Лекция 1

Машинное обучение и анализ данных

Лектор

Елизавета Власова

Практики

Кирилл Захаров
Камила Насибуллина
Михаил Подсытник
Игорь Толстокулаков

Как устроен курс

1

Введение в ML и анализ данных

Типы задач и ML-пайплайн; EDA, статистика, визуализация и снижение размерности.

2

Классическое машинное обучение

Линейные модели, kNN и метрики, деревья, ансамбли, boosting, feature selection и кластеризация.

3

Глубокое обучение и трансформеры

MLP и backprop; кратко CNN; NLP, embeddings, attention, BERT, T5 и GPT.

Оценивание: 100 баллов

60 баллов

Лабораторные работы

Работа с алгоритмами, данными и ML-экспериментами в течение семестра.

20 баллов

Две рубежные работы

Проверка теоретических знаний по материалам курса.

20 баллов

Экзамен

Теоретический экзамен или итоговый практический проект — в зависимости от желаемой оценки.

Лабораторные работы

Тип 1

Алгоритмы и реализация

Лабораторная сдаётся онлайн и проверяется автотестами и, при необходимости, вручную.

$$B_{\text{лаб}} = X \cdot \max(0{,}5; T)$$

X — балл за решение; $T \in [0; 1]$ — результат мини-теста (5 минут, 5 вопросов). Тест не переписывается. При пропуске $T = 0$, поэтому за лабораторную остаётся половина баллов.

Тип 2

Эксперимент и ML system design

Нужно спроектировать и провести эксперимент: выбрать разбиение данных и метрики, сравнить решения, проанализировать результаты и ограничения.

Такие работы сдаются на очной защите.

Оценивание

Две рубежные работы

1

20 октября 2026

Восьмая неделя курса, во время практического занятия.

2

15 декабря 2026

Шестнадцатая неделя курса, во время практического занятия.

Работы выполняются **на бумаге**. Пользоваться конспектами, другими материалами и электронными устройствами нельзя. Проверяются теоретические знания по пройденным темам. Для прохождения каждой контрольной точки нужно набрать не менее **6 из 10 баллов**.

Экзамен

Итоговая форма зависит от оценки, на которую претендует студент.

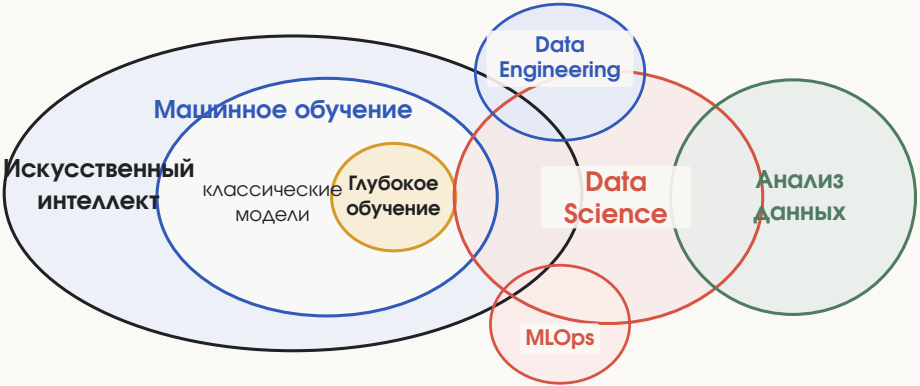
3	**Теоретический экзамен**	Ответить на вопросы по материалам курса.
4	**На выбор**	Сдать теоретический экзамен или выполнить итоговый практический проект.
5	**Итоговый проект**	Выполнить финальный проект в формате мини-курсовой; он заменяет теоретический экзамен.

Искусственный интеллект

Основные понятия, типы задач и области применения.

01

Как связаны AI, ML и Data Science



Области пересекаются по задачам: Data Science может использовать ML, но не сводится к нему; анализ данных также включает методы без обучения моделей.

Где используют искусственный интеллект

Медицина
диагностика и поиск лекарств

Транспорт
восприятие сцены и планирование

Контент
поиск, рекомендации, генерация

Наука и код
AlphaFold, ассистенты разработчика

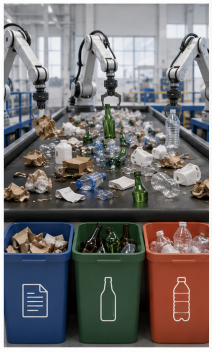
Источник: Google DeepMind: AlphaFold; обзор типовых областей применения AI/ML. Иллюстрация создана для курса.

