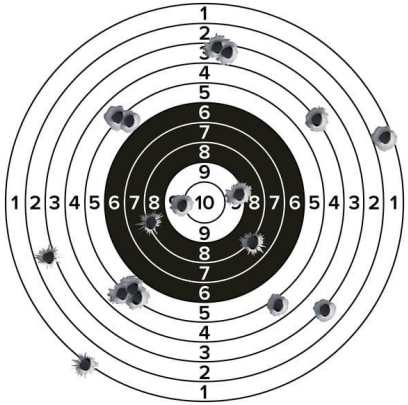
Теория вероятностей

Что такое теория вероятностей?

Случайные события - исходы эксперимента, результат которого невозможно точно предсказать

Теория вероятностей - раздел математики, которая изучает случайные события

Случайная величина — переменная, значения которой представляют собой численные исходы случайного эксперимента



Вероятность

Вероятность — количественная оценка возможности наступления некоторого события. Это величина измеряется от 0 до 1 и обозначается p.

Вероятность вычисляется как **число благоприятных исходов к общему числу исходов**

Вероятность выпадения 6 при бросании одной игральной кости?

Вероятность выпадения орла при одном бросании монеты?

Вероятность вытащить белый шар из урны с 5 белыми и 6 черными шарами?

Вероятность выпадения 7 очков (в сумме) при бросании двух игральных костей?

Зависимые и независимые события

События независимые, если вероятность наступления одного события не зависит от наступления другого

Зависимые наоборот

Выпадение 2 на первой игральной кости и выпадение 5 на второй?

Событие “Выпало 2” и “Выпало четное число”?

Сумма вероятностей одного множества исходов

Сумма вероятностей всех элементарных исходов будет равна 1

Бросок кубика: 6 исходов, вероятность каждого равна ?

P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1

Если событие имеет вероятность 1, то оно достоверное

Если событие имеет вероятность 0, то оно невозможное

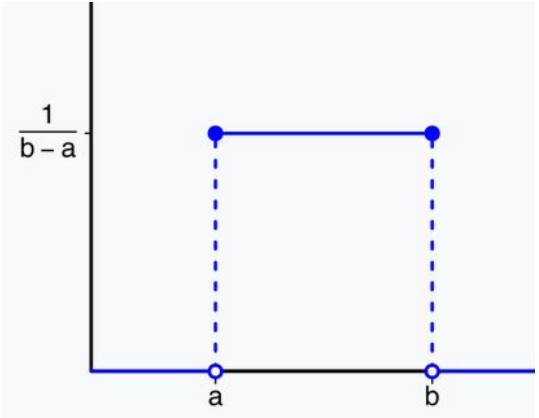
Условная вероятность

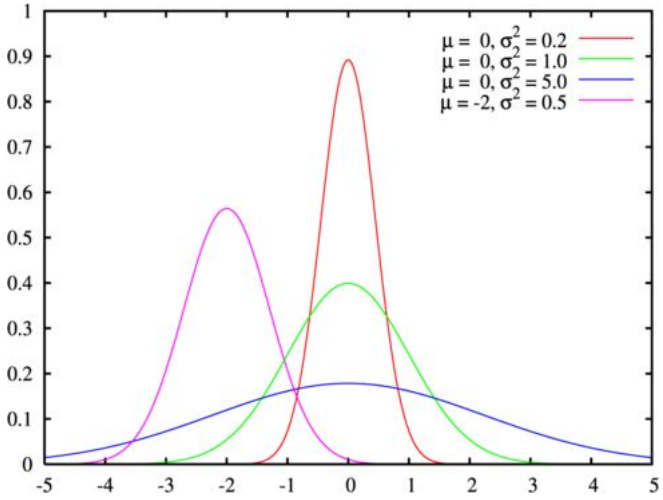
Это вероятность наступления одного события при условии, что наступило другое

Событие “Выпало четное число” и “Выпала 2”. Если наступило первое событие, какова вероятность наступления второго?

