



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

АЛГОРИТМЫ КЛАССИФИКАЦИИ: KNN И РЕШАЮЩИЕ ДЕРЕВЬЯ

Теванян Элен
31.05.2019

Москва, 2019

Задача классификации



КЛАССИФИКАЦИЯ. ПРИМЕРЫ

- Предсказание пола для неизвестного пользователя
- Определение типа документа
- Определение языка документа
- Определение эмоционального окраса отзыва
- Вероятность ухода сотрудника/клиента
- Предсказание типов писем: спам/не спам
- Определение объектов на фотографии
- Оценка состояния человека по ЭЭГ



ПОСТАНОВКА ЗАДАЧИ

- Задача: найти алгоритм по прецедентам, который будет каждому новому объекту приписывать метку класса
- x_i – объект (*потребитель*)
- y_i – целевая переменная (*сегмент*): категориальная переменная
- (x_i, y_i) – прецедент
- Обучающая выборка – набор всех прецедентов



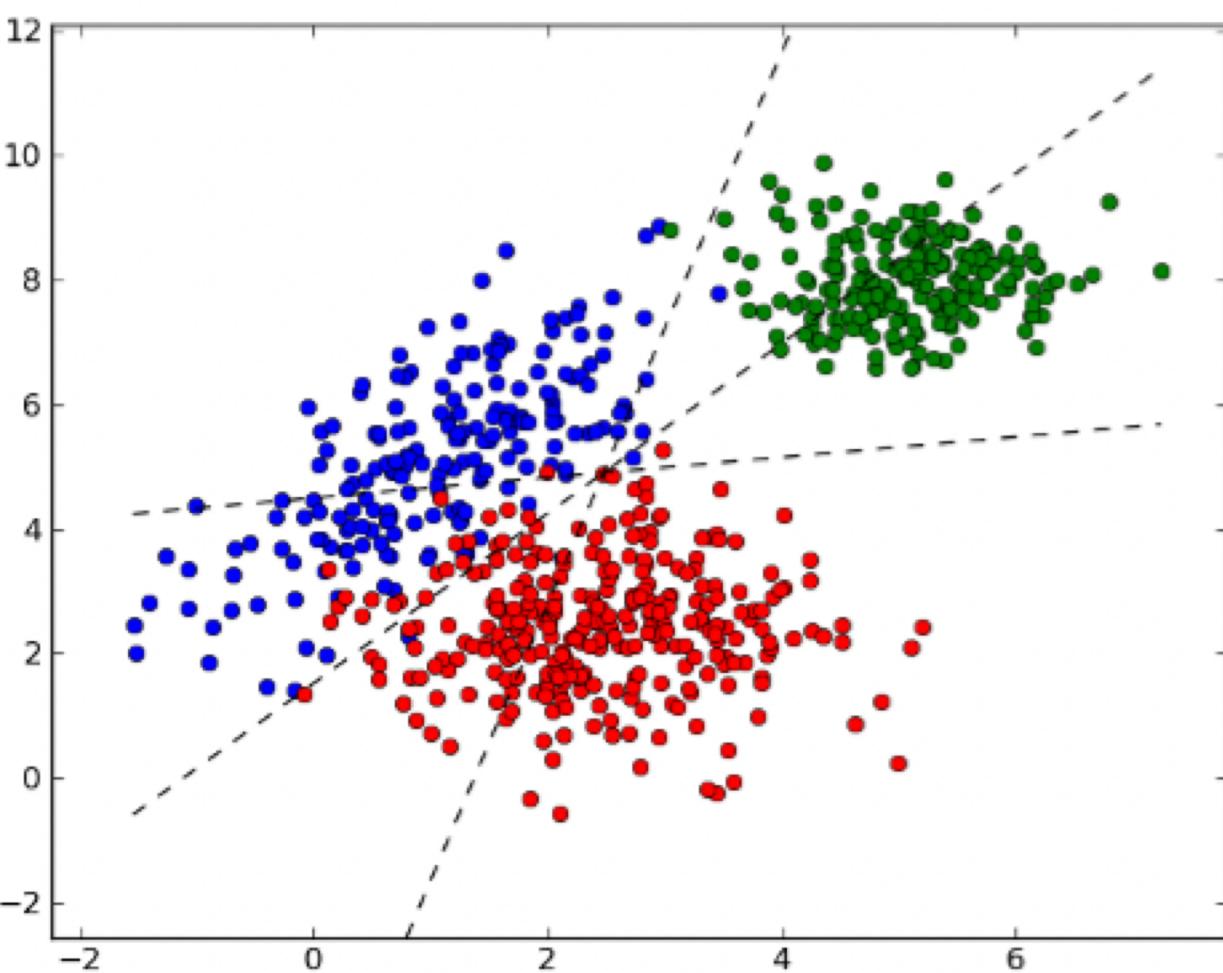
ТИПЫ КЛАССИФИКАЦИИ

- Бинарная: существует только два класса, например: мужчина или женщина



ТИПЫ КЛАССИФИКАЦИИ

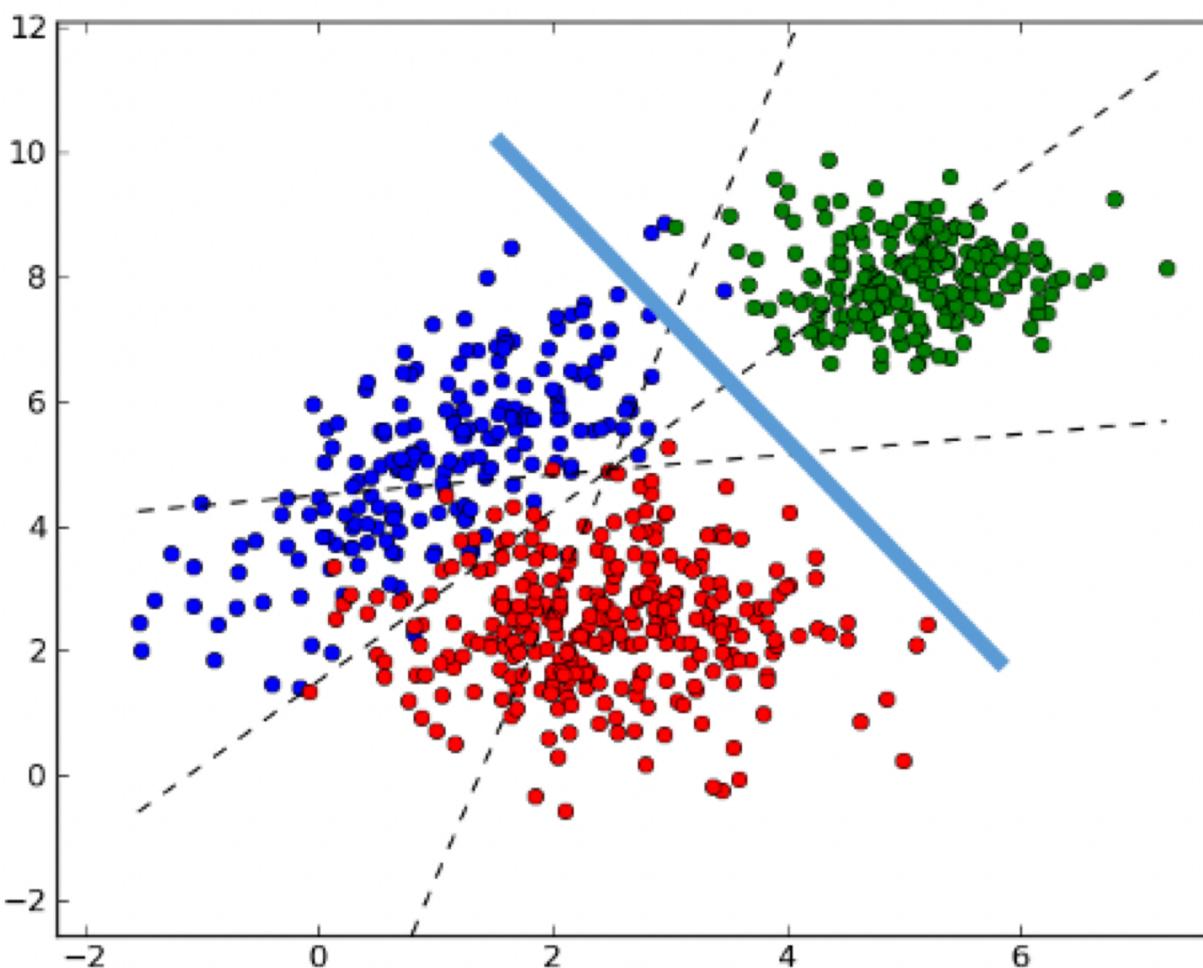
- Многоклассовая: классов несколько, например, три сегмента потребителей: Mass, Affluent, Premium





ТИПЫ КЛАССИФИКАЦИИ

- One-vs-all: тренируем классификатор распознавать один класс против всех остальных. А потом тренируем на остальных данных разделять другие классы.



Алгоритмы классификации



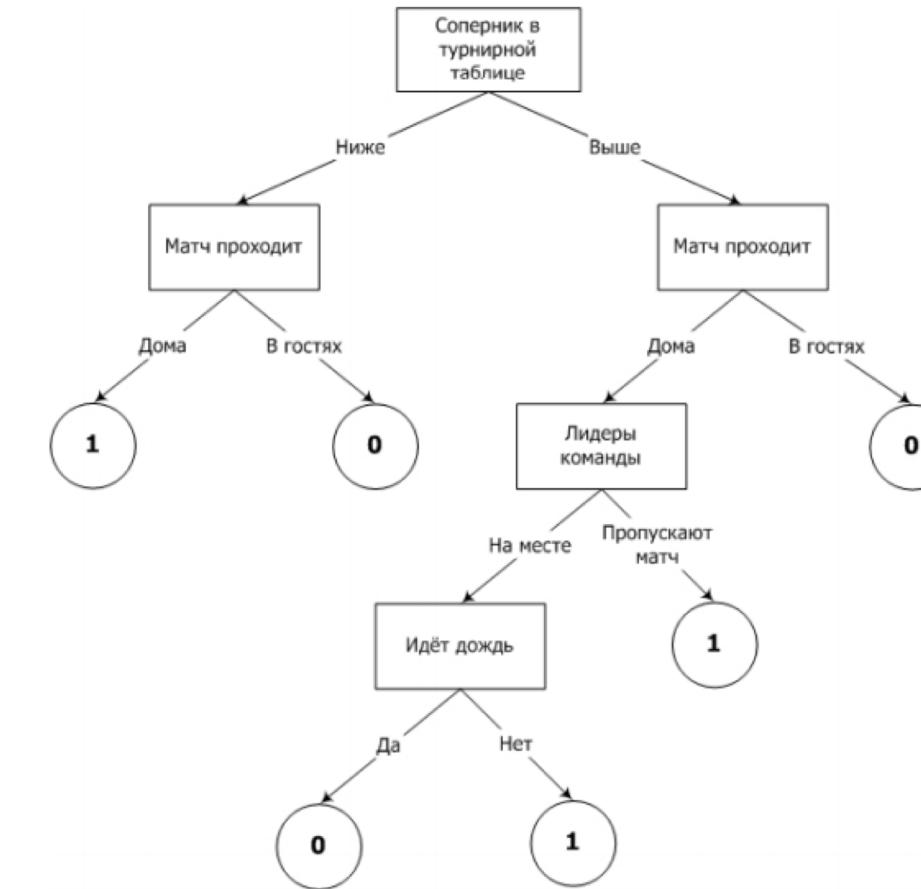
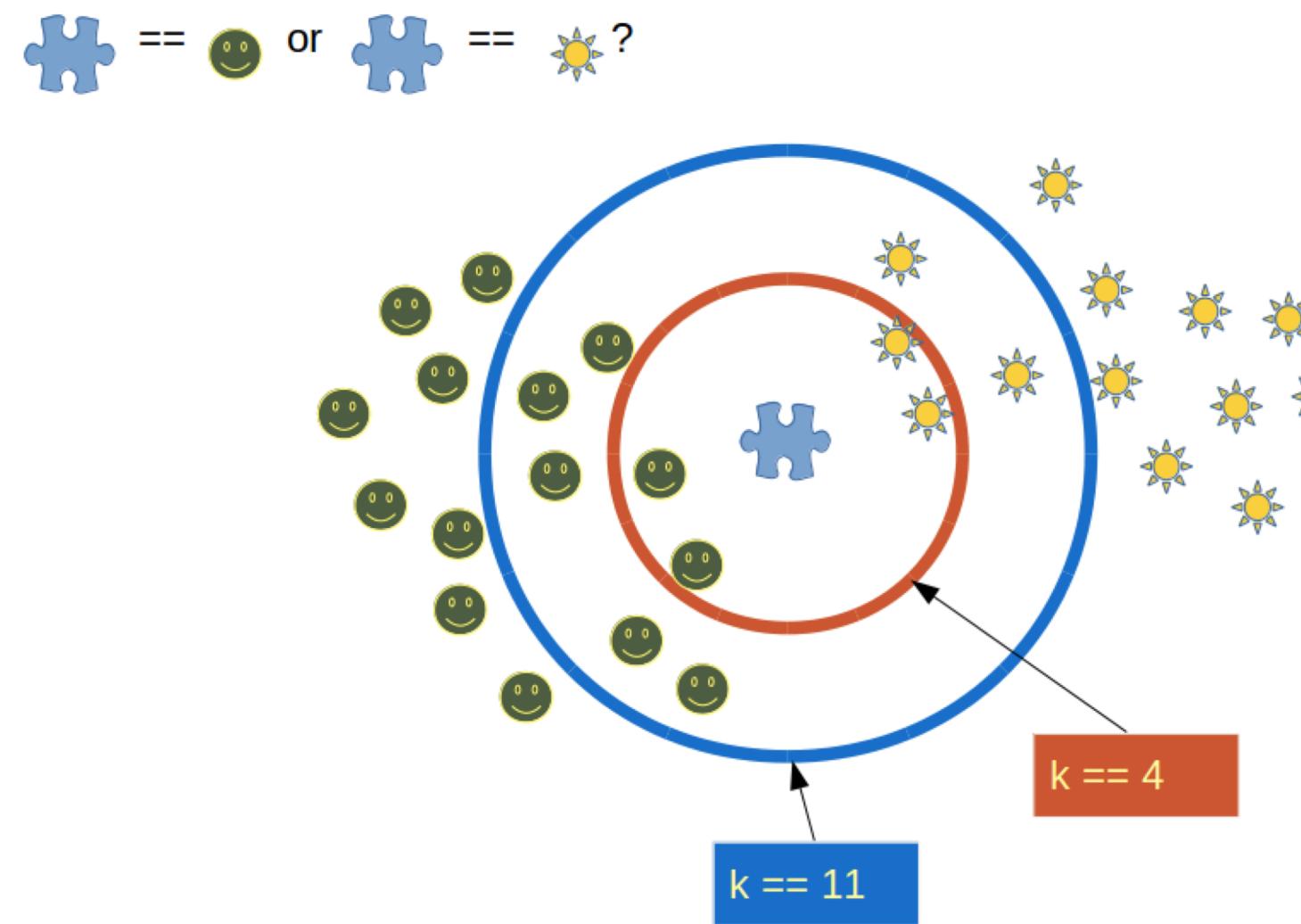
ВВОДНЫЕ

- $Y = \{-1, 1\}$
- -1 – отрицательный класс
- $+1$ – положительный класс
- $a(x)$ – алгоритм, который должен возвращать одно из двух чисел.



РАССМОТРИМ 2 ТИПА АЛГОРИТМОВ

Тип	Метрический	Логический
Представитель	kNN	Решающее дерево





KNN

-
- Метод k-ближайших соседей – метрический, т.е. нужно считать расстояния
 - Дано: (x_i, y_i) – прецеденты
 - Обучение: запоминаем выборку



- Получаем новый объект x
- Сортируем объекты выборки по расстоянию до x

$$\rho(x, x_{(1)}) \leq \dots \leq \rho(x, x_{(\ell)})$$

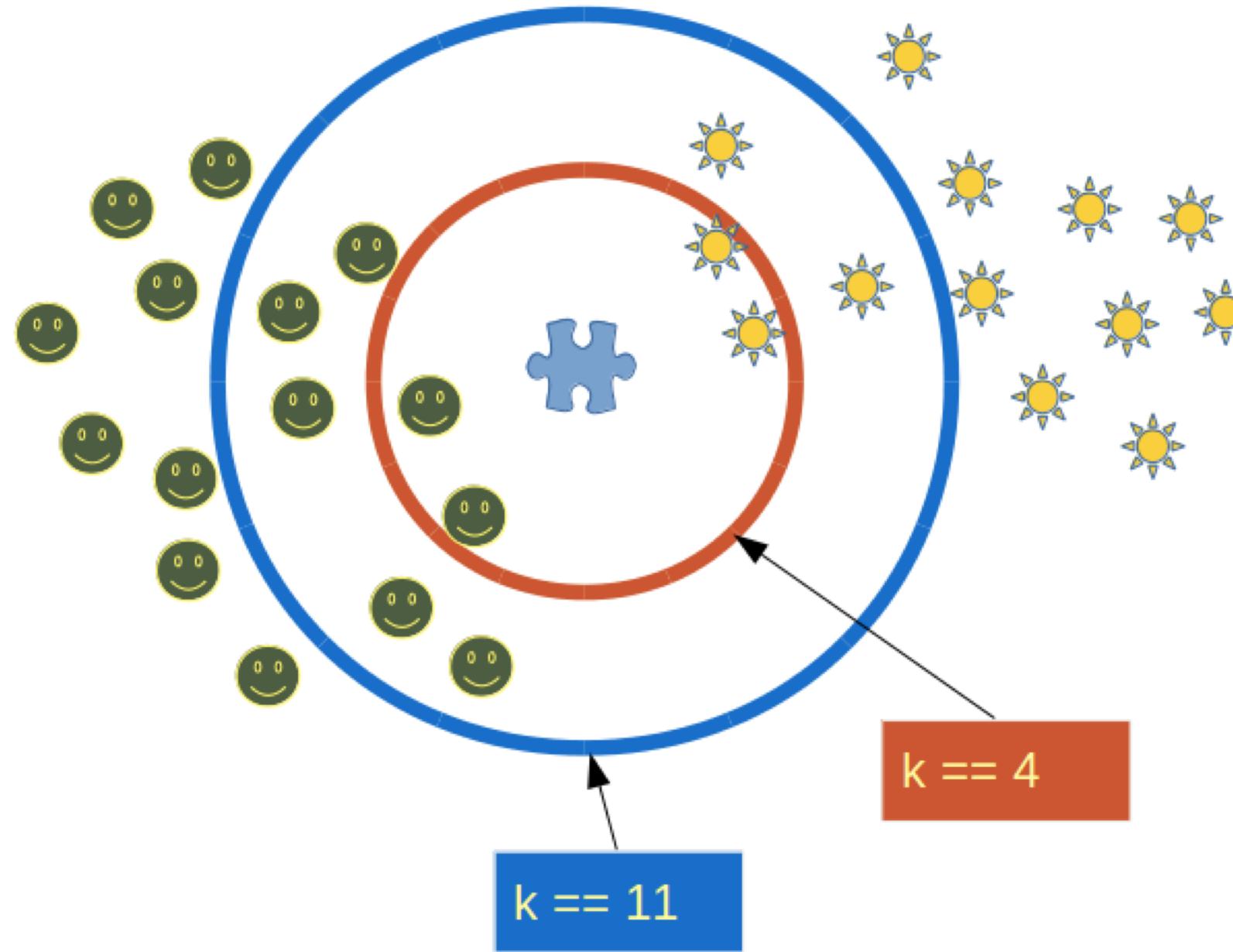
- Присваиваем x класс, наиболее популярный в его окружении

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^{\mathbf{k}} [y_{(i)} = y]$$



KNN

== or == ?



