




Machine Learning

my TRACKER

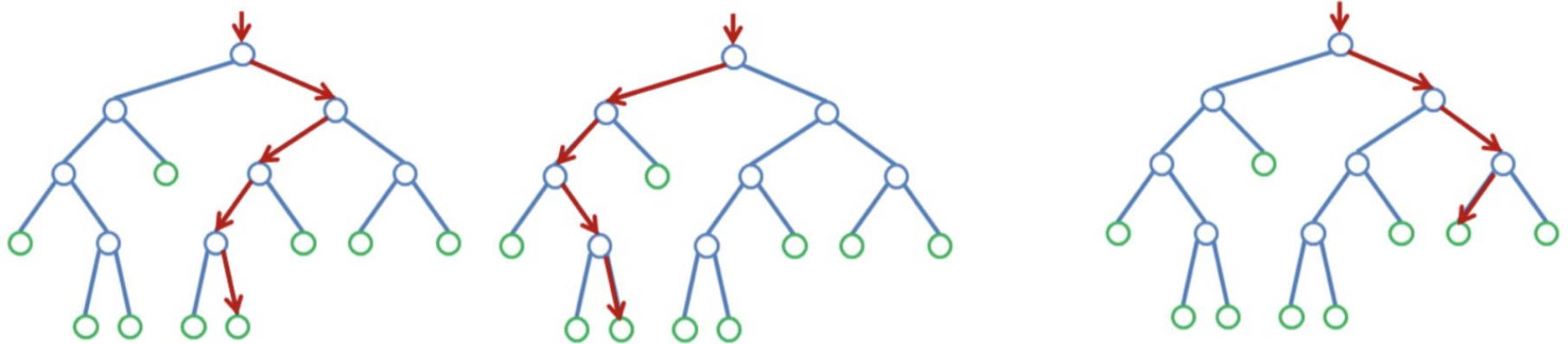


@ mail.ru
group

Композиция алгоритмов

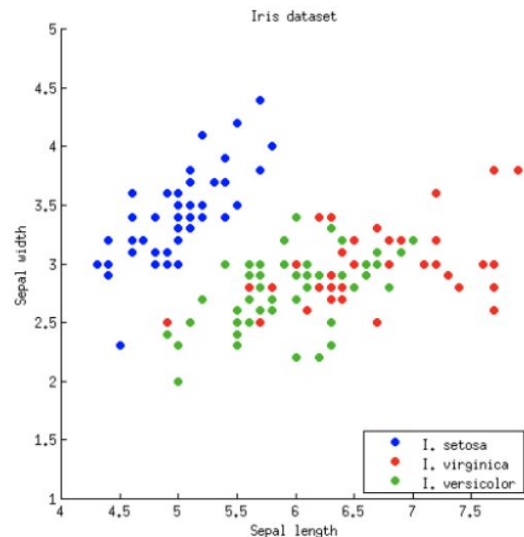


Extremely Randomized Trees



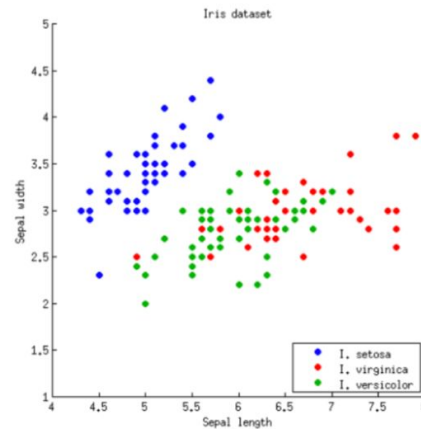
- Проверяем только один признак в узле
- Используем все признаки для построения дерева
- Работают в разы быстрее, чем обычные деревья решений