Закон распределения случайной величины

- нормальный
- равномерный
- экспоненциальное
- и другие





https://ru.wikipedia.org/wiki/Плотность_вероятности

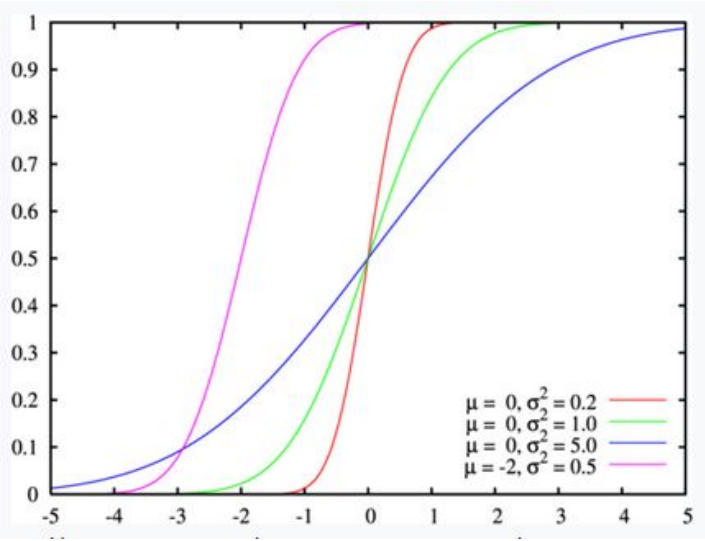
Характеристики

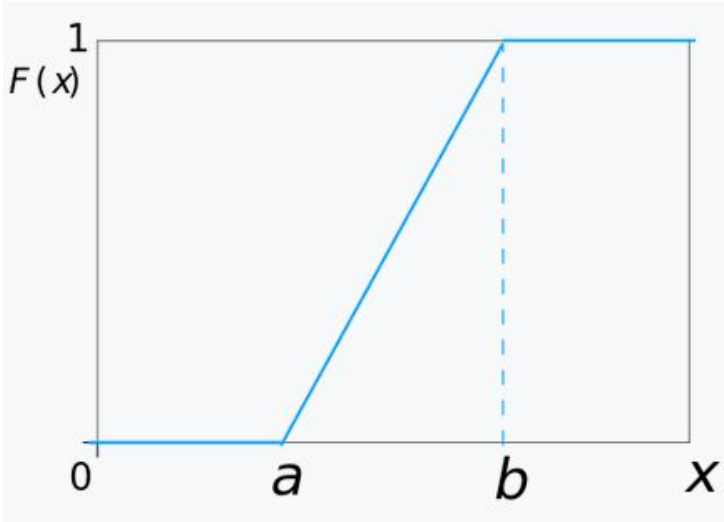
- математическое ожидание
- дисперсия
- функция плотности
- функция распределения
- и другие

Функция распределения

Характеризует распределение случайной величины.

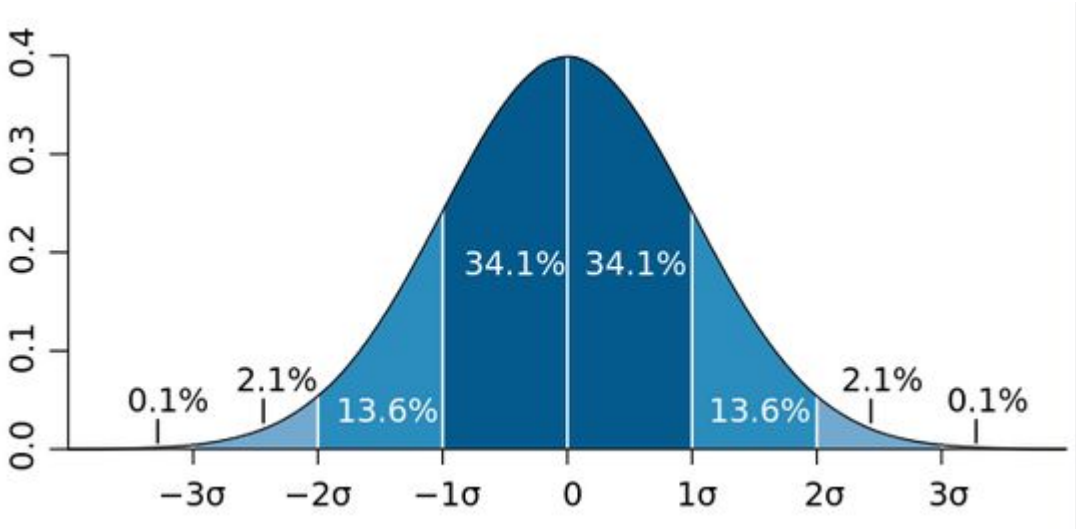
Показывает вероятность того, что случайная величина примет значение, меньшее заданного.