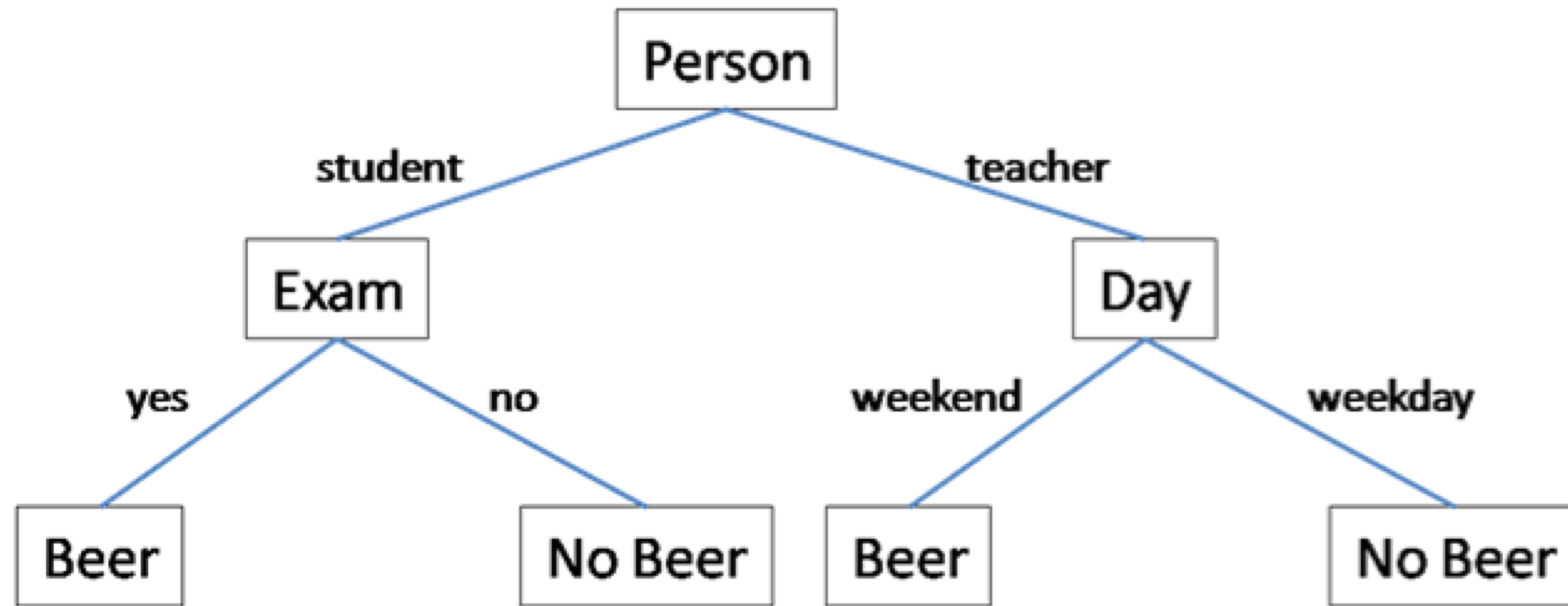


KNN. ПРОБЛЕМЫ И ОСОБЕННОСТИ

- k – гиперпараметр
- Никак не учитывается расстояние до ближайших соседей
- Как такового обучения модели нет

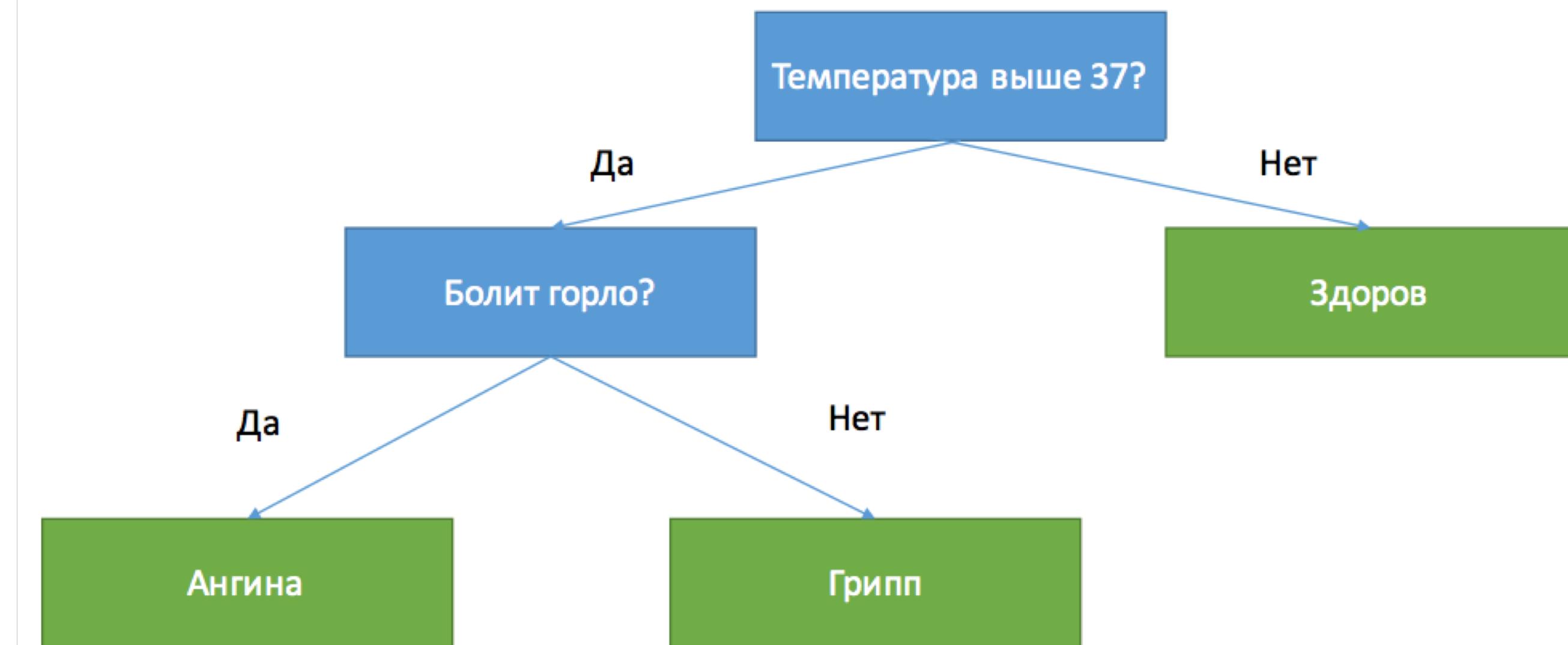


РЕШАЮЩИЕ ДЕРЕВЬЯ





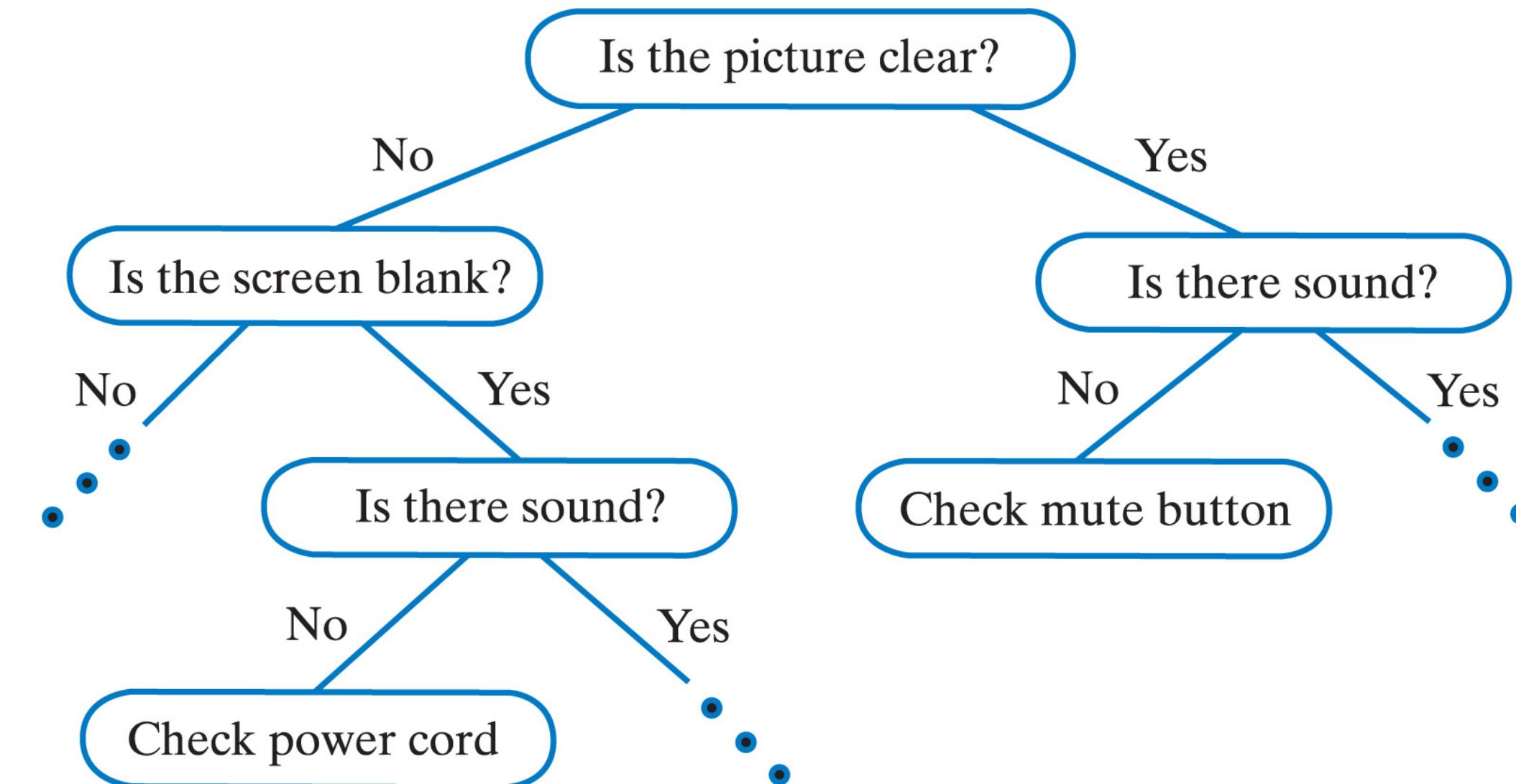
РЕШАЮЩИЕ ДЕРЕВЬЯ





РЕШАЮЩИЕ ДЕРЕВЬЯ

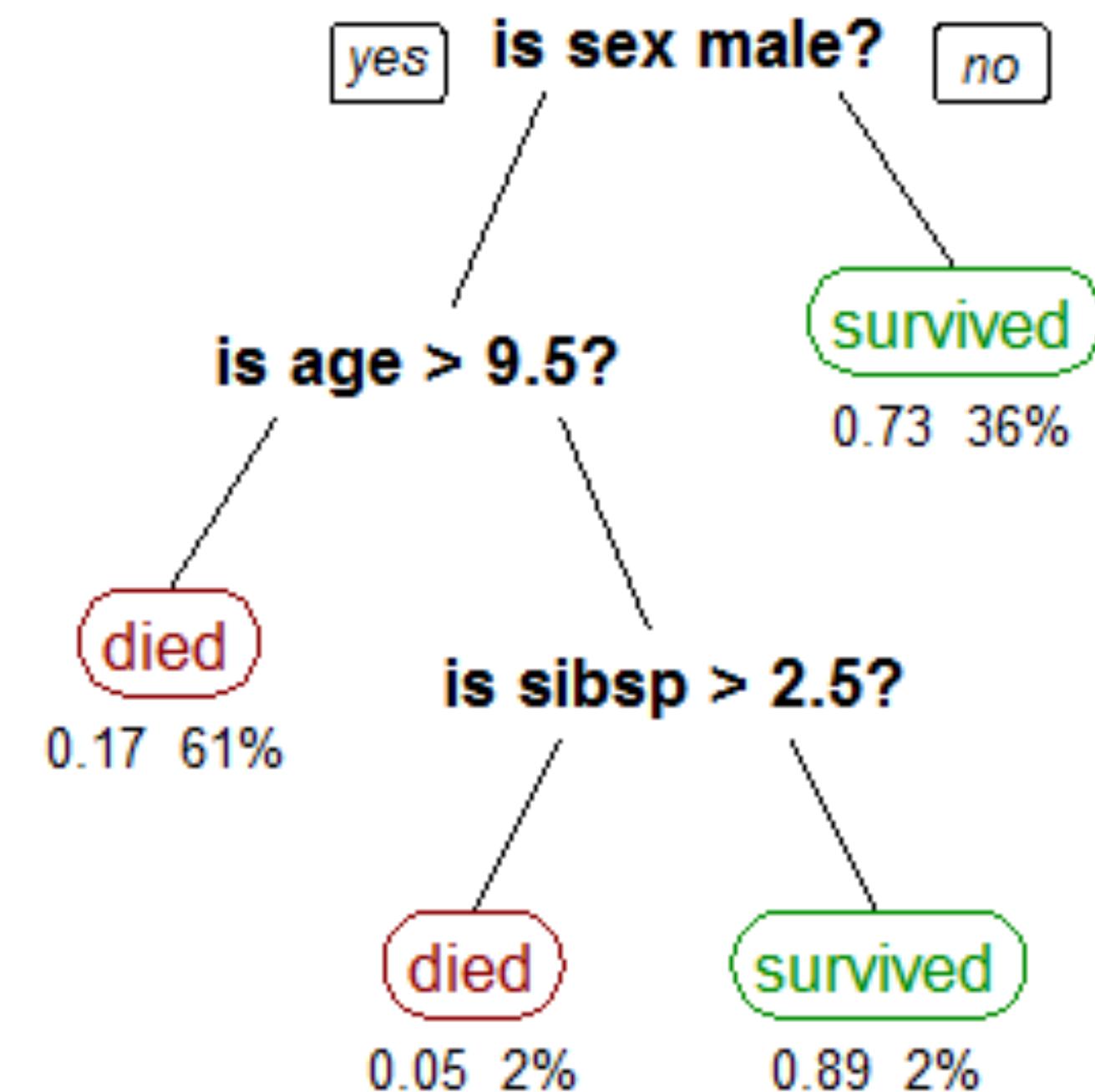
- Состоит из узлов и веток
- Начальный узел – корень дерева
- Конечные узлы – листья





РЕШАЮЩИЕ ДЕРЕВЬЯ

- В каждом узле записано условие
- В зависимости от ответа переходим по одной из веток далее
- Если из узла выходят только по две ветки, то дерево – бинарное
- В листьях записаны метки класса





РЕШАЮЩИЕ ДЕРЕВЬЯ. ТИПЫ УСЛОВИЙ

- Как правило, либо

$$x \leq t$$

- либо

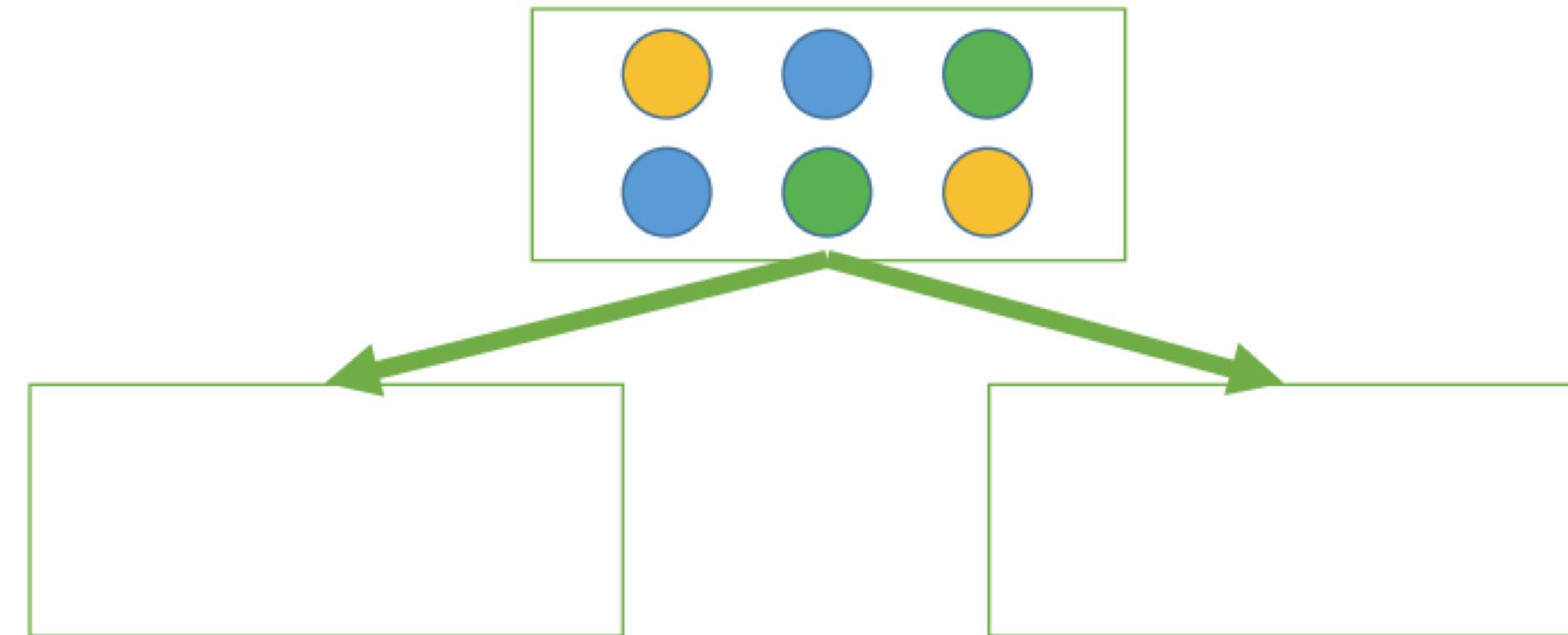
$$x = t$$

- t – какой-то параметр
- [температура ≤ 37 ?]
- [опыт работы == 2?]



РЕШАЮЩИЕ ДЕРЕВЬЯ. ЖАДНЫЙ АЛГОРИТМ

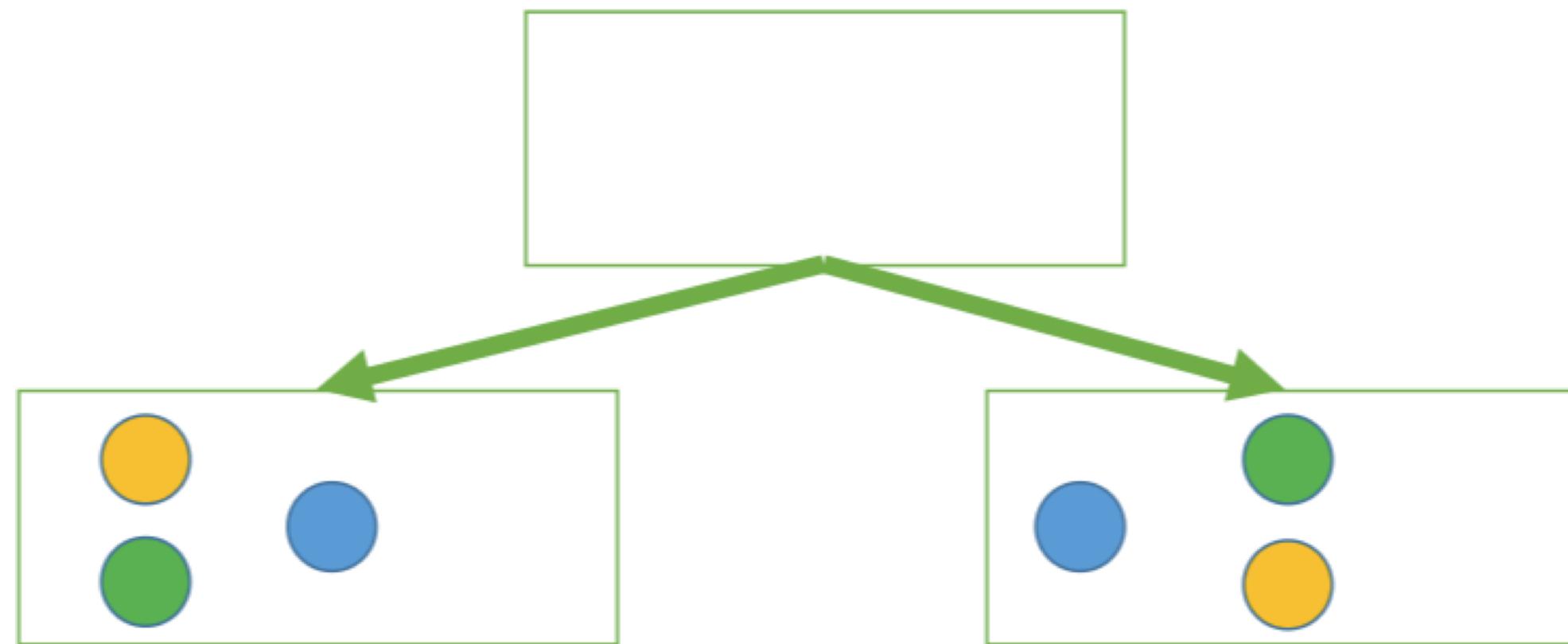
- Все, что называется «жадным», эксплуатирует принцип «Разделяй и властвуй!»
- Как разбить вершину?





