



**Теоретические вопросы информатики, прикладной математики,
компьютерных наук и когнитивно-информационных технологий**

<https://doi.org/10.25559/SITITO.019.202304.857-864>

УДК 004.11.5

Проблема с существующими наборами данных (небольшие объемы, экономия, неоднородность), технический опыт и предложение решения

М. Мабу

Оригинальная статья

ФГАОУ ВО «Российский университет дружбы народов имени Патриса Лумумбы»,
г. Москва, Российская Федерация

Адрес: 117198, Российская Федерация, г. Москва, ул. Миклухо-Маклая, д. 6

mrmabouhmoise@gmail.com

Аннотация

В современном мире наблюдается значительный рост важности и использования данных. Все больше и больше организаций и индивидуальных пользователей зависят от данных для принятия решений и определения своей деятельности. Это связано с развитием информационных технологий, экспоненциальным ростом интернета и массовым проникновением смартфонов и других устройств, которые генерируют и собирают данные. Продолжительное накопление данных приводит к появлению большого количества доступных наборов данных. Они могут включать в себя информацию о пользовательском поведении, медицинские данные, финансовую информацию, социальные данные и многое другое. Все эти данные имеют потенциал для использования в различных областях, включая бизнес, науку, государственное управление и т. д. Однако, несмотря на широкий доступ к данным, многие наборы данных не обладают достаточной надежностью и качеством. Это может быть вызвано различными факторами, такими как ошибки сбора и обработки данных, неполные или противоречивые значения, а также возможности умышленной манипуляции данными. Важно понимать, что отсутствие достоверных данных может иметь серьезные последствия.

В заключение, рост значимости и использования данных в современном мире требует решения проблемы недостоверных наборов данных. Улучшение качества данных является необходимым условием для принятия точных решений и достижения успеха в различных областях деятельности. Дальнейшие исследования по этой теме и разработка новых методов улучшения наборов данных помогут сделать данные надежными и достоверными и обеспечат более эффективное использование данных в современном мире.

Ключевые слова: наборы данных, неполные данные, неточные данные, несогласованные данные, устаревшие данные

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Для цитирования: Мабу М. Проблема с существующими наборами данных (небольшие объемы, экономия, неоднородность), технический опыт и предложение решения // Современные информационные технологии и ИТ-образование. 2023. Т. 19, № 4. С. 857-864. <https://doi.org/10.25559/SITITO.019.202304.857-864>

© Мабу М., 2023



Контент доступен под лицензией Creative Commons Attribution 4.0 License.
The content is available under Creative Commons Attribution 4.0 License.



The Problem with Existing Datasets (Small Volumes, Cost-Efficiency Considerations, Non-Unidimensionality), Technical Expertise, and a Proposed Solution

M. Mabouh

Original article

Peoples' Friendship University of Russia named after Patrice Lumumba,
Moscow, Russian Federation

Address: 6 Miklukho-Maklaya St., Moscow 117198, Russian Federation

mrmabouhmoise@gmail.com

Abstract

In today's world, there has been a significant increase in the importance and use of data. More and more organizations and individuals depend on data to make decisions and guide their activities. This is due to the development of information technology, the exponential growth of the Internet and the massive penetration of smartphones and other devices that generate and collect data. Continued accumulation of data results in many available data sets. This may include information about user behavior, medical data, financial information, social data and more. All this data have the potential to be used in various fields, including business, science, government, etc. However, despite the wide availability of data, many data sets lack sufficient reliability and quality. This can be caused by a variety of factors, such as data collection and processing errors, incomplete or inconsistent values, and the possibility of intentional data manipulation. It is important to understand that the lack of reliable data can have serious consequences.

In conclusion, the increasing importance and use of data in today's world requires addressing the problem of unreliable data sets. Improving data quality is a prerequisite for making accurate decisions and achieving success in various fields of activity. Further research on this topic and the development of new methods to improve data sets will help make the data reliable and valid and ensure more effective use of data in the modern world.

Keywords: Data sets, Incomplete Data, Inaccurate Data, Inconsistent Data, Outdated Data

Conflict of interests: The author declares no conflict of interests.