Функция плотности

Один из способов задания распределения случайной величины. Характеризует вероятность попадания значения случайной величины в определенный интервал

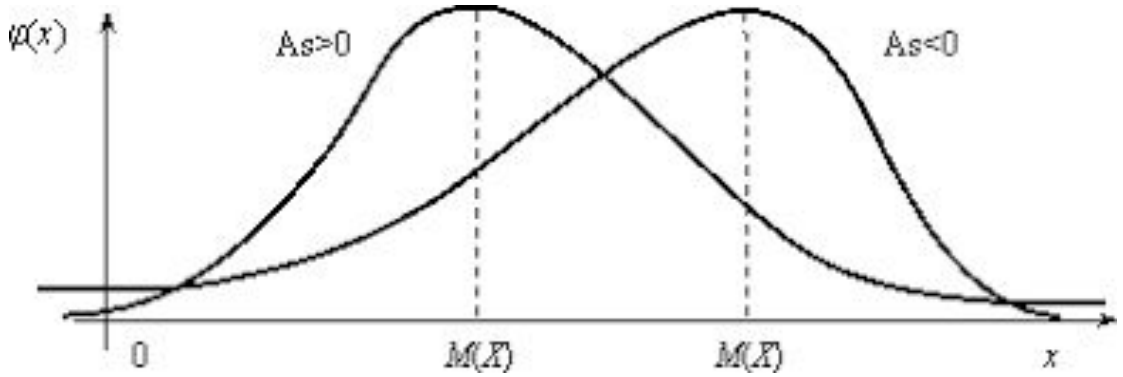


Математическое ожидание

Означает среднее (взвешенное по вероятностям возможных значений) значение случайной величины.

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \, p_i.$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i$$



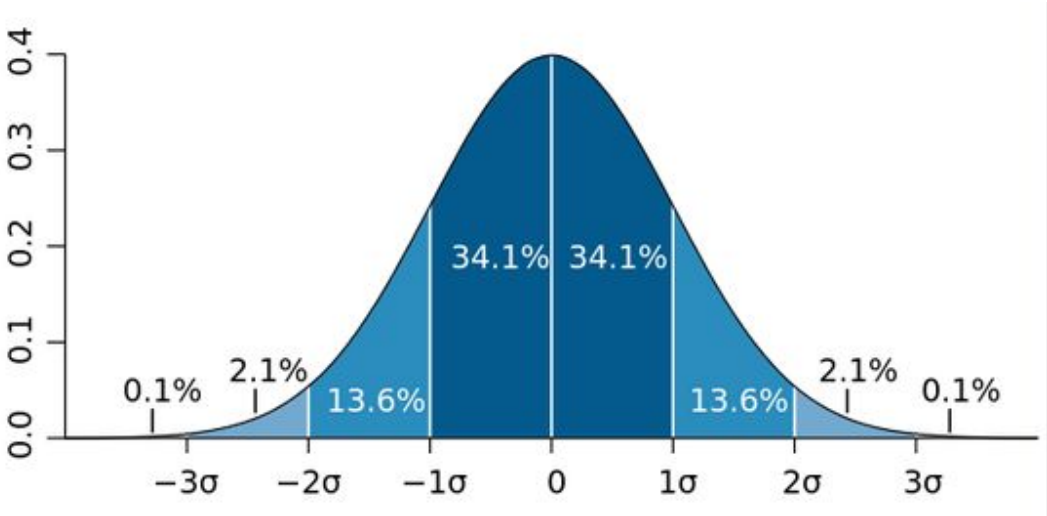
Дисперсия

Мера разброса значений случайной величины относительно её математического ожидания

Корень из дисперсии - это среднеквадратическое отклонение или стандартное отклонение (σ)

$$D[X] = \sum_{i=1}^{n} p_i (x_i - \mathbb{E}[X])^2,$$

Правило 3-х сигм?



