

ПОДГОТОВКА И ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ

ДИСЦИПЛИНА: «ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ»

ЛЕКЦИЯ 2 • МОДУЛЬ 1: ОСНОВЫ DATA MINING И ПОДГОТОВКА ДАННЫХ

ЦЕЛЬ И ЗАДАЧИ ЛЕКЦИИ

Главная цель

Сформировать понимание критической роли качества данных в Data Mining и освоить практические методы очистки, трансформации и подготовки исходного датасета перед применением машинного обучения.

Практические навыки

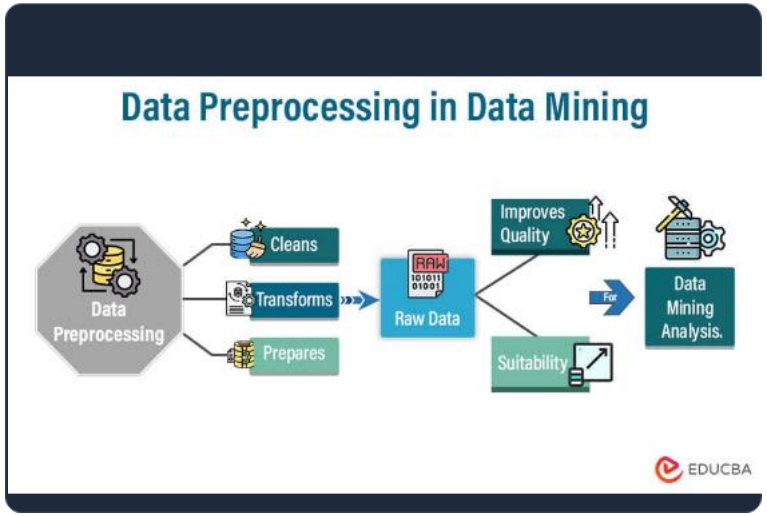
- Выявление и обработка пропусков и выбросов
- Кодирование категориальных признаков
- Масштабирование и нормализация данных
- Отбор ключевых признаков и формирование датасета

ПОЧЕМУ ПОДГОТОВКА ДАННЫХ ВАЖНА?

Принцип Garbage In → Garbage Out: некачественные входные данные неизбежно приведут к некорректным результатам работы алгоритма.

⚠ **Типичные дефекты в сырых данных:**

- Пропущенные значения (клиент 003 — возраст NaN)
- Аномалии и ошибки (клиент 004 — доход -50 000)
- Неявные дубликаты строк (002 и 005)
- Различие форматов названий городов и единиц



ID	Возраст	Доход	Город	Покупки
001	25	250000	Алматы	12
002	31	350000	Астана	8
003	—	280000	Алматы	10
004	29	-50000	Шымкент	7
005	31	350000	Астана	8



КОНВЕЙЕР DATA PREPROCESSING

ПОСЛЕДОВАТЕЛЬНЫЙ ПРОЦЕСС ПРЕВРАЩЕНИЯ СЫРЫХ ДАННЫХ В КАЧЕСТВЕННЫЙ DATASET ДЛЯ DATA MINING

FARABI
KAZAKH NATIONAL UNIVERSITY
1934

ОСНОВНЫЕ ПРОБЛЕМЫ РЕАЛЬНЫХ ДАННЫХ

Пропуски и дубли

Пустые ячейки (NaN, NULL) искажают расчеты, а повторные записи в таблицах смещают итоговую статистику модели.

Ошибки и выбросы

Отрицательный возраст или зарплата 15 млн там, где средняя 300 тыс. Требуется анализ природы аномалии перед удалением.

Форматы и шкалы

Записи «Алматы» и «ALMATY», а также измерения в метрах и сантиметрах воспринимаются компьютерами как разные объекты.