Задачи машинного обучения

Классификация	выбрать один из классов	спам / не спам; дефект на снимке; тема обращения
Регрессия	предсказать число	спрос, цена, время доставки, риск
Кластеризация	найти группы без готовых меток	сегменты пользователей, типы поведения
Рекомендации	ранжировать варианты	фильмы, товары, документы, кандидаты
Генерация	создать новый объект	текст, код, изображение, структура молекулы

Пример 1

Определите тип задачи



Сортировка отходов

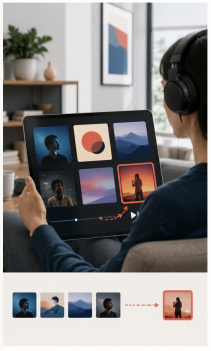
Камера видит объект на конвейере. Нужно выбрать контейнер: *бумага*, *стекло* или *пластик*.

Ответ: классификация

Варианты: классификация · регрессия · кластеризация · рекомендации · генерация

Пример 2

Определите тип задачи



Музыка на вечер

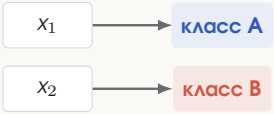
По истории прослушиваний нужно отобрать и упорядочить треки, которые пользователь, вероятно, включит следующими.

Ответ: рекомендации

Варианты: классификация · регрессия · кластеризация · рекомендации · генерация

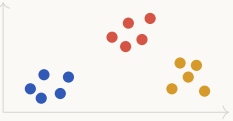
Откуда модель получает сигнал

Supervised



Есть пары «объект — правильный ответ».

Unsupervised



Ответов нет.
Ищем группы и структуру данных.

Self-supervised

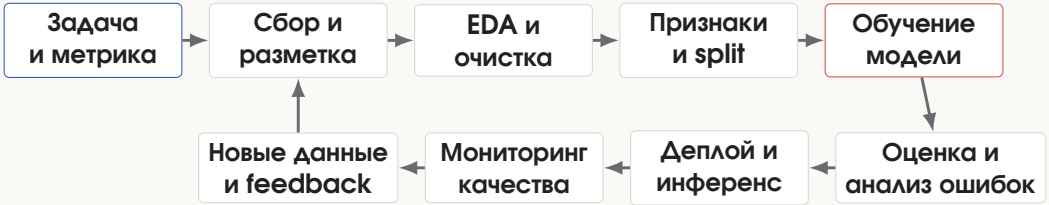
Модель учится на _____.

Текст делят на _____.

Контекст помогает восстановить _____.

Пропущенные слова становятся обучающими целями.

Из каких этапов состоит ML-пайплайн



Обучение модели — один из этапов полного ML-пайплайна.

Источник: Google for Developers, “ML pipelines” и “Production ML systems”.

Обучение, инференс и обратная связь

Обучение

Подготовить модель
Данные → признаки → обучение → валидация → артефакт модели.

Инференс

Выдать предсказание
Запрос → те же преобразования → модель → решение продукта.

Обратная связь

Контролировать качество
Логи → реальные ответы → drift → анализ → переобучение.

Мониторинг охватывает данные, признаки, качество модели и продуктовые метрики.

Источник: Google for Developers, “Production ML systems: monitoring”.

Пример: поиск статьи

Запрос пользователя: «как обнаружить утечку данных»

1. Feedback
Пользователь не кликает по статье и уточняет запрос

▶

2. Новые данные
Добавляем пару «запрос — релевантная статья»