`numpy.random`

https://numpy.org/doc/stable/reference/random/index.html

Параметры у равномерного распределения:

- low - левая граница интервала
- high - правая граница интервала

Параметры у нормального распределения:

- loc - среднее/математическое ожидание
- scale - стандартное отклонение

Элементы статистики

Статистика — наука, в которой излагаются общие вопросы сбора, измерения, мониторинга, анализа массовых статистических данных и их сравнение

Как работать с выборками (Статистика)

Генеральная совокупность — совокупность всех объектов, относительно которых предполагается делать выводы при изучении конкретной задачи

Пример: все больные ОРВИ определенного региона

Выборка — часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом)

Пример: Больные ОРВИ в одной больнице

Оценки для математического ожидания и дисперсии

Оценка математического ожидания или выборочное среднее:

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

Оценка дисперсии:

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

Правило больших чисел: чем больше измерений, тем ближе оценка к теоретическому результату

Способы вычисления в Python

Стандартными средствами:

https://docs.python.org/3/library/statistics.html

`numpy`:

- `mean` - среднее (оценка математического ожидания)
- `std` - стандартное отклонение
- `var` - оценка дисперсии

`pandas`:

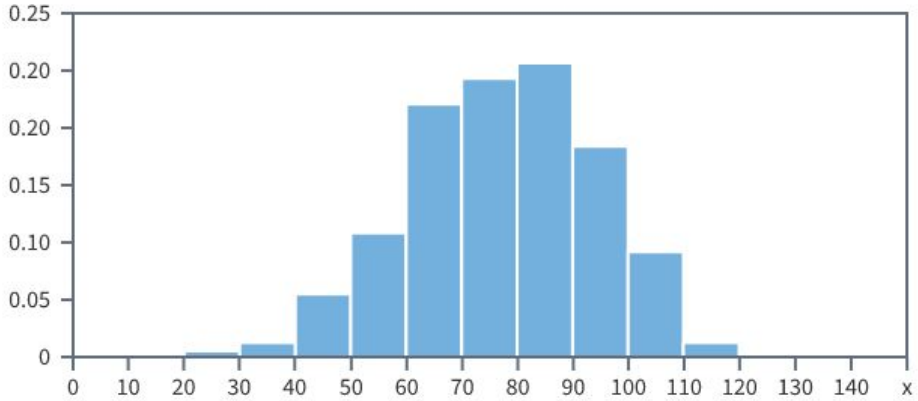
- Отдельно: `mean`, `std`, `var`
- Все вместе: `describe`

Гистограмма

Показывает количество (частоту) объектов попавших в определенный интервал.

По оси X - интервалы

По оси Y - частота или количество



Как построить в Python?

`matplotlib` - `hist`

`seaborn` - `displot`, `histplot`

`pandas` - `hist`

Формула Байеса

Формула Байеса

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

P(A) - вероятность события A

P(B) - вероятность события B

P(B|A) - вероятность наступления B при истинности A

P(A|B) - вероятность события A при наступлении события B

Как работает?

Формула Байеса позволяет переставить причину и следствие: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной

Часто применяется для задачи классификации. Например:

Имеется набор писем: спам и не спам. Подсчитаем для каждого слова вероятность встречи в спаме, количество в спаме ко всему количеству в тексте. Аналогично для слов из не спама. И на основе формулы Байеса делаем вывод куда отнести письмо

Классификация

Задача, в которой имеется множество объектов, разделенных на классы

Дана выборка - множество объектов, для которых известно, к каким классам относятся объекты

Классовая принадлежность остальных объектов неизвестна

Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества

Байесовский классификатор

На основе выборки можно определить:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

- вероятности классов (в формуле P(A))
- условные вероятности того что истинен определенный класс при условии, что признак принял определенное значение (P(B|A))

Далее остается только вычислить произведения и выбрать тот класс, где значение максимально

Ремарка: P(B) (знаменатель) будет одинаков при расчетах, поэтому его можно отбросить

Пример

Даны два класса и один признак, который может принимать значения -1, 0, 1.