ХАРАКТЕРИСТИКИ КАЧЕСТВА ДАННЫХ

Критерий	Суть критерия	Пример нарушения	Статус
Полнота	Доля присутствующих значений в наборе данных	Возраст отсутствует у 5% пользователей	Пропуск
Корректность	Соответствие значений реальным фактам	Значение «Возраст = -10» или «Цена = -500»	Ошибка
Согласованность	Единообразие формата записи значений	«Алматы», «алматы», «г. Алматы», «ALMATY»	Формат
Уникальность	Отсутствие дублирующихся строк и сущностей	Две одинаковые записи покупателя с ID 001	Исправимо

ОБРАБОТКА ПРОПУСКОВ

Клиент	Возраст
A	25
B	31
C	NaN
D	29

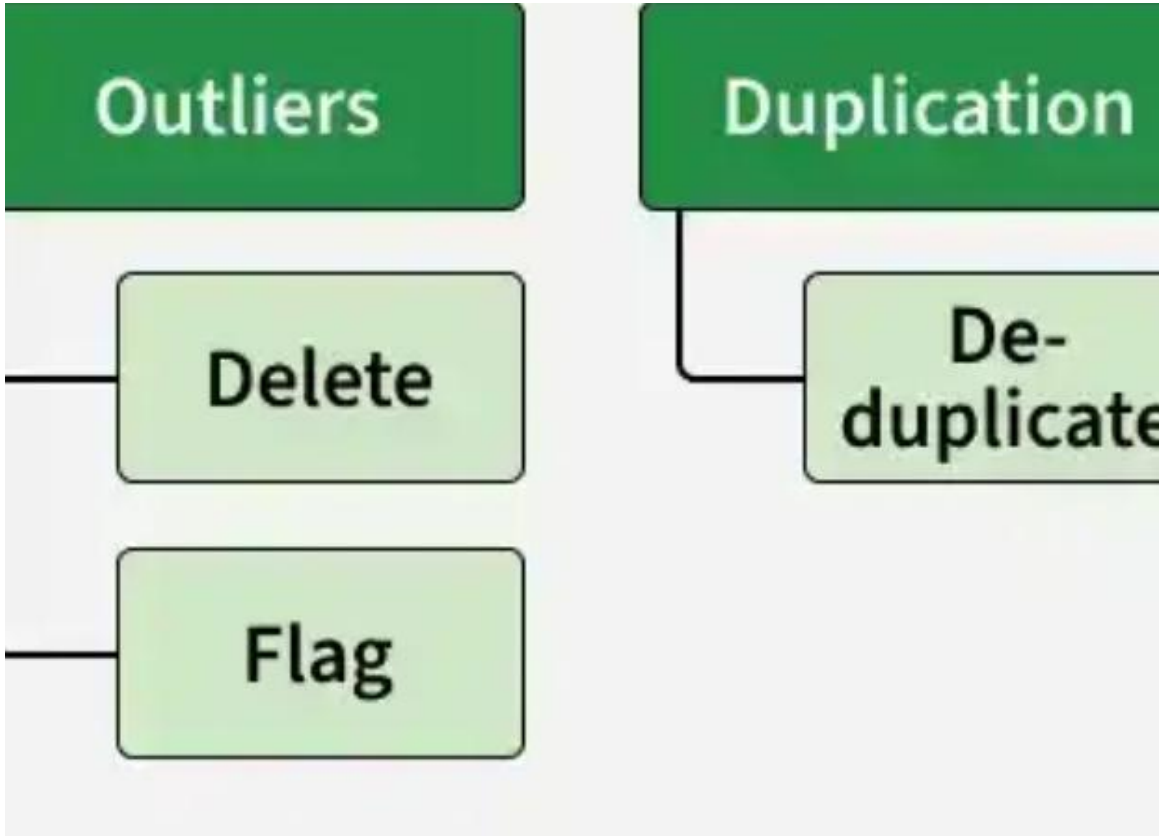
Стратегии восстановления

1. Удаление строк/столбцов: Применяется, если пропусков < 5%. При 30-40% приводит к огромной потере информации.

2. Заполнение средним/медианой: Медиана предпочтительнее для численных признаков с сильными выбросами.

3. Заполнение модой: Использование самого частого значения для категориальных признаков.

