УДК 004.8:629.3

ДАТАСЕТ И ЕГО ИСПОЛЬЗОВАНИЕ

БАРЩЕВСКИЙ Евгений Георгиевич
кандидат технических наук, профессор
БАРЩЕВСКИЙ Георгий Евгеньевич
кандидат технических наук
Государственный университет морского и речного флота имени адмирала С.О. Макарова
г. Санкт-Петербург, Россия

В статье рассмотрены актуальные вопросы, связанные с применением структурированных наборов данных датасетов *для работы с большим количеством информации.*
Ключевые слова: искусственный интеллект, датасеты.

При работе с Big Data используют большое количество информации: вычисляют паттерны поведения, смотрят статистику продаж, обучают ML-модели. Один из вариантов представления информации, который нужен для получения результата датасет [2; 3]. Датасет – структурированный набор обработанных и разложенных по понятным категориям данных.

У датасетов есть три основных варианта применения

Обучение моделей машинного обучения. Программы изучают датасеты, выстраивают закономерности, учатся классифицировать каждый объект.

Научные цели. Учёные используют наборы данных для проверки гипотез, анализа закономерностей, имитации реальных систем.

Бизнес-аналитика. Компании собирают большое количество информации, которую потом можно использовать для поиска правильных решений в бизнесе. Есть несколько видов датасетов. Для примера мы разберём подробнее простую запись, упорядоченную и граф.

Простая запись (flat records). Самый удобный и понятный вид данных. Они просты для понимания и обработки и используются в большинстве видов анализа. Визуально это таблица строк и столбцов. Каждый ряд или строка будет отдельной записью со своим названием, например название продукта, номер ряда в супермаркете или департамента.

Упорядоченные записи (ordered records). В таких датасетах [1] данные имеют строгий порядок, и каждая запись зависит от предыдущей или следующей. Это может быть важно для временных рядов. Например, временные ряды количества заболевших коронавирусом в разных штатах и регионах по всему миру.

Графы (graphs). Это данные, которые выстроены в связанную систему через узлы и рёбра-связи между ними. Используются, когда между разными объектами важны связи (https://neuralinsight.ru/iskusstvennyj-intellekt-itransport/). Например, узлами могут быть люди, места, вещи. Рёбрами – отношения между людьми, дороги между местами, результат комбинации разных вещей между собой. Такие связи могут использоваться в социальных сетях для рекомендаций друзей или при оптимизации маршрута в навигаторе.

Свойства и характеристики датасетов. Dataset можно описать через свойства и характеристики. Свойства дают понимание внутренней структуры, а характеристики помогают оценить содержимое.

Примеры свойств датасета:

– **Структура** объясняет тип датасета – табличный формат, граф, временные ряды.

– **Тип данных** описывает формат значений – числовой, строковый, временной. Пример – столбец «Цена» имеет числовой тип, а «Дата» – временной.

– **Пропуски** говорят об отсутствующих значениях. Например, в 15% случаев у пользователей нет данных о номере телефона или возрасте.

Несколько возможных характеристик

данных:

– **Средние значения** могут считаться поразному, они показывают разные математические показатели усреднённых показателей. Это могут быть среднее арифметическое, медиана, мода значений. Например, средний возраст покупателей — 34 года.

– **Диапазон** показывает минимум и максимум значений. Пример – показатели ежедневных продаж колеблются от 5 000 до 50 000 рублей в день.

– **Распределение данных** говорит о закономерностях, если посмотреть на информацию в общем.

Выводы. Чтобы решить конкретную задачу, необходимо выбрать правильный датасет и подготовить его. Для этого понадобится учесть несколько факторов, в частности, определится с целью использования датасета, объемом датасета, доступом и юридическими ограничениями по использованию данных, структурой и форматом данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Берман Кеннеди* Основы Python для DATA SCIENS. – Изд-во «Питер», 2023. – 230 с.
2. *Никифоров Л.Г.* Математика для DATA SCIENTIST.Анализ данных и математическое моделирование (путеводитель). – Изд-во «Ridero», 2021. – 24 с.
3. *Сергеев Н.* Аналитика и Data Science. Для не-аналитиков и даже 100% гуманитариевита. – Изд-во «Ridero», 2022. – 421 с.

UDC 004.8:629.3

DATASET AND ITS USE

BARSHCHEVSKY Evgeny Georgievich
Candidate of Sciences in Technology, Professor
BARSHCHEVSKY George Evgenievich
Candidate of Sciences in Technology
Admiral Makarov State University of Maritime and Inland Shipping
St. Peterburg, Russia

The article examines current issues related to the use of structured datasets for working with large amounts of information.

Keywords: artificial intelligence, datasets.