Perfect random trees ensembles



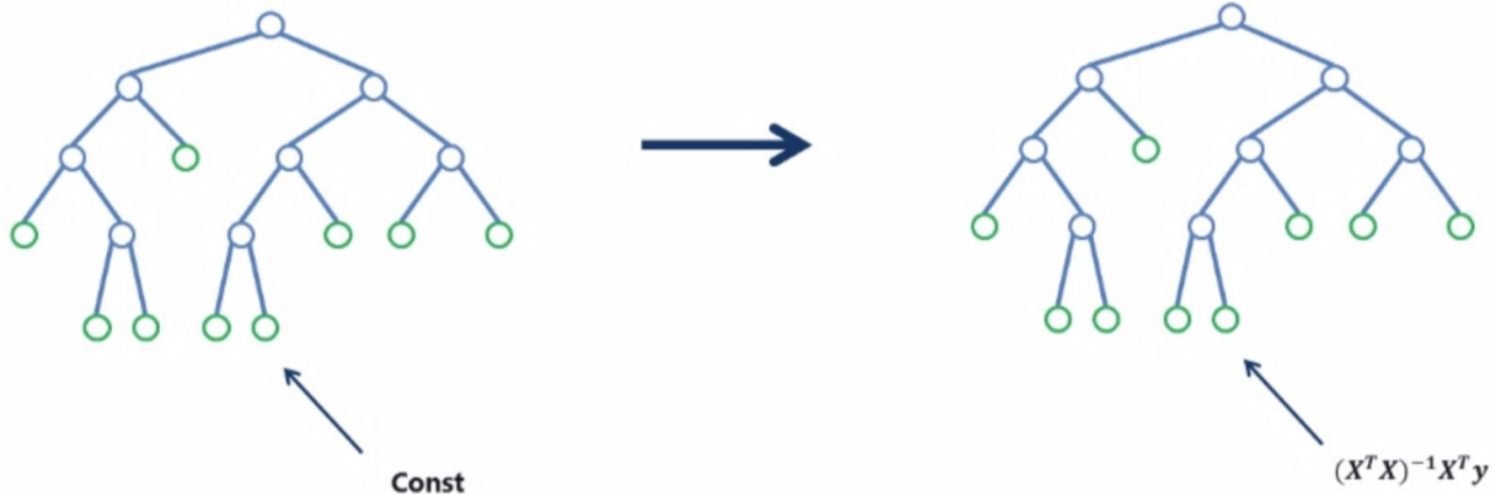
- Быстрое построение деревьев
- Качество сопоставимо с Random Forest
- Работает только с задачами классификации

Дерево решений с линейными признаками (вариант с сеткой)



- Внутри узла перебираем не только существующие признаки, но и добавляем новые линейные признаки, построенные по сетке.

Модельное дерево



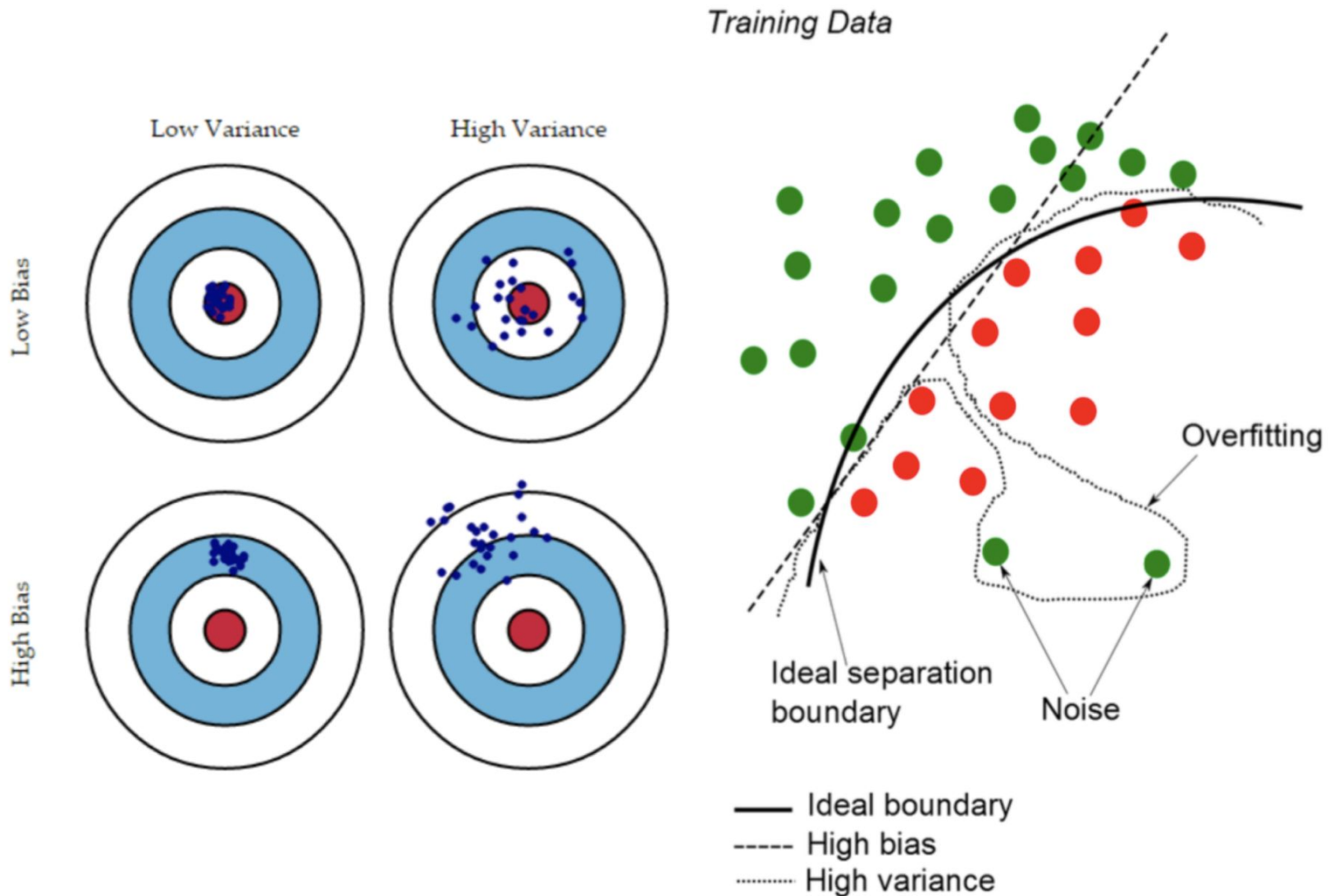
- Поместим в узлы дерева решений другой алгоритм вместо констант