РЕШАЮЩИЕ ДЕРЕВЬЯ. ЖАДНЫЙ АЛГОРИТМ

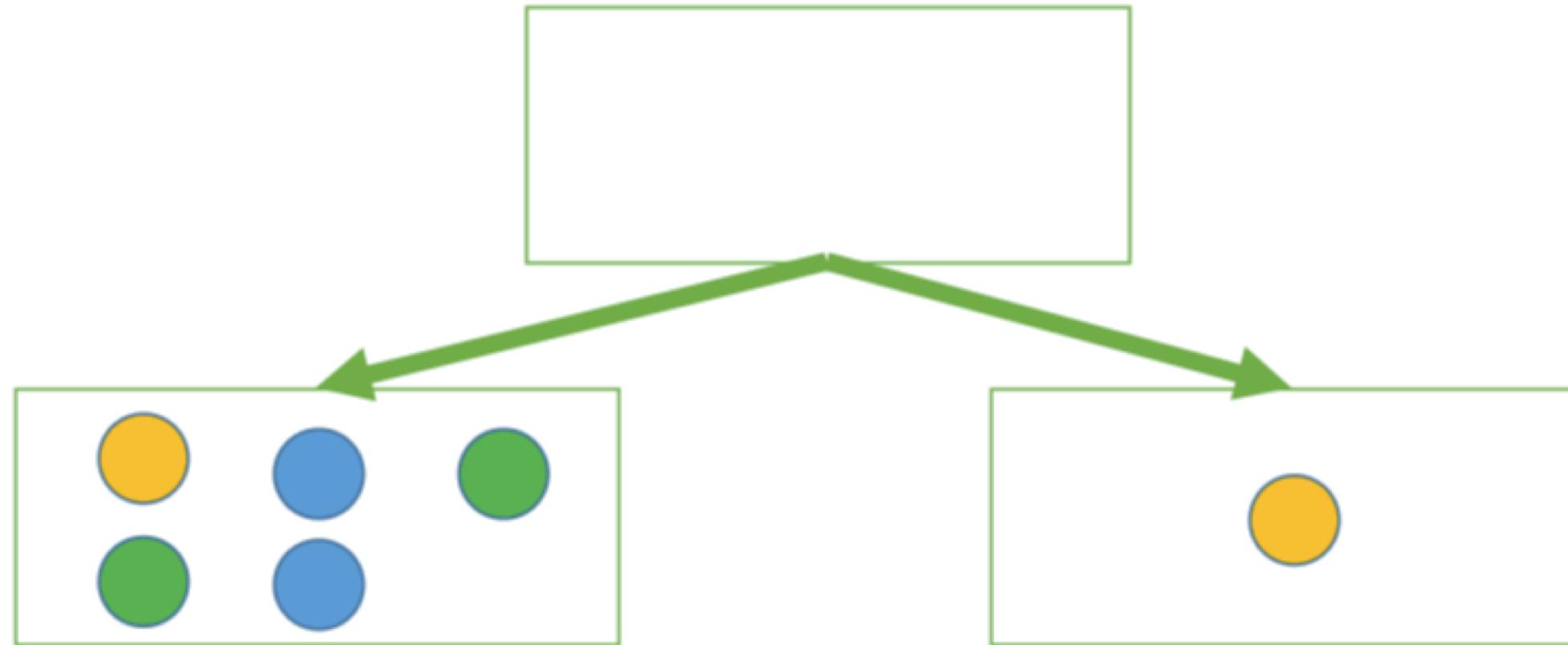
- Как разбить вершину?





РЕШАЮЩИЕ ДЕРЕВЬЯ. ЖАДНЫЙ АЛГОРИТМ

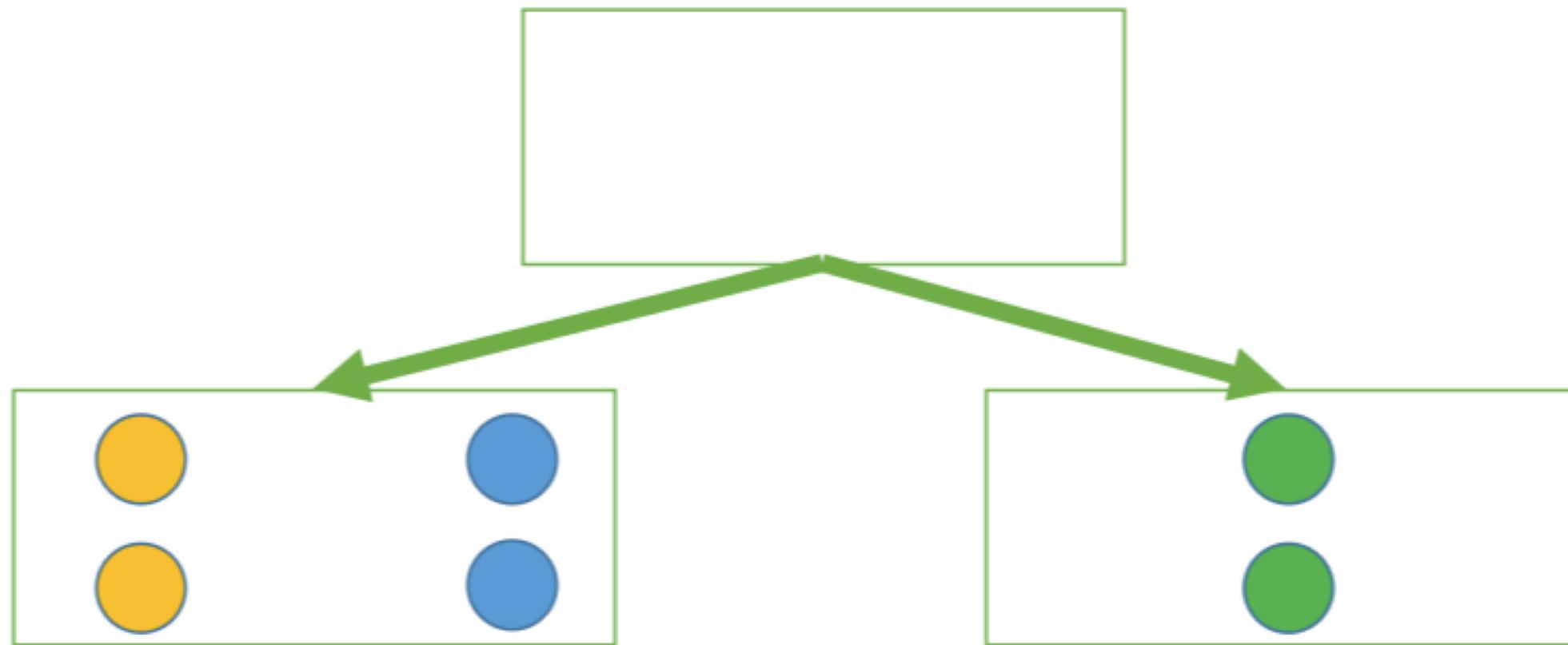
- Как разбить вершину?





РЕШАЮЩИЕ ДЕРЕВЬЯ. ЖАДНЫЙ АЛГОРИТМ

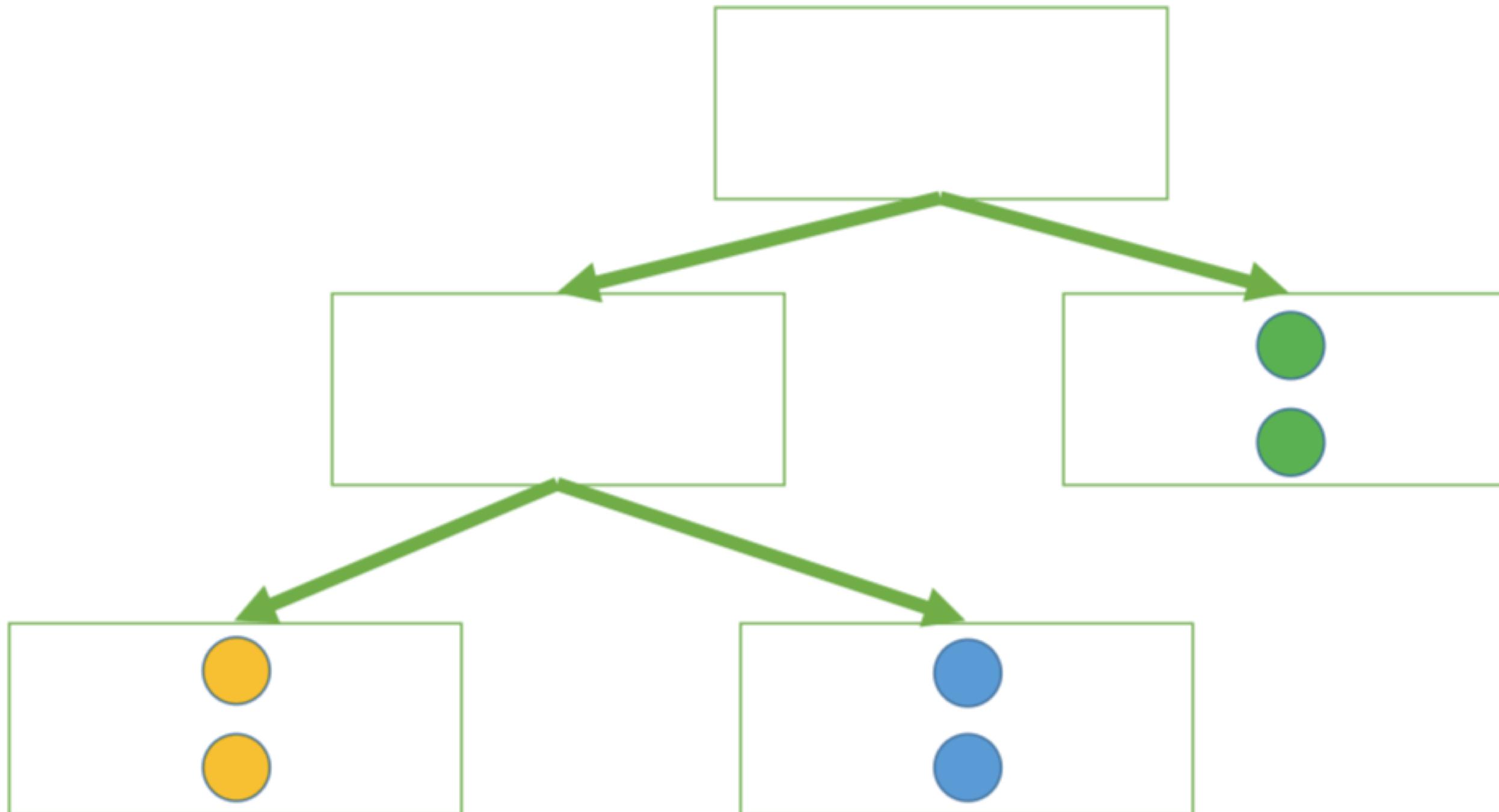
- Как разбить вершину?





РЕШАЮЩИЕ ДЕРЕВЬЯ. ЖАДНЫЙ АЛГОРИТМ

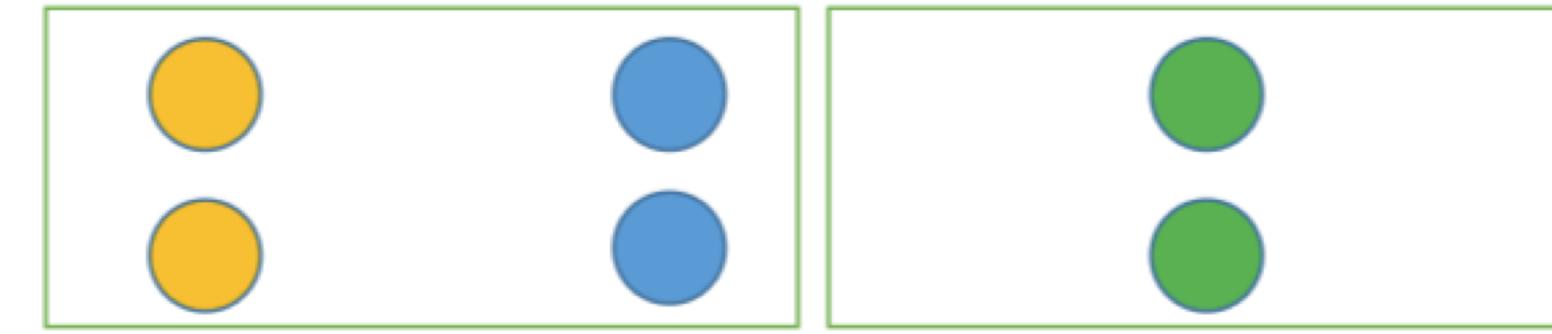
- Как разбить вершину?



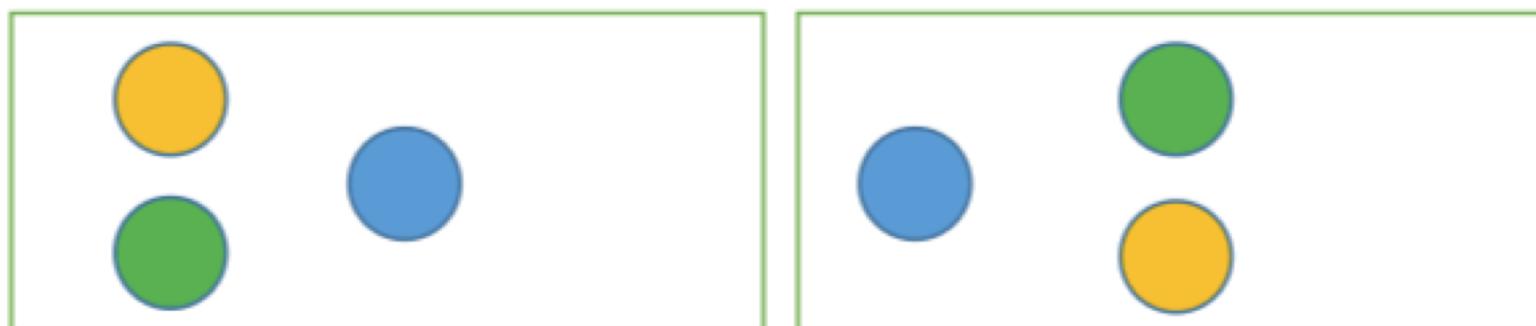


РЕШАЮЩИЕ ДЕРЕВЬЯ. ЖАДНЫЙ АЛГОРИТМ

- Как выбрать?



или





РЕШАЮЩИЕ ДЕРЕВЬЯ. КРИТЕРИЙ ИНФОРМАТИВНОСТИ

- В каждом алгоритме используется энтропийный критерий

$$H(X) = - \sum_{k=1}^K p_k \ln p_k$$

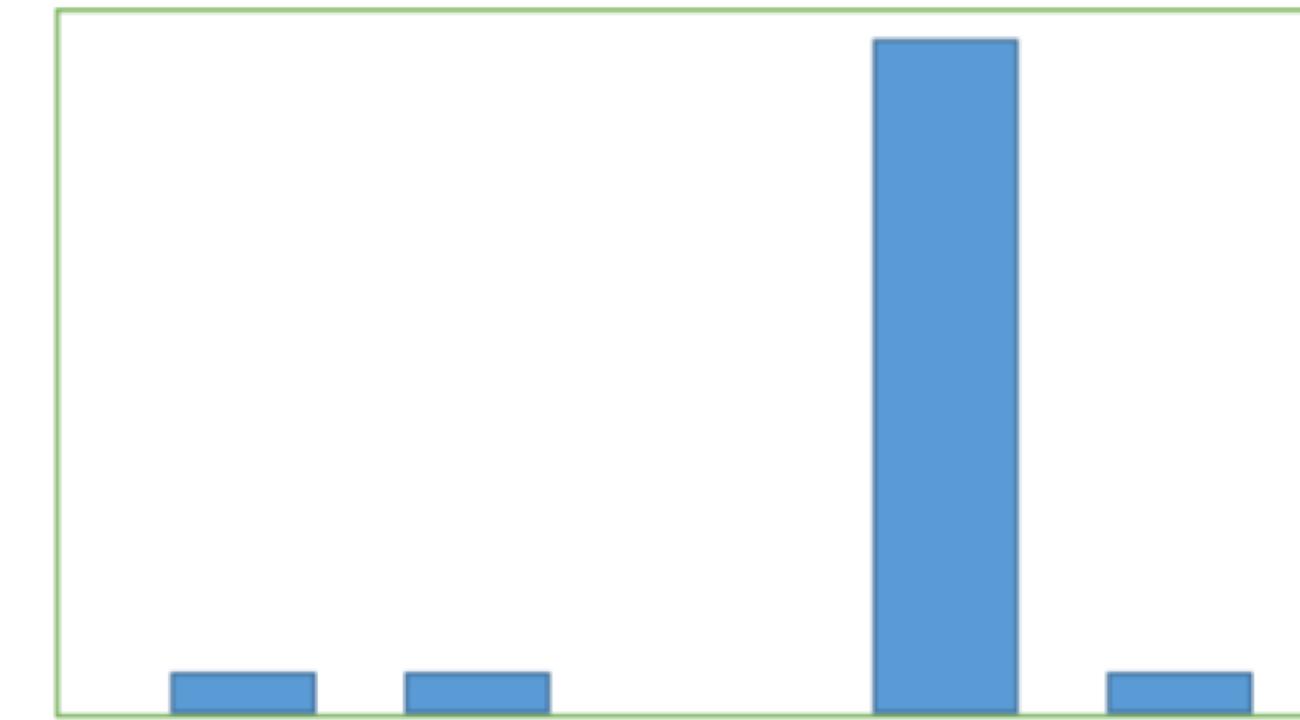
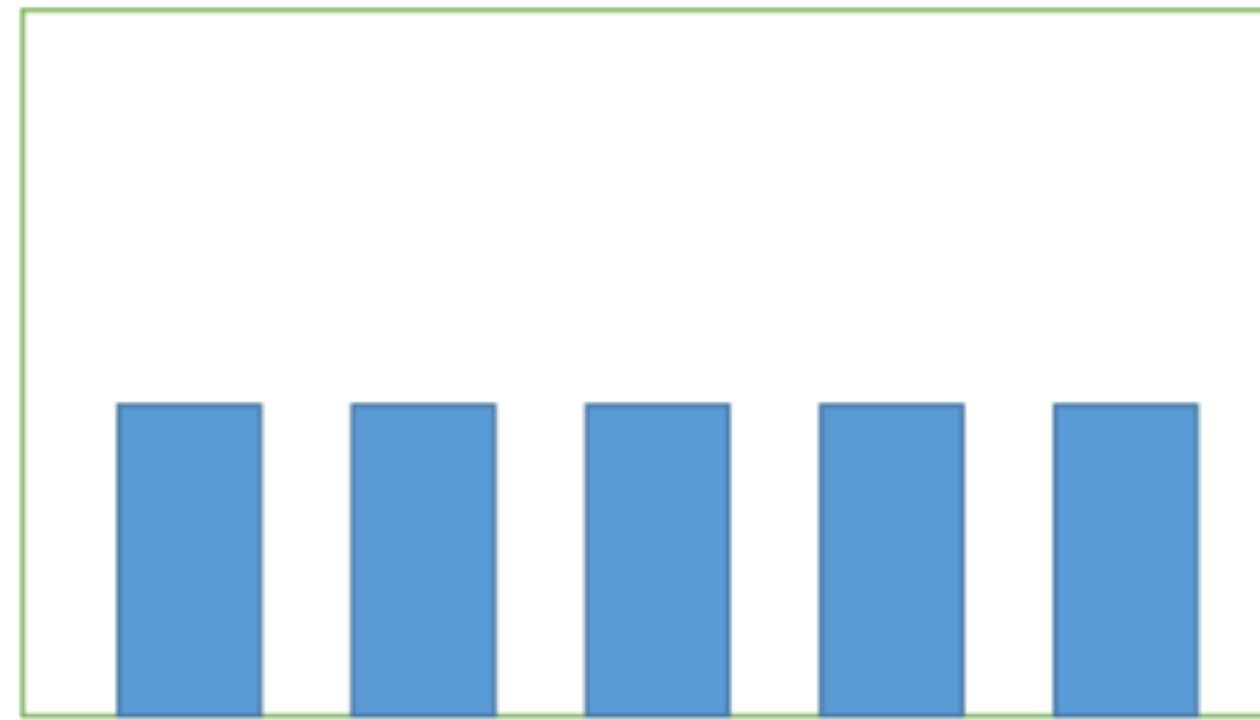
- Доля объектов класса k в выборке X

