ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

П Р Е Д В А Р И Т Е Л Ь Н Ы Й
Н А Ц И О Н А Л Ь Н Ы Й
С Т А Н Д А Р Т
Р О С С И Й С К О Й
Ф Е Д Е Р А Ц И И

ПНСТ

СИНТЕЗ ДАННЫХ

Описание результатов процесса синтеза.

Методика оценки качества

Издание официальное

Москва

Российский институт стандартизации

2026

ПНСТ

Предисловие

1 РАЗРАБОТАН Некоммерческой организацией «Ассоциация Больших Данных» (НО «АБД») и Автономной некоммерческой организацией «Национальный технологический центр цифровой криптографии» (АНО «Национальный технологический центр цифровой криптографии»)

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. № ____

Правила применения настоящего стандарта и проведения его мониторинга установлены в ГОСТ Р 1.16–2011 (разделы 5 и 6).

Федеральное агентство по техническому регулированию и метрологии собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее чем за 4 мес до истечения срока его действия разработчику настоящего стандарта по адресу: 123112 Москва, Пресненская набережная, д. 10, стр. 2, электронная почта: info@rubda.ru и/или в Федеральное агентство по техническому регулированию и метрологии по адресу: 123112 Москва, Пресненская набережная, д. 10, стр. 2.

В случае отмены настоящего стандарта соответствующая информация будет опубликована в ежемесячном информационном указателе «Национальные стандарты» и также будет размещена на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

Содержание

Введение

Синтез данных с использованием искусственного интеллекта (ИИ) представляет собой сложный процесс, требующий балансирования качества и уровня приватности. Процесс синтеза является недетерминированным и предполагает применение современных технологий и методов для получения качественного результата. Синтетические данные, созданные с использованием ИИ, должны не только отражать основные характеристики исходных данных, но и быть полезными, а также соответствовать нормам объяснимости и отсутствия смещенности, вызывать доверие (см. [1]).

Для оценки результатов процесса синтеза данных необходимо использовать различные метрики. Измеряемые метрики позволяют определить степень соответствия синтетических данных исходным данным по основным характеристикам, их применимость для различных задач, а также оценить уровни приватности и минимизации вероятности утечек данных. Важно, чтобы синтетические данные сохраняли статистические свойства исходных данных, одновременно обеспечивая требуемый уровень приватности.

Целью настоящего стандарта является описание критериев качества синтетических данных и методов их оценки. В данном стандарте представлены метрики, которые позволяют оценить качество синтетических данных с точки зрения их соответствия исходным данным, их полезности и непредвзятости. Также стандарт охватывает вопросы устойчивости полностью синтетических данных к возможной утечке данных.

Дополнительно в стандарте рассматриваются вопросы документирования и отчетности по результатам синтеза данных, включая протоколы создания данных, аудит, соответствие внешним стандартам и публикацию отчетов по качеству данных. Такой подход обеспечивает прозрачность и подотчетность процесса синтеза, что является важным фактором для доверия к синтетическим данным и их использования в различных приложениях.

Описываемые ниже методы оценки результатов процессов синтеза нацелены преимущественно на полностью синтетические данные и определены в ПНСТ 1064–2026 Синтез данных. Основные положения. Любые получаемые метрики необходимы для количественной оценки эффективности технологий и не должны подменять требования действующего законодательства (см. [2]).

ПНСТ

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

СИНТЕЗ ДАННЫХ

Описание результатов процесса синтеза. Методика оценки качества

Data synthesis. Description of the synthesis process results. Quality assessment methodology

Срок действия – с __-____-____

до __-____-____

1 Область применения

Настоящий стандарт описывает результаты процесса синтеза структурированных данных, включая методы оценки качества синтетических данных, проверку уровня приватности, а также требования к документированию и отчетности. Стандарт устанавливает подход к методам оценки результата синтеза данных с использованием ИИ, включая оценку качества синтетических данных, их соответствия требованиям обработки информации и полезности для различных задач и приложений. Стандарт применим к организациям любых форм собственности и масштабов, использующим технологии ИИ и синтетические данные.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

Издание официальное

ПНСТ

ГОСТ Р 50922 Защита информации. Основные термины и определения

ГОСТ Р 53114 Защита информации. Обеспечение информационной безопасности в организации. Основные термины и определения

ГОСТ Р 56214 Качество данных. Часть 1. Обзор

ГОСТ Р 56922 Системная и программная инженерия. Тестирование программного обеспечения. Часть 3. Документация тестирования

ГОСТ Р 58143 Информационная технология. Методы и средства обеспечения безопасности. Детализация анализа уязвимостей программного обеспечения в соответствии с ГОСТ Р ИСО/МЭК 15408 и ГОСТ Р ИСО/МЭК 18045. Часть 2. Тестирование проникновения

ГОСТ Р 58412 Защита информации. Разработка безопасного программного обеспечения. Угрозы безопасности информации при разработке программного обеспечения

ГОСТ Р 58606 Системная и программная инженерия. Процесс измерения

ГОСТ Р 58771 Менеджмент риска. Технологии оценки риска

ГОСТ Р 59276 Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения

ГОСТ Р 59334 Системная инженерия. Защита информации в процессе управления качеством системы

ГОСТ Р 59407 Информационные технологии. Методы и средства обеспечения безопасности. Базовая архитектура защиты персональных данных

ГОСТ Р 59900 Системы искусственного интеллекта. Типовые требования к контрольным выборкам исходных данных для испытания систем искусственного интеллекта в образовании

ГОСТ Р 59989 Системная инженерия. Системный анализ процесса управления качеством системы

ГОСТ Р 59991 Системная инженерия. Системный анализ процесса управления рисками для системы

ГОСТ Р 59994 Системная инженерия. Системный анализ процесса гарантии качества для системы

ГОСТ Р 70462.1 Информационные технологии. Интеллект искусственный. Оценка робастности нейронных сетей. Часть 1. Обзор

ГОСТ Р 71438 Информационные технологии. Оценка процессов. Система измерения процессов для оценки их возможностей

ГОСТ Р 71440 Информационные технологии. Оценка процессов. Руководство по определению рисков в процессах

ГОСТ Р ИСО 8000-2 Качество данных. Часть 2. Словарь

ГОСТ Р ИСО 8000-100 Качество данных. Часть 100. Основные данные. Обмен данными характеристик. Обзор

ГОСТ Р ИСО 16269-7 Статистические методы. Статистическое представление данных. Медиана. Определение точечной оценки и доверительных интервалов

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

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

ГОСТ Р ИСО/МЭК 42001 Искусственный интеллект. Система менеджмента

ПНСТ 1064—2026 Синтез данных. Основные положения

П р и м е ч а н и е — 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 термины по ГОСТ Р 50922, ГОСТ Р 53114, ГОСТ Р 56214, ГОСТ Р 58771, ПНСТ 1064—2026, а также следующие термины с соответствующими определениями:

3.1 **восходящий поток информации** (up streaming): Поток синтетических данных, используемых для обучения моделей машинного обучения.

3.2 **нисходящий поток информации** (down streaming): Поток предварительно обученных моделей машинного обучения, используемых для оценки синтетических данных.

3.3

словарь данных (data dictionary): Хранилище всех элементов данных и допустимых значений этих элементов, используемых в спецификациях метаданных ДОИ.

[ГОСТ Р ИСО 26324—2015, пункт 3.3]

Примечание — Словарь данных предназначен для обеспечения единообразного понимания структуры и содержания данных, а также для документирования исходных процессов при анализе, синтезе и обмене информацией.

3.4

технология, улучшающая обеспечение приватности; PET (privacy enhancing technology, PET): Мера и средство контроля и управления приватностью, состоящая из мер, продуктов или сервисов системы информационно-коммуникационных технологий, которые обеспечивают защиту приватности путем уничтожения или сокращения объема персональной идентификационной информации или предотвращения ненужной и (или) нежелательной обработки персональной идентификационной информации без потери функциональности системы информационно-коммуникационных технологий.

[ГОСТ Р ИСО/МЭК 29100—2013, пункт 2.15]

Примечание — PET включают обезличивание, дифференциальную приватность, синтез данных и другие методы защиты.

3.5 **федеративное (совместное) машинное обучение** (shared machine learning): Парадигма машинного обучения, позволяющая агрегировать принадлежащие ряду сторон данные и обеспечивать многостороннюю защиту персональных данных в тех ситуациях, когда различные поставщики данных и вычислительная платформа не доверяют друг другу.

Примечание — Федеративное обучение применяется в ситуациях, когда источники данных не могут быть объединены по причинам, связанным с приватностью, безопасностью или юридическими ограничениями. Алгоритмы могут предусматривать дополнительные меры защиты, включая дифференциальную приватность или защищенные вычисления.

4 Общие положения

4.1 Результат процесса синтеза данных

В рамках ПНСТ 1065–2026 установлены архитектура и процесс синтеза данных с использованием методов ИИ. Основные компоненты процесса включают:

- реальные данные: чувствительные исходные данные, в том числе содержащие персональные данные (ПДн);
- систему синтеза;
- генератор синтетических данных;
- синтетические данные.

Основным результатом процесса синтеза данных являются синтетические данные, для которых могут быть рассчитаны формальные метрики качества и приватности. Упрощенное представление синтеза данных представлено в приложении А ПНСТ 1065–2026 (см. также рисунок 1).

Примечание — Генератор синтетических данных может представлять собой обученную модель, но в ряде случаев включает дополнительные компоненты. При этом как генератор, так и вся система синтеза могут рассматриваться как косвенные результаты процесса в зависимости от целей анализа.

В процессе синтеза данных используются методы обеспечения доверенных технологий ИИ, такие как обезличивание, фильтрация и дифференциальная приватность. В случае распределенного обучения применяется федеративное обучение.

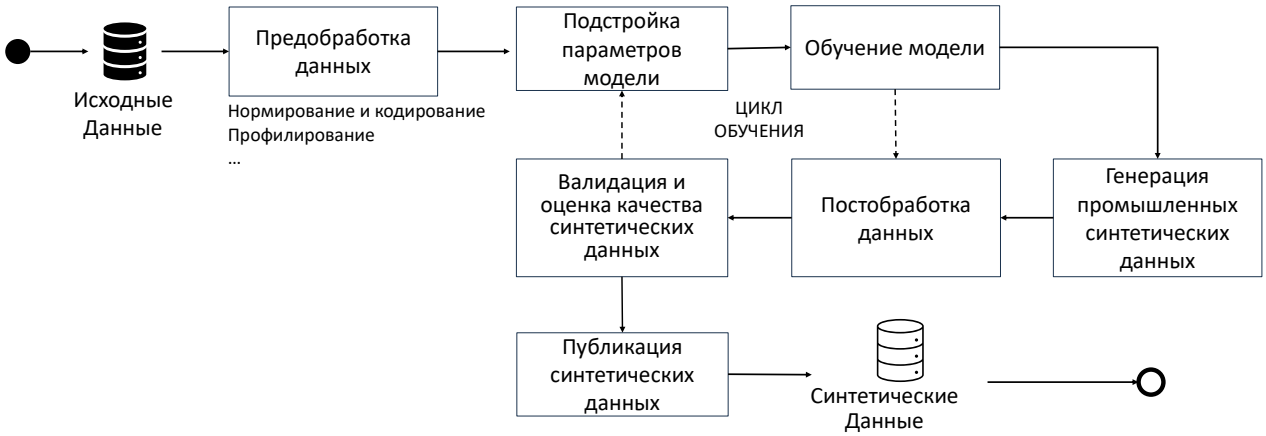

Рисунок 1 — Упрощенный процесс синтеза

Синтетические данные могут применяться в различных сценариях:

- создание обучающих наборов данных для моделей ИИ;
- анализ данных и совместная работа.

Синтетические данные ускоряют исследования, инновации и принятие решений, а также снижают вероятность утечек данных и обеспечивают соответствие правилам обеспечения приватности. В этих целях синтетические данные могут публиковаться как для внутреннего, так и для внешнего использования, как для общего применения, так и для решения специфических задач (например, прогнозирования вероятности утечки информации).

В системах машинного обучения различают восходящий и нисходящий потоки информации. Восходящий поток информации обеспечивает передачу синтетических данных, необходимых для тренировки и адаптации моделей машинного обучения. В свою очередь, нисходящий поток информации представляет собой использование обученных моделей для

последующей обработки и анализа синтетических данных, в частности, для целей их оценки, валидации и проверки.

Применение данных потоков позволяет оптимизировать процесс разработки и эксплуатации систем машинного обучения, особенно в условиях, когда доступ к реальным данным ограничен или сопряжен с рисками безопасности и приватности. При проектировании систем машинного обучения следует обеспечивать четкое разделение и управление этими потоками данных для минимизации рисков и повышения эффективности использования синтетических данных.

Артефакты процесса синтеза данных включают обученную модель, синтетические наборы данных, метаинформацию (конфигурацию параметров обучения), а также подмножества наборов реальных и синтетических данных для формирования наборов данных для обучения и тестирования (включая разделение реального и синтетического наборов на обучающую и тестовую выборки). Также важным результатом является настроенная и работоспособная система синтеза (см. ГОСТ Р 71438). В настоящем стандарте под основными результатами синтеза подразумеваются выходной набор синтетических данных и обученная модель синтеза.