▶

3. Переобучение
Обновляем ранжирование и проверяем метрики

▶

4. Деплой
Выкатываем новую версию поисковой модели

Нужная статья: **18-я позиция** ⟶ **3-я позиция**

Темы курса в контексте ML-пайплайна

1

Задача и данные

постановка · EDA · статистика · визуализация · PCA

Jupyter

pandas

2

Эксперимент

split · cross-validation · метрики · leakage · data validation

NumPy

sklearn

3

Классический ML

линейные · kNN · деревья · boosting · кластеризация

sklearn SHAP

4

DL и NLP

MLP · кратко CNN · embeddings · attention · BERT / T5 / GPT

PyTorch HF

5

ML-система

репозиторий · тесты · serving · monitoring · retraining · RAG / MCP

GitHub

MLflow

Exploratory Data Analysis

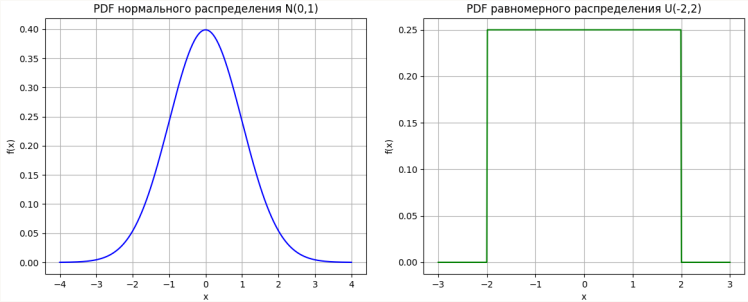
Повторяем базовую статистику и учимся превращать таблицу в проверяемые гипотезы.

02

Функция плотности

Для абсолютно непрерывной случайной величины ξ плотность связана с функцией распределения:

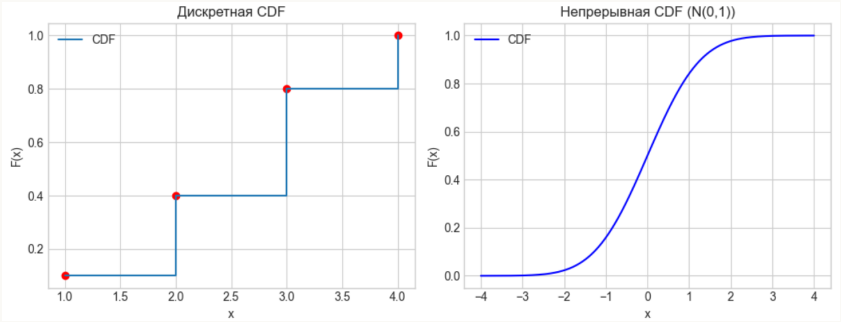
$$f_\xi(x) = F'_\xi(x) \qquad \text{почти всюду.}$$



Функция распределения

Функция распределения случайной величины ξ:

$$F_\xi(x) = \mathbb{P}(\xi \leq x).$$



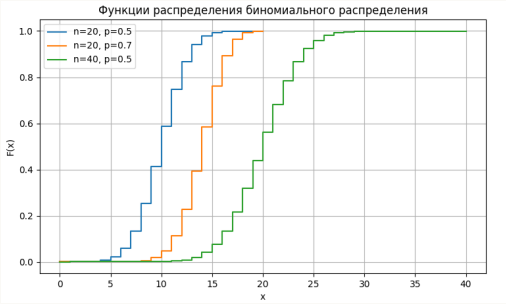
Биномиальное распределение

Число успехов в n независимых испытаниях:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$X \sim B(n, p)$, где p --- вероятность успеха.

$$F(k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}.$$

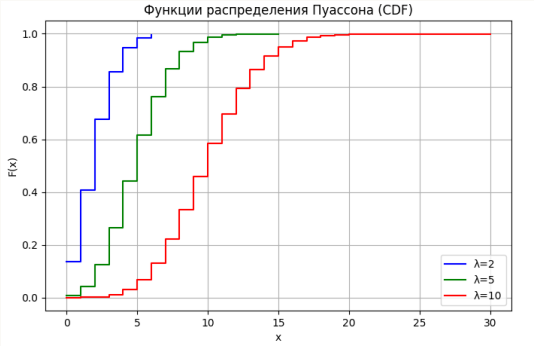


Распределение Пуассона

Число событий на фиксированном интервале при постоянной средней интенсивности:

$$\mathbb{P}(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

Обозначается $\Pi(\lambda)$, где $\lambda > 0$.



Равномерное распределение

Случайная величина ξ имеет равномерное распределение $U(a, b)$, $a < b$, если её плотность постоянна на $[a; b]$.

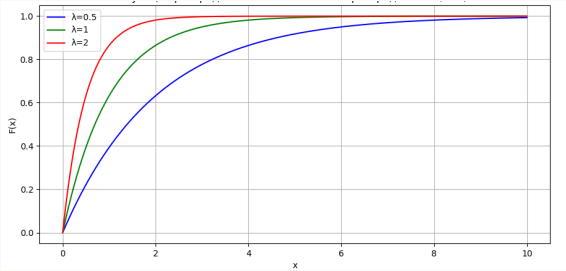
$$f_\xi(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & \text{иначе.} \end{cases}$$

$$F_\xi(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

Показательное распределение

Случайная величина ξ имеет показательное распределение $E(\alpha)$, если её плотность:

$$f_\xi(x) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \qquad \alpha > 0.$$