По выборке определили вероятности классов и условные вероятности

Предположим у нового объекта, класс которого требуется определить, признак равен 0

x_i	-1	0	1	$P(j)$
$p_{x_i\|1}$	0.05	0.8	0.15	0.5
$p_{x_i\|2}$	0.1	0.2	0.7	0.5

Пример

Вычисляем по формуле Байеса

вероятности для каждого класса

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)},$$

x_i	-1	0	1	$P(j)$
$p_{x_i\mid 1}$	0.05	0.8	0.15	0.5
$p_{x_i\mid 2}$	0.1	0.2	0.7	0.5

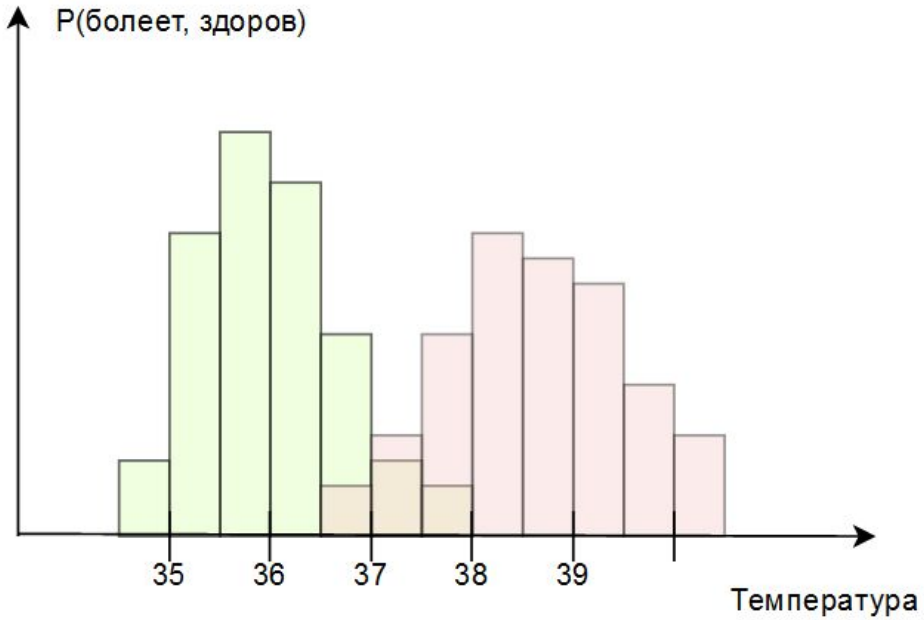
$$P(1 \mid X=0) = \frac{p_{0\mid 1}P(1)}{p_{0\mid 1}P(1) + p_{0\mid 2}P(2)} = \frac{0.8\cdot 0.5}{0.8\cdot 0.5 + 0.2\cdot 0.5} = \frac{0.4}{0.5} = 0.8\ ,$$

$$P(2 \mid X=0) = \frac{p_{0\mid 2}P(2)}{p_{0\mid 1}P(1) + p_{0\mid 2}P(2)} = \frac{0.2\cdot 0.5}{0.8\cdot 0.5 + 0.2\cdot 0.5} = \frac{0.1}{0.5} = 0.2\,.$$