Модели не способны воспроизвести то, чего не было в исходных данных, имеющаяся информация также усваивается с ограниченной точностью (см. [3]). Тем не менее, качественные синтетические данные способны существенно ускорить процессы анализа и обработки, снижая затраты на протяжении всего жизненного цикла разработки систем синтеза.

Синтез данных может рассматриваться как один из подходов обеспечения конфиденциальности, входящий в состав технологий повышения

приватности. При этом уровень приватности достигается не самим фактом наличия синтетических данных, а процессом их генерации, включающим методы управления рисками раскрытия информации.

При использовании сложных генеративных моделей, особенно в многомерных пространствах, затруднена однозначная оценка уровня точности и приватности, поскольку эти параметры могут варьироваться в зависимости от характера данных и архитектуры модели. Например, в задачах, содержащих редкие события или уникальные состояния (например, редкие медицинские диагнозы), существует риск либо искажения статистической значимости, либо утечки информации, относящейся к исходным данным.

В связи с этим рекомендуется использовать методы оценки качества и приватности синтетических данных, адаптированные к конкретному сценарию использования, включая такие параметры, как форма распространения (ограниченный доступ или открытая публикация), способ предоставления (данные или модель), а также доступность параметров синтеза для третьих сторон.

Синтетические данные, создаваемые с применением методов обеспечения приватности (например, дифференциальной приватности), представляют собой новый набор данных, статистически согласованный с исходными данными, но не содержащий прямых копий или трансформированных записей. Такой подход обеспечивает снижение вероятности утечки исходной информации. Вместе с тем использование методов приватности, включая добавление шумов или ограничение точности, может снижать точность последующих моделей и прогнозов (см. [4]), что влечет за собой необходимость формализованной валидации и проверки как самих синтетических данных, так и моделей, их генерирующих. Процесс такой оценки подробно описан в ПНСТ 1065–2026.

4.2 Критерии качества результата процесса

Синтетические данные выдаются обученной моделью, ее генерирующей частью. Базовые свойства качественного генератора синтетических данных включают следующие:

- семантическая корректность, обеспечивающая правдоподобность для синтетических данных и сохранение их структурных свойств;

- уровень приватности, дающий возможность количественной оценки объема информации об исходных данных, которая раскрывается через косвенные измерения в результате выпуска синтетического набора и связанные показатели безопасности (см. ГОСТ Р 53114);

- статистическая точность, отражающая возможность точной количественной оценки статистического сходства;

- полезность, отражающая степень, с которой генератор синтетических данных выполняет цели задачи синтеза для конкретной прикладной задачи.

Общая система оценки качества результатов процесса синтеза может включать критерии качества соответствующей информационной системы, а также оценки качества самих выходных данных, на которые переносятся отдельные свойства генератора. В рамках наиболее широких подходов (см. [4]) используется модель качества систем ИИ, обобщающая аналогичный подход для стандартных автоматизированных систем и программного обеспечения (см. ГОСТ Р ИСО/МЭК 25010). Модель рассматривает качество системы как способность достигать запланированных и желаемых результатов в двух аспектах: качество продукта и качество при использовании.

В контексте синтеза данных необходимо учитывать аспекты применения модели качества: характеристики обученной модели и полученных

данных, показатели функционирования системы, включая процесс обучения, робастность нейронных сетей (см. ГОСТ Р 70462.1), отсутствие угрозы снижения доверия к технологиям синтеза (см. [3]), объяснимость используемой модели ИИ и отсутствие смещенности (см. [5]).

Общая модель качества (см. рисунок 2), применяемая к результату синтеза, наследует общие критерии качества систем ИИ с акцентом на качество выходных данных и обученной модели (см. [6], [7]).

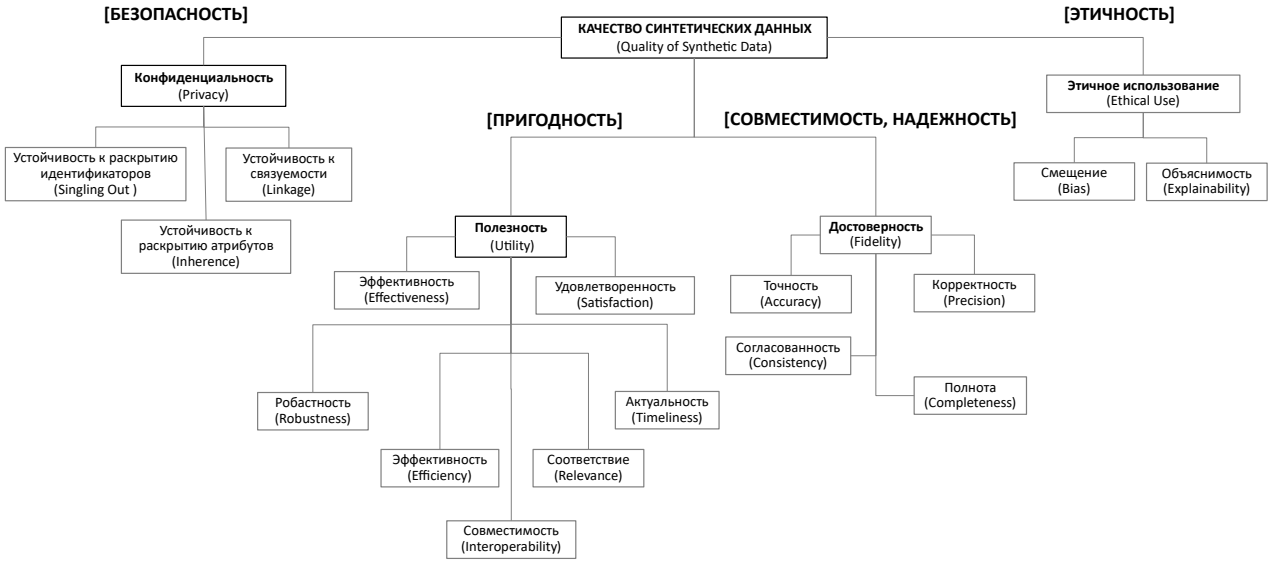

Рисунок 2 – Модель качества результатов процесса синтеза

Построение качественной системы синтеза данных связано с рядом противоречий, поскольку перечисленные выше свойства не всегда совместимы. Не существует универсального алгоритма, способного одновременно обеспечить высокую эффективность, защиту от утечек исходных данных и статистическую точность для любых сценариев применения. Для построения качественных синтетических данных требуется компромисс:

- приватность и точность: повышение статистической точности синтетических данных обычно ведет к снижению уровня приватности, поэтому полностью универсальный набор данных недостижим;

- полезность и точность: при оценке эффективности методов генерации нередко требуется баланс между полезностью и приватностью.

Примечание — Точность не всегда отражает полезность, так как в отдельных случаях возможно снижение точности без потери практической ценности (и наоборот).

5 Методы оценки качества синтетических данных

5.1 Общие положения

Оценка качества синтетических данных представляет собой сложный и многоэтапный процесс, необходимый для подтверждения соответствия данных целям и требованиям их использования (см. ГОСТ Р 59994, ГОСТ Р 56214). Основные аспекты оценки включают такие критерии, как точность, полезность, приватность, уровень смещенности и объяснимость данных, рассмотренные выше.

Методика управления показателями качества на уровне организации должна соответствовать требованиям законодательства, предусматривать оценку рисков, интерпретацию результатов и моделирование возможных угроз, способных негативно повлиять на свойства системы синтеза. Такие меры направлены на обеспечение требуемых характеристик синтетических данных и устойчивости всей системы в соответствии с положениями ГОСТ Р 59991.

Проведенные оценки и измерения позволяют определить, подходят ли синтетические данные для использования в прикладных задачах, включая машинное обучение и аналитику, с учетом требований к доверенным технологиям синтеза и отсутствию смещенности данных.

Оценка качества синтетических данных должна начинаться с анализа исходных данных и определения ключевых характеристик, влияющих на результат. Важную роль при этом играет работа с атрибутами —

как исходными, так и синтетическими (см. [8]). Подготовка данных, согласно ПНСТ 1065–2026, включает нормализацию, очистку, агрегацию, обогащение, трансформацию и профилирование. Атрибуты могут различаться по типу и степени риска повторной идентификации: прямые идентификаторы, квазиидентификаторы и чувствительные атрибуты. Для их выявления и документирования исходных процессов рекомендуется составлять отчет «Словарь данных», пример которого приведен в приложении А.

После завершения подготовки данных должен быть сформирован окончательный тренировочный набор для последующего обучения модели.

Обучение модели представляет собой итеративный процесс оптимизации ее параметров с целью выявления и запоминания закономерностей и структур, содержащихся в исходных данных. Обучение осуществляется в несколько этапов (эпох), в ходе которых модель последовательно обрабатывает тренировочные данные и обновляет свои параметры.

После каждой эпохи должна проводиться оценка модели по заранее определенным метрикам, что позволяет контролировать качество обучения и своевременно выявлять признаки переобучения. Для систематизации проводимых проверок рекомендуется применение методик системного анализа процесса управления качеством системы (см. ГОСТ Р 59989).

После завершения обучения модель должна пройти этап валидации в соответствии с положениями ПНСТ 1065–2026. В ходе валидации качество синтетических данных проверяется на тестовом наборе, который не использовался при обучении модели.

Целью валидации и оценки качества является подтверждение того, что модель способна корректно переносить выявленные статистические закономерности на данные, не входившие в обучающую выборку, а также генерировать синтетические данные, обладающие достаточной точностью, полезностью и устойчивостью к утечкам исходной информации.

Валидация и оценка качества должны включать измерения по ключевым метрикам, таким как точность (см. приложения Б и В), полезность (см. приложение Г), уровень приватности (см. приложение Д), оценка смещенности (см. [5]). Полученные значения позволяют оценить соответствие синтетических данных установленным критериям качества. Проведение валидации является обязательным этапом контроля качества синтетических данных перед их использованием в практических приложениях.

Каждая метрика, используемая для оценки качества синтетических данных, должна иметь установленные на уровне организации пороговые значения, которые определяют допустимые отклонения и служат основанием для оценки соответствия данных установленным требованиям (см. ГОСТ Р 58606).

Выход значения метрики за допустимый порог как в сторону превышения, так и в сторону занижения может свидетельствовать о недостаточном качестве синтетических данных и требует дополнительного анализа или пересмотра параметров генерации.

5.2 Оценка точности

Оценка точности синтетических данных заключается в измерении степени соответствия синтетических данных реальным данным, на основе которых они были созданы. Оценка точности помогает понять,

насколько хорошо синтетические данные сохраняют основные характеристики и распределение оригинальных данных. Состав используемых метрик может варьироваться в зависимости от типа синтетических данных (числовые, категориальные, мультимедийные), их структуры (табличные, неструктурированные) и задач, которые решаются с их помощью (например, классификация или прогнозирование). Рекомендуется использовать набор метрик, приведенный в приложении Б. В зависимости от решаемой прикладной задачи собственно статистическая точность может быть дополнена метриками оценки диверсификации и реалистичности (см. приложение В).

5.2.1 Оценка статистической близости

Статистическая близость представляет собой измерение степени, в которой синтетические данные соответствуют статистическим характеристикам реальных данных. Оценка статистической близости позволяет понять, насколько синтетические данные сохраняют ключевые статистические свойства оригинальных данных, такие как распределение, корреляции и зависимости между признаками. Этот аспект качества синтетических данных важен, поскольку от него зависит возможность использования синтетических данных в тех же приложениях, что и реальные данные.

Основные метрики оценки статистической близости, состав которых может быть адаптирован в зависимости от цели синтеза данных и уровня приватности, приведены в приложении Б. Метрики точности в части оценки диверсификации и реалистичности получаемых результатов приведены в приложении В.

5.2.2 Оценка диверсификации и реалистичности

Метрики оценки диверсификации и реалистичности позволяют оценить, насколько синтетические данные отражают разнообразие и правдоподобие реальных данных. Такие метрики используются для анализа того, насколько синтетические данные сохраняют разнообразие характеристик и выглядят реалистично, особенно в задачах, связанных с изображениями и сложными многомерными данными.

Рассматриваемая группа метрик направлена на количественную оценку реалистичности и разнообразия синтетических данных. Эта группа включает метрики сравнения распределений реальных и синтетических данных, оценки различий в восприятии для табличных данных и изображений, а также метрики обработки естественного языка для анализа реалистичности неструктурированной информации.

Сводка по метрикам оценки полезности (пригодности и соответствия) приведена в приложении Г.

5.3 Оценка полезности

Полезность позволяет оценивать, насколько хорошо синтетические данные могут быть использованы для выполнения конкретных задач**,** например прогнозирования или классификации.

Для оценки полезности (производительности модели) должны использоваться тестовые процедуры, при которых две идентичные по архитектуре модели, соответствующие прикладной задаче, поочередно обучаются на реальных и синтетических данных. После обучения модели оцениваются на общей тестовой выборке, не участвовавшей в обучении.

Результаты предсказаний обеих моделей, полученные при подаче одних и тех же входных данных, подлежат сравнению. Проведенное срав-

нение позволяет определить, насколько синтетические данные сохраняют прикладную значимость и отражают ключевые характеристики исходных данных.

В задачах классификации тестовые модели должны обеспечивать разделение данных на заданные категории или классы. При этом положительные классы интерпретируются как наличие целевой характеристики, а отрицательные — как ее отсутствие.