Нормальное распределение

Случайная величина ξ имеет нормальное распределение, если её плотность:

$$\xi \sim \mathcal{N}(a, \sigma^2), \quad \sigma > 0$$

$$f_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$$

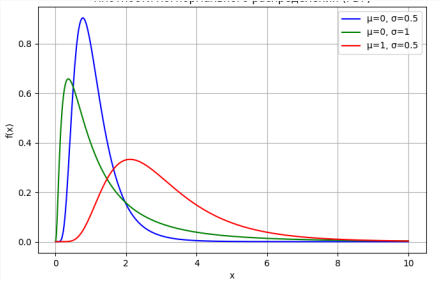
$$F_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-a)^2}{2\sigma^2}\right) dt$$

a --- среднее; σ --- среднеквадратичное отклонение.

log-Нормальное распределение

Если случайная величина X из нормального распределения, то e^X --- из log-нормального распределения.

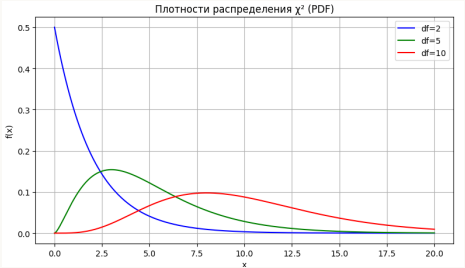
$$X \sim \mathcal{N}(a, \sigma^2), \qquad Y = e^X.$$



Распределение «хи-квадрат»

Распределением «хи-квадрат» $H(n)$ со степенями свободы n называется распределение суммы квадратов независимых стандартных нормальных величин:

$$\chi^2_n = \sum_{i=1}^{n} X_i^2, \qquad X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$



Распределение Стьюдента

Распределение с более тяжёлыми хвостами, чем нормальное:

$$t_k = \frac{X_0}{\sqrt{\chi_k^2/k}}, \qquad X_0 \sim \mathcal{N}(0,1).$$

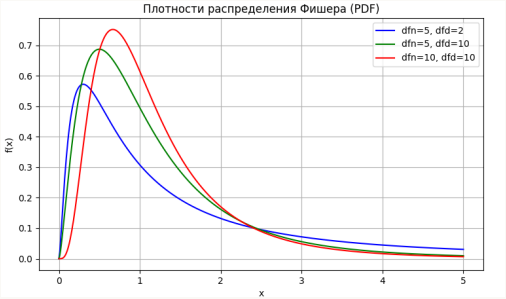
k --- число степеней свободы; X_0 и χ_k^2 независимы.

Распределение Фишера

Распределением Фишера $F(n, m)$ со степенями свободы n и m называется распределение случайной величины:

$$F_{n,m} = \frac{\chi^2_n / n}{\chi^2_m / m}$$

$$\chi^2_n \perp \chi^2_m$$



Математическое ожидание

Дискретный случай: взвешенное по вероятности среднее значение:

$$\mathbb{E}[X] = \sum_x x\mathbb{P}(X = x).$$

Непрерывный случай: взвешивание будем проводить через функцию плотности распределения:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\, dx.$$

Дисперсия, СКО

Дисперсия --- математическое ожидание квадрата отклонения случайной величины от её математического ожидания.

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \mathbb{E}[X^2] - \left(\mathbb{E}X\right)^2.$$

Дисперсия характеризует разброс случайной величины вокруг её математического ожидания. Среднеквадратическое отклонение:

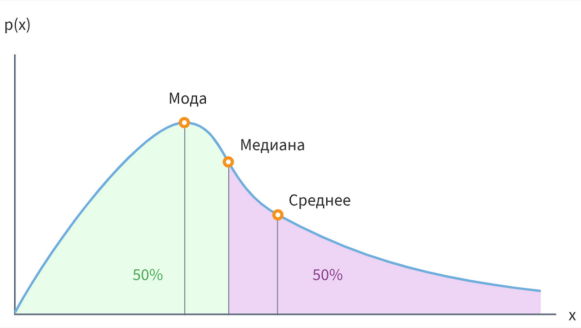
$$\sigma_X = \sqrt{\mathrm{Var}(X)}.$$

Мода, медиана, размах

Медиана --- значение, для которого значение функции распределения равно 0.5. То есть получение числа больше или меньше медианы равновероятно.

