *[14].*

5.2.4 Происхождение данных

Процессы обеспечения качества данных для аналитики и машинного обучения могут быть сложными и состоять из нескольких повторяющихся шагов. 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. Эти сведения могут помочь пользователям данных оценить свою возможность доверять этим данным. Документированные сведения о происхождении данных могут включать:

- сведения об источнике или месте создания данных;
- сведения обо всех процессах, примененных к данным;
- сведения обо всех изменениях, примененных к данным (таких, как статистические преобразования, изменение значений данных);
- сведения обо всех организациях и/или лицах, которые были ответственными хранителями данных, начиная с момента их создания.

Примечание — Для получения дополнительной информации о документированных сведениях о происхождении данных см. *ГОСТ Р 54911*.

5.3 Жизненный цикл данных для аналитики и машинного обучения

5.3.1 Обзор

Аналитика и машинное обучение применяются для получения прогнозов на основе данных, которые организация может использовать для принятия решений. Ввиду этого аналитика и машинное обучение в сильной степени зависят от характеристик данных и характеристик качества данных с точки зрения контекста их использования. В то же время аналитика и машинное обучение осуществляются в рамках одного и того же высокоуровневого жизненного цикла данных. Согласно 5.2.2, требования к качеству данных зависят от целей аналитики и машинного обучения, поэтому требованиями к качеству данных необходимо управлять в соответствии с назначением и жизненным циклом используемых данных.

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

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

- 6-стадийной модели жизненного цикла данных (см. 5.3.2);
- процессов на различных стадиях модели жизненного цикла данных (см. 5.3.3).

5.3.2 Модель жизненного цикла данных

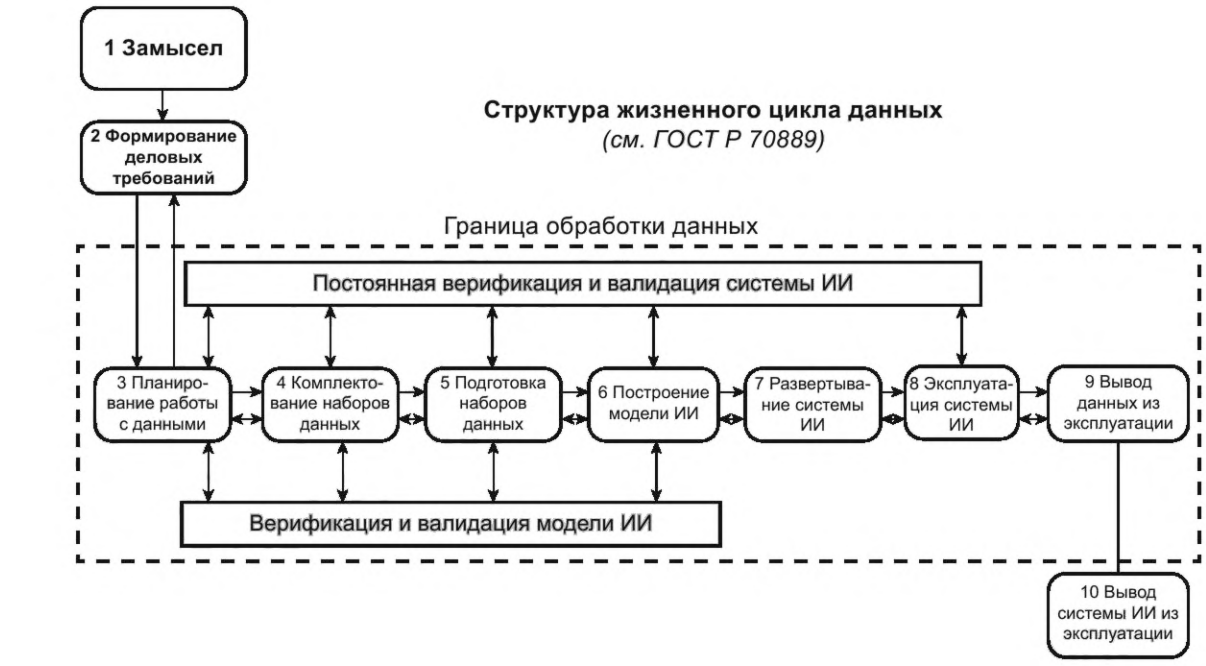
5.3.2.1 Описание модели в целом

Модель жизненного цикла данных для аналитики и машинного обучения, показанная на рисунке 3, сформулирована на основе *ГОСТ Р 70889* и выделяет ряд стадий, которые создают контекст для процессов, используемых в управлении качеством данных. На рисунке 3 однонаправленные стрелки показывают основное направление продвижения по стадиям, а двунаправленные стрелки представляют обратную связь с другими стадиями модели жизненного цикла данных.