В рамках такого подхода для оценки качества модели должны рассчитываться выходные показатели, соответствующие основным категориям классификации, по следующей схеме:

- TP (True Positives) представляет число правильных предсказаний положительных классов;

- TN (True Negatives) дает число правильных предсказаний отрицательных классов;

- FP (False Positives) отражает число неправильно предсказанных положительных классов (ошибочные срабатывания);

- FN (False Negatives) показывает число неправильно предсказанных отрицательных классов (пропущенные случаи).

Метрики, такие как F1-мера (гармоническое среднее между точностью и полнотой), AUC-ROC (площадь под кривой, отражающей способность модели различать классы), MSE (среднеквадратичная ошибка) и pMSE (среднеквадратичная ошибка вероятностей), используют соотношения между показателями TP, TN, FP и FN для количественной оценки того, насколько точно и эффективно модель, обученная на синтетических данных, различает классы.

Наиболее приоритетные метрики оценки качества моделирования включают R^2 (коэффициент детерминации), MSE, F1-метрику, AUC-ROC

и производительность модели. Перечисленные метрики позволяют получить полное представление о том, насколько хорошо модель воспроизводит ключевые характеристики данных, будь то точность предсказаний, различие классов данных или общее соответствие модели задачам.

Выбор метрик должен осуществляться с учетом конкретной прикладной задачи и типа данных. Так, в задачах бинарной классификации наибольшую значимость имеют F1-метрика и AUC-ROC, поскольку они отражают баланс между точностью и полнотой предсказаний, а также оценивают способность модели различать классы.

В задачах регрессии более уместно применение таких метрик, как коэффициент детерминации и среднеквадратичная ошибка, которые позволяют оценить степень соответствия предсказанных значений фактическим. При необходимости учета аспектов приватности рекомендуется использование специализированных метрик, таких как pMSE, позволяющих дополнительно оценивать риски утечек данных при тестировании производительности модели. Рекомендуемые метрики и их характеристики приведены в приложении Г.

5.4 Оценка уровня приватности

5.4.1 Приватность синтетических данных

Приватность синтетических данных важна для предотвращения утечек исходной информации об индивидуальных лицах и данных от несанкционированного доступа, повторной идентификации. Полностью синтетические данные создаются с целью сохранения статистических свойств и характеристик исходных данных, при этом исключая информацию о субъектах данных, и представляют собой данные, отличные от исходных. Однако важно учитывать, что избыточное сохранение статистических

свойств может повысить вероятность утечки исходной информации (восстановления или извлечения чувствительных данных) (см. [9]).

Пример — Если синтетические данные точно воспроизводят распределение доходов в исходном наборе данных, может представиться возможность выделить физических лиц (ФЛ) с высоким уровнем дохода.

Примечания

1 В ПНСТ 1065–2026 рассмотрены архитектурные решения и доверенные технологии синтеза, такие как обезличивание, фильтрация и дифференциальная приватность, направленные на минимизацию вероятности утечки информации.

2 Уровень приватности может снижаться в результате ошибок архитектурного проектирования, недостатков реализации или неверного выбора технических решений (см. ГОСТ Р 58412).

3 Синтетические данные могут быть уязвимы для различных атак (см. ГОСТ Р 58143), направленных на восстановление или извлечение информации. Вероятность успешности таких атак зависит от использования доверенных технологий ИИ и вносит вклад в снижение вероятности утечки исходной информации.

5.4.2 Использование моделей измерения уровня приватности

Для обеспечения приватности данных (полностью синтетических данных) необходимо применять количественные методы оценки, такие как модели измерений уровня приватности и риска (см. [9]). Уровень приватности определяется вероятностью утечки данных. ГОСТ Р ИСО/МЭК 29100 и ГОСТ Р 59334 предлагают методы оценки и управления уровнем доверия в используемых технологиях.

Пример — Уникальность данных в наборе может повышать вероятность утечки исходной информации.

Оценка вероятности утечки включает анализ частоты и редкости данных в наборе (см. ГОСТ Р ИСО/МЭК 29100). Количественная оценка

нарушения уровня доверия к технологиям синтеза основывается на применении математических моделей, установленных в ГОСТ Р 71440.

В контексте настоящего стандарта особое внимание уделяется оценке вероятности утечки исходной информации.

Такая утечка может быть обусловлена уязвимостями в модели синтеза, недостаточной защищенностью методов обработки данных или несовершенством механизмов обеспечения приватности PET (см. ГОСТ Р 50922).

Примечание — Применение статистических методов позволяет предсказать вероятность утечки информации и использовать эти данные для управления вероятностями утечки исходной информации.

Если синтез данных осуществляется с использованием ПДн, то значительное влияние на приватность оказывают контекстные показатели готовности организации-владельца к использованию доверенных технологий ИИ, а вероятность утечки исходных данных в значительной степени обуславливается процессами оператора ПДн, определяемыми согласно ГОСТ Р 59407. Синтез данных должен осуществляться согласно ПНСТ 1065–2026 во внутреннем контуре с использованием общих средств поддержки необходимого уровня приватности. После реализации полностью синтетических данных уровень их приватности, связанный с вероятностью утечки исходных данных, оценивается на основании моделей, описываемых в настоящем стандарте.

Методика предварительного расчета вероятности утечки на основании метрик схожести данных приведена в Д.1. Оценка приватности на основе метрик схожести данных не всегда является надежной, поскольку такие метрики, как различие в распределении данных, не всегда учитывают возможность восстановления исходных данных.

Пример — Данные могут иметь схожее распределение, но содержать уникальные сочетания атрибутов, которые можно использовать для идентификации ФЛ.

Примечание — Метрики схожести не обеспечивают полноценную оценку уровня приватности и могут привести к недооценке вероятности утечки. Не рекомендуется применять оценку уровня приватности синтетических данных только на основании метрик схожести.

Для оценки полностью синтетических данных существуют различные модели, каждая из которых предлагает свои подходы к вычислению метрик для оценки вероятности утечки исходной информации. Эти модели позволяют оценить возможность утечки через вероятность успешности атак на синтетические данные, включая атаки выделения, связывания и вывода.

Теоретическая модель для дифференциальной приватности, реализуемая в случае использования дифференциальной приватности как метода обеспечения уровня доверия к используемым технологиям синтеза с учетом зависимости от количества записей, приведена в Д.2. Комплексная модель, учитывающая различные виды атак на синтетические данные приведена в Д.3.

5.4.3 Ограничения приватности

Синтез данных с использованием методов ИИ, включая глубокое обучение, позволяет достичь баланса между качеством и уровнем приватности данных. Однако для обеспечения должного уровня приватности производитель синтетических данных, согласно ПНСТ 1065–2026, обязан:

- применять обоснованные математические методы обеспечения приватности – PET, такие как дифференциальная приватность;

- в рамках практической реализации контролировать бюджет приватности через учет чувствительности данных;

- учитывать зависимость между объемом синтетических данных и уровнем приватности, которая может варьироваться в зависимости от используемой модели и метода синтеза;

- ограничивать количество наборов данных, генерируемых одной обученной моделью, в соответствии с параметрами методов синтеза, чтобы предотвратить чрезмерную генерацию данных на основе одной и той же модели.

Между количеством строк данных и уровнем приватности существует сложная зависимость, связанная с характеристиками модели и метода синтеза.

Пример — Малое количество строк при большом количестве параметров может привести к переобучению модели, что увеличивает вероятность утечек исходной информации. С другой стороны, слишком большое количество строк может вызвать чрезмерное расходование бюджета приватности, снижая уровень приватности данных.

Для учета этих эффектов рекомендуются:

- оптимизация объема данных: для каждой модели и метода синтеза необходимо определять оптимальный объем данных, регулярно переоценивать модели на предмет переобучения и возможных утечек исходной информации;

- контроль расхода бюджета приватности: при использовании методов, таких как дифференциальная приватность, важно следить за тем, чтобы количество сгенерированных строк не приводило к избыточному расходованию бюджета приватности (см. ПНСТ 1065–2026);

- динамическая настройка параметров синтеза: в процессе синтеза данных параметры генерации целесообразно динамически адаптировать в зависимости от объема данных, например, с увеличением числа строк

повышать уровень приватности для компенсации возросшей вероятности утечек;

- использование регуляризации и контроля параметров: для предотвращения переобучения следует внедрять методы регуляризации в процессе обучения модели, включая ограничение сложности модели, контроль числа параметров и использование проверенных техник регуляризации, таких как L2-регуляризация или раннее прекращение обучения.

Повторное использование модели для генерации синтетических данных также влияет на уровень приватности. В случае многократного синтеза данных с одной модели возможно проведение продвинутых атак на утечки данных в методах ИИ. Для минимизации вероятности утечки исходной информации необходимо:

- в случае применения дифференциальной приватности корректировать параметры методов синтеза, увеличивая бюджет приватности или другие параметры приватности (см. Д.4);

- периодически изменять или перестраивать исходные данные для новой генерации;

- комбинировать различные доверенные технологии ИИ, например обезличивание и дифференциальную приватность;

- ограничивать количество синтетических наборов, генерируемых на одной модели.

5.5 Оценка смещенности данных и объяснимость

Синтетические данные могут использоваться для различных целей, включая тестирование информационных систем и разработку моделей ИИ. Однако если данные не проверены на наличие предвзятости и смещения либо методы их генерации недостаточно прозрачны, это может привести к недоверию со стороны пользователей и регуляторов (см. [10]).

Кроме того, некорректное использование таких данных может повлечь негативные последствия, которые могут рассматриваться как утрата управляемости системой синтеза (см. [11]).

Оценка смещенности синтетических данных — важный аспект обеспечения справедливости, прозрачности и доверия к данным и моделям, созданным на их основе (см. ГОСТ Р 59276).

Оценка смещенности и объяснимости включает следующие ключевые аспекты: проверку на предвзятость, оценку робастности (см. ГОСТ Р 70462.1), объяснимость и прозрачность. Важной частью этого процесса является построение доверия через документирование происхождения данных и методов их генерации (см. [12]).

Оценка смещенности в синтетических данных включает измерение смещения в выборках, моделях и результатах (см. [5]). Метрики для оценки смещения могут включать коэффициенты дисбаланса классов, матрицу ошибок и F1-метрику для различных подгрупп данных. Также применяется анализ перекрестного смещения. Эти метрики помогают выявить и количественно оценить предвзятость, которая может негативно влиять на справедливость моделей.

Прозрачность и объяснимость синтетических данных и моделей, основанных на них, важны для обеспечения доверия. Методы интерпретации, коэффициенты корреляции атрибутов и решений, а также интегрированные градиенты помогают оценить вклад каждого атрибута в конечные решения модели (см. [13]). Эти методы предоставляют доступные объяснения для различных заинтересованных сторон.

Оценка прозрачности в процессах генерации синтетических данных подразумевает раскрытие методов генерации, наличие документации и

доступность исходного кода для анализа. Метрики прозрачности включают уровень раскрытия методов, качество документации процессов и доступность исходного кода и моделей для независимого анализа.

Проведенные оценки смещенности, прозрачности и объяснимости позволяют определить общий уровень доверия к синтетическим данным и моделям. Для общей оценки синтетических данных в разрезе смещенности и объяснимости могут применяться скоринговые методы, начисляющие баллы в зависимости от выявленных смещений, уровня прозрачности и документированности. Эти показатели могут быть конвертированы в индексы доверия к данным, репрезентативности данных и оценки уровня приватности и надежности.

Примечание — В контексте оценки синтетических данных могут применяться как встроенные методы интерпретации результатов (например, анализ взаимосвязей между признаками), так и методы последующего объяснения, используемые после генерации данных. Выбор подхода определяется архитектурой модели синтеза и требованиями к степени объяснимости полученных результатов.

6 Документирование и отчетность

6.1 Протоколы создания данных

Документирование процессов создания и происхождения синтетических данных является ключевым элементом обеспечения их качества, прозрачности и доверия к ним (см. [12]). Без надлежащей документации сложно гарантировать корректность и надежность синтетических данных, а также обеспечить их соответствие требованиям законодательства и стандартов.

Протоколы создания синтетических данных должны использоваться для систематизированного описания и документирования всех этапов их

жизненного цикла — от планирования и проектирования до вывода данных из эксплуатации.

В состав протокола должны входить сведения о применяемых методах синтеза, используемых исходных данных, параметрах и настройках моделей, а также об организациях и ответственных лицах, участвующих в каждом этапе процесса. Основные элементы протокола должны содержать:

- идентификатор набора данных — уникальный идентификатор, присвоенный набору данных;

- методы синтеза — подробное описание методов, алгоритмов и параметров, использованных для создания синтетических данных;

- источники данных — описание и идентификация всех исходных данных, использованных в процессе синтеза;

- ответственные стороны — информация о ФЛ и организациях, участвовавших в создании данных;

- результаты тестирования, валидации и проверки — данные о проведенных тестах и их результатах, подтверждающие качество и соответствие синтетических данных требованиям.

6.2 Аудит и следование стандартам

Следование стандартным процессам и соблюдение установленных норм являются ключевыми элементами обеспечения качества, надежности и прозрачности синтетических данных (см. ГОСТ Р ИСО/МЭК 42001). Стандартизированный подход минимизирует возможность предвзятости, неполноты и непрозрачности данных, что важно для укрепления доверия к данным и моделям, созданным на их основе.