$$p_k = \frac{1}{|X|} \sum_{i \in X} [y_i = k]$$



РЕШАЮЩИЕ ДЕРЕВЬЯ. КРИТЕРИЙ ИНФОРМАТИВНОСТИ

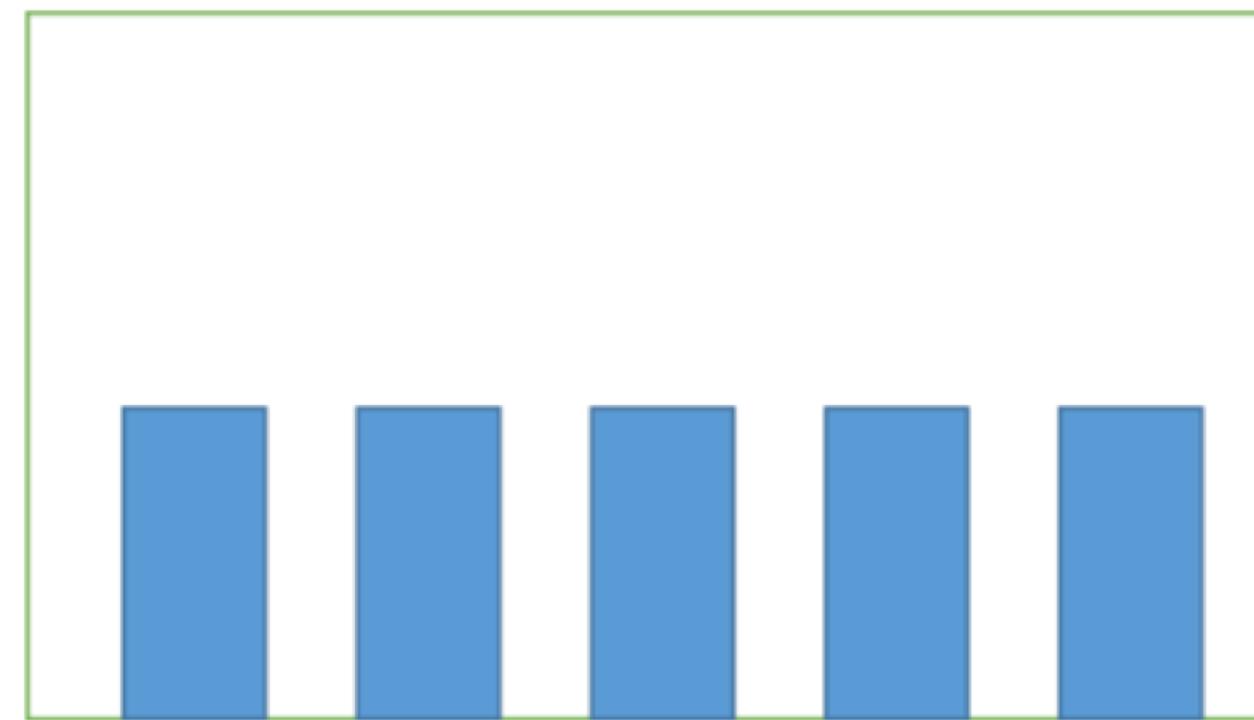
- Энтропия



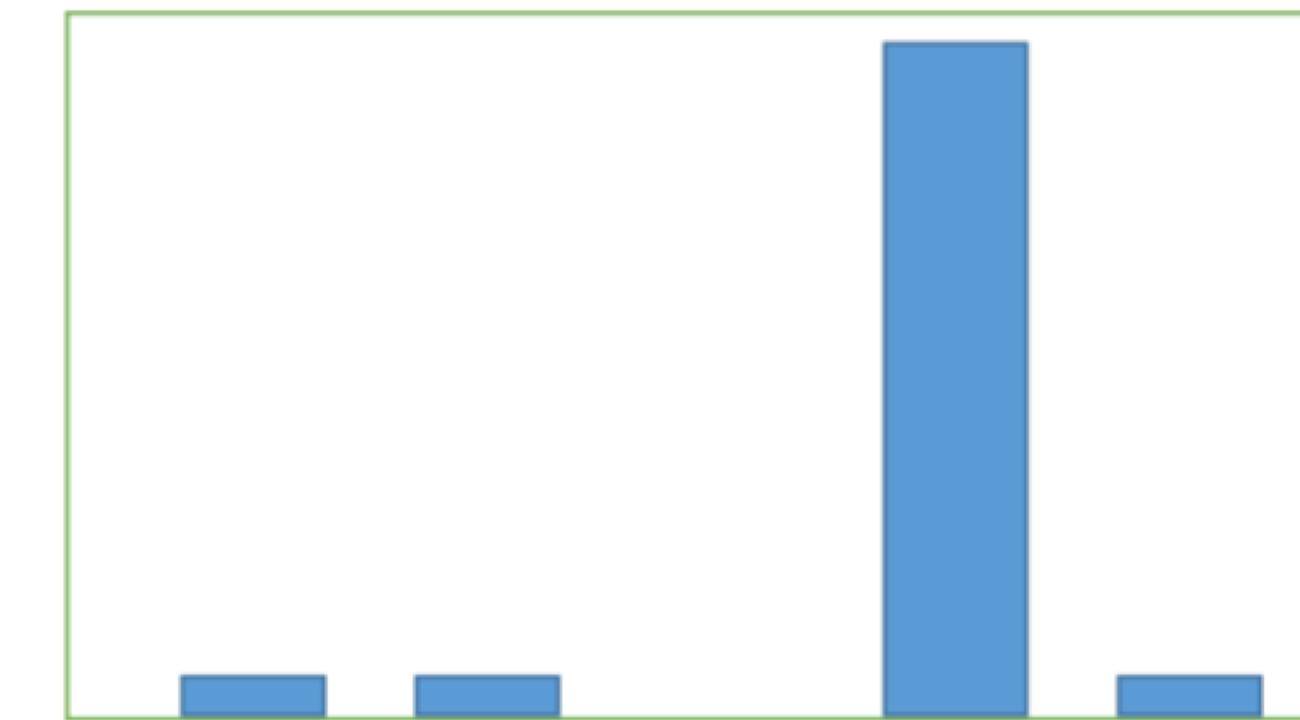


РЕШАЮЩИЕ ДЕРЕВЬЯ. КРИТЕРИЙ ИНФОРМАТИВНОСТИ

- Энтропия – мера неопределенности распределения



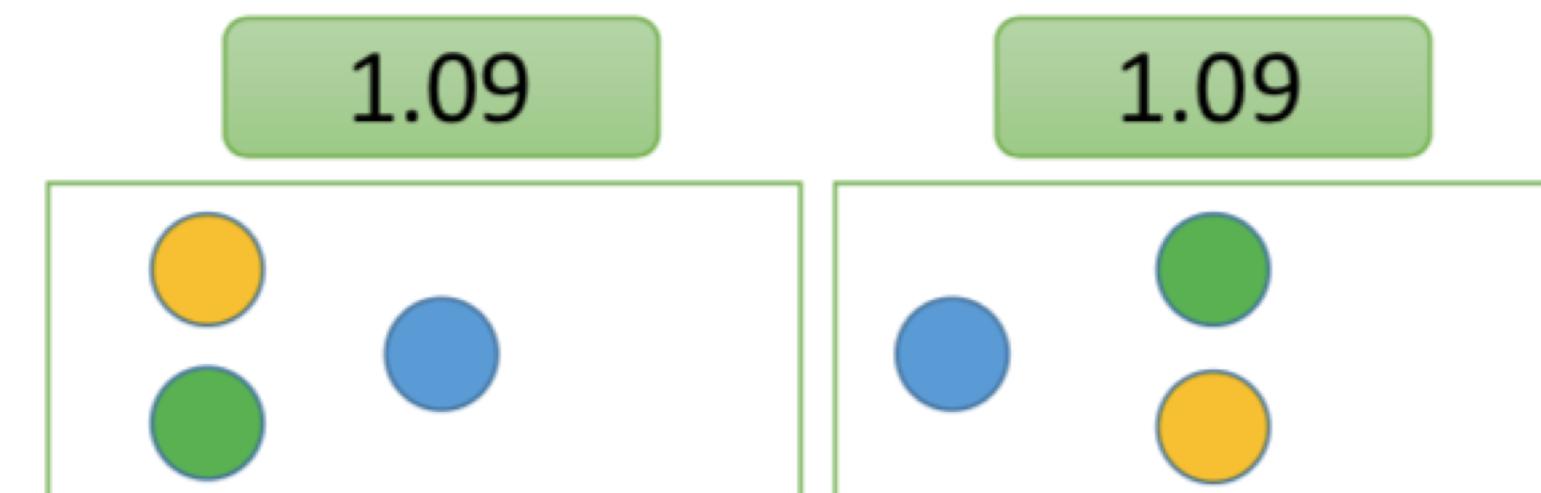
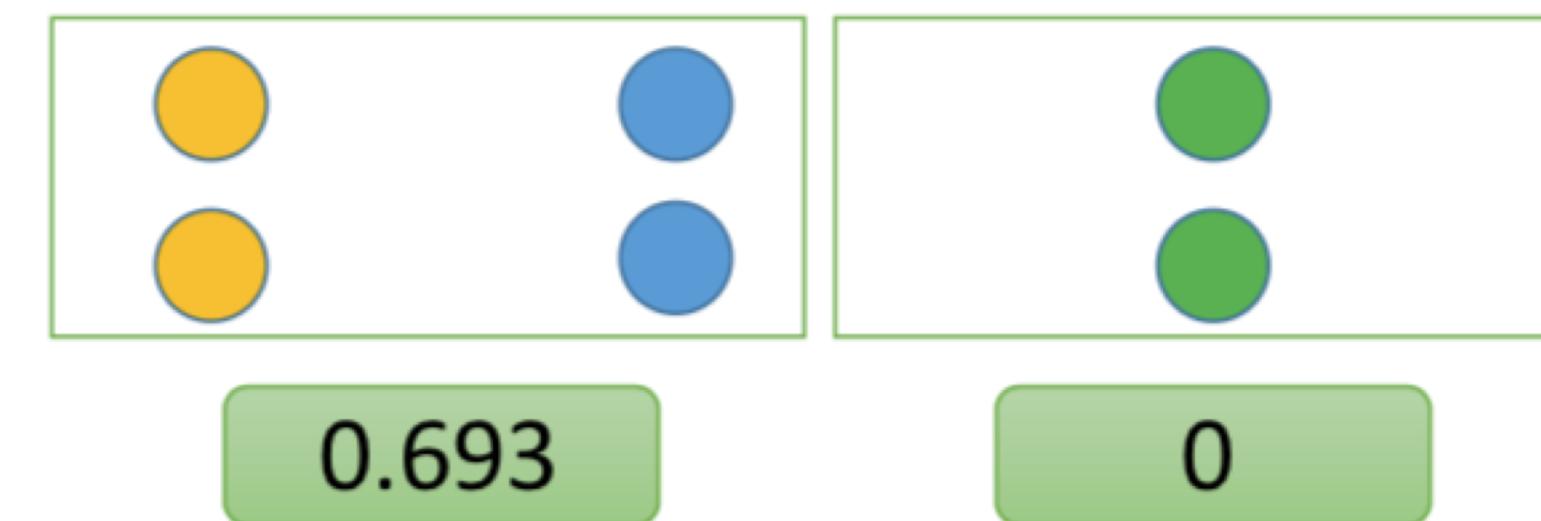
Высокая энтропия



Низкая энтропия



РЕШАЮЩИЕ ДЕРЕВЬЯ. КРИТЕРИЙ ИНФОРМАТИВНОСТИ





РЕШАЮЩИЕ ДЕРЕВЬЯ. КРИТЕРИЙ ОСТАНОВА

- В какой момент прекращать разбиение вершин?
- В вершине один объекты?
- В вершине объекты одного класса?
- Глубина превысила порог?



РЕШАЮЩИЕ ДЕРЕВЬЯ. ЖАДНЫЙ АЛГОРИТМ

1. Поместить все объекты в корневой узел.
2. Начать построение дерева с корневой вершины $m = 1$
3. Если выполнен критерий останова, то закончить алгоритм
4. Найти наилучшее разбиение объектов в выборке по критерию информативности
5. Разбить вершину на две дочерние
6. Повторить шаги 3-6

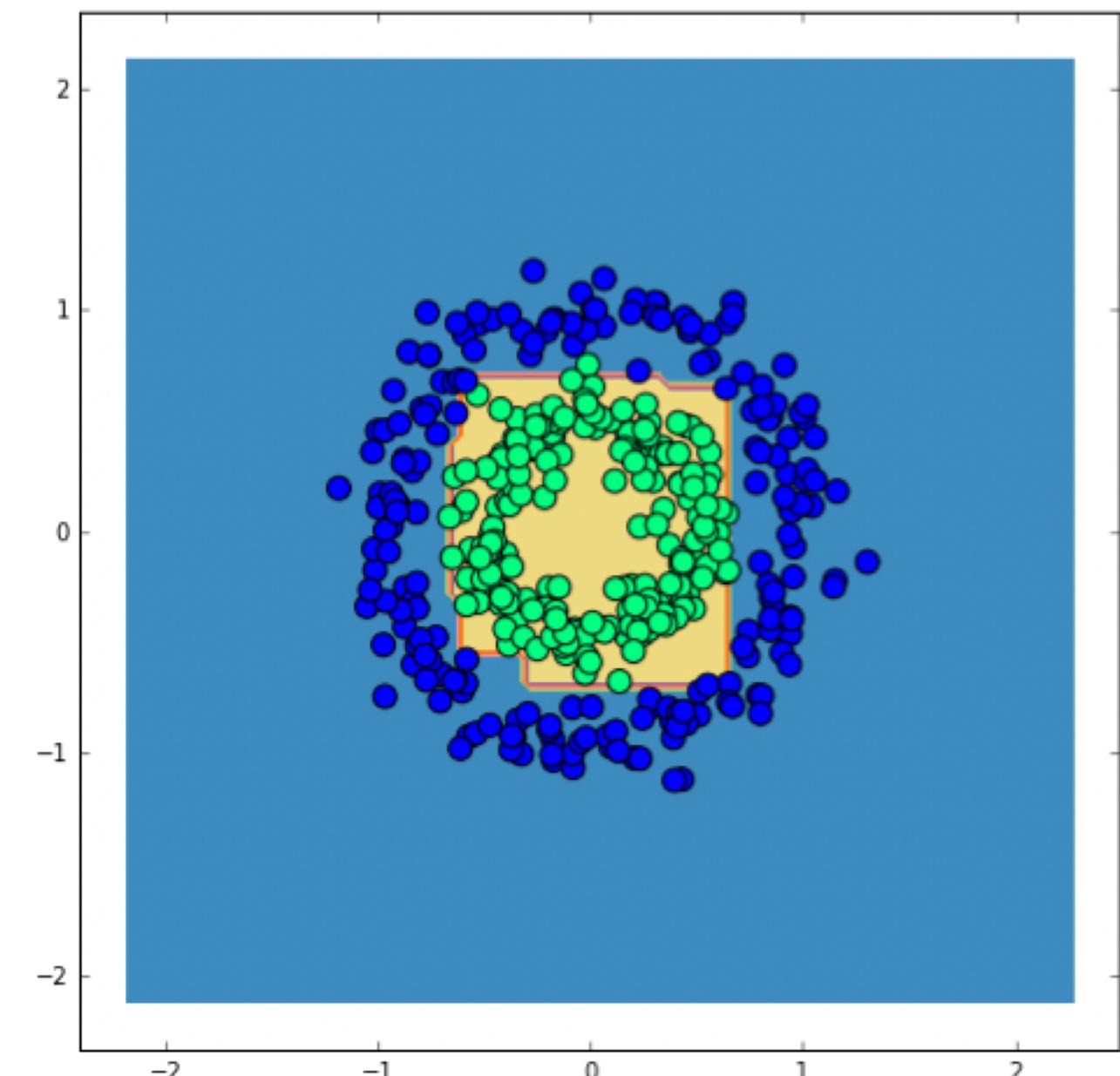
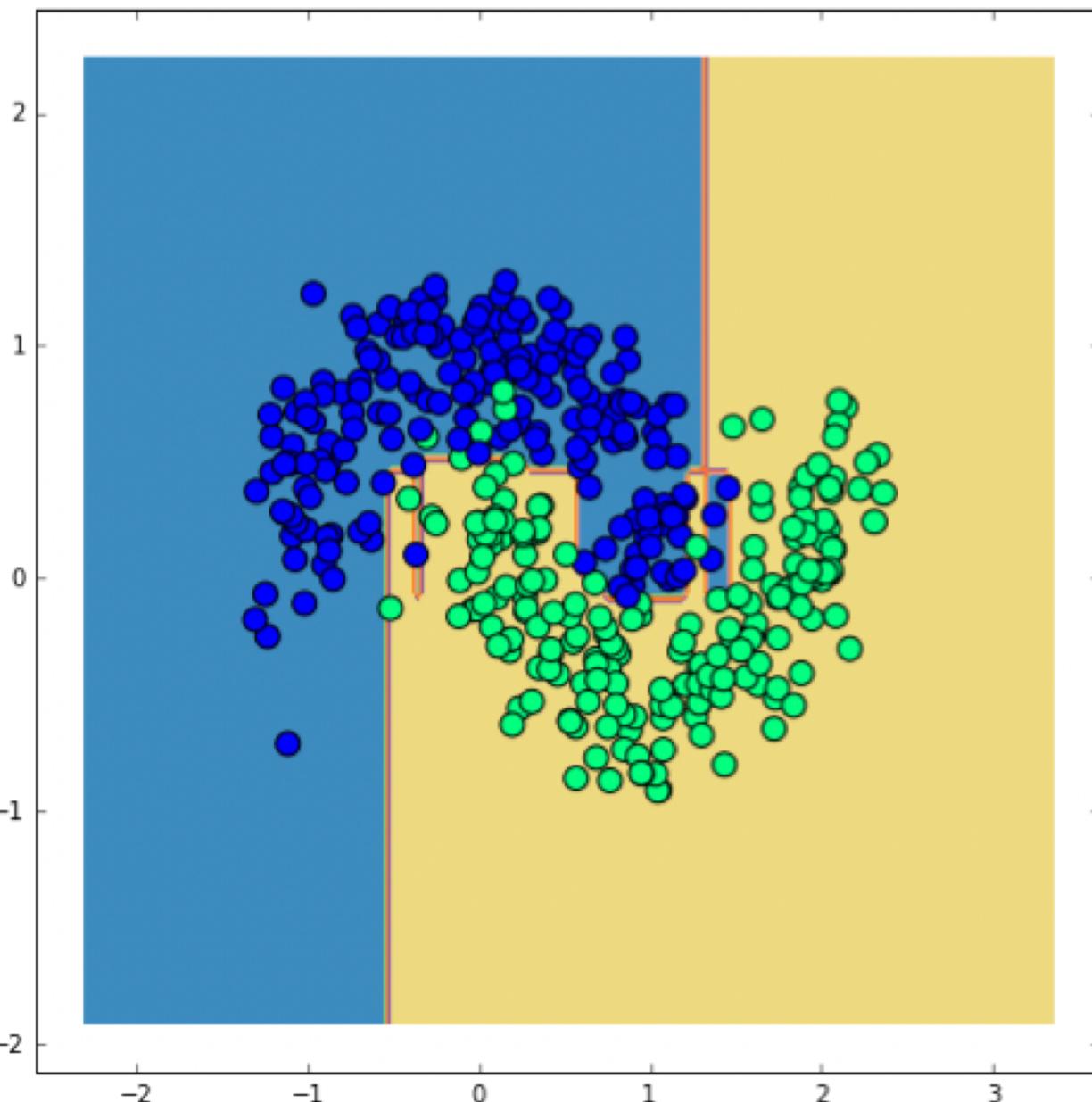
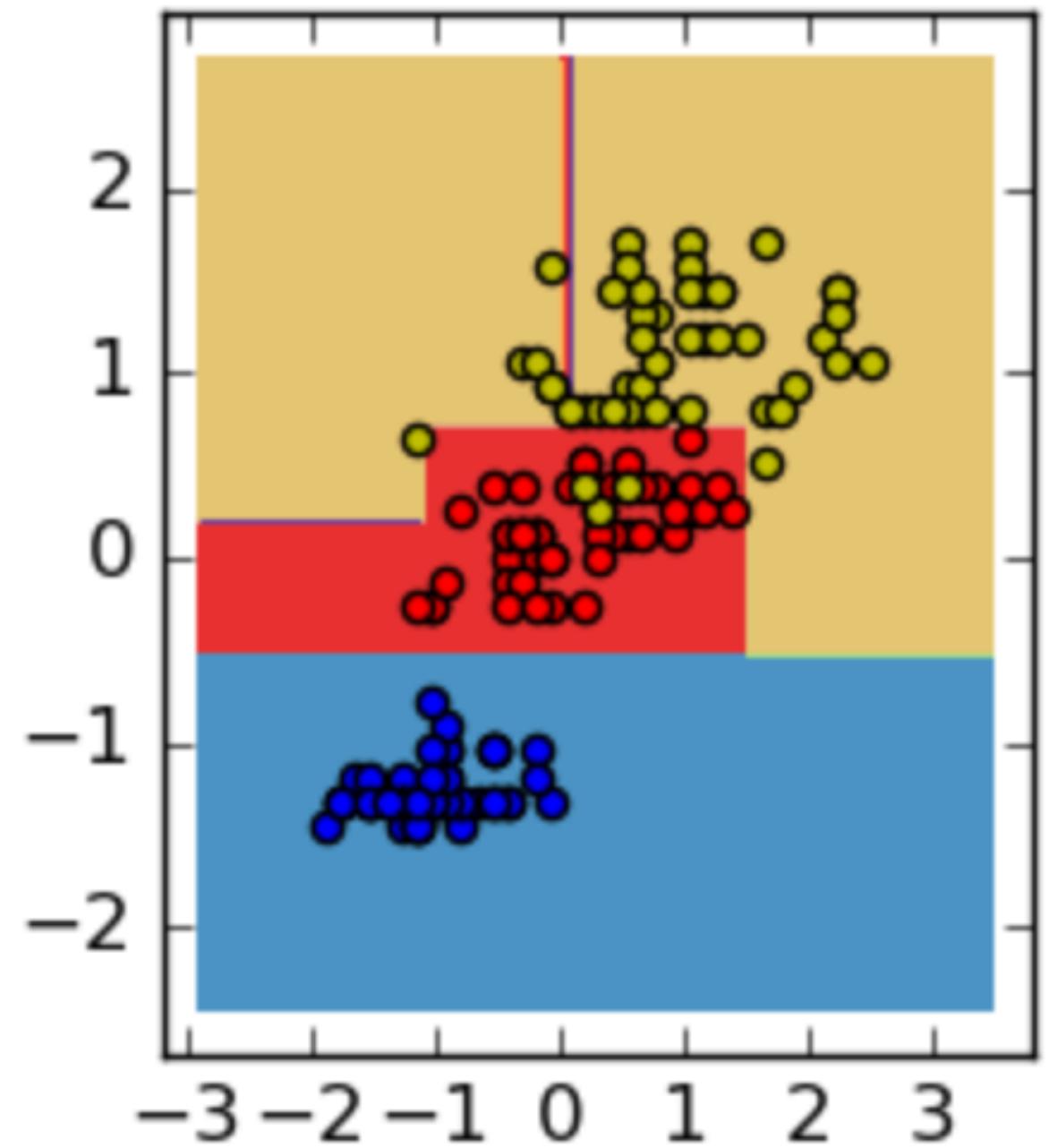