For citation: Mabouh M. The Problem with Existing Datasets (Small Volumes, Cost-Efficiency Considerations, Non-Unidimensionality), Technical Expertise, and a Proposed Solution. *Modern Information Technologies and IT-Education*. 2023;19(4):857-864. <https://doi.org/10.25559/SITITO.019.202304.857-864>



1 Введение

Изучение проблемы с существующими наборами данных представляет собой важный этап для понимания недостатков и искажений, которые могут возникать при использовании таких данных. Наборы данных играют ключевую роль во многих областях, включая исследования, анализ данных и машинное обучение [1]. Исследование проблем с наборами данных позволяет выявить и проанализировать их недостоверность и неполноту [2]. Одной из основных проблем существующих наборов данных является их качество¹ [3]. Качество данных важно, так как оно определяет точность и достоверность результатов, получаемых при их использовании [4]. Недостоверные данные могут привести к неправильным выводам и неверным рекомендациям. Причины возникновения проблем с наборами данных могут быть разнообразными. Одной из основных причин является человеческий фактор [5]. Ошибки и неточности могут возникать при сборе и обработке данных [6]. Неправильная интерпретация и классификация информации также может привести к проблемам с наборами данных. Последствия недостоверных данных могут быть серьезными [7]. Они могут привести к неправильному принятию решений и некачественному анализу данных. В некоторых случаях, особенно в области медицины и финансов, неправильные данные могут иметь серьезные последствия для здоровья и безопасности людей. Оценка качества существующих наборов данных является важным шагом для их улучшения [8]. Критерии оценки качества данных могут включать точность, полноту, надежность и актуальность [9]. Методы проверки достоверности данных могут быть различными и включать в себя анализ погрешностей, сравнение с другими независимыми источниками и использование статистических методов для проверки согласованности данных [10]. Примеры проблемных наборов данных могут варьироваться от ошибок в категоризации до отсутствия актуальности информации [11].

Виды проблем с наборами данных могут включать неполные данные, которые могут возникать из-за пропусков или недостаточного количества информации в наборе. Неполные данные могут быть результатом ошибок в процессе сбора данных или отсутствия доступа к полной информации [12]. В таких случаях набор данных может содержать пропущенные значения или недостаточное количество записей для каждой категории. Неполные данные могут существенно повлиять на анализ и выводы, основанные на этих данных. Недостоверные или неполные значения могут исказить статистические метрики и привести к неверным или неточным результатам и выводам. Кроме того, нехватка данных может ограничивать возможности принятия решений и разработки моделей, основанных на этих данных. Проблемы с неполными данными могут быть вызваны различными факторами,

включая недостаточное финансирование для сбора полной информации, ограничения в доступе к информации, ошибки в процессе сбора данных или сохранения данных в неправильном формате. Некорректно заполненные данные, такие как ошибки ввода, неправильно внесенные значения или неправильная интерпретация данных, также могут привести к неполным наборам данных. Влияние неполных данных может быть значительным. Отсутствие полной информации обо всех переменных и значений может снизить точность статистических оценок и анализа данных [13, 14]. Это может повлиять на предложение политик и принятие решений на основе этих данных. Недостаток данных также может снизить общую надежность и достоверность исследований, основанных на этих данных. Одним из способов решения проблемы с неполными данными может быть сбор дополнительной информации или использование методов заполнения пропущенных значений [15]. Например, можно провести дополнительные исследования или опросы, чтобы получить недостающие данные. Также возможно использование статистических методов, таких как импутация, для заполнения пропущенных значений. В целом, проблемы с неполными данными являются серьезным вызовом для исследователей и аналитиков. Работа с неполными данными требует внимательного анализа и поиска альтернативных способов заполнения пропущенной информации. Контроль и обеспечение качества данных являются ключевыми аспектами при работе с наборами данных, чтобы гарантировать достоверность и полноту информации.