Для достижения высокого качества синтетических данных необходимо учитывать влияние смежных стандартов, регулирующих аспекты работы с большими данными и ИИ.

Пример — [14] описывает эталонную архитектуру больших данных, [12] регулирует требования к происхождению данных. Эти стандарты обеспечивают основу для создания и верификации синтетических данных в соответствии с установленными нормами.

Целью такого аудита является подтверждение того, что синтетические данные соответствуют требованиям к качеству, установленным ограничениям на уровень приватности, а также критериям прозрачности и документированной прослеживаемости происхождения данных.

Основные этапы процесса аудита:

- подготовка: сбор и анализ документации, включая протоколы создания данных, метаинформацию и отчеты о результатах тестирования;

- проверка соответствия: сравнение процессов создания и управления данными с требованиями применимых стандартов (см. [12]);

- проведение тестов: оценка данных и процессов на соответствие требованиям стандарта, включающая тестирование на предмет отсутствия смещений и предвзятости в синтетических данных;

- интервью: встречи с ответственными лицами для подтверждения понимания и соблюдения требований стандартов;

- отчетность: подготовка и представление отчета по результатам аудита, включающего рекомендации по улучшению процессов.

Для оценки зрелости процессов создания и управления синтетическими данными рекомендуется использование моделей зрелости (см. ГОСТ Р ИСО 8000-2). Данные модели оценивают уровень зрелости организации в сфере ИИ, включая аспекты доверия к данным, стратегии,

организационной структуры, технологий и данных. По аналогии с ИИ базовая модель зрелости организации для синтеза данных включает пять уровней:

- начальный уровень: процессы синтеза данных осуществляются спонтанно и несистемно, отсутствует стратегия и управление синтезом данных;

- вовлеченный уровень: организация начинает организовывать работу с ИИ, формируя команды, разрабатывая политику и процедуры;

- определенный уровень: внедряются утвержденные на уровне предприятия подходы, ресурсы и процессы для работы с синтезом данных;

- управляемый уровень: процессы синтеза данных строго следуют утвержденной политике и стандартам, собираются и анализируются метрики для оценки результатов;

- оптимизированный уровень: организация достигает высокого уровня зрелости в работе с синтетическими данными, постоянно улучшая и инновационно развивая свои проекты.

Использование моделей зрелости позволяет выявить слабые места в процессах и наметить направления для их улучшения. Включение оценки зрелости в процесс аудита позволяет более полно оценить соответствие процессов лучшим практикам и стандартам.

6.3 Отчеты о качестве и использовании данных

Отчет о качестве данных является документом, в котором фиксируются результаты оценки качества синтетических данных, включая процесс их валидации и проверки, соответствие установленным стандартам и требованиям. Валидация данных, описанная в ПНСТ 1065–2026, играет

ключевую роль в подтверждении достоверности и пригодности синтетических данных для использования. В данном отчете следует отразить, каким образом синтетические данные соответствуют требованиям, установленным смежными стандартами, такими как ГОСТ Р ИСО 8000-100 (см. также [4]).

Процесс валидации данных включает проверку синтетических данных на соответствие качеству. Валидация охватывает такие аспекты, как достоверность, полнота, точность и соответствие исходным требованиям. Отчет о качестве данных должен содержать результаты валидации, а также описание используемых методологий и инструментов с учетом требований ГОСТ Р 56922.

Неотъемлемой частью отчета по качеству является словарь данных, приведен в приложении А. Отчет о качестве данных должен включать следующие реквизиты:

- дата составления отчета: указывается точная дата;

- данные тестируемого набора: включают уникальный идентификатор (fingerprint) набора данных, краткое описание набора данных, информацию о его хранении;

- информация о системе синтеза: приводится описание используемой системы синтеза данных;

- профайлинг исходных данных: приводится описание исходных данных, использованных для создания синтетических данных;

- выбранный способ реализации доверенных технологий синтеза: приводится описание методов снижения вероятности утечки исходной информации (например, через дифференциальную приватность);

- объем данных и параметры модели: описываются объем данных и ключевые параметры модели.

Основная информация, представленная в отчете, должна включать сведения о валидации и оценке качества синтетических данных, значение полученных метрик:

- оценка точности: включает статистическую близость к исходным данным, диверсификацию и реалистичность;

- оценка полезности: определяет пригодность синтетических данных для различных задач;

- оценка уровня приватности: вычисляет значения вероятности утечки исходной информации;

- оценка смещенности и объяснимости: проверяет отсутствие предвзятости и возможность выделить влияние ключевых признаков (опционально);

- пороговые значения: указываются установленные пороговые значения для каждой метрики.

Шаблон отчета о качестве синтетических данных представлен в приложении Е и содержит полную информацию по валидации и оценке качества, ссылку на источник происхождения, сопровождающую метаинформацию.

Использование синтетических данных охватывает как процессы их первоначального синтеза, так и все последующие генерации, включая процесс публикации данных и выполнение запросов на такие данные со стороны контрагентов. Все этапы использования синтетических данных должны быть задокументированы (протоколированы) в целях обеспечения прозрачности, прослеживаемости и соответствия установленным требованиям. Указанная информация подлежит включению в состав документации по качеству. Протокол использования синтетических данных,

формируемый вручную либо автоматически в составе системы журналирования, должен содержать информацию, обеспечивающую прозрачность и контроль на всех этапах обращения с данными:

- название и идентификатор синтетического набора (включая его контрольную сумму или цифровой отпечаток);

- дату использования и версию набора;

- цель использования и описание процесса доступа к данным (такие как адреса интерфейсов программирования приложений, информацию реализации каналов передачи информации);

- параметры доступа (такие как токены доступа);

- результаты использования (полученные результаты, оценка качества использования, метрики и параметры модели, использующей синтетические данные);

- инциденты и отклонения в данных, ссылки на сопутствующую документацию.

Отчет о качестве данных подлежит регулярному обновлению для обеспечения его актуальности и соответствия текущему состоянию процессов синтеза. Обновление отчета должно осуществляться не реже одного раза в год, а также в случаях, когда в процессе синтеза происходят существенные изменения, включая модификацию моделей, параметров генерации, методов обеспечения приватности PET или архитектурных компонентов системы.

Приложение А
(рекомендуемое)
Словарь данных

В настоящем приложении приводится рекомендуемый процесс ведения словаря данных для работы с исходными данными.

А.1 Общее описание словаря данных

В контексте синтеза данных словарь данных играет ключевую роль, обеспечивая понимание характеристик и свойств каждого атрибута, что необходимо для правильной подготовки, нормализации и профилирования данных.

Словарь данных включает в себя информацию о типах атрибутов, их исходном распределении, методах нормализации, очистки и кодирования, а также о том, как они классифицируются в рамках оценки приватности — например, как прямые идентификаторы, квазиидентификаторы или чувствительные атрибуты. Такое структурированное представление данных позволяет оценивать качество синтетических данных и управлять им на всех этапах обработки.

В рамках подготовки данных в состав словаря данных включается следующая информация о каждом атрибуте:

- атрибут: название атрибута, которое используется в наборе данных;

- тип данных: формат данных (например, целое число, вещественное число, строка);

- класс атрибута: классификация атрибута по его роли в определении уровня приватности данных: прямой идентификатор, квазиидентификатор, чувствительный атрибут;

- исходное распределение: характеристика распределения данных до обработки (например, нормальное, логнормальное);

- метод нормализации: метод, используемый для приведения значений атрибута к сопоставимому масштабу (например, масштабирование значений в диапазон от 0 до 1 или стандартизация относительно среднего и отклонения);

- метод очистки: метод, использованный для очистки данных (например, заполнение медианой, удаление выбросов);

- метод кодирования: метод преобразования категориальных данных в числовую форму, пригодную для обработки моделью (например, замена категорий наборами бинарных признаков или числовыми метками);

- пропуски: процент или количество отсутствующих значений в данных;

- аномалии: процент или количество аномалий, выявленных в данных;

- значение после обработки: результат обработки данных, например нормализованные или закодированные значения;

- комментарий: дополнительные замечания, которые могут помочь понять процесс обработки данных или особенности атрибута.

Возможные классы атрибутов:

- прямой идентификатор: атрибут, который непосредственно идентифицирует субъект;

- квазиидентификатор: атрибут, который может использоваться для идентификации субъекта в комбинации с другими атрибутами;

- чувствительный атрибут: атрибут, раскрытие которого является нежелательным.

А.2 Пример словаря данных

В таблице А.1 приведен пример словаря данных, который может быть настроен в зависимости от процессов организации.

Таблица А.1 — Пример словаря данных для обеспечения качества процесса

Атрибут	Тип данных	Класс атрибута	Исходное распределение	Метод нормализации	Метод очистки	Метод кодирования	Пропуски	Аномалии	Значение после обработки	Комментарий
ID	Целое число	Прямой идентификатор	Уникальное, равномерное	Нет	Нет	Нет	Нет	Нет	Уникальное, без изменений	Идентификатор записи
Возраст	Целое число	Квазиидентификатор	Нормальное	Min-Max	Заполнение медианой	Нет	5 %	Нет	Нормализованное значение в диапазоне [0, 1]	Возраст в годах
Пол	Категориальный	Квазиидентификатор	50 % мужской, 50 % женский	Нет	Категориальная чистка	Двоичное кодирование признаков	Нет	Нет	One-Hot-кодирование (0,1)	Мужской/женский
Доход	Вещественное число	Чувствительный атрибут	Логнормальное	Логарифмическое преобразование	Заполнение средним	Нет	2 %	Значение после логарифмического преобразования	Доход в годовых показателях	Нет
Местоположение	Категориальный	Квази-идентификатор	Группированное по регионам	Нет	Категориальная чистка	Кодирование меток данных	10%	Нет	Закодированное числовое значение	Регион проживания

Окончание таблицы А.1

Атрибут	Тип данных	Класс атрибута	Исходное распределение	Метод нормализации	Метод очистки	Метод кодирования	Пропуски	Аномалии	Значение после обработки	Комментарий
Дата регистрации	Дата	Квазииденти-фикатор	Равномерное	Стандартизация по временному отрезку	Заполнение значением по умолчанию	Векторизация	Нет	1%	Векторное представление даты	Дата первого визита
Транзакции	Целое число	Чувствительный атрибут	Пиковое распределение	Приведение значений к нулевому среднему и единичному стандартному отклонению.	Удаление выбросов	Нет	3%	Нет	Нормализованное значение по z-оценке	Количество транзакций
Кредитный рейтинг	Вещественное число	Чувствительный атрибут	Случайное	Стандартизация по z-оценке	Заполнение медианой	Нет	8%	Нет	Нормализованное значение по z-оценке	Оценка кредитоспособности
Состояние аккаунта	Категориальный	Квази-идентификатор	Преобладающее (активный/неактивный)		Категориальная чистка	Бинарное кодирование	Нет	Нет	кодирование (0,1)	Активный/Неактивный
Координаты	Геопространственный (широта, долгота)	Квази-идентификатор	Кластерное	Геохеширование	Удаление не валидных координат	Геохэш	1%	0,5%	Геохэш с пониженной точностью	Местоположение с точностью до района

Приложение Б

(справочное)

Точность. Метрики оценки статистической близости

В таблице Б.1 приведены рекомендуемые метрики оценки статистической близости, используемые в рамках критерия точности синтетических данных.