РЕШАЮЩИЕ ДЕРЕВЬЯ. РЕЗЮМЕ

1. Восстанавливают сложные закономерности
2. Могут построить сколь угодно сложную поверхность
3. Чем больше глубина — тем сложнее поверхность
4. Склонны к переобучению

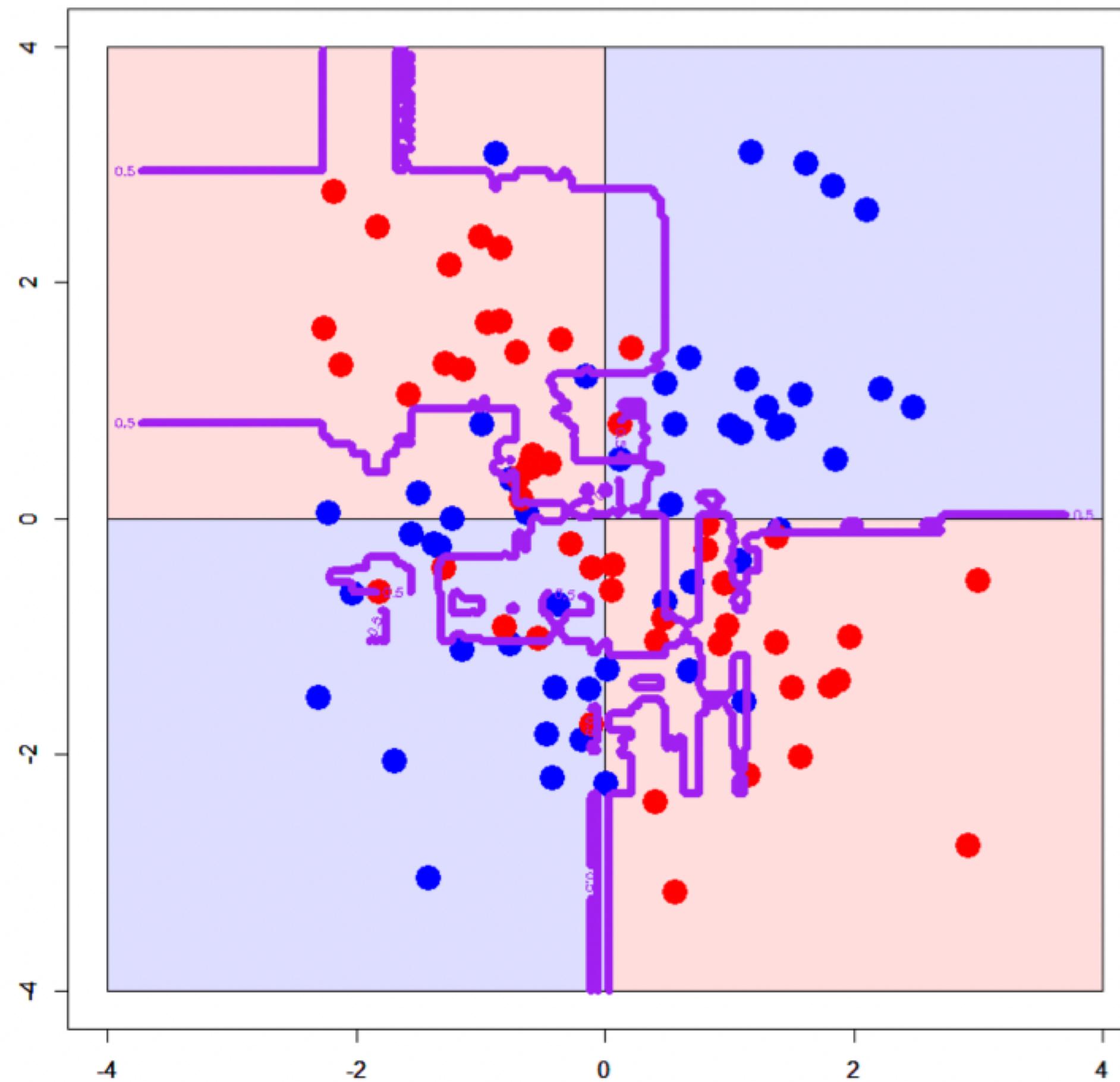


РЕШАЮЩИЕ ДЕРЕВЬЯ. ПЕРЕОБУЧЕНИЕ



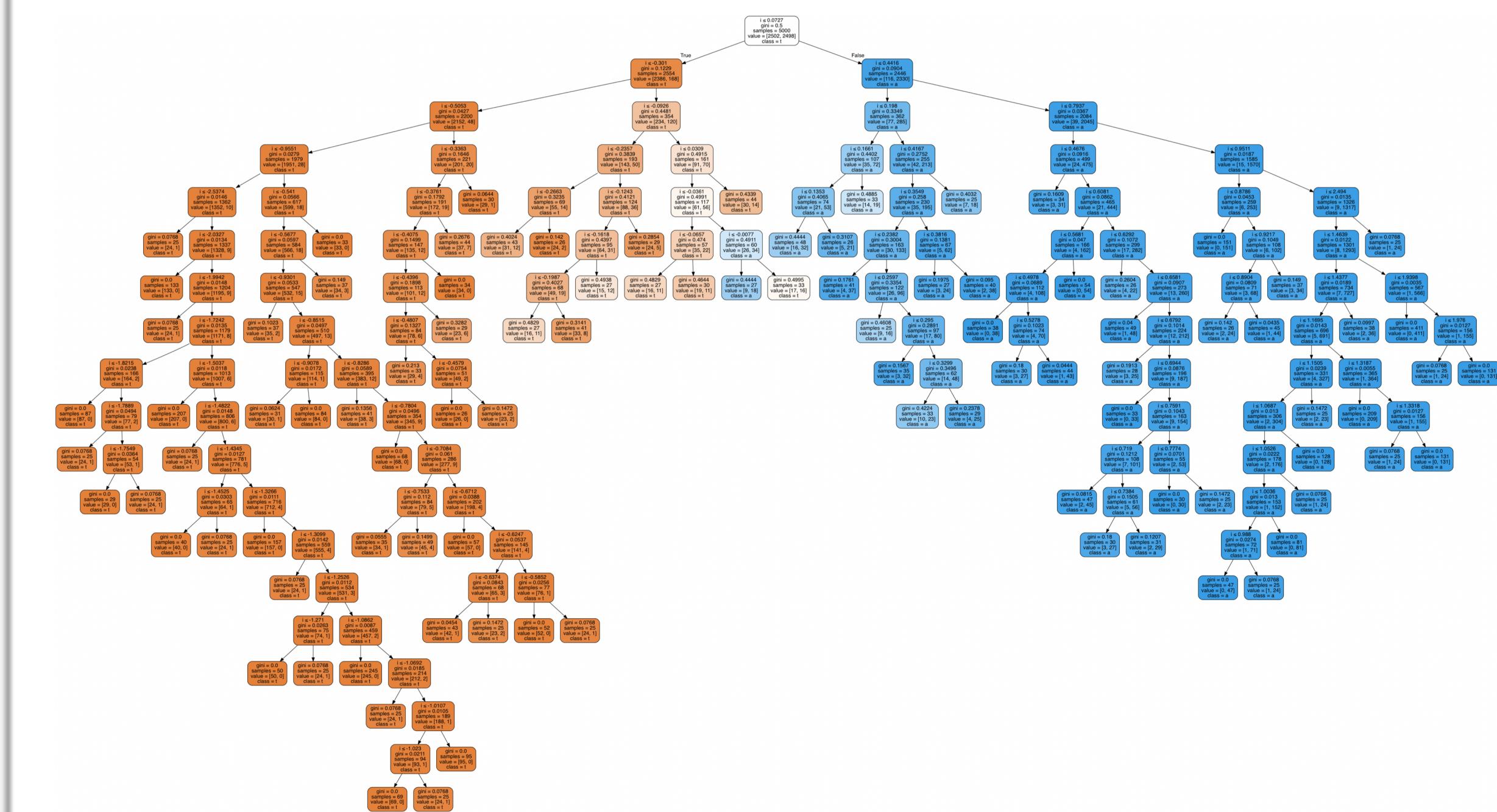
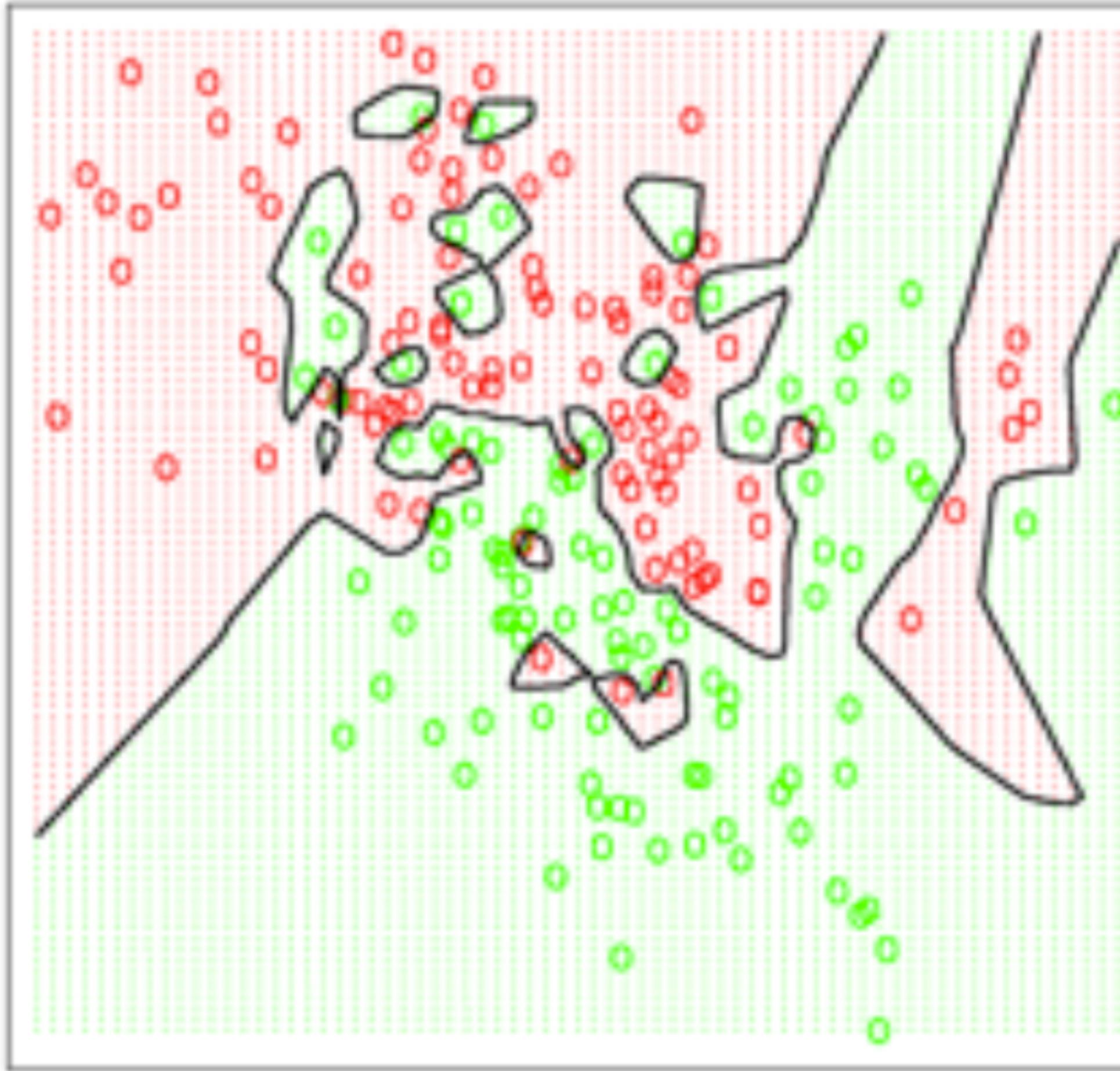


РЕШАЮЩИЕ ДЕРЕВЬЯ. ПЕРЕОБУЧЕНИЕ





РЕШАЮЩИЕ ДЕРЕВЬЯ. ПЕРЕОБУЧЕНИЕ





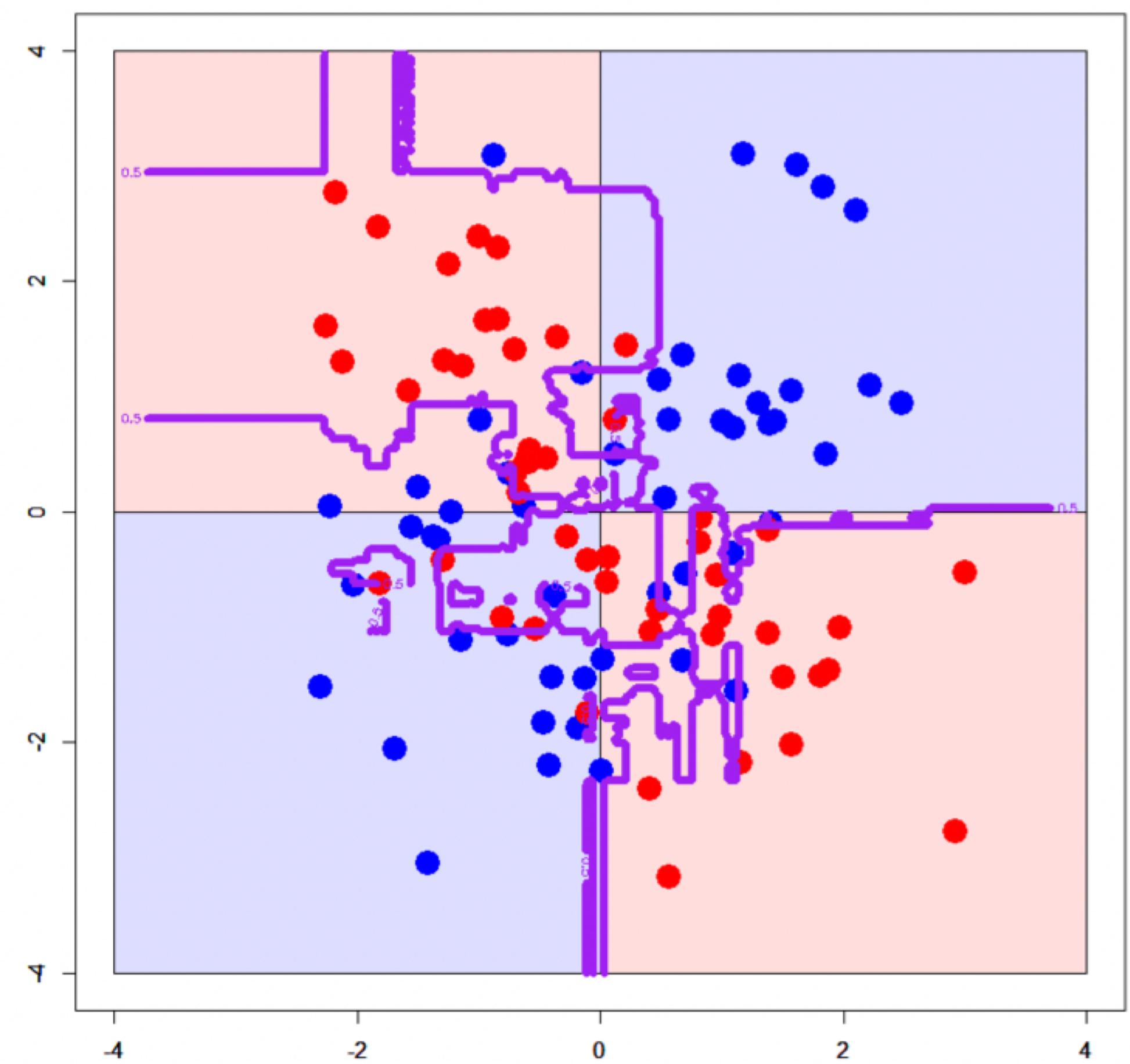
РЕШАЮЩИЕ ДЕРЕВЬЯ. ПЕРЕОБУЧЕНИЕ

1. Дерево может достичь нулевой ошибки на любой выборке
2. Как правило, такое дерево окажется переобученным
3. Выход — ограничивать глубину или число объектов в листе