1.1 Некорректно заполненные данные

Некорректно заполненные данные – одна из основных проблем, с которой можно столкнуться при работе с наборами данных. Эта проблема возникает, когда данные содержат ошибки или неточности в заполнении. Некорректно заполненные данные могут возникать по разным причинам [16]. В некоторых случаях это связано с человеческим фактором, когда пользователь неправильно вводит информацию или допускает опечатки. Это может происходить при ручном заполнении форм, где человек может допустить ошибки при переносе данных. Другая причина – автоматическое считывание и обработка данных, которые могут содержать искажения из-за шума или ошибок в алгоритмах обработки. Некорректно заполненные данные могут иметь серьезные последствия для анализа и принятия решений. Неправильная информация может привести к неверным выводам или искажениям результатов исследования. Например, если в наборе данных присутствуют неправильно указанные возраста или несоответствующие значения, это может исказить анализ связи между возрастом и другими переменными. Некорректные данные также могут привести к неправильным предсказаниям в моделях машинного обучения и ухудшить точность прогнозов. Определение некорректно заполненных данных является важным шагом при оценке качества набора данных. Для этого можно использовать различные

¹ Pipino L. L., Lee Y. W., Wang R. Y. Data quality assessment // Communications of the ACM. 2022. Vol. 45, No. 4. P. 211-218. <https://doi.org/10.1145/505248.506010>



методы проверки, включая визуальный анализ, статистические тесты и алгоритмы обнаружения выбросов. Визуальный анализ позволяет искать явные аномалии или несоответствия в данных, например, с помощью графиков или диаграмм. Статистические тесты могут использоваться для проверки распределения значений переменных и выявления отклонений от ожидаемых паттернов. Алгоритмы обнаружения выбросов помогут выявить аномальные значения, которые могут указывать на некорректно заполненные данные. Для решения проблемы некорректно заполненных данных можно применять различные подходы. Одним из них является использование методов очистки данных, которые позволяют автоматически исправить ошибки или удалить некорректные значения [17]. Для этого можно применять алгоритмы заполнения пропущенных значений или методы интерполяции для исправления неточностей. Другим подходом является обучение моделей машинного обучения на основе правильно заполненных данных и использование их для предсказания и исправления некорректных значений [17]. Примеры успешного применения методов по улучшению наборов данных с некорректно заполненными данными включают использование алгоритмов машинного обучения для автоматической обработки и исправления ошибок. Например, некорректные значения в таблице данных могут быть предсказаны с использованием модели, обученной на основе правильно заполненных данных. Это позволяет автоматически исправить ошибки и улучшить качество данных. В заключение, некорректно заполненные данные являются одной из серьезных проблем, с которыми сталкиваются исследователи и аналитики при работе с наборами данных. Они могут исказить результаты и привести к неправильным выводам. Оценка качества данных и применение методов исправления ошибок помогут улучшить точность анализа и принятия решений на основе наборов данных.

1.2 Причины возникновения проблем

Причины возникновения проблем с наборами данных, связанных с неправильным сбором информации, оказывают значительное влияние на качество и достоверность данных. Ошибки в процессе сбора данных могут возникать по различным причинам и варьировать своей природой. Одной из причин возникновения проблем с наборами данных является неправильная настройка инструментов для сбора информации. Некорректные настройки могут привести к неправильной интерпретации данных или их искажению. Например, неправильные настройки инструментов для сбора данных могут привести к ошибкам в записи числовых значений, неправильной классификации данных или потере части информации. Другой причиной неправильного сбора данных может быть человеческий фактор. Ошибки, допущенные при ручном вводе данных или при обработке информации, могут привести к искажению данных [18]. Человеческая неосторожность или незнание требований к сбору данных могут привести к ошибкам в

структурировании и систематизации информации. Также можно выделить отсутствие четких стандартов и инструкций для сбора данных как причину возникновения проблем. В отсутствие четкого руководства, операторы могут иметь разное понимание того, какой информации необходимо собирать, как ее записывать и классифицировать. Это может привести к непоследовательности и неоднородности данных, что затрудняет их дальнейшую обработку и анализ. Еще одной важной причиной неправильного сбора данных является неправильное документирование процесса сбора информации. Отсутствие точных и подробных инструкций по сбору данных может привести к тому, что операторы не смогут собрать информацию в нужном объеме или не смогут правильно интерпретировать данные, что исказит эффективность набора данных. Таким образом, неправильный сбор данных может происходить по разным причинам, от неправильных настроек инструментов до ошибок, допущенных вручную. Эти проблемы могут привести к недостоверным искаженным данным, что в свою очередь может оказать негативное влияние на результаты анализа и принимаемые на их основе решения. Поэтому крайне важно обращать особое внимание на правильный сбор данных и разработку соответствующих стандартов и инструкций для этого процесса.