Таблица Б.1 — Метрики оценки статистической близости

Метрика	Описание	Расчет	Оценки
Тест Колмогорова-Смирнова (KS-statistic)	Оценивает различие между распределениями синтетических и реальных данных. Одномерные числовые данные, преимущественно структурированные данные	$D_{KS} = \sup_x \lvert F_{n1}(x) - F_{n2}(x) \rvert$, где $F_{n1}(x)$ — кумулятивное распределение синтетических данных; $F_{n2}(x)$ — кумулятивное распределение реальных данных. Требует построения кумулятивных функций и нахождения их максимального различия	Диапазон значений: [0,1], где 0 — полное совпадение распределений (нет различий); 1 — максимальное различие между распределениями (полное несовпадение). Рекомендуемый порог: $D_{KS} \leq 0{,}1$ для признания синтетических данных достаточно близкими
Расстояние Вассерштейна (Wasserstein Distance)	Измеряет расстояние между распределениями синтетических и реальных данных (в смысле стоимости преобразования одного распределения в другое). Как для одномерных, так и многомерных данных, подходит для медиаресурсов.	$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x,y)^p d\gamma(x,y) \right)^{1/p}$, где $\Gamma(\mu,\nu)$ – совокупность всех мер с маргинальными распределениями параметров μ, ν (M,d) – метрическое пространство мер Радона (локально конечных и регулярных); γ – конкретная вероятностная мера из множества $\Gamma(\mu,\nu)$, по которой берётся инфимум Заключается в оценке минимальной транспортной стоимости для преобразования одного распределения в другое	Диапазон значений: [0, ∞). Рекомендуемый порог: низкие значения

Продолжение таблицы Б.1

Метрика	Описание	Расчет	Оценки
Корреляция Пирсона (Pearson Correlation)	Оценивает линейную корреляцию между синтетическими и реальными данными	$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$. Учитывает расчет ковариации и стандартных отклонений двух переменных	Диапазон значений: [-1, 1]. Рекомендуемый порог: > 0,7 между *P* и *M*, между *Q* и *M* соответственно
Логарифмически-преобразованный корреляционный индекс (Log-Transformed Correlation Score)	Измеряет корреляцию между логарифмически преобразованными данными. Числовые данные с широким диапазоном значений или лог-нормальные распределения (текст при векторизации)	$r_{log} = \frac{\sum(\log x_i - \log \bar{x})(\log y_i - \log \bar{y})}{\sqrt{\sum(\log x_i - \log \bar{x})^2}\sqrt{\sum(\log y_i - \log \bar{y})}}$. Включает логарифмическое преобразование данных и расчет корреляции между преобразованными значениями	Диапазон значений: [-1, 1]. Рекомендуемый порог: > 0,7
Дивергенция Кульбака-Лейблера (KL Divergence)	Измеряет различие (относительную энтропию) между распределениями синтетических и реальных данных. Может применяться к любым данным, представляемым как вероятностные распределения (широко используется для текстовых данных, например при оценке сходства распределений слов и изображений)	$D_{KL}(P\|\|Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$, где $P(x)$ — вероятность события x в распределении P; $Q(x)$ — вероятность события x в распределении Q; X — множество возможных событий. Требует расчет вероятностных распределений и их относительное различие	Диапазон значений: [0, ∞), где 0 — полное совпадение распределений *P* и *Q*. ∞ — максимальное различие, где распределение *Q* не перекрывает *P*. Рекомендуемый порог: для хорошего соответствия синтетических данных реальным значение KL-дивергенции должно быть как можно ближе к 0. Значения выше 0,1 могут указывать на значительное различие между распределениями

Продолжение таблицы Б.1

Метрика	Описание	Расчет	Оценки
Дивергенция Йенсена (Jensen-Shannon Divergence, JSD)	Измеряет различие между распределениями синтетических и реальных данных. Подходит для текстовых данных и изображений, особенно когда необходимо симметричное и сглаженное измерение различий между распределениями (например, распределений слов или признаков)	$D_{JS}(P\|\|Q) = \frac{1}{2}D_{KL}(P\|\|M) + \frac{1}{2}D_{KL}(Q\|\|M)$, где $D_{KL}(P \parallel M)$ и $D_{KL}(Q \parallel M)$ — *KL*-дивергенции между P и M, между Q и M соответственно; $M = \frac{1}{2}(P + Q)$ — среднее распределение между P и Q. Представляет собой симметричную и сглаженную версию *KL*-дивергенции, что делает ее более стабильной и интерпретируемой	Диапазон значений: [0;$\ln 2$], где 0 — полное совпадение *P* и *Q* $\ln 2 \approx 0{,}693$ — максимальное различие, где распределения полностью различны. Рекомендуемый порог: < 0,1 (высокая степень схожести); 0,1–0,2 допустимо для некоторых приложений
Многомерное расстояние Хеллинджера (Multivariate Hellinger Distance)	Измеряет различие между распределениями синтетических и реальных данных в многомерном пространстве. Может применяться к текстовым данным (например, для оценки схожести распределений тем) и изображениям, представляемым в виде многомерных признаков	$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum(\sqrt{p_i} - \sqrt{q_i})^2}$. Расчет различий между многомерными вероятностными распределениями	Диапазон значений: [0, 1]. Рекомендуемый порог: < 0,2
Расстояние Бхаттачарьи (Bhattacharyya Distance)	Измеряет перекрытие между двумя вероятностными распределениями, интерпретируется как мера сходства (в том числе текст и изображения)	$D_B(P,Q) = -\ln\left(\sum\sqrt{P(x)Q(x)}\right)$. Расчет вероятностей и их среднегеометрическое значение	Диапазон значений: [0, ∞). Рекомендуемый порог: низкие значения

Продолжение таблицы Б.1

Метрика	Описание	Расчет	Оценки
Статистический тест Стьюдента (Student's T-test for the comparison of means)	Оценивает, являются ли средние значения двух групп (реальных и синтетических данных) статистически значимыми. Применим для числовых данных	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, где $\bar{x}_1$,$\bar{x}_2$ — средние значения групп; s_1^2, s_2^2 — дисперсии групп; n_1,n_2 — размеры выборок	Диапазон: зависит от степени свободы. Рекомендуемый порог: p-value < 0,05 — вероятность получения наблюдаемого результата
U-тест Манна-Уитни (Mann-Whitney U-Test)	Непараметрический тест, который оценивает различие между двумя выборками. Применим для числовых данных, когда распределения не нормальны	Вычисляется путем ранжирования объединенной выборки и оценки различия между рангами	Диапазон: зависит от размера выборки. Рекомендуемый порог: p-value < 0,05
Критерий хи-квадрат (Chi-squared Test)	Оценивает, есть ли значительное различие между ожидаемыми и наблюдаемыми частотами категориальных данных	$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, где O_i — наблюдаемые частоты (в реальных данных); E_i — ожидаемые частоты (частоты появления в синтетических данных)	Диапазон: зависит от размера выборки. Рекомендуемый порог: p-value < 0,05
Максимальное среднее отклонение (Maximum Mean Discrepancy, MMD)	Оценивает различие между двумя выборками путем сравнения их средних значений в многомерных пространствах	Использует функции ядра (функции нелинейного разделения данных, используемые в машинном обучении) для оценки различий. В качестве функций ядра могут быть применены функции вида: $K(x, y) = x \cdot y$ – линейное ядро или $K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ – гауссово (RBF) ядро	Зависит от задачи

Окончание таблицы Б.1

Метрика	Описание	Расчет	Оценки
Сходство на основе гистограммы (Histogram-based similarity)	Измеряет сходство между распределениями синтетических и реальных данных путем сравнения их гистограмм. Подходит для числовых и категориальных данных	Вычисляется как разница между гистограммами, например, по среднему абсолютному отклонению	Диапазон: [0, 1]. Рекомендуемый порог: стремится к 1

Примечания

1 Все указанные оценки для рекомендуемых порогов носят эмпирический характер, приводят приблизительные значения на основании экспертных мнений и могут уточняться для конкретных задач.

2 В дополнение к метрике сходства на основе гистограммы для более детального анализа соответствия распределений синтетических данных реальным могут использоваться метрики усредненных ошибок, такие как усредненные ошибки медианы, стандартного отклонения и среднего. Эти метрики рекомендуется применять в случаях, когда требуется детализированная оценка отклонений в статистических характеристиках данных.

3 Для обеспечения статистической схожести и сравнимости синтетических данных с реальными рекомендуется выбирать метрики в зависимости от типа данных. Для непрерывных признаков приоритетными являются тест Колмогорова-Смирнова, расстояние Вассерштейна и корреляция Пирсона. Для категориальных данных следует использовать критерий Хи-квадрат (c^2) и U-тест Манна-Уитни. В случаях с многомерными распределениями рекомендуется применять многомерное расстояние Хеллинджера, расстояние Бхаттачарьи и максимальное среднее отклонение.

Приложение В
(справочное)
Точность. Метрики оценки реалистичности синтетических данных

В таблице В.1 приводятся рекомендуемые метрики оценки реалистичности синтетических данных.

Таблица В.1 — Рекомендуемые метрики оценки реалистичности синтетических данных

Метрика	Описание	Расчет	Оценки
Расстояние Фреше между инцепциями (Frechet Inception Distance, FID)	Оценивает визуальное сходство между реальными и синтетическими изображениями	$FID = \|\mu_r - \mu_s\|^2 + Tr(\Sigma_r + \Sigma_s - 2(\Sigma_r\Sigma_s)^{1/2}$, где μ_r, μ_s — средние значения признаков реальных и синтетических данных соответственно; Σ_r,Σ_s — ковариационные матрицы признаков реальных и синтетических данных; Tr — след матрицы, сумма элементов на главной диагонали	Диапазон: [0, ∞). Рекомендуемый порог: < 10
Потери восприятия (Perceptual Loss, Perceptual Similarity)	Оценивает визуальное сходство на основе признаков, извлеченных из промежуточных слоев нейронной сети. Визуально схожие изображения будут иметь малые значения потерь восприятия	Вычисляется путем сравнения промежуточных выходных данных нейронной сети (например, Inception-v3). Разница между признаками реальных и синтетических данных оценивается по $L2$-норме (Евклидово расстояние)	Диапазон: [0, ∞). Рекомендуемый порог: стремится к 0
Индекс структурного сходства (Structural Similarity Index, SSIM)	Оценивает структурное сходство между двумя изображениями, учитывая яркость, контраст и текстуру. Важен для задач, где требуется сохранить восприятие изображения человеком	$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\delta_{xy} + C_2)}{(\mu_x{}^2 + \mu_y{}^2 + C_1)(\sigma_x{}^2 + \sigma_y{}^2 + C_2)}$, где σ_x^2,σ_y^2 — дисперсии яркости для двух изображений; σ_{xy} — ковариация яркости между изображениями; C_1, C_2 — малые константы для стабилизации деления при слабом контрасте	Диапазон: [-1, 1]. Рекомендуемый порог: > 0,8 для высококачественных изображений

Продолжение таблицы В.1

Метрика	Описание	Расчет	Оценки
Корреляция признаков	Оценивает, насколько хорошо сохраняются корреляции между признаками в синтетических данных по сравнению с реальными. Полезна для задач с многомерными данными	Рассчитывается путем построения корреляционных матриц для реальных и синтетических данных и их последующего сравнения (например, через корреляцию Пирсона)	Диапазон: [-1, 1] Рекомендуемый порог:> 0,8 для близкого сходства
Анализ корреляционных матриц	Оценивает соответствие всех корреляций между признаками в синтетических данных по сравнению с реальными. Позволяет увидеть, насколько синтетические данные сохраняют взаимозависимости	Формируется матрица корреляций для всех признаков, затем сравниваются корреляционные матрицы синтетических и реальных данных с использованием нормы Фробениуса: $\|A\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|a_{ij}\right\|^2}$, Применяется к разности корреляционных матриц реальных и синтетических данных PCD- (Pairwise Correlation Distance, CorrDistance) тест: $PCD\ (X_R, X_S) = \|Corr(X_R) - Corr(X_S)\|_F$	Диапазон: [0, ∞). Рекомендуемый порог: > 0,1 для близкого сходства
Согласованность	Оценивает, насколько синтетические данные сохраняют логические и структурные зависимости между признаками, например, во временных рядах или текстах	Оценивается путем анализа взаимосвязей между данными в синтетическом наборе и сравнения их с реальными зависимостями. Может использоваться логическое или статистическое моделирование	Диапазон: зависит от задачи. Рекомендуемый порог: максимально близкое соответствие реальным данным

Продолжение таблицы В.1

Метрика	Описание	Расчет	Оценки
Log-кластер метрика	Оценивает схожесть скрытых структур через кластеризацию данных. Проверяет, насколько синтетические данные следуют тем же кластерам, что и реальные данные	$U_c(X_R, X_C) = \ln\left(\frac{1}{G}\sum_{j=1}^{G}\left(\frac{n_j^R}{n_j} - c\right)^2\right)$, где G — количество кластеров; n_j^R — количество точек в кластере j для реальных данных; n_j — общее количество точек в кластере j (для объединенного набора данных); c — некоторая константа или центр кластеризации	Диапазон: [0, ∞). Рекомендуемый порог: 0,1 для близкого соответствия кластерных структур
Анализ важности признаков (Feature Importance Analysis)	Оценивает, насколько синтетические данные сохраняют важность признаков, которая используется моделью машинного обучения	Определяется через различие значений важности признаков между моделями, обученными на реальных и синтетических данных. Обычно используются такие алгоритмы, как случайный лес или градиентный бустинг	Чем меньше разница, тем лучше синтетические данные сохраняют важность признаков

Продолжение таблицы В.1

Метрика	Описание	Расчет	Оценки
Анализ репрезентативности (Population Stability Index, PSI)	Метрика измеряет отклонения в распределении синтетических данных по сравнению с реальными, особенно полезна для обнаружения смещений (drift). PSI измеряет отклонение распределений данных по каждому признаку между двумя наборами (синтетическими и реальными)	$PSI = \sum_{i=1}^{n} (p_i - q_i) * \ln\left(\frac{p_i}{q_i}\right)$, где n — количество бинов (групп) или интервалов, на которые разделены данные для оценки распределений; p_i — доля наблюдений в синтетических данных, попадающих в i-й бин; q_i — доля наблюдений в реальных данных, попадающих в i-й бин; $\ln\left(\frac{p_i}{q_i}\right)$ — натуральный логарифм отношения долей синтетических и реальных данных в i-м бине	Диапазон: [0, ∞). Рекомендуемый порог: < 0,1 для близкого соответствия распределений
Инвертированный коэффициент силуэта (Inverted Silhouette Score)	Метрика используется для оценки качества кластеризации данных, определяя, насколько хорошо кластеры разделены и однородны	Рассчитывается как инвертированный коэффициент силуэта $S_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$, где $S(i)$ — значение коэффициента силуэта для точки i; $a(i)$ — среднее расстояние от точки i до всех других точек в том же кластере; $b(i)$ — среднее расстояние от точки i до точек в ближайшем кластере (т. е. минимальное среднее расстояние до точек в других кластерах). $ISS = 1 - \frac{1}{N}\sum_{i=1}^{N} S(i)$, где *ISS* — инвертированный коэффициент силуэта (inverted silhouette score); *N* — количество точек в наборе данных	Диапазон: [0, 2]. Рекомендуемый порог: < 0,5 для близкого соответствия кластеризаций

Окончание таблицы В.1

Метрика	Описание	Расчет	Оценки
Анализ доли наблюдений с высокими значениями расстояния Махаланобиса (Mahalanobis Distance Analysis)	Метрика оценивает, насколько хорошо синтетические данные воспроизводят многомерные корреляционные структуры	Определяется путем расчета расстояния Махаланобиса для каждой точки в синтетических данных относительно реальных данных. $D_M(x) = \sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$, где x — вектор признаков точки; μ — вектор средних значений признаков для набора данных; Σ — ковариационная матрица набора данных; Σ^{-1} — обратная ковариационная матрица	Диапазон: [0, ∞). Рекомендуемый порог: < 1 для близкого соответствия распределений
Примечание — Для обеспечения сравнимости и качества синтетических данных рекомендуется выбирать метрики в зависимости от типа данных и задач. Визуальные метрики, такие как FID, Perceptual Loss и SSIM, приоритетны для изображений, особенно в задачах, связанных с генеративными моделями. Для анализа структурных зависимостей и кластеризации данных предпочтительны такие метрики, как Feature Correlation, Correlation Matrix Analysis, Log-cluster Metric и Inverted Silhouette Score, которые оценивают сохранение корреляций и кластерных структур. В задачах машинного обучения, где важно сохранить значимость признаков и репрезентативность данных, рекомендуется использовать Feature Importance Analysis и PSI. Эти метрики помогут достичь высокой точности и реалистичности синтетических данных в различных сценариях применения.			