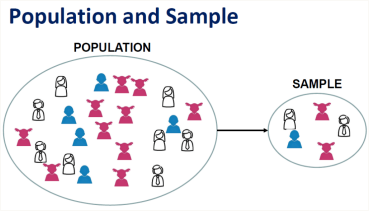
Мода --- самое вероятное значение.

Размах --- разность между максимальным и минимальным значением.

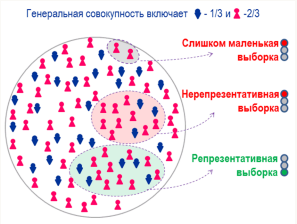


Генеральная совокупность и репрезентативность

Генеральная совокупность --- все объекты, о которых мы хотим сделать вывод. **Выборка** --- наблюдаемая часть этой совокупности.



Выборка репрезентативна, если отражает важные свойства и распределение генеральной совокупности.



Репрезентативность. Пример

Генеральная совокупность --- вес вообще всех родившихся за последние три года щенков пуделя.

Выборка --- опросили 10 заводчиков и собрали информацию о весе их щенков.

Является ли выборка репрезентативной?

Характеристики выборок

Выборочным средним называется величина:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Выборочной дисперсией называется величина:

$$D^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

Исправленной выборочной дисперсией называется величина:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}D^*.$$

EDA: признаки и пропуски

Числовые → категориальные → пропуски → практические примеры.

03

До графиков нужно понять, что именно лежит в таблице

01

Единица наблюдения

Что означает одна строка? Как она появилась?

02

Схема данных

Что означают столбцы, типы и единицы измерения?

03

Качество

Есть ли пропуски, дубли и невозможные значения?

04

Распределения и связи

Что типично, а что требует отдельной проверки?

EDA и очистка данных образуют цикл.

Найденная проблема возвращает нас к исходной таблице и к процессу сбора данных.

Источник: UC Berkeley Data 100, *Data Cleaning and Exploratory Data Analysis.*

Тип определяется смыслом, а не только dtype

order_id	temperature_c	items	postal_code	rating	delivery_speed
A-001	18.4	1	101000	5	express
A-002	21.0	3	190000	4	standard
A-003	19.7	0	101000	2	standard
A-004	20.2	2	443000	3	pickup

Числовые
temperature_c --- непрерывный; items --- дискретный.

Номинальные
postal_code и delivery_speed --- категории, хотя код может выглядеть числом.

Порядковые
rating имеет естественный порядок, но разность между оценками не обязана быть равной.

Источник: Google Machine Learning Crash Course, *Numerical Data* и *Categorical Data*.

Один универсальный график для всех столбцов не работает

Вопрос	Данные	Первый график	Сводка
Как распределено время ответа?	одна числовая	гистограмма / ECDF	медиана, квантили, хвосты
Какие категории встречаются чаще?	одна номинальная	bar plot	частоты и доли
Связаны ли цена и площадь?	две числовые	scatter plot	корреляция, условные сводки
Отличается ли цена между районами?	числовая + категория	boxplot / violin plot	групповые медианы

Цвет, порядок и масштаб осей должны отражать смысл переменной, а не случайный порядок значений в таблице.

Источник: UC Berkeley Data 100, *Visualization I*; Google Machine Learning Crash Course, *Numerical Data*.

Дискретные и непрерывные значения анализируем по-разному

Дискретные

Счётные значения: число покупок, ошибок, членов семьи.

- таблица частот и столбчатая диаграмма;
- доля нулей и редких значений;
- медиана, квантили, размах.

Непрерывные

Измерения: рост, температура, время ответа, концентрация.

- гистограмма, KDE и ECDF;
- среднее, медиана и квантили;
- boxplot, хвосты и потенциальные выбросы.

Числовой dtype не гарантирует числовой смысл: ID и коды категорий нельзя анализировать как числа.

Одинаково необычные точки могут требовать разных решений

Ошибка измерения или ввода --- сверяем с источником и ограничениями домена.

Редкая, но плотная группа может быть корректным отдельным режимом.

Низкая локальная плотность --- повод проверить контекст и влияние на модель.

Источник: Google ML Crash Course, *Numerical Data: First Steps*; scikit-learn, *Comparing anomaly detection algorithms on toy datasets.*

Описываем частотами, а не расстояниями

Признак	Примеры значений
house_color	red, blue, white
street_name	Oak St, Maple Ave
is_spam	yes, no
device_type	mobile, desktop, tablet

0, 1, 2 могут быть только метками. Из такого кода не следует, что blue ``между'' red и white.