Недостаточный контроль качества данных – это одна из главных причин возникновения проблем с существующими наборами данных [5]. В процессе сбора и обработки данных, особенно в больших объемах, часто происходят ошибки и неточности, которые могут существенно исказить результаты и оказывать влияние на принимаемые решения. Когда контроль качества данных недостаточен, не всегда возможно обнаружить и устранить ошибки и неточности. Это может быть связано с неполными или некорректными методами проверки, отсутствием достаточных ресурсов или ограничениями в доступности источников. Одной из распространенных проблем является недостоверность данных, которая может возникать в результате их некорректного сбора или оценки. Например, при автоматическом сборе данных из различных источников могут возникать проблемы с точностью и актуальностью информации. Также недостоверность данных может возникать из-за ошибок в процессе их обработки, например, при использовании некорректных алгоритмов или неправильных методов агрегации. Другой причиной недостаточного контроля качества данных является отсутствие систематического подхода к их проверке. В некоторых случаях, когда данные собираются из разных источников или по разным критериям, может отсутствовать единая методика для оценки их достоверности и точности.

2 Актуальность

Актуальность данных является одним из ключевых критериев оценки качества наборов данных [4]. Устаревшие данные могут нанести серьезный ущерб при проведении анализа, прогнозировании или



принятии важных решений. Поэтому важно поддерживать актуальность данных на достаточно высоком уровне. Актуальность данных подразумевает, что информация в наборе данных является временно релевантной и соответствует текущим требованиям и ожиданиям пользователей. Актуальные данные обладают определенной свежестью и отражают наиболее актуальную информацию по теме исследования. Одной из причин утраты актуальности данных может быть их долгосрочное обновление или отсутствие механизмов автоматической актуализации.

3 Примеры проблемных наборов данных

3.1 Набор данных с неполнотой, когда отсутствуют важные атрибуты для анализа

Набор данных с неполнотой, когда отсутствуют важные атрибуты для анализа, является проблемой, с которой сталкиваются многие исследователи и аналитики данных. В таких наборах данных может быть нехватка информации, необходимой для полного понимания и анализа явления или явления, подлежащего исследованию. Отсутствие важных атрибутов может существенно ограничить возможности проведения анализа и получения значимых результатов [12]. Например, при изучении демографических данных может оказаться, что отсутствует информация о доходах определенных групп населения. Это может существенно повлиять на способность исследователей выявить связи между доходами и другими факторами, такими как образование или здоровье. Неполнота данных может возникнуть по разным причинам. В некоторых случаях, данные могут быть собраны не полностью из-за недостатка ресурсов, времени или финансирования. В других случаях данные могут быть неполными из-за ошибок сбора или хранения. Например, при сборе данных может потребоваться определенная экспертиза, и если такие эксперты не доступны или не могут полностью заполнить атрибуты, данные будут неполными. Последствия использования неполных наборов данных могут быть серьезными. Несмотря на ограничения, с данными все равно строятся модели и принимаются решения, которые могут иметь влияние на людей и общество в целом. Неполные данные могут приводить к искажению результатов и принятию неправильных решений. Отсутствие важных атрибутов может изменить картину явления или явления и привести к неправильным выводам. Для решения проблемы с неполными наборами данных можно предпринять несколько подходов. Во-первых, можно провести дополнительное исследование и собрать недостающие атрибуты. Это может быть затратным и трудоемким процессом, однако он может привести к получению более полной и достоверной информации. Во-вторых, можно использовать методы заполнения пропущенных значений. Существуют различные подходы к заполнению пропущенных значений, включая использование средних, медианы или моды значений других атрибутов, или прогнозирование пропущенных значений с использованием моделей машинного обучения [19-22]. Также можно применить методы

анализа данных, которые учитывают неполные данные. Например, можно использовать методы множественного импутации, которые позволяют оценивать пропущенные значения на основе имеющейся информации. Пример успешного применения методов по улучшению неполных наборов данных включает исследование клинических данных, где отсутствие информации о некоторых показателях здоровья может быть заполнено при помощи прогнозирования на основе других доступных атрибутов. В заключение, неполные наборы данных представляют существенную проблему для аналитиков и исследователей. Однако, существуют различные подходы и методы для улучшения качества данных и заполнения пропущенных значений. Оптимальный подход зависит от конкретной задачи и доступных ресурсов. Дальнейшие исследования и развитие темы помогут усовершенствовать методы работы с неполными наборами данных и повысить достоверность результатов и анализа.