5.3.2.2 Стадия 1: формирование требований к данным

Стадия формирования требований к данным включает:

- определение того, какие данные требуются для проекта аналитики или машинного обучения;
- проверку наличия данных для проекта аналитики или машинного обучения;
- конкретизацию модели качества данных посредством включения в нее релевантных характеристик качества данных.



Реализация модели жизненного цикла данных для стандартов серии *ГОСТ Р 71484*

Обозначения:

Жизненный цикл данных для аналитики и машинного обучения

Итерация

Прямой путь при разработке

Путь с обратной связью

Рисунок 3 — Модель жизненного цикла данных для аналитики и машинного обучения

5.3.2.3 Стадия 2: планирование работы с данными

Стадия планирования работы с данными обеспечивает соответствие данных, которые будут использоваться, требованиям, сформулированным на стадии формирования требований к данным, поддерживая при этом цели использующего эти данные проекта аналитики или машинного обучения. Данная стадия включает:

- разработку архитектуры данных (т. е. определение полной природы и охвата необходимых данных, а также способов их использования);

- оценку усилий, необходимых для сбора и подготовки данных для проекта аналитики или машинного обучения. Эта оценка может учитывать любую необходимую реструктуризацию данных, время на передачу и/или сбор данных и построение модели качества данных для проекта аналитики или машинного обучения.

5.3.2.4 Стадия 3: комплектование наборов данных

Стадия комплектования наборов данных включает в себя сбор данных, которые используются в проекте аналитики или машинного обучения. Оперативные и ретроспективные данные собираются из

источников, определенных на стадии планирования работы с данными. В зависимости от требований проекта аналитики или машинного обучения, эти данные могут поступать в виде потока или пакетами. Данная стадия включает:

- защиту персональных данных субъектов данных и обеспечение безопасности данных;
- тестирование и, при необходимости, улучшение методов сбора данных;
- измерение качества данных и, при необходимости, повышение качества данных. Это потенциально позволит уменьшить нагрузку на пользователей данных и снизить риск возникновения по ходу дальнейшей обработки несогласованности в данных вследствие различных преобразований.

Примечание — Для получения дополнительной информации о стадии комплектования наборов данных см. *ГОСТ Р 70889*.

5.3.2.5 Стадия 4: подготовка наборов данных

На стадии подготовки наборов данных собранные данные преобразуются в форму, пригодную для использования в проекте аналитики или машинного обучения. Данная стадия играет важную роль в обеспечении соответствия требованиям к качеству данных и может быть выполнена повторно в зависимости от результатов процесса аналитики или производительности обученной модели МО. В зависимости от выявленных требований к качеству данных эта стадия может включать следующие опциональные процессы:

- трансформация данных: преобразование данных из одного представления в другое;
- валидация данных: обеспечение корректности данных на основе валидации характеристик качества данных, таких как правильность, значимость, безопасность и защита персональных данных;
- очистка данных: обнаружение неаккуратных или недостающих данных и корректировка 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. Выборка может выполняться как с возвращением (возвратом), так и без возвращения;
- создание признаков: создание новых атрибутов, позволяющих извлекать основную информацию из данных более эффективно, чем исходные атрибуты;
- отбор признаков: уменьшение размерности данных за счет использования подмножества доступных признаков;
- обогащение: связывание различных источников данных и добавление дополнительного контекста к данным;
- разметка и аннотирование данных: обучающие, валидационные и тестовые данные для машинного обучения с учителем требуют наличия значений для одной или нескольких целевых переменных; разметка данных представляет собой процесс присвоения значений целевым переменным, если эти значения не присутствовали в скомплектованных наборах данных.

Примечания

1 Для различных задач машинного обучения могут потребоваться дополнительные уникальные процессы подготовки данных.

2 Для получения дополнительной информации о подготовке данных см. *ГОСТ Р 70889* и *ГОСТ Р 71484.4*.

5.3.2.6 Стадия 5: предоставление данных

На стадии предоставления данных подготовленные данные передаются для использования в проекте аналитики или машинного обучения. На данной стадии производительность аналитики или обученной модели МО оценивается на предмет соответствия требованиям. Если результаты анализа или производительность модели машинного обучения не соответствуют ожиданиям, то могут быть выполнены следующие шаги:

- для аналитики и машинного обучения устанавливается, в какой степени обучающие данные и/или алгоритм являются основной причиной несоответствия;
- создатель данных и распорядитель данными информируются о проблемах качества данных, выявленных на стадии предоставления данных (например, создателю данных и распорядителю данными можно сообщить о проблемах с качеством данных, которые отрицательно влияют на производительность модели машинного обучения); создатели данных и распорядители данными могут использовать

такую информацию для повышения качества данных на начальных стадиях обработки в интересах будущих пользователей данных;

- улучшить качество данных, повторно пройдя стадии со 2 по 4.