Что проверяем

- частоты и доли категорий;
- редкие и новые значения;
- опечатки и дублирующиеся названия;
- связь категории с целевой переменной;
- различие между пропуском и категорией ``unknown''.

Источник: Google Machine Learning Crash Course, *Categorical Data* и *One-hot Encoding*.

Анализ категориальных переменных

Методы машинного обучения работают с числовыми значениями. Как сделать из категориальной переменной числовую?

color	color_code
red	0
green	1
blue	2

Почему это плохо?

Анализ категориальных переменных

One hot encoding --- это метод представления категориальных данных в виде бинарных векторов, где каждая уникальная категория кодируется отдельным разрядом: значение категории обозначается 1 в соответствующем разряде, а все остальные разряды принимают значение 0.

color	is_blue	is_green	is_red
red	0	0	1
green	0	1	0
blue	1	0	0

Анализ порядковых переменных

Порядковые переменные принимают значения с естественным порядком. При этом расстояния между значениями часто неизвестны и не всегда равны.

ID	Name	Education Level	Income Range
1	John Smith	High School	$20k-$40k
2	Alice Brown	Bachelor's	$40k-$60k
3	Bob White	Master's	$60k-$80k
4	Carol Green	PhD	$80k-$100k
5	Dave Blue	Bachelor's	$40k-$60k
6	Emma Black	High School	$20k-$40k

Пропуск — это информация о процессе сбора данных

Значение может отсутствовать по разным причинам, и от причины зависит стратегия обработки.

Случайная ошибка

Датчик не записал измерение; анкета потеряла поле.

Зависит от данных

Ответ чаще пропускают участники определённой группы.

Часть явления

Отсутствие анализа означает, что врач не счёл его нужным.

Сначала спрашиваем «почему нет значения?», и только потом «чем заполнить?».

Источник: UC Berkeley Data 100, *Data Cleaning and Exploratory Data Analysis.*

Стратегия зависит от причины пропуска

Возможные решения

- удалить строки, если пропусков мало и они действительно случайны;
- числовой признак заполнить медианой или модельной оценкой;
- категориальный признак заполнить модой или отдельной категорией;
- добавить индикатор пропуска, если сам факт отсутствия информативен;
- оставить NaN, если алгоритм явно умеет с ним работать.

После обработки проверяем

- не исчезла ли редкая группа;
- не изменились ли распределения;
- не возник ли искусственный сигнал;
- одинаков ли процесс на train и inference.

Источник: UC Berkeley Data 100; scikit-learn User Guide, *Imputation of missing values*.

Импутация обучается только на train

Правильно

Медиана, мода или параметры модели импутации вычисляются только по train.

Data leakage

Заполнять пропуски до разбиения — значит использовать информацию из test.

04

EDA на двух датасетах

Сначала Wine без пропусков, затем Penguins с числовыми, категориальными и незаполненными значениями.

Wine dataset: химический профиль трёх сортов вина

178
наблюдений

13
числовых признаков

3
класса

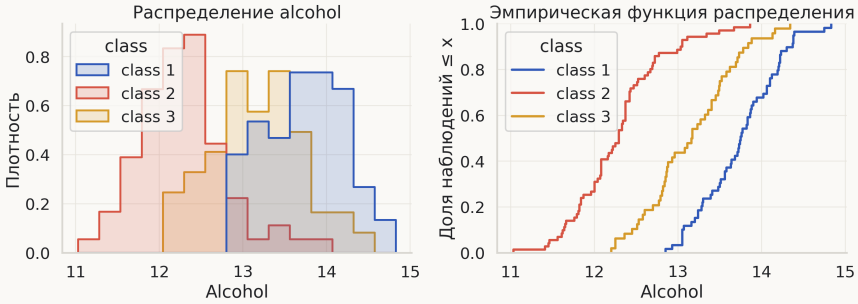
0
пропусков

Объект
Образец вина из одного региона Италии.

Признаки
Результаты химического анализа: alcohol, malic acid, flavanoids, color intensity, proline и др.