Приложение Г
(справочное)
Полезность. Метрики оценки эффективности

В данном приложении приводятся рекомендуемые метрики оценки полезности. Набор метрик полезности зависит от целевой задачи и уточняется на этапе планирования процесса генерации синтетических данных. В таблице Г.1 приводятся рекомендуемые метрики полезности.

Таблица Г.1 — Рекомендуемые метрики оценки полезности

Метрика	Описание	Расчет	Оценки
Коэффициент детерминации R^2 (Coefficient of Determination)	Оценивает, насколько хорошо модель объясняет вариацию зависимой переменной на основе синтетических данных. Применение — Числовые данные, регрессия	$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$, где y_i — реальные значения; $\hat{y}_i$ — предсказанные значения; $\bar{y}_i$ — среднее значение	Диапазон: [0, 1]. Рекомендуемый порог: > 0,7
Среднеквадратичная ошибка (Mean Squared Error, MSE)	Измеряет среднее квадратичное отклонение между предсказанными и реальными значениями. Применение — Числовые данные, регрессия	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, где y_i — реальные значения; $\hat{y}_i$ — предсказанные значения.	Диапазон: [0, ∞). Рекомендуемый порог: зависит от задачи, стремится к 0
Средняя абсолютная ошибка (Mean Absolute Error, MAE)	Измеряет среднее абсолютное отклонение между предсказанными и фактическими значениями, предоставляя более интерпретируемую меру ошибки	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert$, где y_i — реальные значения; $\hat{y}_i$ — предсказанные значения модели	Диапазон: [0, ∞). Рекомендуемый порог: ≤ 10 %–20 % от диапазона значений целевой переменной; при нормализованных данных — MAE ≤ 0,1

Продолжение таблицы Г.1

Метрика	Описание	Расчет	Оценки
Точность моделирования (Accuracy)	Измеряет, насколько точно модель или алгоритм предсказывает правильные результаты на основе данных. Применение — Числовые и категориальные данные в задачах классификации	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, Требует значения истинных положительных (*TP*), истинных отрицательных (*TN*), ложных положительных (*FP*) и ложных отрицательных (*FN*) прогнозов	Диапазон: [0, 1]. Рекомендуемый порог: > 0,9
Корректность (Precision)	Измеряет долю правильных положительных предсказаний среди всех предсказанных положительных исходов. Применение — Важна для задач, где ложные срабатывания недопустимы	$Precision = \frac{TP}{TP + FP}$, где *TP* — истинно положительные предсказания (True Positives); *FP* — ложноположительные предсказания (False Positives)	Диапазон: [0, 1]. Рекомендуемый порог:> 0,7 для задач, где важна точность
F1-метрика	Гармоническое среднее между точностью и полнотой, оценивает баланс между ними. Применение — Классификация, бинарные и многоклассовые задачи	$F1 = 2 \cdot \frac{Precision - Recall}{Precision + Recall}$, где $Precision = \frac{TP}{TP+FP}$; $Recall = \frac{TP}{TP + FN}$.	Диапазон: [0, 1]. Рекомендуемый порог: > 0,7
Площадь под кривой ROC (AUC-ROC) (Area Under the ROC Curve)	Оценивает способность модели различать классы. Применение — Классификация, бинарные и многоклассовые задачи	Расчет площади под ROC-кривой, которая строится по координатам: (True Positive Rate, False Positive Rate)	Диапазон: [0, 1]. Рекомендуемый порог: > 0,8

Окончание таблицы Г.1

Метрика	Описание	Расчет	Оценки
Приватная среднеквадратичная ошибка (Private Mean Squared Error, pMSE)	Вариант MSE, учитывающий приватность данных, чтобы минимизировать утечки данных. Применение — Числовые данные, регрессия с учетом приватности	Аналогично MSE, но с добавлением шумов или иных механизмов для обеспечения приватности - PET: $pMSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i + Z)^2,$ где y_i — реальные значения, $\hat{y}_i$ — предсказанные значения, Z — добавляемый шум	Диапазон: [0, ∞). Рекомендуемый порог: зависит от задачи, стремится к 0
Расстояние Хаусдорфа (Hausdorff Distance)	Измеряет максимальное расстояние между двумя множествами точек, зачастую используется для оценки пространственных данных. Применение — Пространственные данные, изображения	$\mathrm{d}_H(A,B) = \max\left\{\sup_{a\in A} inf_{b\in B}\|a-b\|, \sup_{b\in B} inf_{a\in A}\|a-b\|\right\},$ где *A* и *B* – множества точек	Диапазон: [0, ∞). Рекомендуемый порог: зависит от задачи, стремится к 0

Примечания

1 Для оценки полезности синтетических данных рекомендуется выбирать метрики в зависимости от целей моделирования и типов задач. Метрики, такие как коэффициент детерминации R^2, MSE и MAE, приоритетны для задач регрессии, где важна точность предсказаний. Метрики точности моделирования (Accuracy), корректности (Precision) и F1-метрика рекомендованы для классификационных задач, особенно когда важно учитывать баланс между точностью и полнотой. Для задач, связанных с оценкой общей производительности моделей, AUC-ROC позволяет измерить способность модели различать классы. В сценариях, где требуется учитывать приватность, предпочтительно использование приватной MSE (pMSE). Расстояние Хаусдорфа полезно в задачах, где важно оценить максимальное отклонение между предсказанными и реальными значениями, например в задачах на основе изображений или геометрических данных.

2 Если синтетические данные создаются для конкретной задачи с заданной целевой переменной, метрики полезности следует рассчитывать исключительно для этой переменной, чтобы обеспечить точность и релевантность оценки. В случаях, когда целевая переменная не определена или данные предполагается использовать для различных целей, рекомендуется оценивать полезность по всем признакам. Это позволяет получить общую оценку пригодности данных для потенциальных задач и избежать искажения результатов.

Приложение Д

(справочное)

Метрики и модели оценки риска утечки данных

Д.1 Предварительная оценка уровня приватности

В качестве предварительной оценки уровня приватности можно использовать метрики, ориентированные на анализ индивидуальных записей и их схожести, для выявления потенциально уязвимых записей. К этим метрикам относятся метрики DCR и NNDR, описание которых приведено в таблице Д.1.

Таблица Д.1 — Рекомендуемые метрики на основе близости

Метрика	Описание	Расчет	Оценки
Минимальное расстояние до ближайшей записи (Distance to Closest Record, DCR)	Метрика измеряет расстояние от каждой синтетической записи до ближайшей реальной записи. Чем меньше это расстояние, тем выше вероятность, что синтетические данные будут реалистичными	$DCR_i = \min_j\left(dist(x_i^s, x_j^r)\right)$, где x_i^s — *i*-я синтетическая запись; x_j^r — ближайший сосед (первая ближайшая реальная запись); $dist$ — выбранная метрика расстояния (например, евклидово расстояние, манхэттенское расстояние и т. д.)	Рассчитать равенство распределений DCR между реальными данными (отложенными, зарезервированными за пределами обучения) и синтетическими, между реальными данными (выборка для обучения генератора) и синтетическими (см. также ГОСТ Р 59900)

Окончание таблицы Д.1

Метрика	Описание	Расчет	Оценки
NNDR (Nearest Neighbor Distance Ratio)	Метрика измеряет отношение расстояния до ближайшего соседа к расстоянию до второго ближайшего соседа. Помогает определить, насколько точно синтетические данные воспроизводят структуру реальных данных	$NNDR_i = \frac{dist(x_i^s, x_j^r)}{dist(x_i^s, x_k^r)}$, где x_i^s — *i*-я синтетическая запись; x_j^r — ближайший сосед (первая ближайшая реальная запись); x_k^r — второй ближайший сосед (первая ближайшая реальная запись); $dist$ — выбранная метрика расстояния	Реалистичность: значения ближе к 1. Кластеризация: значения ближе к 0

Примечания

1 Расчет расстояний осуществляется с привлечением метрик Евклидова расстояния, косинусного расстояния или аналогичных.

2 Для расчета расстояний рекомендуется использовать квази-идентификаторы и тестовый набор данных, отличный от того, который использовался для обучения.

3 Использование метрик DCR и NNDR обычно дополняется более комплексными методами оценки на основе моделей оценки утечки данных для обеспечения всестороннего анализа приватности данных.

Д.2 Модель дифференциальной приватности

Дифференциальная приватность — это математический подход к обеспечению приватности данных, который гарантирует, что выход алгоритма незначительно изменяется при добавлении или удалении одной записи из набора данных. Основная идея механизма дифференциальной приватности состоит в том, что при анализе или обработке данных выходной результат не должен существенно зависеть от наличия или отсутствия любого отдельного элемента, тем самым снижая вероятность утечки данных о любом конкретном субъекте. В ПНСТ 1065–2026 приводятся основные сведения о реализации дифференциальной приватности.

Для алгоритмов с дифференциальной приватностью, выполняющих запросы на наборе данных, гарантируется, что для всех возможных наборов данных D_1 и D_2, различающихся лишь одной записью, и для всех множеств выходных данных S выполняется неравенство

$$Pr[\mathcal{M}(D_1) \in S] \leq e^{\epsilon} \cdot Pr[\mathcal{M}(D_2) \in S] + \delta, \qquad (Д.1)$$

где $Pr[\mathcal{M}(D_1) \in S]$ — вероятность того, что алгоритм $\mathcal{M}$ при работе с набором данных D_1 выдаст результат, принадлежащий множеству S;

$Pr[\mathcal{M}(D_2) \in S]$ — вероятность того, что алгоритм $\mathcal{M}$ при работе с набором данных D_2 выдаст результат, принадлежащий множеству S;

ϵ — параметр приватности, контролирующий степень изменений результатов при изменении одной записи в наборе данных;

δ — параметр, представляющий вероятность того, что ϵ — гарантия не будет выполнена.

Параметры (ϵ, δ) образуют бюджет приватности: количественную меру уровня приватности, определяемую параметрами ϵ и δ. Параметр ϵ измеряет степень, на которую результаты алгоритма могут изменяться при изменении одной записи в наборе данных, а δ указывает на вероятность того, что гарантии приватности могут быть нарушены. Чем меньше значения ϵ и δ, тем выше уровень приватности, обеспечиваемый алгоритмом. Бюджет приватности связан как с объемом синтезируемых данных, так и с возможностью многократного использования генератора для синтеза данных, поскольку такой бюджет продолжает расходоваться после обучения модели.

Контроль бюджета приватности осуществляется через механизм управления, который следит за накопленным расходом ϵ и δ при выполнении множественных запросов, чтобы минимизировать вероятность утечки данных.

Через данные параметры можно формально определить вероятность утечки данных P_{data} (вероятность того, что изменение в одной записи вызовет изменение в выходном значении алгоритма).

$$P_{data} \approx 1 - e^{\epsilon}. \qquad (Д.2)$$

Таблица Д.2 — Метрики дифференциальной приватности

Метрика	Определение	Расчет	Оценки
ϵ (эпсилон)	Мера максимального влияния на результат алгоритма, вызванного изменением одной записи в наборе данных. Чем меньше значение ϵ, тем выше уровень приватности	Вычисляется на основе модели алгоритма и данных; определяется заданным бюджетом приватности	$0 < \epsilon < \infty$ Высокая приватность: $\epsilon \leq 0{,}1$. Средняя приватность: $0{,}1 < \epsilon \leq 1$. Низкая приватность: $\epsilon > 1$
δ (дельта)	Вероятность того, что ϵ-гарантия не будет выполнена. Чем меньше δ, тем выше уровень приватности, тем меньше вероятность утечки данных	Определяется заданным бюджетом приватности; может быть близким к 0	$0 \leq \delta \leq 1$ Высокая приватность: $\delta \leq 10^{-5}$. Средняя приватность: $10^{-5} < \delta \leq 10^{-3}$. Низкая приватность: $\delta > 10^{-3}$

Примечания

1 Метрики ϵ и δ являются ключевыми параметрами в обеспечении уровня приватности при обработке данных. Они помогают настроить баланс между приватностью и полезностью данных в зависимости от конкретных задач и сценариев использования.