Примечание — Иногда единственным способом добиться приемлемой производительности модели машинного обучения является использование других данных;

- повторно проводится анализ или перестраивается модель машинного обучения.

5.3.2.7 Стадия 6: вывод данных из эксплуатации

На стадии вывода данных из эксплуатации данные вместе с метаданными могут быть сохранены или заархивированы для будущего использования. В некоторых случаях может потребоваться уничтожение данных либо их возврат распорядителю. Вместе с архивированными данными также должны быть сохранены требования к данным и сведения об условиях, в которых используются данные для системно-зависимых данных.

5.3.3 Сквозные процессы, выполняемые на ряде стадий

5.3.3.1 Общие положения

На рисунке 4 показаны процессы, которые должны выполняться на нескольких стадиях жизненного цикла данных для аналитики и машинного обучения.

Рисунок 4 — Процессы, выполняемые на ряде стадий

Оперативное и стратегическое управление качеством данных и отслеживание происхождения данных описаны в 5.2.

5.3.3.2 Безопасность данных

Набор данных должен храниться защищенным образом на всех стадиях жизненного цикла данных, чтобы обеспечить его доступность авторизованным лицам и процессам, а также отсутствие несанкционированных изменений в данных. Несанкционированные изменения в наборе данных сами по себе могут привести к неправильным результатам при использовании моделей машинного обучения и выполнении других аналитических задач.

Примечание — Для получения дополнительной информации о безопасности данных см. *[17]* и взаимосвязанные с ним международные стандарты.

5.3.3.3 Защита персональных данных

Наборы данных, используемые для машинного обучения и аналитики, могут содержать персональные данные, которые должны защищаться в соответствии с применимыми требованиями на всех стадиях жизненного цикла данных. Для удаления персональных данных могут быть использованы методы деидентификации (анонимизации, обезличивания), однако эксплуатационные данные, используемые для прогнозирования в отношении отдельных людей, могут по-прежнему содержать персональные данные.

Примечание — Для получения дополнительной информации о защите персональных данных см. *[17]*.

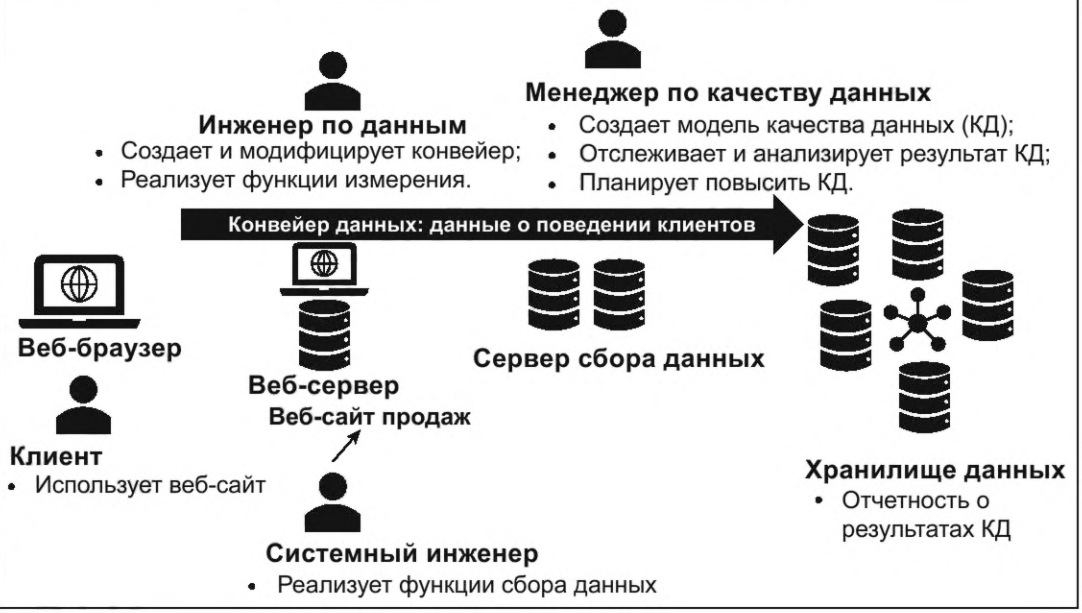
Приложение А
(справочное)

Примеры и сценарии обеспечения качества данных для аналитики и машинного обучения

В настоящем приложении представлены примеры и сценарии, относящиеся к качеству данных для аналитики и машинного обучения (например, традиционные статистические методы, машинное обучение или глубокое обучение).