КРИТЕРИЙ ОСТАНОВА

1. Все объекты в вершине относятся к одному классу
2. Простое условие
3. Но приводит к переобучению



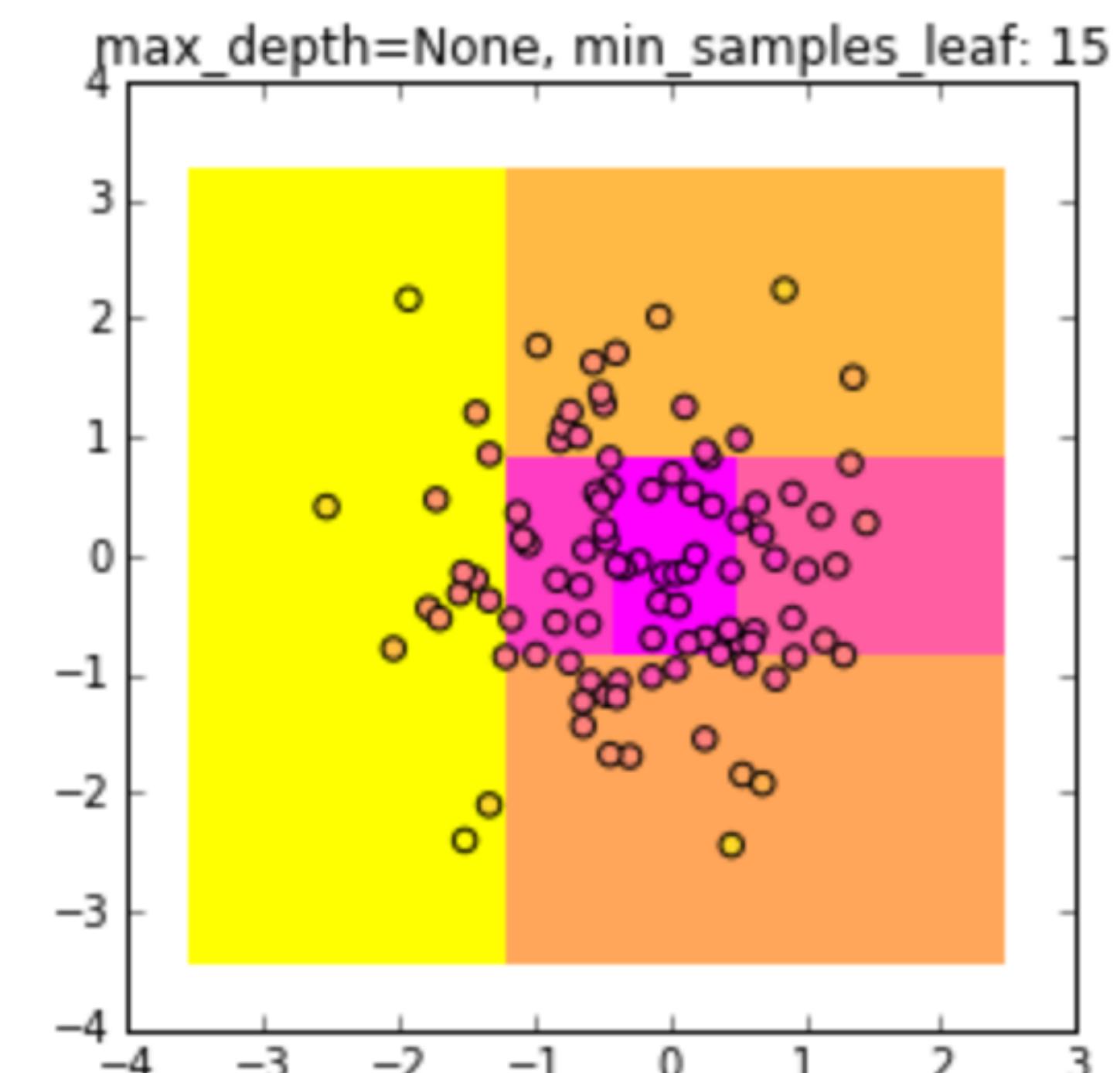
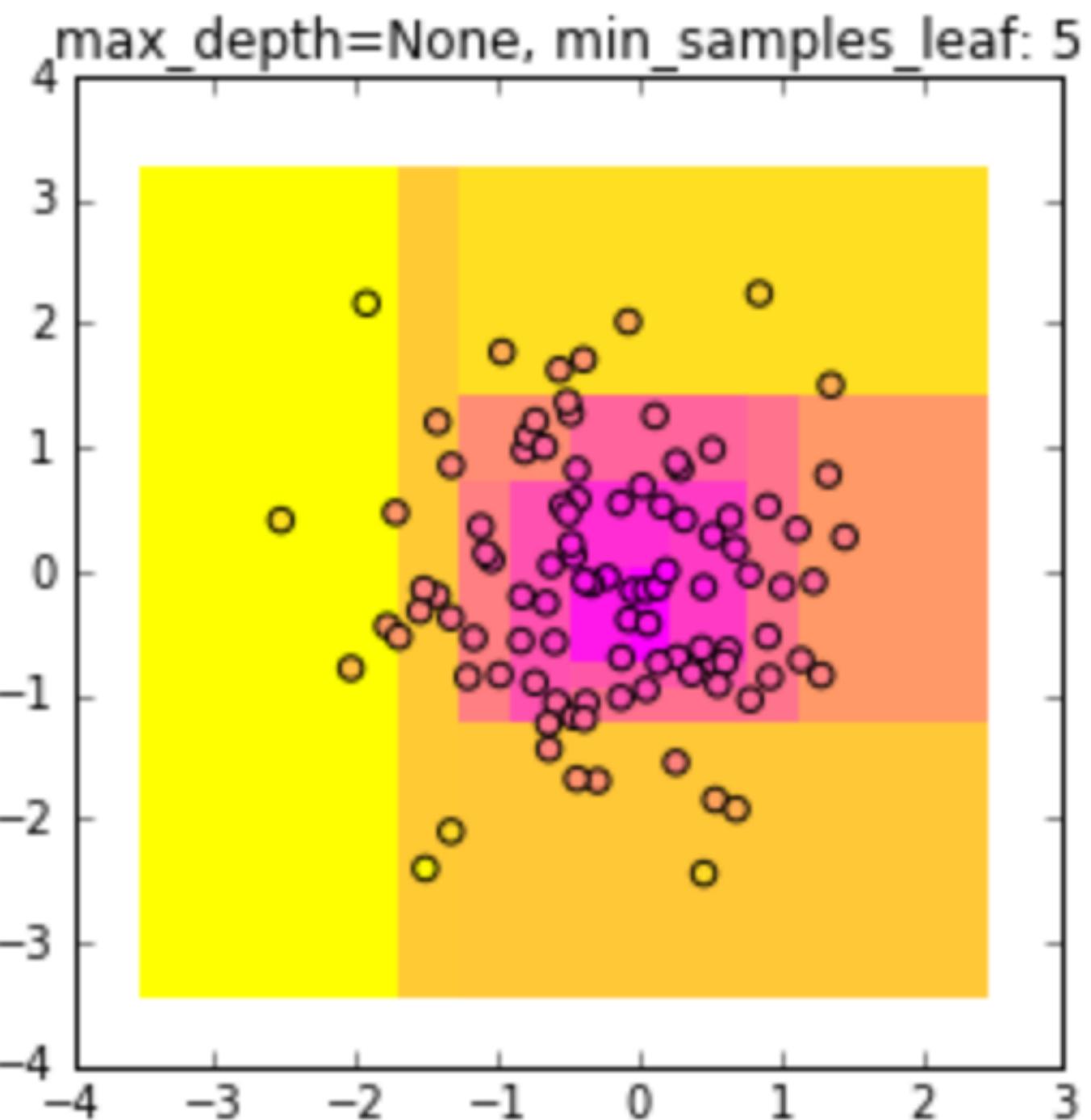
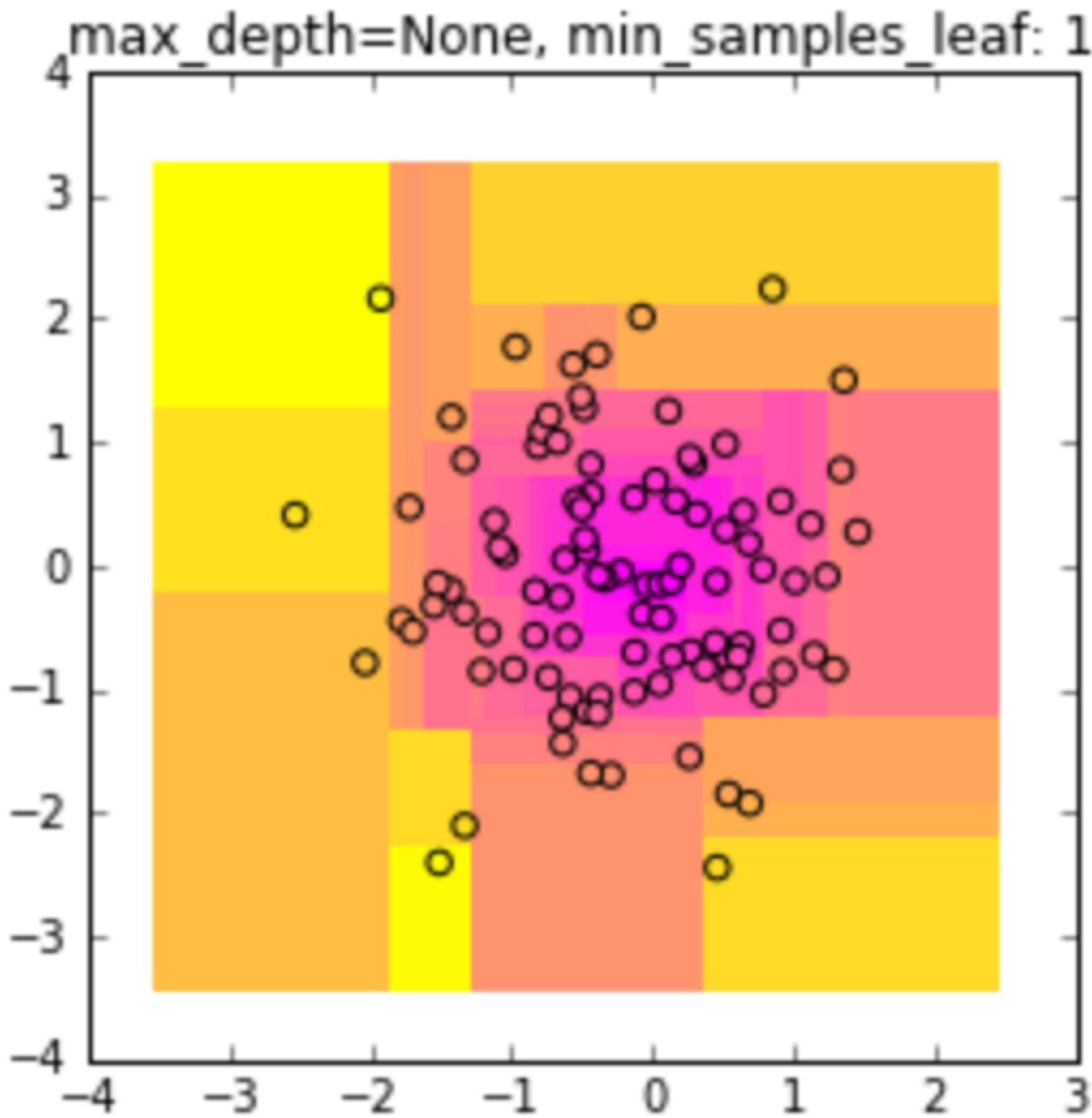


КРИТЕРИЙ ОСТАНОВА

1. В вершину попало $\leq n$ объектов
2. При $n = 1$ получаем максимально переобученные деревья
3. n должно быть достаточно, чтобы построить надёжный прогноз
4. Рекомендация: $n = 5$

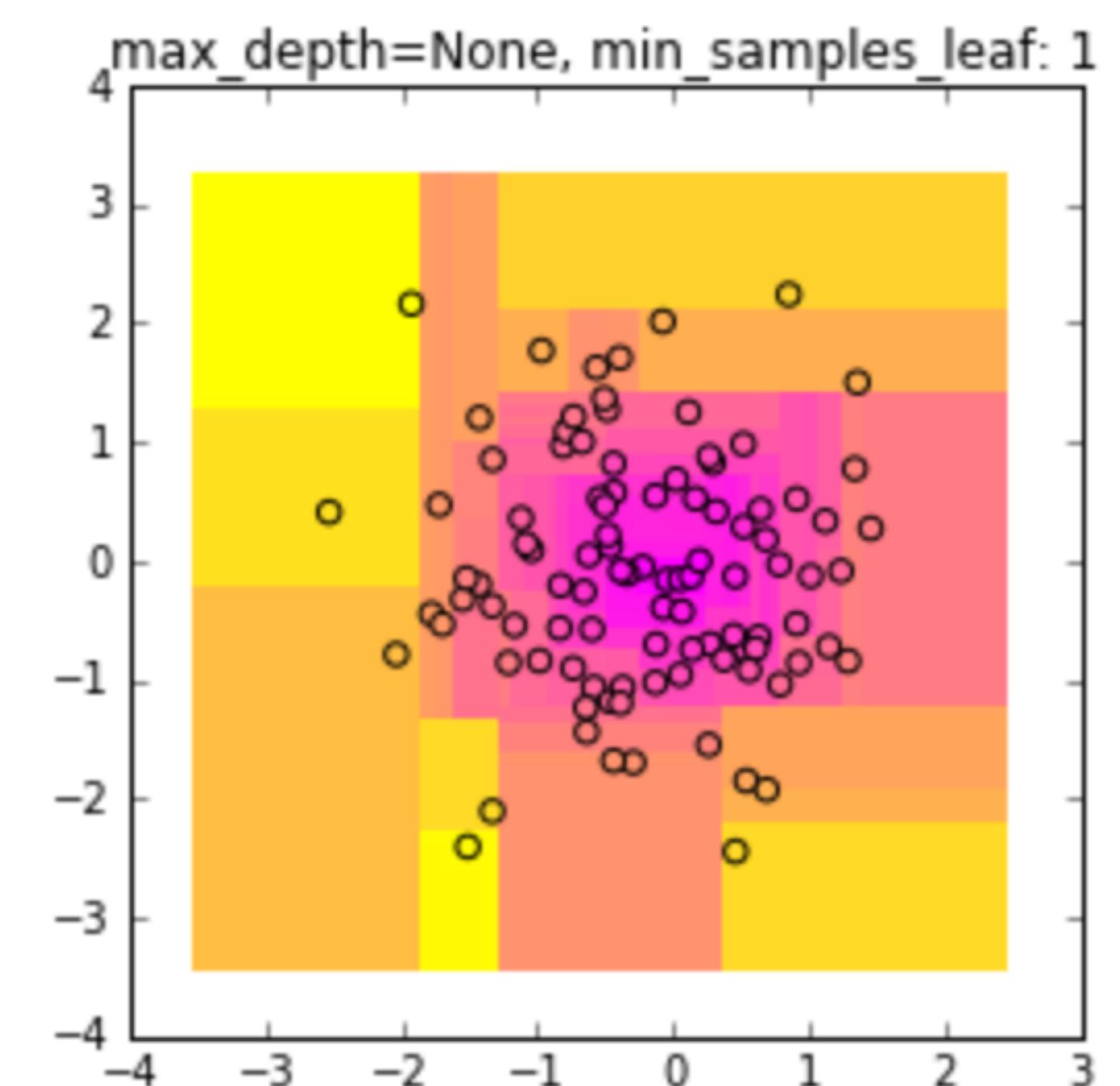
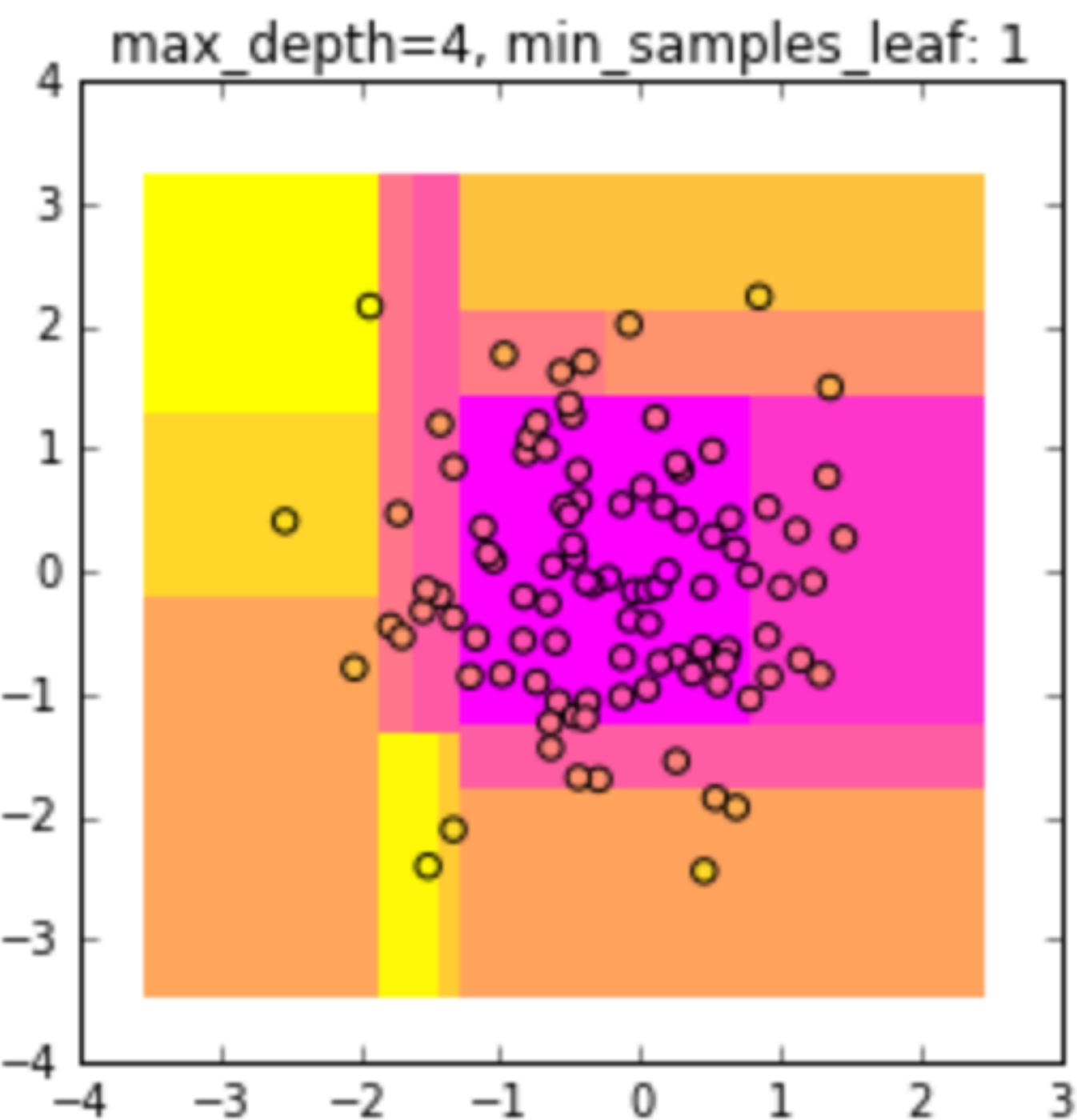
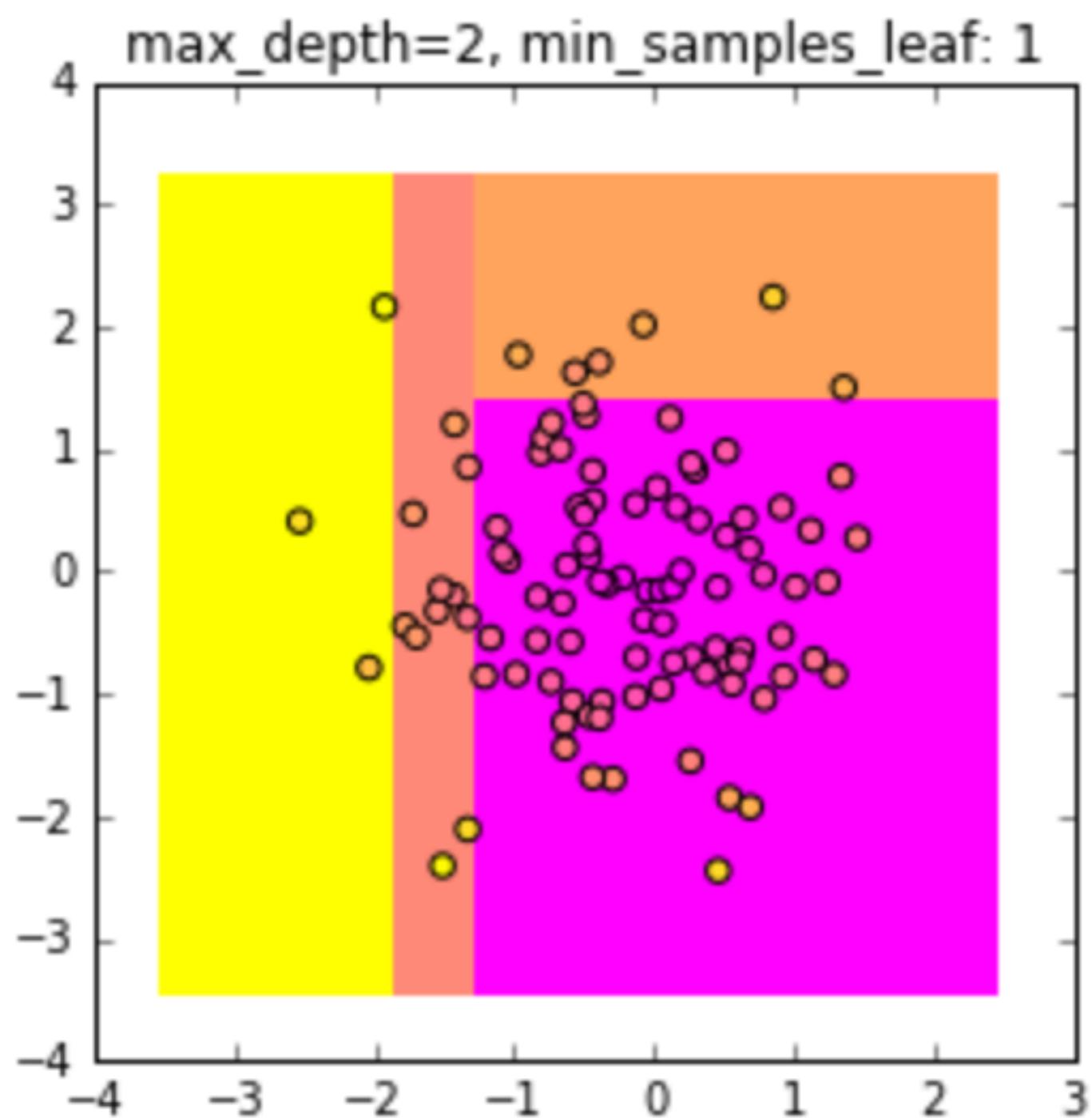


