* **C4.5** — усовершенствованная версия алгоритма ID3, в которую добавлена возможность работы с пропущенными значениями атрибутов (по версии издания Springer Science в 2008 году алгоритм занял 1-е место в топ-10 наиболее популярных алгоритмов Data Mining).
* **CART (Classification and Regression Tree)** — алгоритм обучения деревьев решений, позволяющий использовать как дискретную, так и непрерывную целевую переменную, то есть решать как задачи классификации, так и регрессии. Алгоритм строит деревья, которые в каждом узле имеют только два потомка.

**Преимущества алгоритма**

Рассмотрев основные проблемы, возникающие при построении деревьев, было бы несправедливо не упомянуть об их достоинствах:

* быстрый процесс обучения;
* генерация правил в областях, где эксперту трудно формализовать свои знания;
* извлечение правил на естественном языке;
* интуитивно понятная классификационная модель;
* высокая точность предсказания, сопоставимая с другими методами анализа данных (статистика, нейронные сети);
* построение непараметрических моделей.

**Недостатки:**

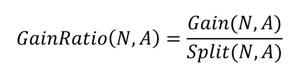
* Деревья решений чувствительны к шумам во входных данных. Небольшие изменения обучающей выборки могут привести к глобальным корректировкам модели, что скажется на смене правил классификации и интерпретируемости модели.
* Разделяющая граница имеет определенные ограничения, из-за чего дерево решений по качеству классификации уступает другим методам.
* Возможно переобучение дерева решений, из-за чего приходится прибегать к методу «отсечения ветвей», установке минимального числа элементов в листьях дерева или максимальной глубины дерева.
* Сложный поиск оптимального дерева решений: это приводит к необходимости использования эвристики типа жадного поиска признака с максимальным приростом информации, которые в конечном итоге не дают 100-процентной гарантии нахождения оптимального дерева.
* Дерево решений делает константный прогноз для объектов, находящихся в признаковом пространстве вне параллелепипеда, который охватывает не все объекты обучающей выборки.

**Области применения**

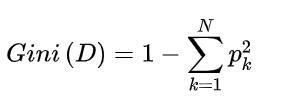
Модули для построения и исследования деревьев решений входят в состав большинства аналитических платформ. Они являются удобным инструментом в [системах поддержки принятия решений](https://wiki.loginom.ru/articles/decision-support-system.html) и интеллектуального анализа данных.

Деревья решений успешно применяются на практике в следующих областях:

* **Банковское дело.** Оценка [кредитоспособности](https://wiki.loginom.ru/articles/creditworthiness.html) клиентов банка при выдаче кредитов.
* **Промышленность.** Контроль за качеством продукции (выявление дефектов), испытания без разрушений (например, проверка качества сварки) и т.д.
* **Медицина.** Диагностика заболеваний.
* **Молекулярная биология.** Анализ строения аминокислот.
* **Торговля.** Классификация клиентов и товаров.

Единственное различие между C4.5 и ID3 - это критерии выбора атрибутов разделения. Чтобы решить проблему, заключающуюся в том, что ID3 отдает предпочтение атрибутам с большим количеством возможных значений, C4.5 вводит коэффициент передачи информации и использует принцип максимальной скорости набора информации для выбора текущего атрибута разделения узла.  
Используйте атрибут A для разделения узла NСкорость сбора информации:  
  
Предполагая, что атрибут A имеет v различных значений дискретных атрибутов, атрибут A можно использовать для разделения узла N на v дочерних узлов {N1, N2,…, Nv}. После разделения узла N с атрибутом AЭнтропия раздела：  


Алгоритмы ID3 и C4.5 основаны на информационной энтропии для выбора узлов разделов, которые в основном используются для задач классификации. Дерево решений CART называется деревом классификации и регрессии (Дерево классификации и регрессии), могут использоваться как задачи классификации, так и регрессии. Разница между CART и двумя предыдущими алгоритмами заключается в том, что, когда каждый узел выносит суждение, рассматривается только случай двух классификаций, даже если можно получить несколько значений (например, есть три цветных шара, ID3 и C4. 5 напрямую разделены на три подкатегории, а КОРЗИНА может быть разделена только на определенный цвет, а затем оценена, когда она разделена один раз), на самом деле это двоичное дерево.

Стандарт классификации CART заключается в использовании индекса Джини для разделения:  
Индекс Джини：

1. Определения в алгоритме DBSCAN:

Чтобы понять DBSCAN более подробно, давайте углубимся в это.***Основная концепция алгоритма DBSCAN состоит в том, чтобы найти области высокой плотности, которые отделены друг от друга областями низкой плотности.***Итак, как мы измеряем плотность региона? Ниже приведены 2 шага -

* Плотность в точке P: количество точек в окружности радиуса*EPS (ϵ)*из точки P*,*
* Плотная область: для каждой точки в кластере окружность с радиусом ϵ содержит как минимум минимальное количество точек (*MinPts*).

2. Этапы алгоритма DBSCAN:

С определениями выше, мы можем пройти через шаги алгоритма DBSCAN, как показано ниже:

1. Алгоритм начинается с произвольной точки, которая не была посещена, и информация о его окрестности извлекается из параметра..
2. Если этот пункт содержит*MinPts*в ϵ окрестности начинается формирование кластера. В противном случае точка помечается как шум Эта точка может быть позже найдена в ϵ-окрестности другой точки и, таким образом, может стать частью кластера. Здесь важна концепция достижимости плотности и точек, связанных плотностью.
3. Если точка найдена как центральная точка, то точки в окрестности also также являются частью кластера. Таким образом, все точки, найденные в ϵ окрестности, добавляются вместе с их собственной ϵ окрестностью, если они также являются центральными точками.
4. Вышеописанный процесс продолжается до тех пор, пока кластер, связанный плотностью, не будет найден полностью.
5. Процесс возобновляется с новой точкой, которая может быть частью нового кластера или помечена как шум.

Из приведенных выше определений и алгоритмов вы можете догадаться*два из самых больших недостатков алгоритма DBSCAN*,

* Если в базе данных есть точки данных, которые образуют кластеры различной плотности, то DBSCAN не удается хорошо кластеризовать точки данных, поскольку кластеризация зависит от ϵ и*MinPts*параметр, они не могут быть выбраны отдельно для всех кластеров.
* Если данные и функции не так хорошо понятны специалисту в области, то настройте ϵ и*MinPts*может быть сложно и, возможно, потребуется сравнение для нескольких итераций с различными значениями ϵ и*MinPts*,

**a/b**

Выбор размера выборки

Мощность теста (1 - β) - это вероятность обнаружения статистической разницы между группами в нашем тесте, когда разница действительно присутствует. Обычно по соглашению устанавливается на 0,8 (если вам интересно, вот дополнительная информация о статистической мощности).

Альфа-значение (α) - критическое значение, которое мы установили ранее равным 0,05.

Размер эффекта - насколько большой, по нашему мнению, будет разница между коэффициентами конверсии.

Сбор и подготовка данных

Убираем повторяющие записи

Разделяем датасет на группы

Первое, что мы можем сделать, это вычислить базовую статистику, чтобы получить представление о том, как выглядят наши образцы.