



Виталий Гульчеев | ИИ

**Секреты датасетов:**  
**практическое руководство по  
анализу и обработке данных**

**Искусственный Интеллект**  
**Виталий Александрович Гульчеев**  
**Секреты датасетов:**  
**практическое руководство по**  
**анализу и обработке данных**

*[http://www.litres.ru/pages/biblio\\_book/?art=69217987](http://www.litres.ru/pages/biblio_book/?art=69217987)  
SelfPub; 2023*

**Аннотация**

"Секреты датасетов: практическое руководство по анализу и обработке данных" представляет собой всеобъемлющий и доступный ресурс для специалистов и начинающих исследователей данных. Книга охватывает ключевые аспекты работы с датасетами, начиная с источников данных, форматов и структур, и заканчивая предобработкой, анализом и визуализацией. Она предоставляет примеры работы с датасетами с использованием популярных языков программирования и библиотек, таких как Python, R, pandas и dplyr. Автор делится опытом и лучшими практиками по балансировке данных, аугментации, разделению датасета на обучающую, валидационную и тестовую выборки, а также исследовательскому анализу данных. Книга также освещает важные этические аспекты сбора данных и обработки персональных данных. Это

практическое руководство подходит для всех, кто хочет улучшить свои навыки в работе с датасетами и получить ценные знания о современных подходах к анализу данных.

# Содержание

|   |    |
|---|----|
| Глава 1: Введение в датасеты                                | 6  |
| 1.2 Важность датасетов в анализе данных и машинном обучении | 8  |
| Глава 2: Источники датасетов                                | 11 |
| 2.1 Общедоступные ресурсы и базы данных                     | 11 |
| 2.2 Создание собственного датасета                          | 13 |
| 2.3 Этические аспекты сбора данных                          | 15 |
| Конец ознакомительного фрагмента.                           | 17 |

# Виталий Гульчеев, Искусственный Интеллект Секреты датасетов: практическое руководство по анализу и обработке данных

*Добро пожаловать в "Секреты датасетов:  
практическое руководство по анализу и обработке  
данных"!*

*В эпоху больших данных возможность грамотно  
работать с датасетами становится все более  
ценной и востребованной. В этой книге мы  
рассмотрим широкий спектр тем, связанных  
с датасетами, чтобы помочь вам научиться  
извлекать полезную информацию из сырых данных  
и применять эти знания в реальной жизни.*

*Мы начнем с основных понятий, таких  
как форматы и структуры данных, а затем  
перейдем к более продвинутым темам, таким как  
предобработка, анализ и визуализация данных. Вам  
предоставятся практические примеры и наработки  
на основе популярных языков программирования и  
библиотек, таких как Python и R, что позволит вам  
быстро освоить материал и начать применять его*

на практике.

*Это вступление – лишь начало вашего пути в мир датасетов и анализа данных. Надеемся, что эта книга станет для вас полезным инструментом и надежным путеводителем в процессе освоения этой увлекательной области знаний. Приятного чтения и успешного обучения!*

*Автор выражает надежду на развитие культуры качественного анализа данных в России. По его мнению, технологическое развитие во многом зависит от искусственного интеллекта, который должен быть обучен на основе точных и качественных данных.*

*Виталий Гульчеев*

# **Глава 1: Введение в датасеты**

## **1.1 Определение и основные понятия**

Датасет (от англ. dataset, «набор данных») – это структурированная коллекция данных, используемая для анализа, обработки или обучения моделей машинного обучения. Датасет состоит из наблюдений (экземпляров) и признаков (характеристик), которые описывают каждое наблюдение. В контексте машинного обучения наблюдения называются объектами, а признаки – переменными или атрибутами.

Рассмотрим пример датасета с информацией о погоде:

| День       | Температура | Влажность | Осадки |
|------------|-------------|-----------|--------|
| 01.01.2023 | -5          | 80        | снег   |
| 02.01.2023 | 2           | 75        | дождь  |
| 03.01.2023 | 1           | 60        | облака |

В данном примере каждая строка – это наблюдение (день), а столбцы – признаки (температура, влажность и осадки). В зависимости от типа данных признаки могут быть числовыми, категориальными или текстовыми.

## **1.2 Важность датасетов в анализе данных и машинном обучении**

Датасеты играют ключевую роль в анализе данных и машинном обучении, поскольку они являются основой для получения новых знаний и создания прогнозных моделей. Без качественных данных невозможно построить эффективные модели и получить точные результаты.

Важность датасетов в анализе данных:

Описательный анализ: датасеты позволяют выявить основные статистические закономерности, связи и зависимости между переменными.

Визуализация: с помощью датасетов можно создавать графические представления данных, что упрощает понимание сложных закономерностей и динамики изменений.

Поддержка принятия решений: анализ датасетов позволяет получить информацию, необходимую для принятия обоснованных решений на основе данных.

Важность датасетов в машинном обучении:

Обучение моделей: датасеты используются для обучения моделей машинного обучения, которые могут выполнять задачи классификации, регрессии, кластеризации и другие. Обучение моделей на качественных данных позволяет достичь высокой точности и обобщающей способности.

Валидация и тестирование: разделение датасета на обуча-



ющую, валидационную и тестовую выборки позволяет оценить качество модели, ее способность предсказывать результаты на новых данных, а также избежать переобучения.

Оптимизация гиперпараметров: с использованием датасетов можно настраивать гиперпараметры моделей для улучшения их производительности и точности.