3.2 Набор данных с неточностью, содержащий ошибки и неточные значения

Набор данных с неточностью, содержащий ошибки и неточные значения, представляет собой серьезную проблему при анализе данных. Такие неправильные значения могут исказить результаты и привести к неверным выводам [16]. Неточности могут возникнуть по разным причинам, как человеческий фактор, так и ошибки в процессе сбора и обработки данных. Ошибки в наборе данных могут быть разного характера. Это могут быть опечатки, неверные значения, неправильные форматы данных и другие неточности. Неточные значения могут возникать из-за недостаточной проверки данных перед включением их в набор или из-за неточностей в источниках данных. Такие проблемы с наборами данных могут привести к некорректным результатам и искажению выводов при анализе. Исследователи и аналитики, использующие такие данные, могут сделать ошибочные заключения и принять неверные решения на основе неточных значений. Ошибки в наборах данных могут быть особенно проблематичными в ситуациях, где эти данные используются для принятия важных решений, например, в медицине или финансовой сфере. Для решения проблемы с неточностью в наборах данных важно проводить качественную проверку данных перед их включением в набор. Это может включать автоматическую проверку на наличие ошибок и неточностей, а также ручную проверку данных аналитиками и экспертами. Одним из подходов к улучшению качества данных является использование методов машинного обучения. Алгоритмы машинного обучения могут быть обучены на правильных значениях данных для определения и исправления неточностей в наборах данных. Это может быть особенно полезно при больших объемах данных, где ручная проверка является трудоемкой задачей. Примеры успешного применения методов по улучшению наборов данных с неточными значениями могут включать использование алгоритмов обучения с подкреплением для автоматической корректировки



ошибок или применение алгоритмов классификации для идентификации и удаления неточных значений [24]. В заключение, проблема с наборами данных, содержащих неточности и ошибки, является серьезной проблемой в анализе данных. Неточные значения могут исказить результаты и привести к неправильным выводам. Проверка и исправление неточностей в наборах данных является важной задачей для обеспечения достоверности и надежности анализа данных. Использование алгоритмов машинного обучения может быть эффективным подходом к улучшению качества данных и устранению неточностей. Дальнейшие исследования должны быть направлены на разработку более эффективных методов и инструментов для проверки и улучшения качества наборов данных.

3.3 Набор данных с несогласованностью, где значения атрибутов противоречат друг другу

Набор данных с несогласованностью, где значения атрибутов противоречат друг другу, является одной из распространенных проблем, с которыми сталкиваются исследователи и аналитики данных. Такие наборы данных создают сложности при анализе и использовании, поскольку информация, содержащаяся в них, не соответствует друг другу или является противоречивой [23]. Причины возникновения несогласованности в значениях атрибутов в наборе данных могут быть разнообразными. Во-первых, это может быть связано с ошибками при сборе данных, когда разные источники информации вносят различные значения для одного и того же атрибута. Например, если рассматриваемый атрибут – это год выпуска продукта, разные источники могут указывать разные значения для этого атрибута. Во-вторых, несогласованность данных может быть обусловлена изменениями в структуре набора данных. Например, если атрибут «страна» изменяет свои значения в разные периоды времени (например, из-за политических изменений или изменения географических границ), то значения атрибута в разных записях могут не совпадать. Такие несогласованности в значениях атрибутов создают серьезные проблемы при анализе набора данных и использовании его для принятия решений. Например, если аналитик пытается определить частоту продажи продукта по годам, то несогласованность в значениях года выпуска продукта может существенно исказить результаты анализа. Для борьбы с проблемой несогласованности данных существуют различные подходы и методы. Одним из способов решения этой проблемы является проведение качественной предобработки данных. Это включает в себя процесс проверки и очистки данных, а также выявление и обработку несогласованностей. Например, можно провести процесс исправления значений атрибутов, приводя их к единому стандарту, или же удалить записи с противоречивыми значениями. Также для решения проблемы несогласованности данных можно использовать методы машинного обучения. Решение этой проблемы требует проведения качественной предобработки данных, использования методов машинного обучения и обращения к примерам

успешного применения этих методов в различных областях. Поэтому, правильная обработка несогласованности данных играет важную роль в обеспечении достоверности и качества анализа данных.