КОЛ-ВО ОБЪЕКТОВ В УЗЛЕ



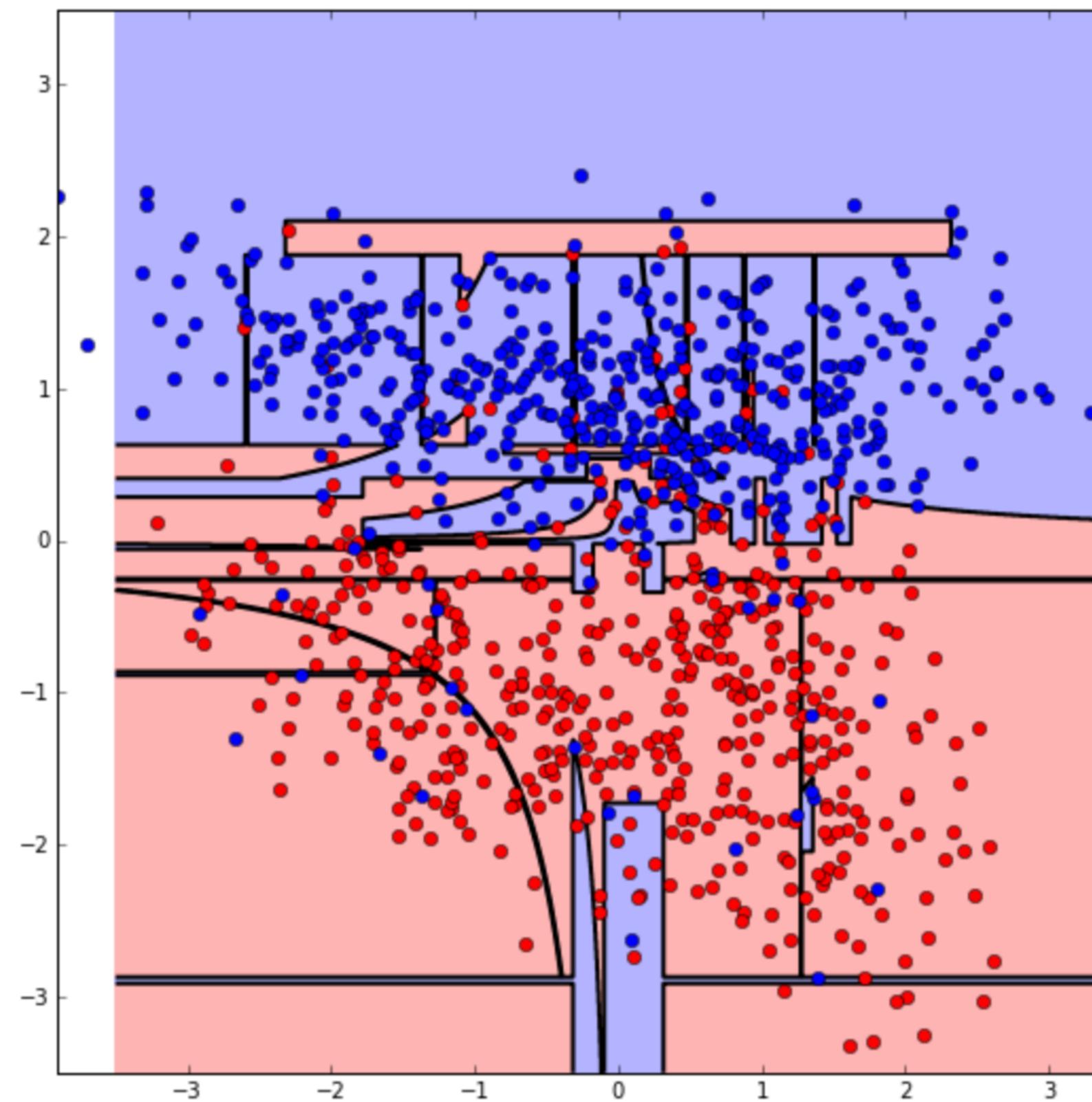
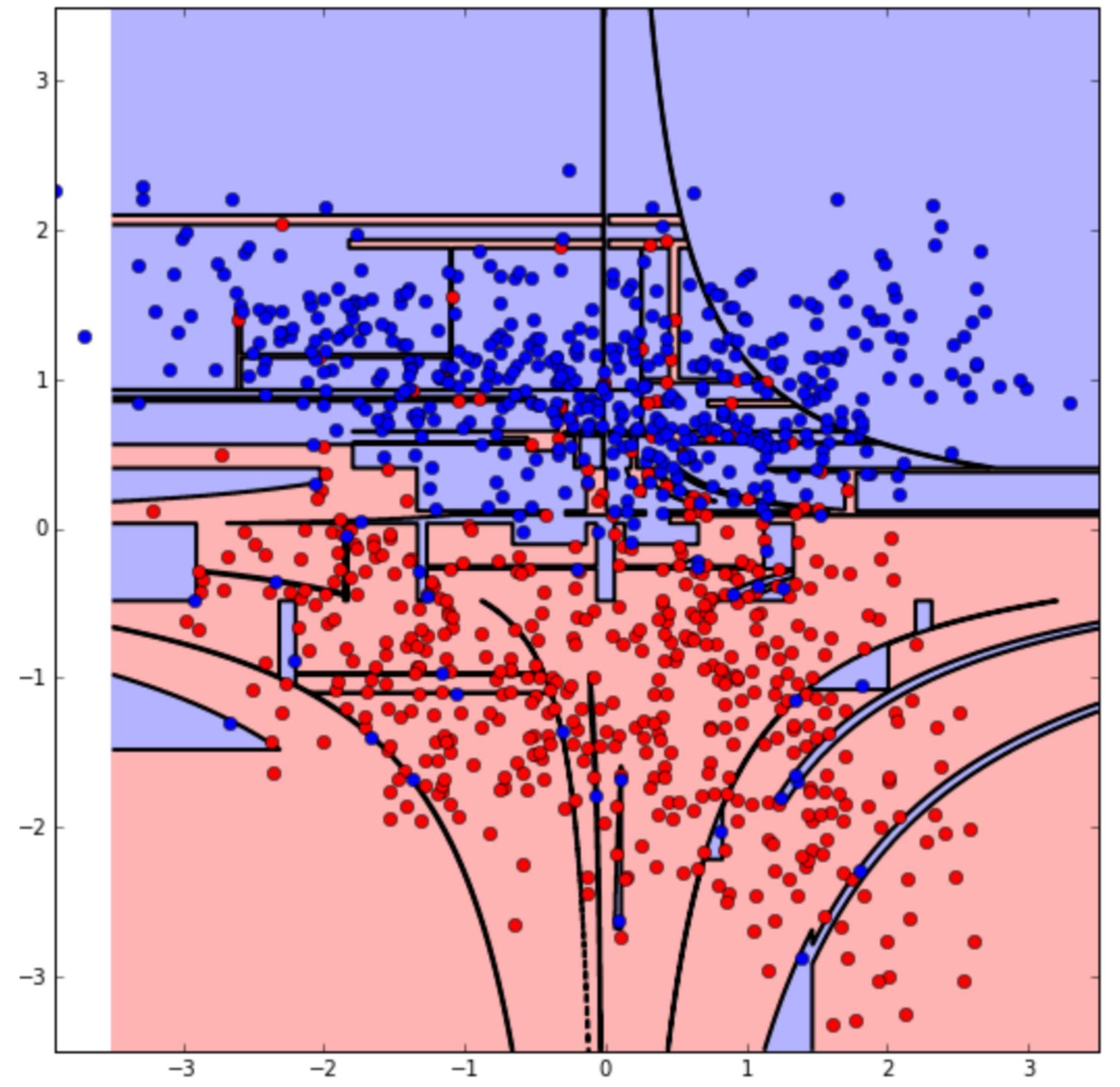


ОГРАНИЧЕНИЕ НА ГЛУБИНУ





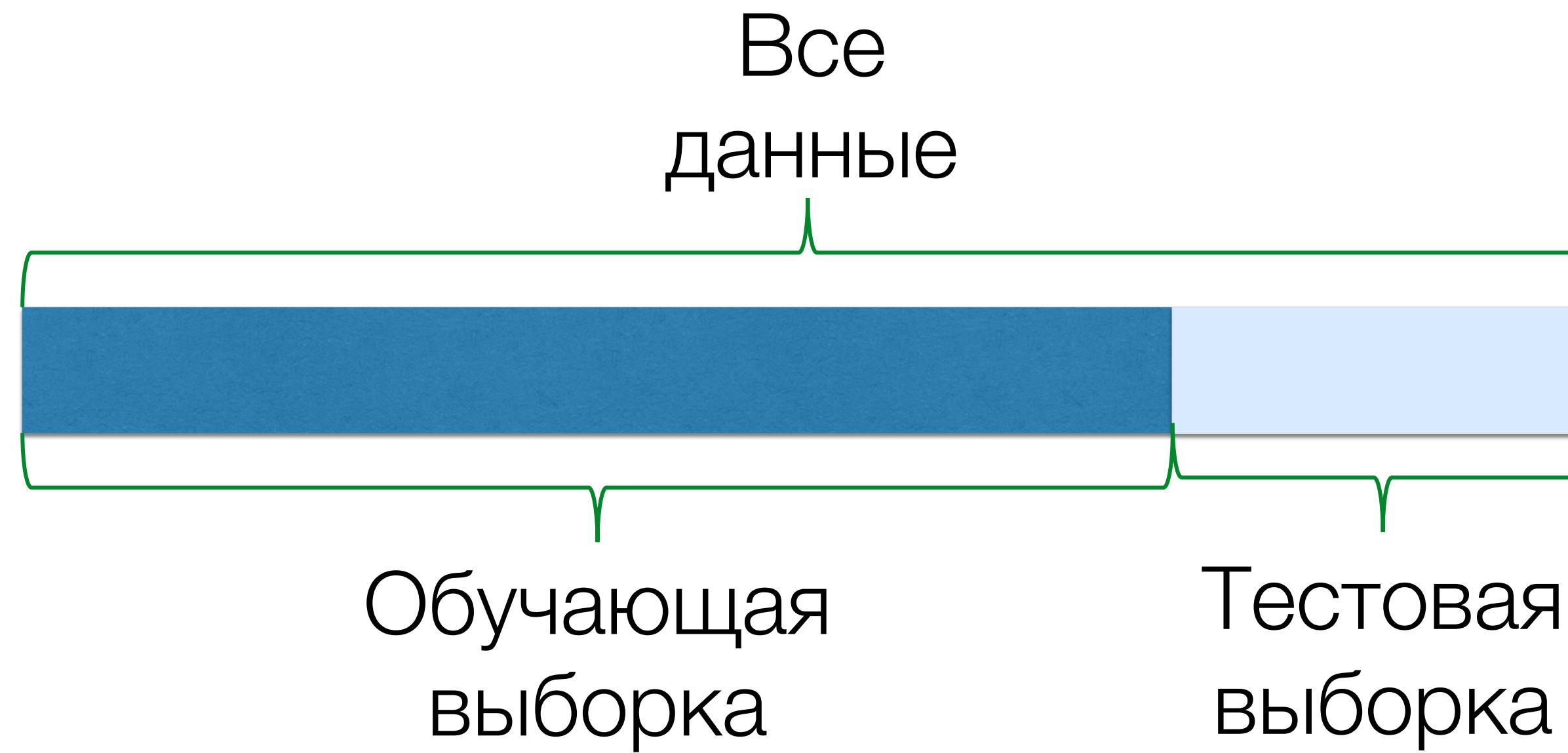
НЕУСТОЙЧИВОСТЬ ДЕРЕВЬЕВ



Валидация



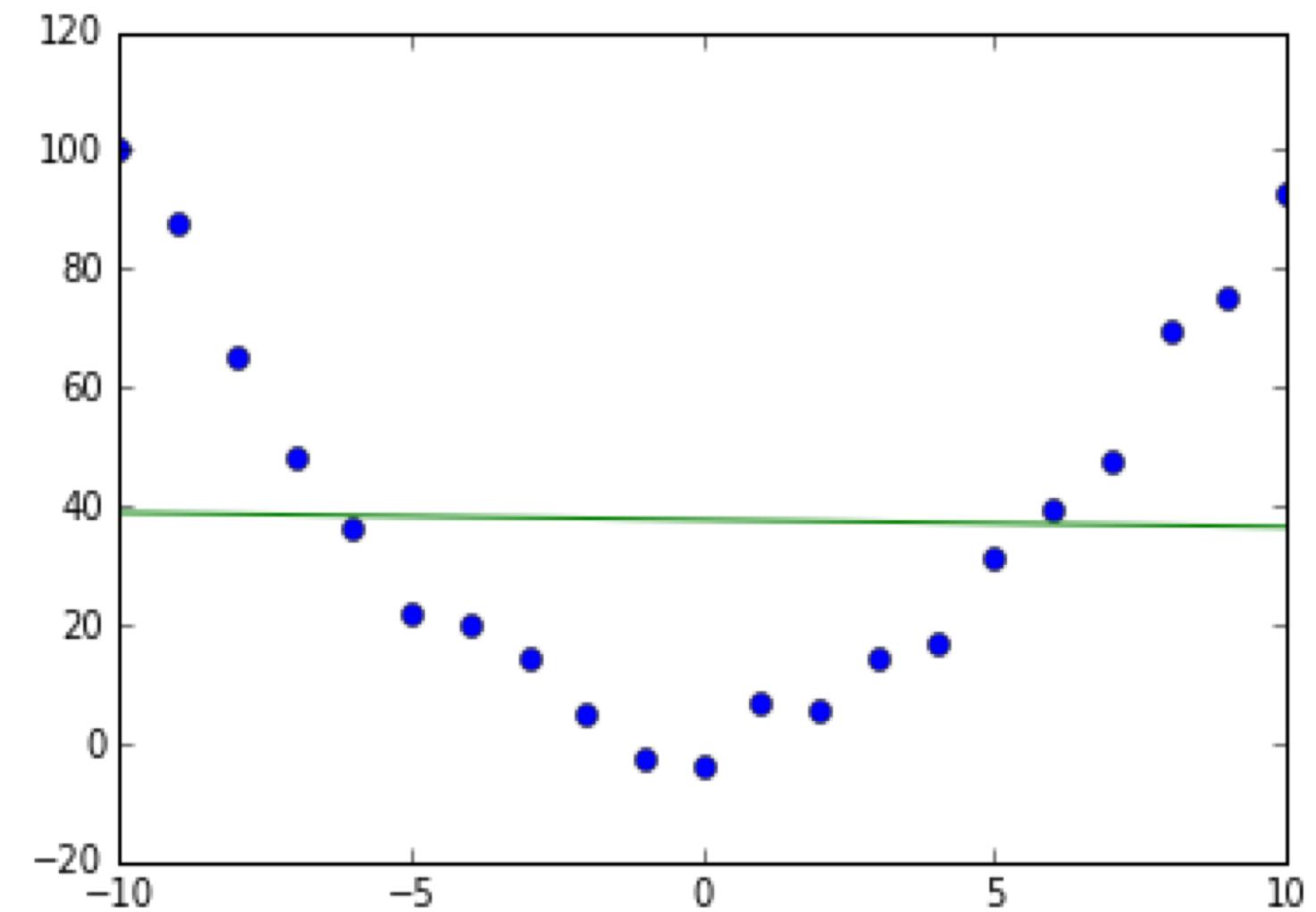
КАК НАЙТИ ЛУЧШЕЕ РЕШЕНИЕ?