2 Параметры ϵ и δ корерктируются в зависимости от объема данных. Чем больше записей в наборе данных, тем меньшим должен быть допустимый бюджет приватности ϵ, чтобы поддерживать тот же уровень приватности. В таблице Д.2 низкий уровень приватности применим к наборам данных с объемом от несколько сотен до тысяч записей, средний уровень – к наборам до десятков тысяч записей, высокий – для крупных наборов в сотни тысяч и более

записей. Важно учитывать это при реализации алгоритмов с дифференциальной приватностью. С учетом зависимости от объема данных формула (Д.2) может быть скорректирована:

$$P \approx 1 - e^{-\epsilon(n)} = 1 - e^{-\frac{\epsilon_0}{\sqrt{n}}} \ , \qquad (Д.3)$$

где *n* — количество записей в наборе данных;

ϵ_0 — базовый уровень є при определенном объеме данных.

Примерные диапазоны значений ϵ_0:

- очень строгая приватность: ϵ_0 от 0,01 до 0,1 — используется в случаях, где критически важно учитывать чувствительность данных (например, медицинские данные, ПДн);

- умеренная приватность: ϵ_0 от 0,1 до 1 — подходит для большинства корпоративных и исследовательских данных, где балансируются приватность и точность;

- слабая приватность: ϵ_0 от 1 до 10 — используется, когда точность данных более важна, чем их уровень приватности, например для некоторых бизнес-аналитических задач, а также при использовании нечувствительных исходных данных.

3 Параметры ϵ и δ предоставляют теоретические гарантии приватности, но их практическая реализация требует тщательного управления бюджетом приватности, что делает необходимым использование надежных механизмов контроля, таких как «приватный бухгалтер». Приводимые оценки носят рекомендательный характер.

Д.3 Комплексная модель синтетических данных «Утечка информации»

Приложение приводит основные алгоритмы оценки вероятности успешности популярных атак.

Вероятность утечки данных может быть оценена с помощью событийной модели (см. [9]) – совокупный уровень приватности синтетических данных является взвешенной суммой вероятности успеха допускаемых атак (см. ГОСТ Р 58771). Общую вероятность потери приватности синтетическими данными Pr рассчитывают по формуле

$$Pr = \sum_{i=1}^{n} wa_i \cdot Pa_i \ , \qquad (Д.4)$$

где Pa_i – вероятность успеха по одному из векторов атак;

wa_i – вес соответствующего вектора атак.

ПНСТ

Распространенная модель оценки уровня приватности синтетических данных (комплексная модель утечки данных) сводит общую формулу к сумме вероятностей атак выделения P_{singl}, вывода P_{inf} и связывания P_{link}:

$$Pr = w_1 \cdot P_{singl} + w_2 \cdot P_{inf} + w_3 \cdot P_{link} . \quad (Д. 5)$$

Выбор весов w_i может быть осуществлен исходя из анализа архитектуры системы синтеза, важности атак и условия нормирования:

$$\sum_{i=1}^{n} w_i = 1 . \quad (Д. 6)$$

Оценка вероятности отдельных атак на синтетические данные может быть выполнена через так называемую оценку сверху, или максимальный показатель вероятности утечки данных. В данном подходе предполагается, что вероятность успешного проведения атаки максимальна при наихудшем сценарии. Например, если злоумышленник обладает полной информацией о синтетических данных и использует все доступные методы для извлечения информации, вероятность атаки оценивается как максимальная. Этот подход полезен для создания «наихудшего» сценария и разработки доверенных технологий синтеза, рассчитанных на максимальные вероятности негативных событий. Примером такой оценки является оценка вероятности утечки данных на основе распределения данных и их уникальности.

Оценку вероятности P_k k-й угрозы синтетических данных с объемом n-записей и распределением F_s рассчитывают по формуле

$$P_k = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{F_s} . \quad (Д. 7)$$

Другим подходом является использование доверительного интервала Вильсона (см. ГОСТ Р ИСО 16269-7) для более точной оценки вероятности атаки. Доверительный интервал Вильсона используется для оценки наблюдаемой вероятности успеха события на основе наблюдаемых данных $\hat{p}$ (например, искомая оценка вероятности атаки выделения) и рассчитывается по формуле

$$\hat{p} = \frac{\hat{p} + \frac{z^2}{2n}}{1 + \frac{z^2}{n}} \pm \frac{z}{1 + \frac{z^2}{n}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}, \quad (Д.8)$$

где n — общее количество наблюдений (например, количество строк в наборе данных);

z — критическое значение для желаемого уровня доверия (например, 1,96 для 95 %-го доверительного интервала).

Атака на выделение предполагает идентификацию отдельного индивида в синтетических данных. Если уникальные характеристики или комбинации атрибутов, такие как возраст, пол и почтовый индекс, сохраняются в синтетических данных, злоумышленник может выделить эти данные и привязать их к конкретному лицу. Например, если в исходных данных только один человек в определенном возрастном диапазоне и с определенным почтовым индексом, а синтетические данные повторяют эти характеристики, данная комбинация может позволить злоумышленнику идентифицировать этого человека в синтетических данных.

Атрибуты, позволяющие совместно выделить (идентифицировать) субъекта данных, называются квазиидентификаторами. Получив через них идентифицируемую персону, потенциальный злоумышленник может узнать дополнительные сведения, хранящиеся в чувствительных атрибутах, таких как уровень зарплаты, состояние здоровья и т. п. В условиях синтеза данных полностью синтетические данные поддерживаются одной или несколькими доверенными технологиями, описанными в ПНСТ 1065–2026, например обезличиванием, фильтрацией, добавлением шума (искажением) с использованием дифференциальной приватности.

В рамках доверенных технологий может использоваться строгая ϵ-дифференциальная приватность, которая может быть в ряде случаев упрощена до (ϵ, δ)-дифференциальной приватности (приближенной дифференциальной приватности): $(\epsilon, \delta) - \mathrm{DP}$, гарантирует, что изменение одного элемента в наборе данных несущественно изменяет результат запроса, тем самым уменьшая вероятность выделения индивидуальных данных. Значение ϵ контролирует уровень приватности (бюджет приватности); чем меньше ϵ, тем выше уровень приватности, но возможно ухудшение полезности данных. Значение δ характеризует вероятность того, что приватность может быть нарушена:

$$P_{sngl} \approx \delta = 1 - e^{-\epsilon}. \quad (Д.9)$$

Например, для дифференциальной приватности с параметром $\epsilon = 0{,}1$ теоретическая вероятность выделения записи $P_{sngl} \approx 0{,}0951$. Если не применяется дифференциальная приватность, количество выделяемых записей определяется теорией *k*-анонимности. В рамках k-анонимности уровень доверия к технологии обеспечивается тем, что каждое сочетание квазиидентификаторов встречается не менее k раз и вероятность может быть рассчитана по формуле доверительного интервала Вильсона (Д.8).

Примечание — Синтетические данные могут неточно отражать реальные распределения, что может повлиять на измерения *k*-анонимности. Если синтетические данные слишком однородны или им не хватает разнообразия, *k*-анонимность может быть неэффективной метрикой. Расчет *k*-анонимности может быть выполнен только при отсутствии гарантий приватности с помощью методов на уровне модели, таких как дифференциальная приватность.

Д.4 Атака на связывание и атака на членство

Атака на связывание предполагает соединение синтетических данных с внешними источниками для идентификации индивидов. Злоумышленники могут использовать квазиидентификаторы, такие как возраст, пол, почтовый индекс для связывания синтетических данных с реальными. При наличии совпадений между синтетическими и реальными данными идентификация становится возможной. Проверка возможности связывания синтетических данных с исходным набором данных, на основании которого и были построены синтетические данные, позволяет оценить успешность атаки связывания как ограничение сверху, максимальную вероятность – $\sup P_{link}$. Успешность такой атаки может свидетельствовать о недостаточном уровне приватности при генерации синтетических данных. В случаях когда синтетические данные очень похожи на реальные, вероятность успешной атаки на связывание может приближаться к вероятности атаки на членство, поскольку обе атаки направлены на выявление связи между синтетическими и исходными данными.

В условиях дифференциальной приватности вероятность атаки на принадлежность к множеству ограничена значением параметра ϵ. По мере уменьшения ϵ снижается вероятность возможного определения принадлежности, а также уменьшается

разница между вероятностями атак на связывание и атакой на членство. Такую зависимость можно объяснить тем, что более низкое значение ϵ обеспечивает больший уровень приватности, снижая вероятность успешного связывания записей. В общем виде вероятность успешности атаки на членство может быть ограничена сверху вероятностью атаки на связывание с добавлением параметра ϵ :

$$P(MIA) \leq \sup P_{link} + \epsilon\,, \qquad (Д.10)$$

где *P(MIA)* – вероятность атаки на связывание;

$\sup P_{link}$ – верхняя возможная оценка успешности атаки связывания;

ϵ – малая добавка, показывающая расхождение.

Используя данное неравенство, можно провести числовые оценки для атаки связывания и атаки на членство, например, с использованием следующего алгоритма:

- кластеризация: исходные и синтетические данные могут быть разделены на кластеры на основе комбинаций значений квазиидентификаторов, при этом выбор алгоритма кластеризации (например, *k*-средних, иерархическая кластеризация или другие методы) зависит от структуры данных и аналитических целей;

- анализ l -разнообразия: анализ l-разнообразия предполагает выявление внутри каждого кластера групп с низким разнообразием чувствительных атрибутов, при этом малое значение параметра l может указывать на повышенный риск утечки данных и снижение уровня приватности;

- векторное представление: каждая запись представляется в виде вектора признаков, включающих квазиидентификаторы и другие важные атрибуты, для последующего сравнения;

- сравнение косинусной схожести: выполняется сравнение между векторами записей исходных и синтетических данных с использованием косинусной схожести, высокие значения схожести могут указывать на потенциальное связывание записей:

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|} \leq treshold(\theta)\,, \qquad (Д.11)$$

где θ – угол между векторами записей;

$threshold$ – заданный порог;

$\vec{a}$ и $\vec{b}$ – сравниваемые векторы;

- оценка супремума вероятности атаки на связывание: производится подсчет количества уникально связываемых записей (записей, для которых вероятность связывания превышает установленный порог), это количество может быть выражено как процент от общего числа записей и использоваться для оценки супремума вероятности атаки на связывание ($\sup P_{link}$).

Примечания

1 Алгоритм приводится как пример оценки на основе кластеризации и может быть заменен при практическом применении. Результаты анализа могут существенно зависеть от выбранного алгоритма кластеризации и его параметров. Важно учитывать, что различные методы могут давать разные результаты и выбор подходящего метода должен основываться на специфике данных. При применении с высокоразмерными данными может быть проведена адаптация для снижения размерности, например, с использованием метода главных компонент PCA.

2 Выбор значения параметра l критичен для идентификации групп с высокой вероятностью утечки данных. Необходимо тщательно подбирать это значение в зависимости от чувствительности данных. Установка порога для косинусной схожести является важным шагом, который может влиять на конечные результаты. Неправильный выбор порога может привести либо к недооценке, либо к переоценке вероятности утечки данных.

Д.5 Атака на вывод

Атака на вывод (Inference Attack) предполагает извлечение чувствительной информации о лицах или группах на основе корреляций в синтетических данных. Если существует сильная корреляция между атрибутами в реальных и синтетических данных, злоумышленник может попытаться сделать вывод об исходной чувствительной информации. Например, если известно, что в реальных данных определенная возрастная группа в большинстве случаев страдает от определенного заболевания, и это отражено в синтетических данных, злоумышленник может попытаться сделать аналогичный вывод.

Оценка успешности атаки на вывод может быть сделана с использованием условной энтропии по атрибутам синтетического набора данных:

$$H(Y|X) = -\sum_{y\in Y}\sum_{x\in X} p(y,x)\log\frac{p(y,x)}{p(x)}\,, \qquad (Д.12)$$

где X — квазиидентификаторы;

Y — чувствительные атрибуты;

$p(y, x)$ — совместное распределение вероятностей;

$p(x)$ — маргинальное распределение вероятностей (показывает, с какой вероятностью данное значение X встречается в данных в целом, независимо от значений Y).

Рассчитав условную вероятность для группы атрибутов, можно непосредственно оценить вероятность успешной атаки вывода:

$$P_{inf} = 1 - \frac{H(Y|X)}{H_{\max}}, \quad (Д.13)$$

где $H_{\max}$ — максимальная возможная энтропия (например, $H(Y)$ — энтропия Y, если X не учитывается). Низкая условная энтропия указывает на высокую вероятность успешной атаки, и, наоборот, высокая энтропия указывают на низкую вероятность успешного вывода.