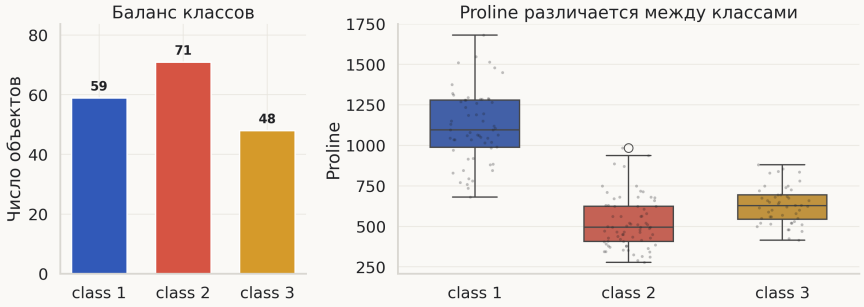
Источник: UCI Machine Learning Repository, Wine dataset; версия из `sklearn.datasets.load_wine`.

Гистограмма и ECDF отвечают на разные вопросы



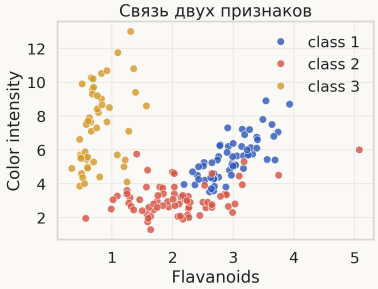
Гистограмма показывает форму и локальные пики; ECDF позволяет напрямую читать долю наблюдений ниже выбранного значения.

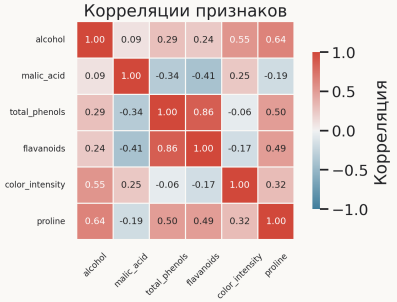
Перед сравнением групп проверяем их размер и разброс



Классы немного несбалансированы, а распределения proline заметно различаются — это уже гипотеза для будущей модели.

Связи между признаками видны до обучения модели





Высокая корреляция не означает причинность, но помогает заметить дублирующий сигнал и структуру классов.

Промежуточные выводы EDA

- структура таблицы согласована: 178 объектов, 13 числовых признаков, пропусков и дублей нет;
- классы различаются по ряду признаков, но имеют разные размеры;
- часть признаков сильно коррелирует — масштаб и избыточность потребуют внимания;
- в нескольких столбцах есть потенциальные выбросы;
- следующие решения: split, масштабирование, baseline и корректная метрика.

Результат EDA — набор гипотез и список рисков для следующего этапа эксперимента.

Второй практический пример

Palmer Penguins: разные типы признаков и реальные пропуски

344
наблюдения

7
признаков

3
вида

19
пропущенных значений

Числовые признаки
Размеры клюва и крыла, масса тела.

Категориальные признаки
Вид, остров и пол. Пропуски сосредоточены в поле *sex*.

Источник: Palmer Station Antarctica LTER; Horst, Hill, Gorman, *palmerpenguins*.

Сначала проверяем таблицу и смысл столбцов

Первые строки после минимальной очистки

	species	island	bill length	bill depth	flipper	mass	sex
0	Adelie	Torgersen	39.1	18.7	181	3750	male
1	Adelie	Torgersen	39.5	17.4	186	3800	female
2	Adelie	Torgersen	40.3	18.0	195	3250	female
4	Adelie	Torgersen	36.7	19.3	193	3450	female
5	Adelie	Torgersen	39.3	20.6	190	3650	male
6	Adelie	Torgersen	38.9	17.8	181	3625	female
7	Adelie	Torgersen	39.2	19.6	195	4675	male

344
наблюдения

7
признаков

19
пропущенных значений

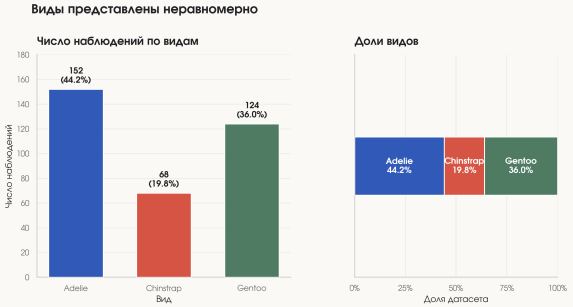
3
категориальных признака

Схема совпадает с исходным примером: четыре числовых и три категориальных признака, без дополнительного столбца year.

