

Искусственный интеллект

КАЧЕСТВО ДАННЫХ ДЛЯ АНАЛИТИКИ И МАШИННОГО ОБУЧЕНИЯ

Часть 4

Структура процесса управления качеством данных

Artificial intelligence.
Data quality for analytics and machine learning.
Part 4. Data quality process framework

Дата введения — 2025—01—01

1 Область применения

Настоящий стандарт устанавливает общие организационные подходы, используемые независимо от типа, размера или характера организации, для обеспечения качества данных для обучения и оценки в области аналитики и машинного обучения. Стандарт включает в себя руководство по процессу управления качеством данных для:

- машинного обучения с учителем;
- машинного обучения без учителя;
- машинного обучения с частичным привлечением учителя;
- аналитики.

Настоящий стандарт применим к обучающим и тестовым данным, которые поступают из различных источников, включая сбор и комплектование данных, подготовку данных, разметку данных, оценку и использование данных. Настоящий стандарт не определяет конкретные сервисы, платформы или инструменты.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ Р 54995 Телевидение вещательное цифровое. Требования к кодированию аудио и видеосигналов для приложений вещания, основанных на транспортных потоках MPEG-2

ГОСТ Р 59926—2021 (ИСО/ИЕС TR 20547-2:2018) Информационные технологии. Эталонная архитектура больших данных. Часть 2. Варианты использования и производные требования

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

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

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 ИСО 2859-1 Статистические методы. Процедуры выборочного контроля по альтернативному признаку. Часть 1. Планы выборочного контроля последовательных партий на основе приемлемого уровня качества

ГОСТ Р ИСО/МЭК 17826 Информационные технологии. Интерфейс управления облачными данными (CDMI)

ГОСТ Р ИСО/МЭК 19794-5 Автоматическая идентификация. Идентификация биометрическая. Форматы обмена биометрическими данными. Часть 5. Данные изображения лица

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 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 следующие термины с соответствующими определениями:

3.1

аутсорсинг (outsourcing): Любая работа (или процесс), передаваемая организацией для исполнения внешней организации.

[ГОСТ Р 56398—2015, пункт 3.14]

3.2 **автономная аннотация (stand-off annotation)**: Аннотация, охватывающая различные слои первичных данных и сериализуемая в документе, отделенном от документа, который содержит первичные данные.

Примечание — См. [1], пункт 2.7.

3.3

служба облачных вычислений (cloud service): Одна или более возможностей, предоставляемых через облачные вычисления, вызываемых посредством определенного интерфейса.

[ГОСТ ISO/IEC 17788—2016, пункт 3.2.8]

3.4 **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, упомянутого в данных, описанного ими, либо явно или неявно связанного 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. [2], пункт 3.2.

3.5

ограничивающий прямоугольник (bounding box): Прямоугольная область, содержащая аннотируемый объект.

[ГОСТ Р 70268.2—2022, пункт 3.3]

3.6 **сегментация (segmentation)**: Процесс отделения интересующих объектов от их окружения.