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

Таблица А.1 — Сбор и хранение данных

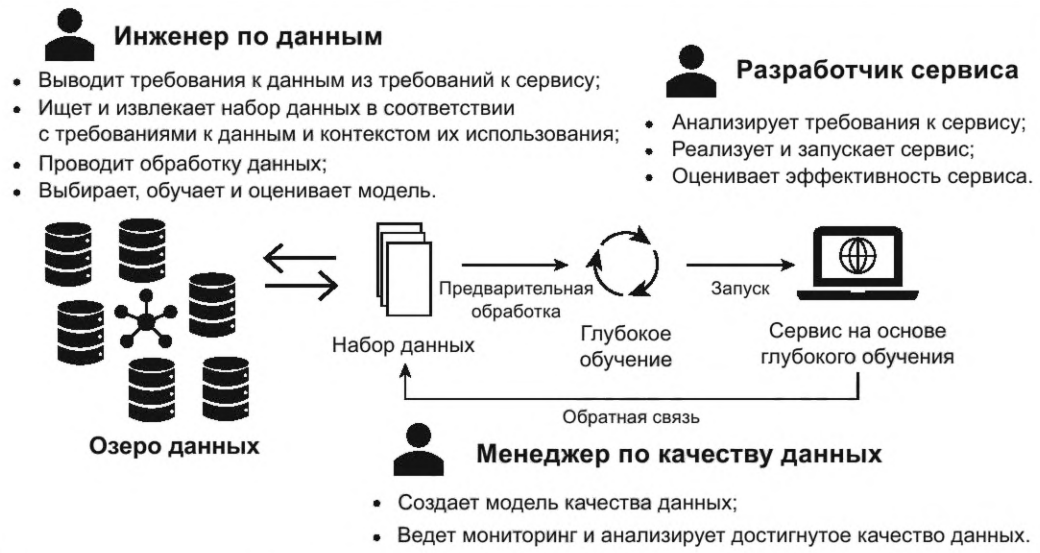
Название	Сбор и хранение данных с использованием конвейера данных в хранилище данных
Описание	Компания, занимающаяся онлайн-продажами, планирует собирать на своем веб-сайте данные о закономерностях поведения клиентов и использовать их в маркетинговых целях. С этой целью создается конвейер данных для сбора данных о связанных с действиями клиентов событиях и их размещения в хранилище данных. Эти данные затем используются для различного рода анализа. Для этого: - шаг 1: (системный инженер) реализует функции сбора данных о событиях, связанных с клиентом; - шаг 2: (менеджер по качеству данных) создает универсальную модель качества данных для собранного набора данных; - шаг 3: (инженер по данным) создает конвейер для предварительной обработки данных, включающей очистку данных, ETL-обработку (от англ. extract, transform, load — «извлечение, преобразование, загрузка»); сегментацию данных в блоки по времени и размещение их в хранилище данных; - шаг 4: (менеджер по качеству данных, инженер по данным) реализует функции измерения для каждой характеристики качества данных на основе модели качества данных, созданной на шаге 2, и добавляет эти 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 удовлетворяет требованиям, то (менеджер по качеству данных) анализирует причину и предлагает план по повышение качества данных; - шаг 7: (инженер по данным) модифицирует конвейер данных во исполнение плана, разработанного на шаге 6; - шаг 8: когда качество данных достигает желаемого уровня, (инженер по данным) завершает конвейер данных. Таким образом, шаги 1—8 будут при необходимости повторяться, пока качество данных не достигнет желаемого уровня, после чего данный цикл будет завершен
Рисунок	

Окончание таблицы А.1

Название	Сбор и хранение данных с использованием конвейера данных в хранилище данных
Вопросы качества данных и технические аспекты	- построение модели качества данных для данных и наборов данных общего назначения; - реализация функций измерения качества данных; - подготовка отчетности о качестве данных во время сбора и хранения данных для поддержки: - анализа причин в случае, когда качество данных не удовлетворяет требованиям; - повышения качества данных; - хранение информации о качестве данных в метаданных и в каталоге данных

В таблице А.2 описан сценарий управления качеством данных в ходе итераций стадий жизненного цикла данных.