Пропуски проверяем не только по количеству

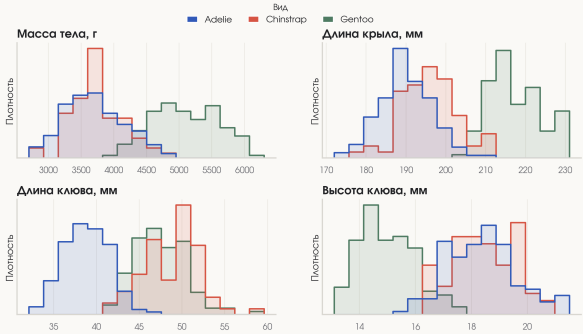
Девять строк содержат только пропуск пола, ещё две строки не содержат ни пола, ни числовых измерений.

Виды представлены неравномерно



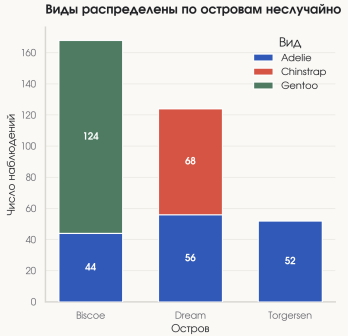
Частоты и доли категорий проверяем до разбиения данных и выбора метрики.

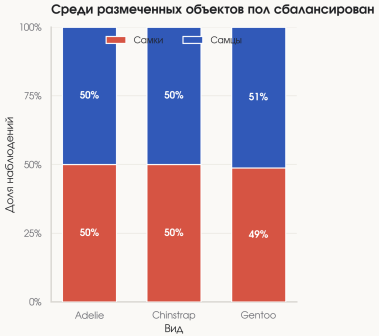
Распределения числовых признаков зависят от вида



Масса и длина крыла хорошо отделяют Gentoo, а геометрия клюва помогает различать Adelie и Chinstrap.

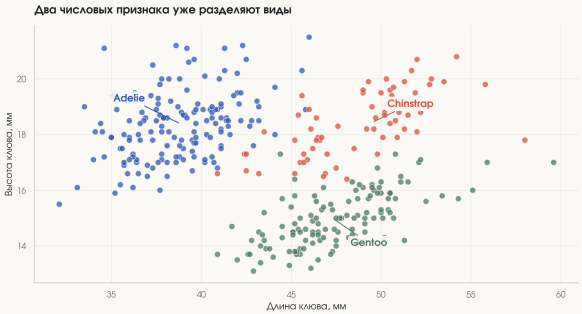
После частот проверяем связи категорий





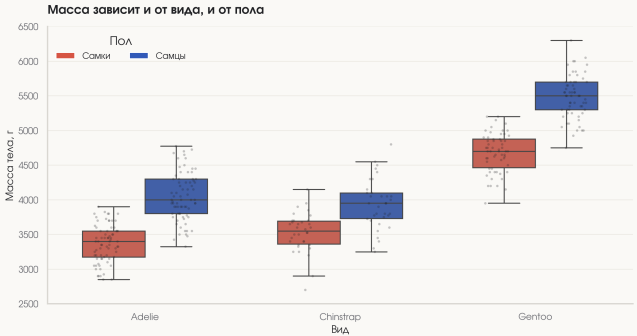
Вид связан с островом, поэтому случайное разбиение может скрыть географический сдвиг данных.

Связь двух признаков показывает структуру классов



Одномерные распределения перекрываются, но совместное пространство длины и высоты клюва уже разделяет виды.

При сравнении групп учитываем сразу несколько факторов



Масса зависит и от вида, и от пола: сравнение без стратификации смешало бы два разных эффекта.

Что мы узнали до обучения модели

- пропуски редкие, но их причина и распределение по группам требуют проверки;
- виды несбалансированы и неслучайно распределены по островам;
- числовые признаки имеют разные масштабы и формы распределений;
- двумерные связи дают более ясную структуру, чем отдельные гистограммы;
- пол влияет на массу и должен учитываться при интерпретации;
- следующий шаг — зафиксировать split, fit препроцессинга только на train и построить baseline.

EDA не «доказывает» гипотезу, а подсказывает, что проверить в честном эксперименте.

Ко(т)нец презентации

На следующей лекции:

Нормализация данных,
статистические гипотезы
и корреляции

Теоретический минимум

- AI, ML и Data Science; основные типы задач
- ML-пайплайн: от постановки задачи до инференса и обратной связи
- Функция распределения, плотность и ключевые распределения
- Математическое ожидание, дисперсия, медиана, мода и выборочные оценки
- Генеральная совокупность, выборка и репрезентативность
- EDA числовых, категориальных и порядковых признаков
- Выбросы и пропуски: диагностируем причину, затем обрабатываем
- Препроцессинг обучается только на train