- Задача регрессии
- $y = f(x) + \epsilon$
- $\epsilon = \mathcal{N}(0, \sigma_\epsilon)$
- Ищем $\hat{f}(x)$ аппроксимирующую $f(x)$. Мы можем оценить матожидание среднеквадратичной ошибки для некоторой точки x_0

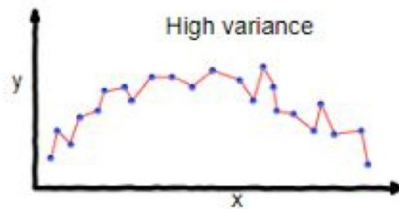
$$Err(x_0) = E[(y - \hat{f}(x_0))^2]$$

$$Err(\vec{x}) = \text{Bias}(\hat{f})^2 + \text{Var}(\hat{f}) + \sigma^2$$

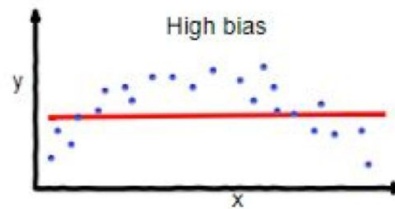
The bias-variance decomposition



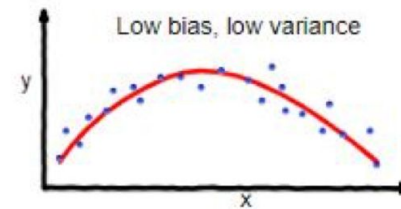