4. Advanced Imputation: KNN Imputer, интерполяция и модели машинного обучения.



УСТОЙЧИВОСТЬ: СРЕДНЕЕ VS МЕДИАНА

100

Выброс в ряду: 20, 22, 23, 24, 100

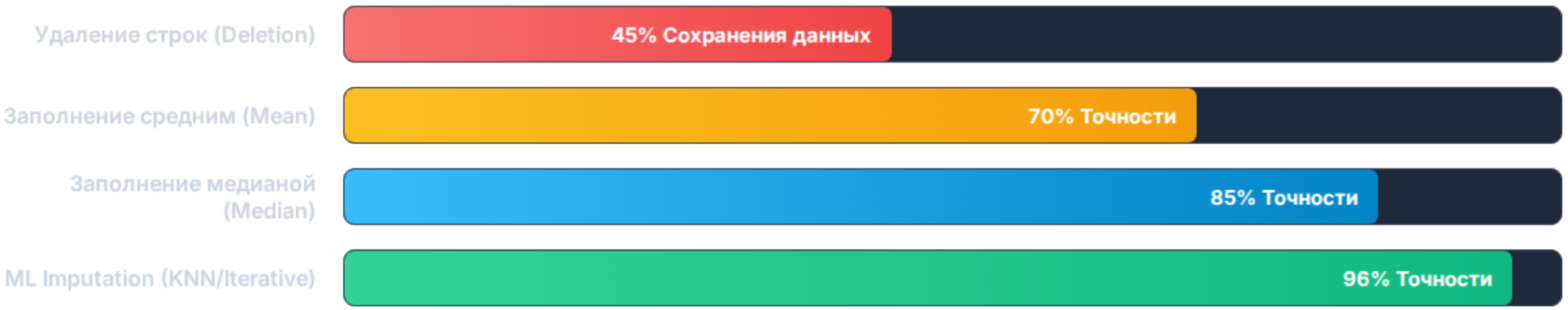
Влияние аномалий на статистику

При наличии резких выбросов обычное среднее арифметическое сильно смещается (в примере среднее = 37.8, что не отражает основную группу).

Медиана (Median) = 23.0

Медиана устойчива к выбросам, поэтому именно она используется для импутации доходов, цен и возраста при асимметричных распределениях.

ЭФФЕКТИВНОСТЬ МЕТОДОВ ИМПУТАЦИИ



Продвинутые методы (KNN Imputer) учитывают взаимосвязи между всеми признаками объекта, обеспечивая наивысшую точность восстановления данных.

МАСШТАБИРОВАНИЕ ПРИЗНАКОВ

Min-Max Normalization

Приводит значения признака к фиксированному диапазону [0, 1].

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Чувствителен к выбросам. Подходит для нейросетей и алгоритмов, требующих жестких границ.

Z-Score Standardization

Преобразует данные так, чтобы среднее было equal 0, а станд. отклонение equal 1.

$$X_{stand} = \frac{X - \mu}{\sigma}$$

Устойчив к аномалиям. Необходим для метрических алгоритмов (KNN, K-Means, PCA, SVM).

КОДИРОВАНИЕ КАТЕГОРИАЛЬНЫХ ПРИЗНАКОВ

Label Encoding

Присваивает каждой категории целое число (напр., Алматы=0, Астана=1, Шымкент=2). Применяется для порядковых данных (Ordinal).

One-Hot Encoding

Создает бинарные столбцы (0 или 1) для каждого значения. Исключает ложные математические приоритеты между категориями.

Binary Encoding

Переводит категории в бинарный код. Оптимален для категориальных признаков с огромным количеством уникальных значений (High Cardinality).

ВОПРОСЫ И ОБСУЖДЕНИЕ

КАЧЕСТВО ПОДГОТОВКИ ДАННЫХ ОПРЕДЕЛЯЕТ УСПЕХ ВСЕГО DATA MINING ПРОЕКТА.

ДИСЦИПЛИНА: ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ • ЛЕКЦИЯ 2

