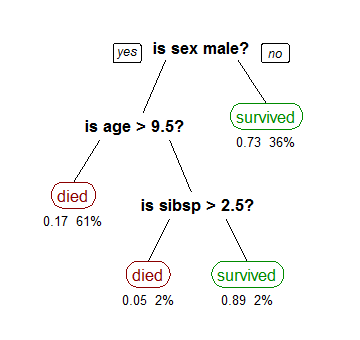
**Дерево принятия решений** (также может называться деревом классификации или регрессионным деревом) — средство поддержки [принятия решений](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BE%D1%80%D0%B8%D1%8F_%D0%BF%D1%80%D0%B8%D0%BD%D1%8F%D1%82%D0%B8%D1%8F_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B9), использующееся в [статистике](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0) и [анализе данных](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85) для прогнозных моделей. Структура дерева представляет собой «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения [целевой функции](https://ru.wikipedia.org/wiki/%D0%A6%D0%B5%D0%BB%D0%B5%D0%B2%D0%B0%D1%8F_%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F), а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение.



Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.

Типология деревьев

Деревья решений, используемые в [Data Mining](https://ru.wikipedia.org/wiki/Data_Mining" \o "Data Mining), бывают двух основных типов:

* Дерево для классификации, когда предсказываемый результат является классом, к которому принадлежат данные;
* Дерево для регрессии, когда предсказываемый результат можно рассматривать как вещественное число (например, цена на дом, или продолжительность пребывания пациента в больнице).

Упомянутые выше термины впервые были введены [Брейманом](https://ru.wikipedia.org/wiki/%D0%91%D1%80%D0%B5%D0%B9%D0%BC%D0%B0%D0%BD,_%D0%9B%D0%B5%D0%BE" \o "Брейман, Лео) и др.[[2]](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D0%BF%D1%80%D0%B8%D0%BD%D1%8F%D1%82%D0%B8%D1%8F_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B9#cite_note-autogenerated1-2) Перечисленные типы имеют некоторые сходства (рекурсивный алгоритмы построения), а также некоторые различия, такие, как критерии выбора разбиения в каждом узле.[[2]](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D0%BF%D1%80%D0%B8%D0%BD%D1%8F%D1%82%D0%B8%D1%8F_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B9#cite_note-autogenerated1-2)

Некоторые методы позволяют построить более одного дерева решений (ансамбли деревьев решений):

1. [Бэггинг](https://en.wikipedia.org/wiki/Bootstrap_aggregating) над деревьями решений, наиболее ранний подход. Строит несколько деревьев решений, неоднократно интерполируя данные с заменой ([бутстреп](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B1%D1%83%D1%82%D1%81%D1%82%D1%80%D1%8D%D0%BF" \o "Статистический бутстрэп)), и в качестве консенсусного ответа выдаёт результат голосования деревьев (их средний прогноз);[[3]](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D0%BF%D1%80%D0%B8%D0%BD%D1%8F%D1%82%D0%B8%D1%8F_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B9#cite_note-3)
2. **Классификатор «Случайный лес»** основан на бэггинге, однако в дополнение к нему случайным образом выбирает подмножество признаков в каждом узле, с целью сделать деревья более независимыми;
3. [Бустинг](https://en.wikipedia.org/wiki/Boosting_(machine_learning)) над деревьями может быть использован для задач как регрессии, так и классификации.[[4]](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D0%BF%D1%80%D0%B8%D0%BD%D1%8F%D1%82%D0%B8%D1%8F_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B9#cite_note-4) Одна из реализаций бустинга над деревьями, алгоритм [XGBoost](https://en.wikipedia.org/wiki/Xgboost" \o "en:Xgboost), неоднократно использовался победителями соревнований по анализу данных.
4. «Вращение леса» — деревья, в которых каждое дерево решений анализируется первым применением метода главных компонент (PCA) на случайные подмножества входных функций.[[5]](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D0%BF%D1%80%D0%B8%D0%BD%D1%8F%D1%82%D0%B8%D1%8F_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B9#cite_note-5)