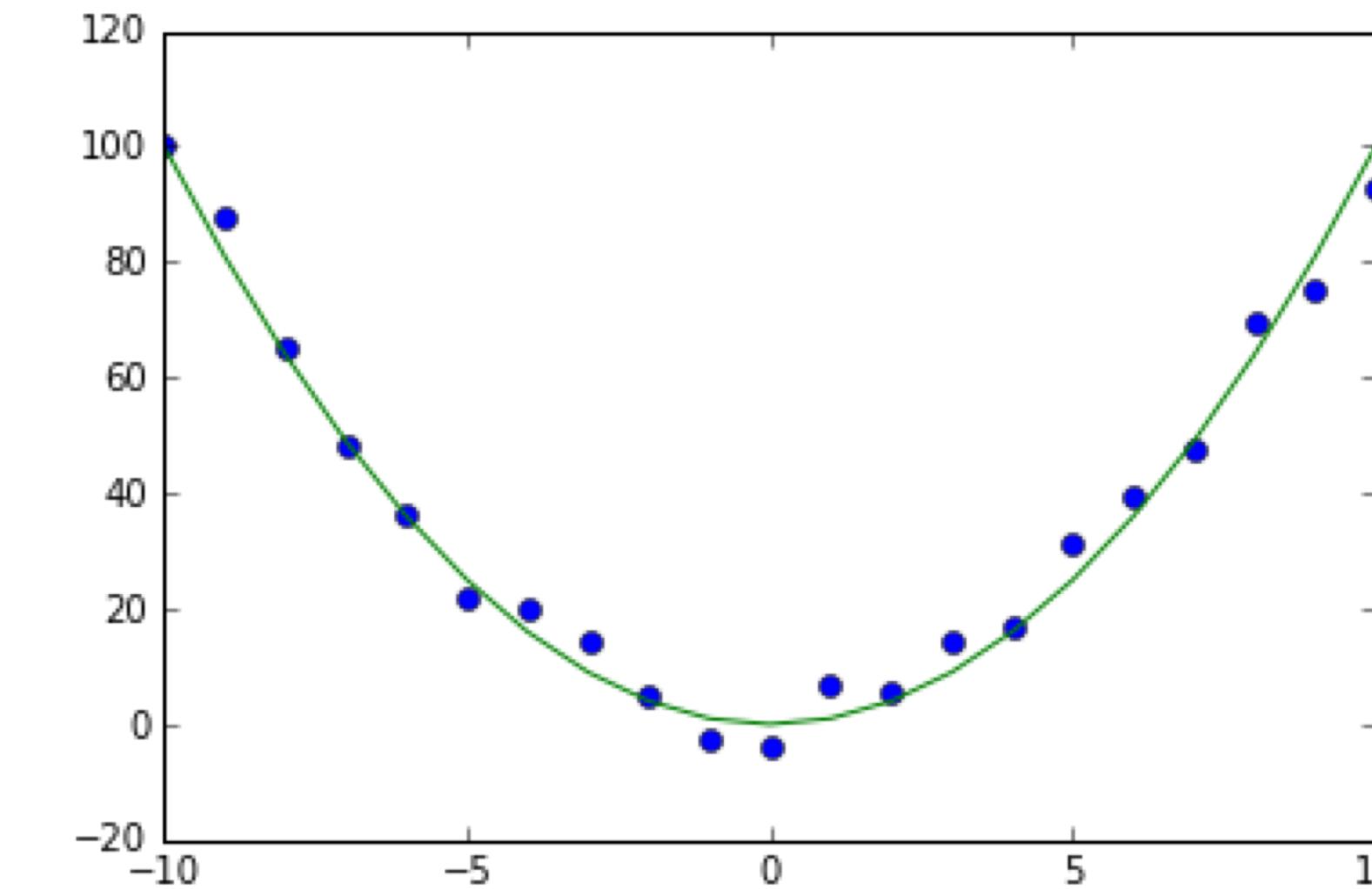
- Алгоритм обучается на обучающей выборке
- Алгоритм тестируется на тестовой выборке (валидационной)



ОБОБЩАЮЩАЯ СПОСОБНОСТЬ АЛГОРИТМА



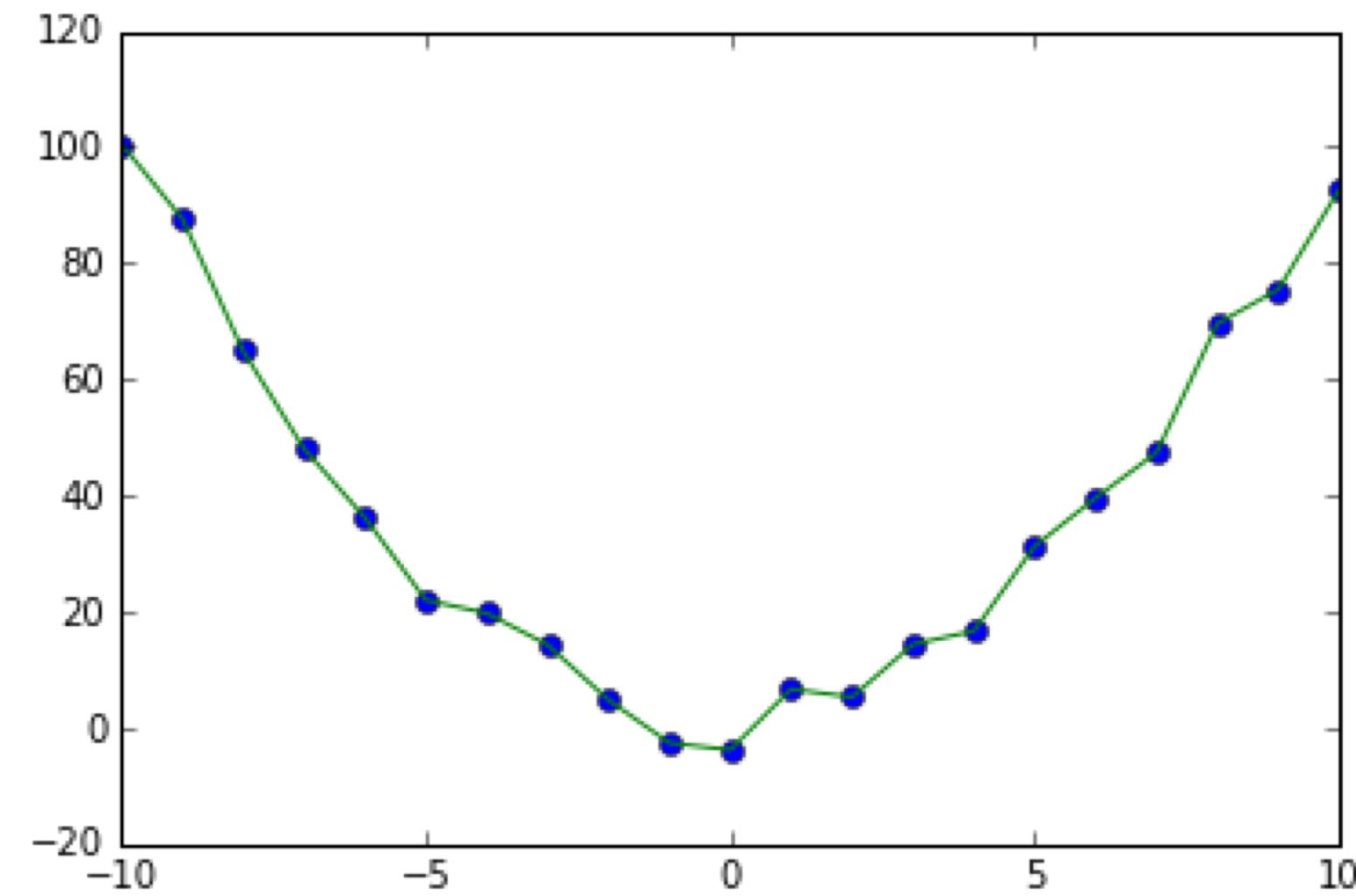
$$E = 22.868$$



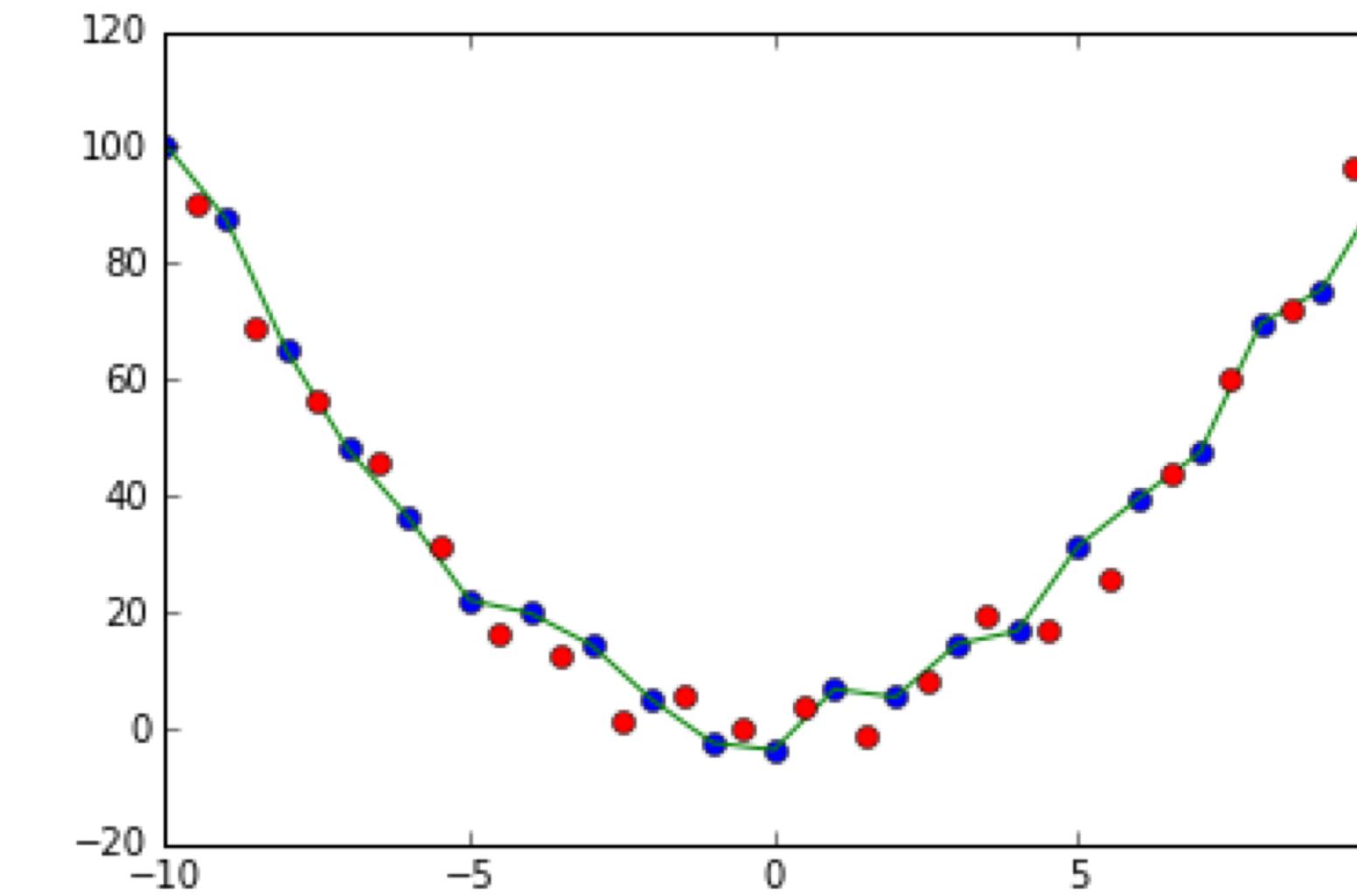
$$E = 2.960$$



ОБОБЩАЮЩАЯ СПОСОБНОСТЬ АЛГОРИТМА



$$E = 0$$



$$E = 3.912$$

Переобучение, как и недообучение – враги наши



СТРАТЕГИИ БОРЬБЫ С НЕОЖИДАННОСТЯМИ

- Отложенная выборка
- Много отложенных выборок
- Кросс-валидация



ОТЛОЖЕННАЯ ВЫБОРКА

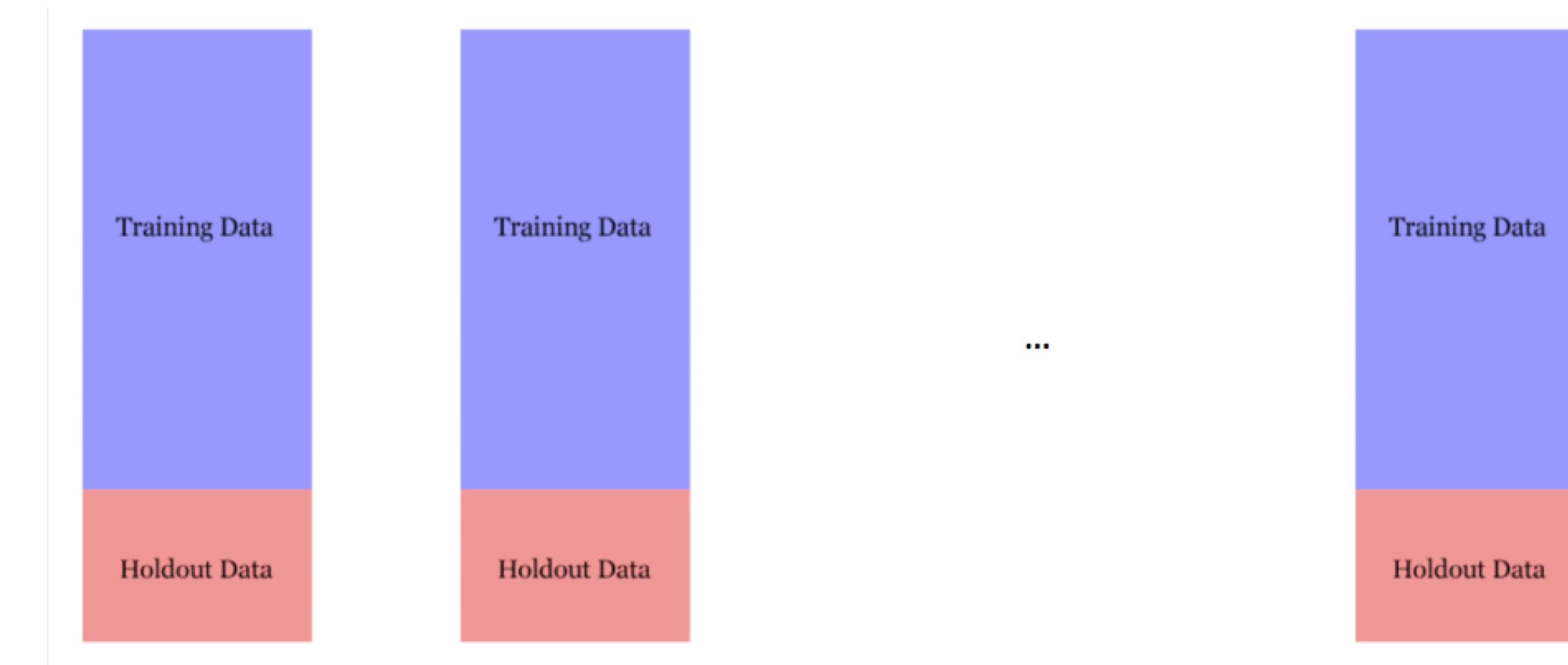
- Пропорции 70/30, 80/20, 0.632/0.368





МНОГО ОТЛОЖЕННЫХ ВЫБОРОК

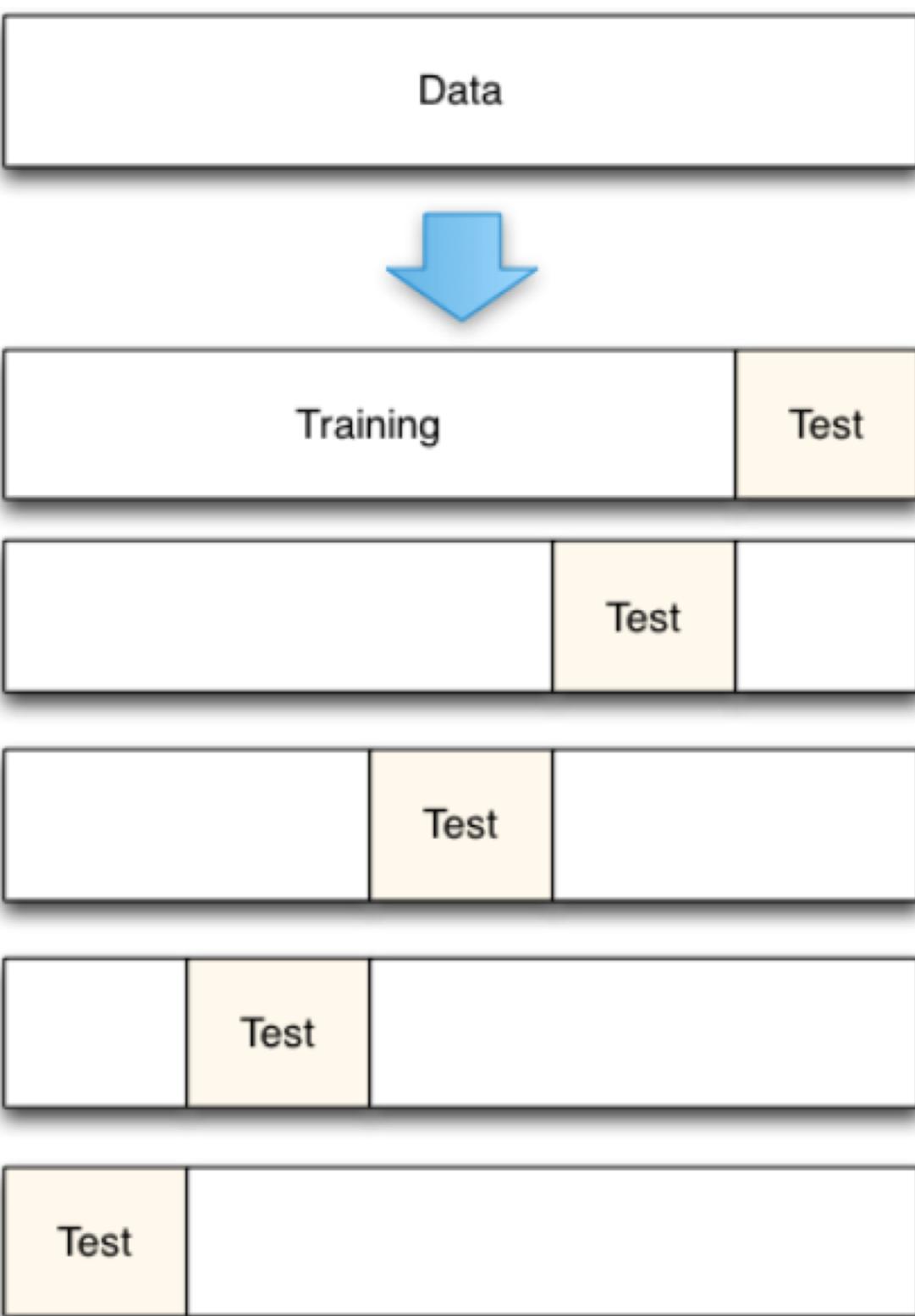
- Пропорции 70/30, 80/20, 0.632/0.368





КРОСС-ВАЛИДАЦИЯ (СКОЛЬЗЯЩИЙ КОНТРОЛЬ)

- k-fold – k-фолдов/блоков



Отдельно благодарим Е.Соколова за материалы к курсу «Машинное обучение» майнора АИД



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