Сравнение различных моделей: датасеты позволяют сравнивать разные алгоритмы машинного обучения, выбирая наиболее подходящий для конкретной задачи.

Пример использования датасета для задачи машинного обучения:

Предположим, что у нас есть датасет с информацией о пациентах, и нашей задачей является предсказание наличия диабета на основе набора признаков, таких как возраст, индекс массы тела (ИМТ) и уровень глюкозы.

Для этого мы можем использовать алгоритмы классификации, такие как логистическая регрессия или случайный лес. Мы разделим датасет на обучающую, валидационную и тестовую выборки, обучим модель на обучающей выборке и проверим ее качество на валидационной выборке. Затем мы проведем оптимизацию гиперпараметров и, наконец, оценим качество модели на тестовой выборке.

В заключение, датасеты являются неотъемлемой частью анализа данных и машинного обучения. Качественные датасеты позволяют получать точные результаты, создавать эффективные модели и выявлять новые закономерности. Важ-

но уделить внимание предобработке и очистке данных, а также выбору подходящих методов и алгоритмов для конкретной задачи.

# Глава 2: Источники датасетов

## 2.1 Общедоступные ресурсы и базы данных

Существует множество источников, где можно найти готовые датасеты для анализа данных и машинного обучения. Некоторые популярные ресурсы и базы данных включают:

Kaggle (<https://www.kaggle.com/>): платформа для соревнований по анализу данных и машинному обучению, которая предлагает большое количество датасетов на различные темы, включая финансы, здравоохранение и технологии.

UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>): один из старейших репозиторий датасетов, содержащий сотни датасетов для задач машинного обучения, включая классификацию, регрессию и кластеризацию.

Google Dataset Search (<https://datasetsearch.research.google.com/>): поисковик от Google, который позволяет найти датасеты, размещенные на различных веб-сайтах и порталах.

Data.gov (<https://www.data.gov/>): официальный портал правительства США, предоставляющий доступ к датасетам

на различные темы, такие как экономика, здравоохранение, образование и климат.

Европейский портал открытых данных (<https://www.europeandataportal.eu/>): портал, содержащий датасеты от различных стран Европейского союза.

Пример использования датасета с Kaggle: предположим, что вы хотите проанализировать данные о продажах видеоигр. На Kaggle есть датасет "Video Game Sales" (<https://www.kaggle.com/gregorut/videogamesales>), который содержит информацию о продажах видеоигр, платформах, жанрах и рейтинге.

## 2.2 Создание собственного датасета

В некоторых случаях готовых датасетов может быть недостаточно, и вам придется создать свой собственный датасет. Некоторые способы сбора данных:

**Веб-скрапинг:** сбор данных с веб-сайтов с использованием инструментов и библиотек, таких как BeautifulSoup и Scrapy для Python. Веб-скрапинг позволяет извлекать информацию с веб-страниц и преобразовывать ее в структурированный формат, например таблицу.

**API (Application Programming Interface):** использование API предоставляет доступ к данным из различных сервисов и платформ, таких как социальные сети, погодные сервисы и финансовые платформы. API обычно возвращает данные в формате JSON или XML, которые можно преобразовать в структурированный формат и добавить в свой датасет.

**IoT-устройства и датчики:** сбор данных с помощью датчиков, встроенных в различные устройства и системы, такие как смартфоны, автомобили и промышленное оборудование. Эти данные могут быть использованы для анализа и прогнозирования поведения устройств, определения аномалий и оптимизации процессов.

**Опросы и анкеты:** сбор данных с помощью анкетирования пользователей или экспертов, чтобы получить качественные и количественные оценки по определенным вопросам или

проблемам.

Пример создания собственного датасета с использованием веб-скрапинга: предположим, что вы хотите собрать данные о стоимости жилья в вашем городе. Вы можете использовать веб-скрапинг для сбора информации о ценах, местоположении, площади и других параметрах с сайтов по недвижимости.

## 2.3 Этические аспекты сбора данных

Сбор данных может иметь этические последствия, особенно когда данные связаны с личной информацией людей. Некоторые ключевые этические аспекты, которые следует учитывать при сборе данных, включают:

**Защита конфиденциальности:** соблюдение конфиденциальности пользователей, собирая только те данные, которые необходимы для вашей задачи. Обезличивание данных, скрывая личную информацию и уникальные идентификаторы, может помочь обеспечить приватность пользователей.

**Согласие на сбор данных:** получение разрешения от пользователей или владельцев данных перед сбором и использованием данных. Это может быть особенно важно при использовании веб-скрапинга или API, так как некоторые сайты и сервисы могут иметь ограничения на использование данных.

**Недискриминация:** избегание сбора и использования данных, которые могут привести к дискриминации или неравному обращению с определенными группами пользователей.

**Прозрачность:** информирование пользователей о целях сбора данных, методах обработки и хранения, а также о том, как их данные будут использоваться. Это важно для создания доверия и уважения к личной информации пользователей.

**Компетентность и ответственность:** обеспечение правиль-

ного и аккуратного сбора данных, а также надлежащего их использования. Необходимо избегать намеренного искажения результатов, основанных на данных, и следить за актуальностью данных, чтобы обеспечить точность анализа и прогнозов.



# Конец ознакомительного фрагмента.

Текст предоставлен ООО «ЛитРес».

Прочитайте эту книгу целиком, [купив полную легальную версию](#) на ЛитРес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.