Пример — X представляет собой возрастную группу, а Y — наличие заболевания. Если условная энтропия $H(Y \mid X)$ низка, это означает, что знание возраста значительно увеличивает вероятность правильного вывода о наличии заболевания, что повышает вероятность атаки на вывод.

Примечания

1 Примененный для оценки метод исходит из предположения о том, что вероятность успешной атаки вывода линейно зависит от условной энтропии, что может не выполняться в сложных многомерных распределениях.

2 Метод предполагает наличие определенных статистических свойств данных, таких как корреляция между публичными и чувствительными атрибутами. Если эти свойства нарушаются (например, если данные плохо сбалансированы или содержат выбросы), расчеты могут быть неточными.

3 Условная энтропия дает оценку только одного аспекта вероятности наступления негативного события — неопределенности в предсказании одного атрибута на основе другого. Этот подход не учитывает другие возможные методы атак или более сложные зависимости между атрибутами, которые могут быть использованы злоумышленниками.

Д.6 Оценка вероятности утечки данных через продвинутые атаки

Д.6.1 Расчет вероятности утечки данных через атаку на членство

Атака на членство представляет собой тип атаки, в которой злоумышленник пытается определить, была ли конкретная запись данных использована для обучения модели. Такая атака представляет серьезную угрозу приватности, и хотя применение доверенных технологий, таких как дифференциальная приватность, существенно снижает риск успешной атаки, полностью исключить эту угрозу может быть сложно из-за ограничений практической реализации синтеза. Атака работает, анализируя предсказания модели и выявляя различия в поведении модели по отношению к данным, которые были использованы для обучения, и новым данным.

Основная схема атаки на членство:

- обучение модели: целевая модель обучается на некотором наборе данных, который включает чувствительные данные, связанные с исходными записями;

- доступ злоумышленника: злоумышленник имеет доступ к предсказаниям модели, но не знает, какие данные были использованы для обучения;

- тестирование предсказаний: злоумышленник вводит записи в модель, анализируя, как она реагирует на определенные данные, чтобы выявить возможное участие этих данных в обучении.

Для усиления атак на членство зачастую используется вспомогательная техника теневых моделей, подход который атакующий использует для симуляции поведения целевой модели. Теневые модели обучаются на других данных, имитируя процесс обучения целевой модели. Такие модели затем используются для изучения того, как целевая модель ведет себя на данных, которые могли быть использованы для ее обучения.

Процесс использования теневых моделей:

- злоумышленник создает несколько теневых моделей, обученных на наборах данных, похожих на исходный обучающий набор целевой модели;

- теневые модели обучаются распознавать, присутствует ли запись в данных, на которых они были обучены;

- предсказания теневых моделей анализируются для выявления различий между тем, как модель реагирует на данные, которые были частью обучения, и на данные, которые не были включены в обучение;

- на основе анализа этих различий проводится атака на членство целевой модели.

Многократное использование одной и той же модели для генерации синтетических данных увеличивает вероятность успешной атаки на членство, так как каждый новый синтез данных может предоставить злоумышленнику больше информации для анализа предсказаний и поведения модели. Чем больше экземпляров синтетических данных сгенерировано, тем выше вероятность, что злоумышленник сможет точно определить, какие записи были частью исходного набора данных.

Формальное описание модели атаки с использованием теневых моделей может быть построено следующим образом: $\mathcal{M}$ — это целевая модель, обученная на наборе данных D. Цель MIA — определить, принадлежит ли определенная запись x_i обучающему набору *D*. В таком случае:

- злоумышленник создает несколько теневых моделей $\mathcal{M}_s$, которые обучаются на аналогичных наборах данных D_s, пытаясь смоделировать поведение целевой модели $\mathcal{M}$;

- теневые модели $\mathcal{M}_s$ обучаются распознавать, является ли запись x частью их обучающего набора данных. Для каждой теневой модели

$$\mathcal{M}_s(x) = \{0; 1\}, \qquad (Д.14)$$

где 0 – запись отсутствует в наборе;

1 – запись присутствует в наборе;

- для целевой модели $\mathcal{M}$, на основе ее предсказаний $\mathcal{M}(x)$, злоумышленник строит конструкцию

$$H_0: x \in D \ \cap H_1: x \notin D \ , \qquad (Д.15)$$

где H_0 и H_1 представляют две гипотезы: данные принадлежат обучающему набору или нет.

Когда модель обучается на конкретной записи, ее предсказания на этой записи становятся более точными и уверенными по сравнению с новыми, неизвестными записями. Если модель сталкивается с записью, на которой она была обучена, выходные вероятности будут значительно отличаться (более уверенные предсказания) по сравнению с новыми данными.

Если предсказания модели на одну и ту же запись x сильно отличаются от других предсказаний (например, повышенная уверенность модели в предсказании класса для этой записи), это может указывать на то, что эта запись была частью обучающего набора.

Вероятность утечки данных можно выразить через вероятность успешной MIA-атаки как долю успешных попыток злоумышленника правильно идентифицировать, была ли конкретная запись данных использована для обучения модели. За n принимают общее количество проверяемых записей, за s — количество успешных атак (когда злоумышленник правильно определил, использовалась ли запись в обучении). В таком случае вероятность успешной атаки P_{MIA} можно выразить:

$$P_{MIA} = \frac{s}{n} \ . \qquad (Д.16)$$

Указанная вероятность оценивает успешность атаки как среднее количество правильных предсказаний среди всех попыток. Чем выше данная вероятность, тем выше вероятность утечки данных через атаки на членство. Общий расчет вероятности все еще может осуществляться по формуле Д.5.

Д.6.2 Расчет вероятности утечки данных для атаки с инверсией модели

Инверсия модели дает тип атаки, при которой злоумышленник использует доступ к модели (или к ее предсказаниям) для восстановления исходных данных, на которых эта модель была обучена. Эта атака общая для всех методов с использованием ИИ и требует доступа к модели. В случае синтетических данных злоумышленник может попытаться восстановить исходные данные, на которых была обучена модель, генерирующая синтетические наборы. Такие атаки особенно опасны, если модель обучена на чувствительных исходных данных или если синтетические данные слишком точно отражают исходные данные.

Алгоритм применения атаки с инверсией модели:

- доступ к модели: злоумышленник получает доступ к модели, которая использовалась для генерации синтетических данных (либо доступ к синтетическим данным и предсказаниям модели);

- запросы к модели: злоумышленник выполняет серию запросов к модели, изменяя входные данные с целью понять, как изменяются выходные значения;

- построение инверсии: на основе полученных предсказаний злоумышленник пытается восстановить исходные данные, обучая вспомогательную модель или выполняя обратную оптимизацию, чтобы минимизировать разницу между предсказанными и реальными выходами модели;

- восстановление данных: злоумышленник восстанавливает наиболее вероятные исходные данные, используя инверсию модели.

Вероятность успешной атаки инверсии модели можно выразить через вероятность восстановления исходных данных, используя собранные предсказания модели. За P_{inv} принимают вероятность успешного восстановления исходных данных, которая зависит от точности модели, количества сделанных запросов q, и чувствительности модели к изменениям данных σ. Вероятность успешной атаки P_{inv} может быть описана как

$$P_{inv} = f(q, \sigma) \ , \qquad (Д.17)$$

где $f(q, \sigma)$ — функция, которая учитывает количество запросов q и чувствительность модели σ: чем больше запросов сделано и чем выше чувствительность модели к небольшим изменениям данных, тем выше вероятность успешной инверсии и восстановления исходных данных.

Примечание — Атака с инверсией модели может быть исключена при размещении модели в зашифрованном виде в закрытом доступе (см. ПНСТ 1065–2026).

Приложение Е

(рекомендуемое)

Шаблон отчета о качестве

Отчет о качестве синтетических данных

Версия: _____

Отчет №: _____ Организация: _____

Дата составления: _____ Автор(ы): _____

Ответственное лицо: _____

Информация о данных:
Наименование набора данных: _____ Размер данных:
Наименование файла: _____ _____
Уникальный идентификатор (Fingerprint) данных:
_____ Формат данных:

Информация о хранении: _____ _____

Количество записей:

_____

Уровень доступа:

_____

1 ВВЕДЕНИЕ
1.1 Цель тестирования
(Описание цели тестирования, например: «Цель тестирования — зафиксировать результаты оценки качества синтетических данных, их соответствие установленным стандартам и требованиям»)
1.2 Область применения
(Описание области применения – для каких целей или бизнес-процессов создан синтетический набор данных)

1.3 Нормативные ссылки
(Список внутренних и внешних стандартов, нормативных документов и процессов, которым отвечает проверка качества синтетических данных)

2 ОПИСАНИЕ ТЕСТИРУЕМОГО НАБОРА ДАННЫХ
2.1 Исходные данные
2.1.1 Словарь данных (включается непосредственно в данный отчет или приводится как приложение к отчету)

2.1.2 Основные характеристики
(Информация о характеристиках исходных данных, профайлинге, атрибутах)

2.2 Система синтеза данных
(Описание системы синтеза данных, архитектуры генератора, маркировка модели, версия)

2.3 Синтетические данные
(Описание синтетического набора, включая его размер, методы кодирования, информацию о найденных аномалиях, …)

3 ОПИСАНИЕ МЕТОДОВ ОБЕСПЕЧЕНИЯ УРОВНЯ ПРИВАТНОСТИ
(Описание методов и доверенных технологий синтеза, включая уровень выбранных параметров (шума, агрегации), порядка применения, используемых техник фильтрации и т. п.)

4 РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ
4.1 Оценка точности
(Описание метрик и их значений по точности/fidelity: статистическая близость, диверсификация данных, реалистичность данных, выводы по полученным результатам)

4.2 Оценка полезности
(Описание применимости для задач, проведенных измерений, выводы, сравнения с альтернативными методами)

4.3 Оценка уровня приватности
(Описание метрик, порядка расчета, полученных значений, выводы о вероятности утечки данных)

4.4 Оценка смещенности и объяснимость
(Качественная или количественная оценка смещенности и объяснимости, включая описание найденных смещений, ограничения выборки, замечания об ограничениях на использование, ссылки на документацию, лицензии, описание использованных методов для объяснения модели)

5 ВЫВОДЫ
5.1 Общие выводы
(Синтетические данные соответствуют/не соответствуют установленным целям, имеющимся стандартам и требованиям)

5.2 Рекомендации
(Рекомендуется провести следующие действия: _____)

Лицо, ответственное за проверку качества:
_____

Дата проверки: _____

Подпись: _____ (_____)

Библиография

[1] Национальная стратегия развития искусственного интеллекта на период до 2030 года (утверждена Указом Президента Российской Федерации от 15 февраля 2024 г. № 124)

[2] Федеральный закон от 27 июля 2006 г. № 152-ФЗ «О персональных данных»

[3] ISO/IEC TR 5469:2024 Искусственный интеллект. Функциональная безопасность и системы искусственного интеллекта
(Artificial intelligence — Functional safety and AI systems)

[4] ИСО/МЭК 25059:2023 Программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Модель качества для систем на основе искусственного
[Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems]

[5] ISO/IEC TR 24027:2021 Информационные технологии. Искусственный интеллект (ИИ). Смещенность в системах ИИ и при принятии решений с помощью ИИ
[Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making]

[6] ISO/IEC TS 25058:2024 Системная и программная инженерия. Требования и оценка качества систем и

программной продукции (SQuaRE). Руководство по оценке качества систем искусственного
[Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems]

[7] Financial Conduct Authority (FCA). Using Synthetic Data in Financial Services. Financial Conduct Authority, 2024

[8] ИСО/МЭК 5259-3:2024 Искусственный интеллект. Качество данных для аналитики и машинного обучения (ML). Часть 3. Требования и руководство по менеджменту качества данных
[Artificial Intelligence — Data Quality for Analytics and Machine Learning (ML) Part 3: Data Quality Management Requirements and Guidelines]

[9] ИСО/МЭК 23894:2023 Информационная технология. Искусственный интеллект. Руководство по менеджменту риска (Information technology — Artificial intelligence — Guidance on risk management)

[10] United Nations. A Framework for Ethical AI at the United Nations. United Nations, 2021

[11] ISO/IEC TS 8200:2024 Информационная технология. Искусственный интеллект. Управляемость автоматизированных систем искусственного интеллекта

(Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems)

[12] Рекомендация МСЭ-Т Y.3602(2018) Большие данные. Функциональные требования к источнику данных (Big data — Functional requirements for data provenance)

[13] ISO/IEC TR 24368:2022 Информационные технологии. Искусственный интеллект. Обзор этических и социальных проблем
(Information technology — Artificial intelligence — Overview of ethical and societal concerns)

[14] ISO/IEC TR 20547-5:2018 Информационные технологии. Эталонная архитектура больших данных. Часть 5. Дорожная карта стандартов
(Information technology Big data reference architecture — Part 5: Standards roadmap)

УДК 004.89:004.9:519.68:006.354 ОКС 35.020

Ключевые слова: синтез данных, результаты процесса синтеза данных, методика оценки качества

Руководитель организации-разработчика

Ассоциация больших данных
наименование организации

Исполнительный директор	_____	**А.В. Нейман**
должность	личная подпись	инициалы, фамилия

Руководитель разработки

_____	_____	
должность	личная подпись	инициалы, фамилия

Исполнитель:

_____	_____	
должность	личная подпись	инициалы, фамилия