3.4 Набор данных с устаревшими значениями, который не отражает текущую реальность

Набор данных с устаревшими значениями, который не отражает текущую реальность, является одной из серьезных проблем, с которыми сталкиваются исследователи и аналитики. Такие наборы данных не только устарели, но и могут привести к неправильным выводам и недостоверным результатам исследований [7]. Причины возникновения данной проблемы могут быть разнообразными. Например, процессы сбора и обновления данных могут быть неэффективными или неправильно организованными. Также, изменения в окружающей среде или в самой предметной области могут привести к тому, что набор данных, который ранее был актуальным, теперь устарел и не отражает текущей реальности. Последствия использования устаревших наборов данных могут быть достаточно серьезными. Например, они могут привести к неправильным выводам и неправильным решениям, что в свою очередь может нанести ущерб бизнесу или обществу в целом. Кроме того, использование устаревших данных может привести к потере доверия пользователей и клиентов, что может негативно сказаться на репутации организации. Для решения проблемы с устаревшими значениями в наборах данных необходимы соответствующие подходы и методы.

Для решения проблемы с существующими наборами данных требуется применение различных подходов и методов, которые позволят улучшить качество этих данных. Одним из основных подходов является проведение процесса очистки данных, в ходе которого устраняются ошибки, пропуски, выбросы и другие недостатки, которые могут повлиять на достоверность информации. Очистка данных может включать в себя применение различных алгоритмов и методов, таких как удаление дубликатов, заполнение пропущенных значений, фильтрация аномалий и выбросов. Также важным шагом является стандартизация данных, которая позволяет привести их к общему формату и согласовать значения. Однако, для эффективного решения проблемы с наборами данных необходимо не только провести очистку, но и осуществить проверку достоверности и качества данных. Для этого можно применить различные критерии и методы, такие как проверка на наличие ошибок, несоответствий и противоречий с использованием специальных алгоритмов и аналитических инструментов. Кроме того, использование алгоритмов машинного обучения может значительно повысить качество данных. Например, можно применить алгоритмы автоматической классификации и кластеризации для определения категорий и группировки данных. Также возможно использование алгоритмов обработки естественного языка для извлечения и структурирования информации из текстовых данных. Для успешного улучшения наборов данных можно обратиться к примерам успешного применения



методов. Например, исследования показали, что использование алгоритмов машинного обучения и статистических моделей позволяет снизить ошибки и повысить точность классификации данных. Также можно упомянуть примеры использования методов обработки естественного языка для анализа и классификации текстовых данных, что позволяет получить более структурированную и полезную информацию. В заключение, решение проблемы с существующими наборами данных требует комплексного подхода, включающего проведение очистки данных, проверку достоверности и качества данных, а также использование алгоритмов машинного обучения. Однако, необходимо учитывать, что каждый набор данных уникален и требует индивидуального подхода. Дальнейшие исследования и развитие данной темы могут включать создание новых методов и инструментов для улучшения качества данных и обеспечения их достоверности.

4 Обсуждение и заключение

В ходе данной работы была проведена систематическая аналитическая работа, посвященная проблеме существующих наборов данных. Были выявлены различные проблемы, связанные с недостоверностью данных, и проанализированы их причины и последствия. Также была проведена оценка качества существующих наборов данных с использованием определенных критериев и методов. Исследование продемонстрировало, что существуют различные проблемы с наборами данных, такие как отсутствие актуальности, неполные данные, несогласованные форматы и дублирование информации. Проблемы возникают по разным причинам, включая недостаточные ресурсы, неправильную сегментацию данных и недостаточно жесткие контрольные механизмы. Последствия использования недостоверных данных могут быть серьезными и негативно сказываться на результатах исследований или на принятии подходящих решений. Некорректные данные могут привести к неверным выводам, ошибочным прогнозам или даже серьезным потерям для бизнеса