Правило 37 градусов

Зеленый - здоров

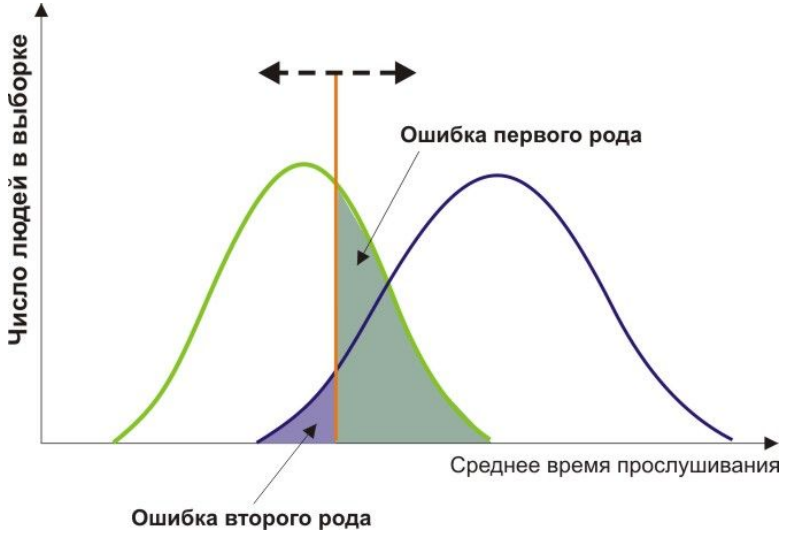
Красный - болен



Ошибки

Ошибка первого рода - ложноположительное срабатывание

Ошибка второго рода - ложноотрицательное срабатывание



	Реальность		
Предсказание		Гипотеза верна	Гипотеза ложна
	Мы приняли гипотезу	Верное решение	Ошибка первого рода
	Мы отвергли гипотезу	Ошибка второго рода	Верное решение

Байесовский классификатор в Python

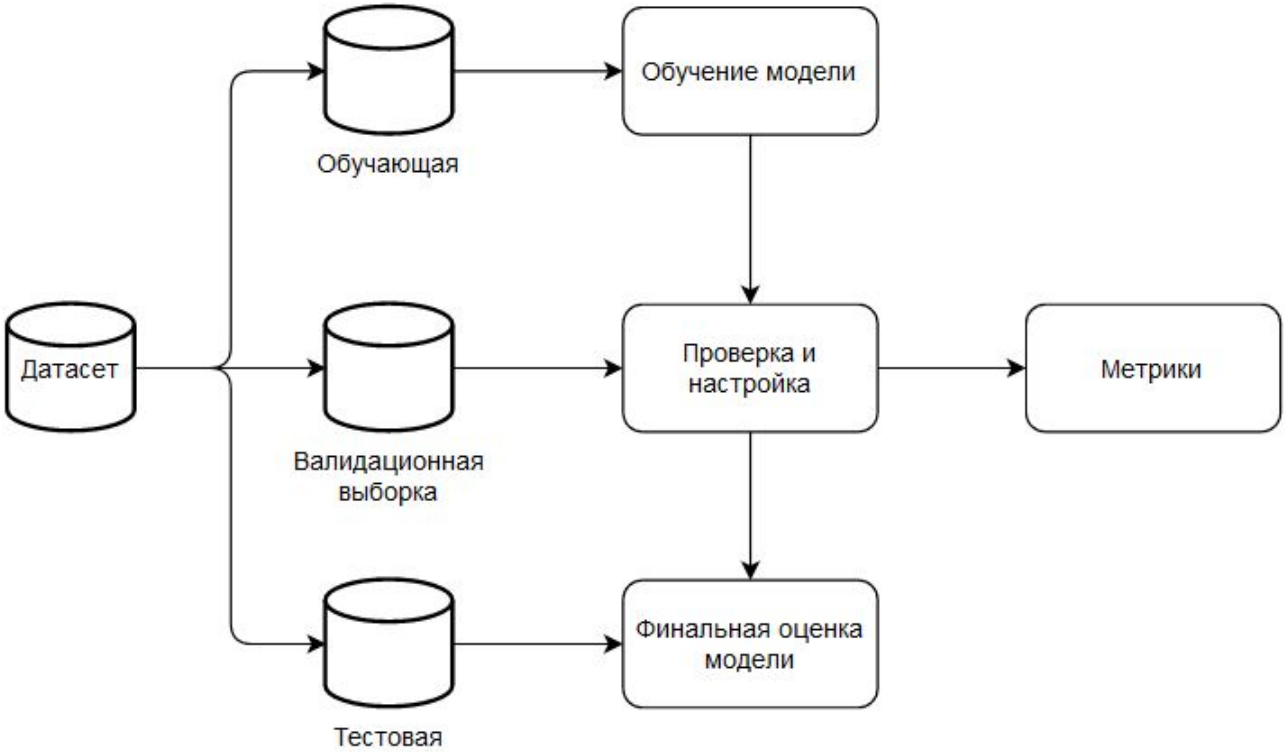
В `sklearn`

https://scikit-learn.ru/1-9-naive-bayes/

Туториал:

https://www.datacamp.com/tutorial/naive-bayes-scikit-learn

Рабочий процесс



https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Достоинства наивного Байеса

Высокая скорость работы (даже на больших наборах данных)

Требуется небольшой объем обучающих данных

Недостатки наивного Байеса

Предполагает независимость признаков, что может быть не всегда верно