Таблица А.2 — Управление качеством данных

Название	Управление качеством данных в ходе итераций стадий жизненного цикла данных
Описание	Компания стремится разработать сервис, который будет рекомендовать каждому пользователю персонализированную рекламу посредством использования глубокого обучения. Для этого: - шаг 1: (инженер по данным, разработчик сервиса) анализирует требования к сервису и выводит из них требования к данным; - шаг 2: (менеджер по качеству данных) создает модель качества данных в соответствии с требованиями к данным; - шаг 3: (инженер по данным) получает доступ к «озеру данных» и извлекает полезные данные в соответствии с требованиями к данным; - шаг 4: (инженер по данным) проводит обработку данных (включая повышение их качества), преобразуя их к форме, необходимой для глубокого обучения; - шаг 5: (инженер по данным) выбирает, обучает и оценивает модель; - шаг 6: (разработчик сервиса) реализует и запускает сервис, оценивает его эффективность. Опционально, с учетом измеренной эффективности сервиса, определяются дополнительные направления улучшения модели, и, исходя из этого, повторно выполняются шаги со 2 по 6
Рисунок	
Вопросы качества данных и технические аспекты	- поиск и извлечение набора данных на основе требований к данным и контекста их использования; - мониторинг качества данных и принятие решения о необходимости повторного измерения качества данных; - анализ воздействия обработки данных на измеренное значение других характеристик качества данных, включенных в модель качества данных

Приложение ДА
(справочное)

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

Таблица ДА.1

Обозначение ссылочного национального стандарта	Степень соответствия	Обозначение и наименование ссылочного международного стандарта
ГОСТ Р 71476—2024 (ИСО/МЭК 22989:2022)	MOD	ISO/IEC 22989:2022 «Информационные технологии. Искусственный интеллект. Концепции и терминология искусственного интеллекта»

Примечание — В настоящей таблице использовано следующее условное обозначение степени соответствия стандарта:

- MOD — модифицированный стандарт.

Библиография

[1] *ИСО/МЭК FDIS 5259-5 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 5. Структура стратегического управления качеством данных (Artificial intelligence — Data quality for analytics and machine learning — Part 5: Data quality governance framework)*

[2] *ИСО/МЭК CD TR 5259-6 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 6. Структура визуализации качества данных (Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 6: Visualization framework for data quality)*

[3] *ПНСТ 838—2023/ИСО/МЭК 23053:2022 Структура описания систем искусственного интеллекта (ИИ), использующих машинное обучение*

[4] *ИСО/МЭК ISO/IEC 23751:2022 Информационная технология. Облачные вычисления и распределенные платформы. Рамочное соглашение об обмене данными (DSA) (Information technology — Cloud computing and distributed)*

[5] *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]*

[6] *ИСО/МЭК 25030:2019 Системная и программная инженерия. Требования и оценка качества систем и программной продукции (SQuaRE). Концепция требований качества [Systems and software engineering — Systems and software quality requirements and evaluation (SQuaRE) — Quality requirements framework]*

[7] *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]*

[8] *Руководство ИСО/МЭК 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)]*

[9] *ИСО/МЭК 11179-33:2023 Информационная технология. Регистры метаданных (РМД). Часть 33. Метамодель для регистрации набора данных (Information technology — Metadata registries (MDR) — Part 33: Metamodel for data set registration)*

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

[11] *ISO/TR 21965:2019 Информация и документация. Управление записями в корпоративной архитектуре (Information and documentation — Records management in enterprise)*

[12] *ИСО/МЭК 11179-1:2023 Информационная технология. Регистры метаданных (РМД). Часть 1. Основные положения (Information technology — Metadata registries (MDR) — Part 1: Framework)*

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

[14] *ИСО/МЭК 20547-3:2020 Информационные технологии. Эталонная архитектура больших данных. Часть 3. Эталонная архитектура (Information technology — Big data reference architecture — Part 3: Reference architecture)*

[15] *ИСО/МЭК 38505 (все части) Информационная технология. Стратегическое управление данными. (Information technology — Governance of IT — Governance of data)*

[16] *ПНСТ 843—2023 (ИСО/МЭК 38507:2022) Информационные технологии. Стратегическое управление информационными технологиями. Последствия влияния стратегического управления при использовании искусственного интеллекта организациями (Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations)*

[17] *ИСО/МЭК 27701:2019 Методы и средства обеспечения защиты. Расширение ИИСО/МЭК 27001 и ИСО/МЭК 27002 в отношении менеджмента персональной информации. Требования и руководящие указания (Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines)*

УДК 004.01:006.354 ОКС 35.020

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

Редактор *З.А. Лиманская*
Технический редактор *В.Н. Прусакова*
Корректор *Р.А. Ментова*
Компьютерная верстка *Л.А. Круговой*

Сдано в набор 31.10.2024. Подписано в печать 08.11.2024. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 2,79. Уч.-изд. л. 2,12.

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