или организации [1]. Однако в данной работе были предложены решения проблемы существующих наборов данных. Основными подходами к улучшению качества данных являются увеличение актуальности данных, стандартизация формата данных, повышение контроля качества и усиление сотрудничества в области обмена данными [17]. Кроме того, было выявлено, что использование алгоритмов машинного обучения может быть эффективным инструментом для обработки данных и выявления потенциальных проблем. В рамках работы были представлены примеры успешного применения методов по улучшению наборов данных. Они демонстрируют, что правильное использование современных технологий и подходов может значительно улучшить качество данных и основанные на них аналитические процессы. В заключение стоит отметить, что данная работа позволяет сделать следующие выводы по исследованию проблемы с существующими наборами данных. Во-первых, проблемы с наборами данных являются актуальной и серьезной проблемой для многих организаций и научных исследователей. Во-вторых, оценка качества данных является важным этапом в исследовательском процессе и требует использования определенных критериев и методов. В-третьих, улучшение качества данных возможно с помощью различных подходов и технологий, включая использование алгоритмов машинного обучения. Направлениями дальнейших исследований и развития темы могут быть углубленный анализ проблем данных в конкретных сферах, разработка новых методов и алгоритмов для улучшения качества данных, а также создание инструментов и руководств для оценки качества данных и решения проблемы с существующими наборами данных [10], [25]. В целом, данная работа позволяет лучше понять проблему с существующими наборами данных, выделить основные причины и последствия проблем, а также предложить некоторые решения для улучшения качества данных. Тема имеет большую практическую важность и является актуальной для множества областей и отраслей, где используются данные для принятия решений и научных исследований.

References

1. Gong Y., Liu G., Xue Y., Li R., Meng L. A survey on dataset quality in machine learning. *Information and Software Technology*. 2023;162:107268. <https://doi.org/10.1016/j.infsof.2023.107268>
2. Whang S.E., Roh Y., Song H., Lee J.G. Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. *arXiv:2112.06409*. 2021. <https://doi.org/10.48550/arXiv.2112.06409>
3. Batini C., Scannapieco M. Data and Information Quality: Dimensions, Principles and Techniques. *Data-Centric Systems and Applications*. Cham: Springer; 2016. 500 p. <https://doi.org/10.1007/978-3-319-24106-7>
4. Cai L., Zhu Y. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*. 2015;14:2. <https://doi.org/10.5334/dsj-2015-002>
5. Cho H., Lee S. Data Quality Measures and Efficient Evaluation Algorithms for Large-Scale High-Dimensional Data. *arXiv:2101.01441*. 2021. <https://doi.org/10.48550/arXiv.2101.01441>
6. Abedjan Z., Chu X., Deng D., Fernandez R.C., Ilyas I.F., Ouzzani M., Papotti P., Stonebraker M., Tang N. Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*. 2016;9(12):993-1004. <https://doi.org/10.14778/2994509.2994518>
7. Haug A., Zachariassen F., Van Liempd D. The costs of poor data quality. *Journal of Industrial Engineering and Management*. 2011;4(2):168-193. <https://doi.org/10.3926/jiem.2011.v4n2.p168-193>
8. Tsai C.-W., Lai C.-F., Chao H.-C., Vasilakos A.V. Big data analytics: a survey. *Journal of Big Data*. 2015;2:21. <https://doi.org/10.1186/s40537-015-0030-3>