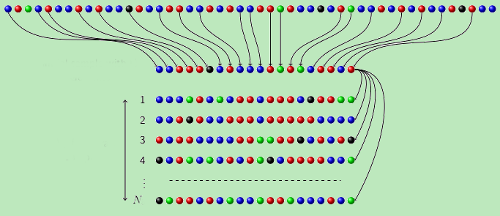
**Статистический бутстрэп** (*бутстреп*, *бутстрэппинг*, [англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *bootstrap*, *bootstrapping*) — практический компьютерный метод исследования распределения статистик [вероятностных распределений](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9), основанный на многократной генерации выборок [методом Монте-Карло](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%9C%D0%BE%D0%BD%D1%82%D0%B5-%D0%9A%D0%B0%D1%80%D0%BB%D0%BE) на базе имеющейся выборки[[1]](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B1%D1%83%D1%82%D1%81%D1%82%D1%80%D1%8D%D0%BF#cite_note-1)

Предложен в [1977 году](https://ru.wikipedia.org/wiki/1977_%D0%B3%D0%BE%D0%B4_%D0%B2_%D0%BD%D0%B0%D1%83%D0%BA%D0%B5) [Брэдли Эфроном](https://ru.wikipedia.org/wiki/%D0%AD%D1%84%D1%80%D0%BE%D0%BD,_%D0%91%D1%80%D1%8D%D0%B4%D0%BB%D0%B8) (первая публикация относится к [1979 году](https://ru.wikipedia.org/wiki/1979_%D0%B3%D0%BE%D0%B4_%D0%B2_%D0%BD%D0%B0%D1%83%D0%BA%D0%B5)[[2]](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B1%D1%83%D1%82%D1%81%D1%82%D1%80%D1%8D%D0%BF#cite_note-.D0.AD.D1.84.D1.80.D0.BE.D0.BD.E2.80.941979.E2.80.94.E2.80.94-2)). Суть метода состоит в том, чтобы по имеющейся выборке построить [эмпирическое распределение](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D0%BE%D1%80%D0%BE%D1%87%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5). Используя это распределение как теоретическое распределение вероятностей можно с помощью датчика псевдослучайных чисел сгенерировать практически неограниченное количество псевдовыборок произвольного размера, например, того же как у исходной. На множестве псевдовыборок можно оценить не только анализируемые статистические характеристики, но и изучить их вероятностные распределения. Таким образом, например, оказывается возможным оценить дисперсию или [квантили](https://ru.wikipedia.org/wiki/%D0%9A%D0%B2%D0%B0%D0%BD%D1%82%D0%B8%D0%BB%D1%8C) любой статистики независимо от её сложности. Данный метод является методом непараметрической статистики.

**Метод имитации статистического выбора**

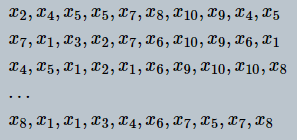
Начиная с середины прошлого века инициированные развитием вычислительных технологий вышли в свет работы (Эфрон Б., Тьюки Дж. и др.), посвященные так называемому ресэмплингу — генерации дополнительных выборок из уже имеющихся. Одним из таких подходов, распространенных в век компьютерных технологий, является bootstrap-метод, или метод имитации статистического выбора.

Суть  метода заключается в формировании множества выборок, на основе случайного выбора с повторениями. Если о законе распределения выборки нет никаких априорных сведений, а получить оценки его характеристики все-таки необходимо, то bootstrap-метод может быть весьма полезным инструментом. Рассмотрим его работу на примере вычисления среднего значения.



Допустим, имеется набор x1,x2,…,x10 значений, на основе которых необходимо оценить среднее.  На языке "математической статистики" данная задача звучит так: *имеется выборка из некоторого неизвестного закона распределения; необходимо оценить математическое ожидание  этого распределения и найти доверительные границы  оценки.*

Метод виртуальных выборок (bootstrap-метод) заключается в том, что на основе исходного набора x1,x2,…,x10 с использованием процедуры случайного выбора с повторениями формируются наборы вида:



Таким образом, создается множество виртуальных выборок.  Вычисленные средние значения для каждой из таких виртуальных выборок будут являться в статистическом смысле оценками искомого математического ожидания.

Допустим мы сформировали 1000 выборок и нашли для каждого набора среднее значение. Обозначим их m1,m2,…,m1000. Поскольку количество элементов в каждой из выборок одинаково, найденные средние значения представляют собой выборку из некоторого закона распределения вероятностей, который достаточно хорошо представлен (имеется 1000 выборочных значений!).  Медиана полученной выборки может служить оценкой неизвестного математического ожидания, а соответствующие ей проценитили (если задаться доверительным уровнем вероятности) будут границами искомого доверительного интервала.