9. Kwon O., Lee N., Shin B. Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*. 2014;34(3):387-394. <https://doi.org/10.1016/j.ijinfomgt.2014.02.002>
10. Le Quy T., Roy A., Iosifidis V., Zhang W., Ntoutsis E. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*. 2022;12(3):e1452. <https://doi.org/10.1002/widm.1452>
11. Imran M., Qayyum F., Kim D.-H., Bong S.-J., Chi S.-Y., Choi Y.-H. A Survey of Datasets, Preprocessing, Modeling Mechanisms, and Simulation Tools Based on AI for Material Analysis and Discovery. *Materials*. 2022;15(4):1428. <https://doi.org/10.3390/ma15041428>
12. Emmanuel T., Maupong T., Mpoeleng D., Semong T., Mphago B., Tabona O. A survey on missing data in machine learning. *Journal of Big Data*. 2021;8:140. <https://doi.org/10.1186/s40537-021-00516-9>
13. Van Buuren S. Flexible Imputation of Missing Data. 2nd ed. Boca Raton: CRC Press; 2018. 444 p. <https://doi.org/10.1201/9780429492259>
14. Miao X., Wu Y., Chen L., Gao Y., Yin J. An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering*. 2023;35(7):6630-6650. <https://doi.org/10.1109/TKDE.2022.3186498>
15. Kazijevs M., Samad M.D. Deep Imputation of Missing Values in Time Series Health Data: A Review with Benchmarking. *arXiv:2302.10902*. 2023. <https://doi.org/10.48550/arXiv.2302.10902>
16. Abedjan Z., Golab L., Naumann F. Profiling relational data: A survey. *The VLDB Journal*. 2015;24:557-581. <https://doi.org/10.1007/s00778-015-0389-y>
17. Chu X., Ilyas I.F., Krishnan S., Wang J. Data Cleaning: Overview and Emerging Challenges. In: Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16). New York, NY, USA: Association for Computing Machinery; 2016. p. 2201-2206. <https://doi.org/10.1145/2882903.2912574>
18. Fatima A., Nazir N., Khan M.G. Data Cleaning in Data Warehouse: A Survey of Data Pre-processing Techniques and Tools. *International Journal of Information Technology and Computer Science*. 2017;9(3):50-61. <https://doi.org/10.5815/ijitcs.2017.03.06>
19. Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data. 2nd ed. John Wiley & Sons, Inc.; 2002. 408 p. <https://doi.org/10.1002/9781119013563>
20. Stekhoven D.J., Bühlmann P. MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118. <https://doi.org/10.1093/bioinformatics/btr597>
21. Zhang Z. Missing data imputation: focusing on single imputation. *Annals of Translational Medicine*. 2016;4(1):9. <https://doi.org/10.3978/j.issn.2305-5839.2015.12.38>
22. Wang Z., Akande O., Poulos J., Li F. Are deep learning models superior for missing data imputation in large surveys? Evidence from an empirical comparison. *arXiv:2103.09316*. 2021. <https://doi.org/10.48550/arXiv.2103.09316>
23. Krishnan S., Haas D., Franklin M.J., Wu E. Towards reliable interactive data cleaning: a user survey and recommendations. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA '16). New York, NY, USA: Association for Computing Machinery; 2016. Article number: 9. <https://doi.org/10.1145/2939502.2939511>
24. Abedjan Z., Golab L., Naumann F. Profiling relational data: a survey. *The VLDB Journal*. 2015;24:557-581. <https://doi.org/10.1007/s00778-015-0389-y>
25. Chung Y., Krishnan S., Kraska T. A Data Quality Metric (DQM): How to Estimate the Number of Undetected Errors in Data Sets. *arXiv:1611.04878*. 2016. <https://doi.org/10.48550/arXiv.1611.04878>

Поступила 08.09.2023; одобрена после
рецензирования 06.10.2023; принята к публикации
30.11.2023.

Submitted 08.09.2023; approved after reviewing
06.10.2023; accepted for publication 30.11.2023.

Об авторе:

Мабу Моисе Эрмани, аспирант кафедры математического моделирования и искусственного интеллекта факультета физико-математических и естественных наук, ФГАОУ ВО «Российский университет дружбы народов имени Патриса Лумумбы» (117198, Российская Федерация, г. Москва, ул. Миклухо-Маклая, д. 6), **ORCID:** <https://orcid.org/0000-0002-1337-0188>, mrmabouhmoise@gmail.com

Автор прочитал и одобрил окончательный вариант рукописи.

About the author:

Moise Hermann Mabouh, Postgraduate Student of the Department of Mathematical Modeling and Artificial Intelligence, Faculty of Science, Peoples' Friendship University of Russia named after Patrice Lumumba (6 Miklukho-Maklaya St., Moscow 117198, Russian Federation), **ORCID:** <https://orcid.org/0000-0002-1337-0188>, mrmabouhmoise@gmail.com

The author has read and approved the final manuscript.