The bias-variance decomposition



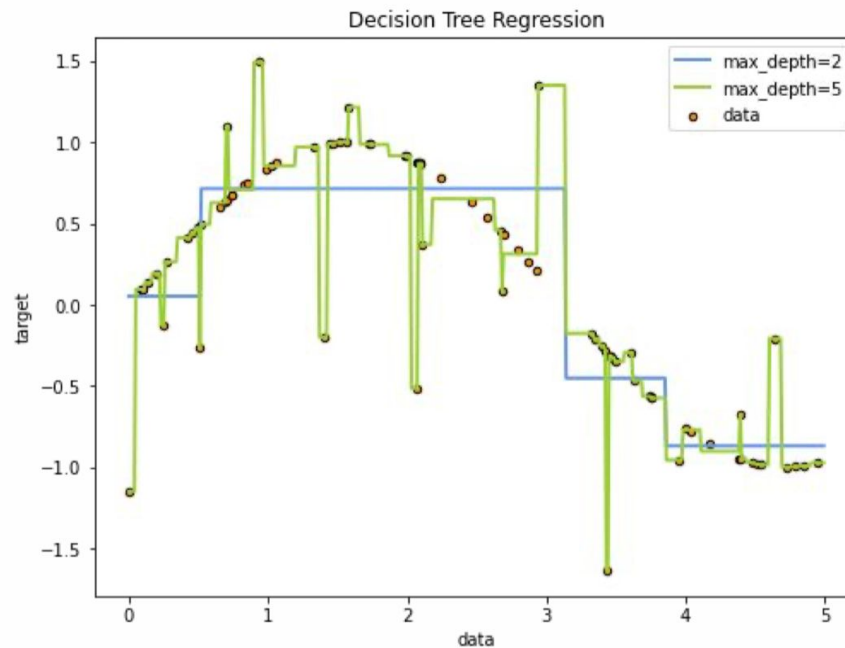
overfitting



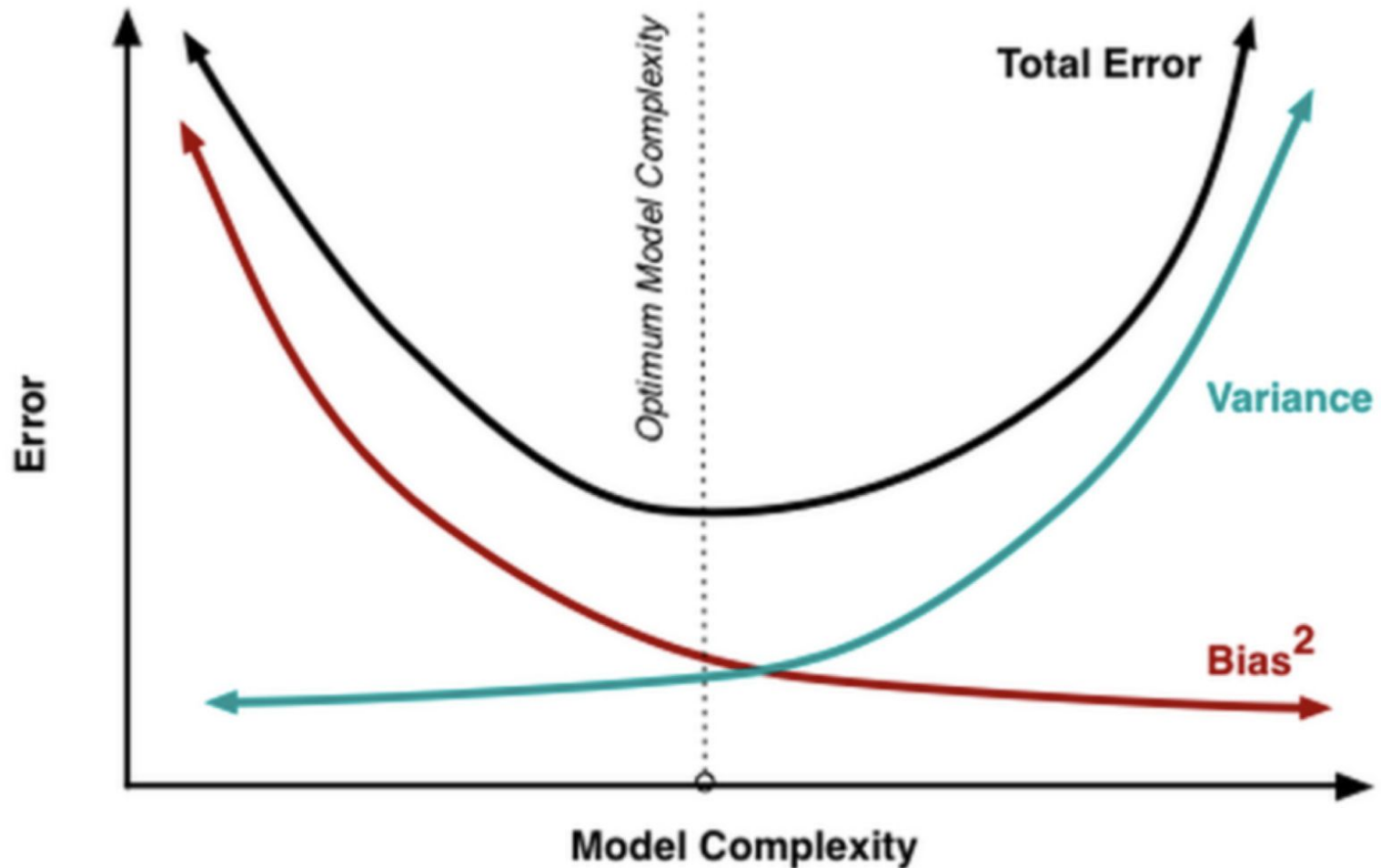
underfitting



Good balance



The bias-variance decomposition



Теорема Кондорсе «о жюри присяжных» (1784).

Если каждый член жюри присяжных имеет независимое мнение, и если вероятность правильного решения члена жюри больше 0.5, то тогда вероятность правильного решения присяжных в целом возрастает с увеличением количества членов жюри и стремится к единице.

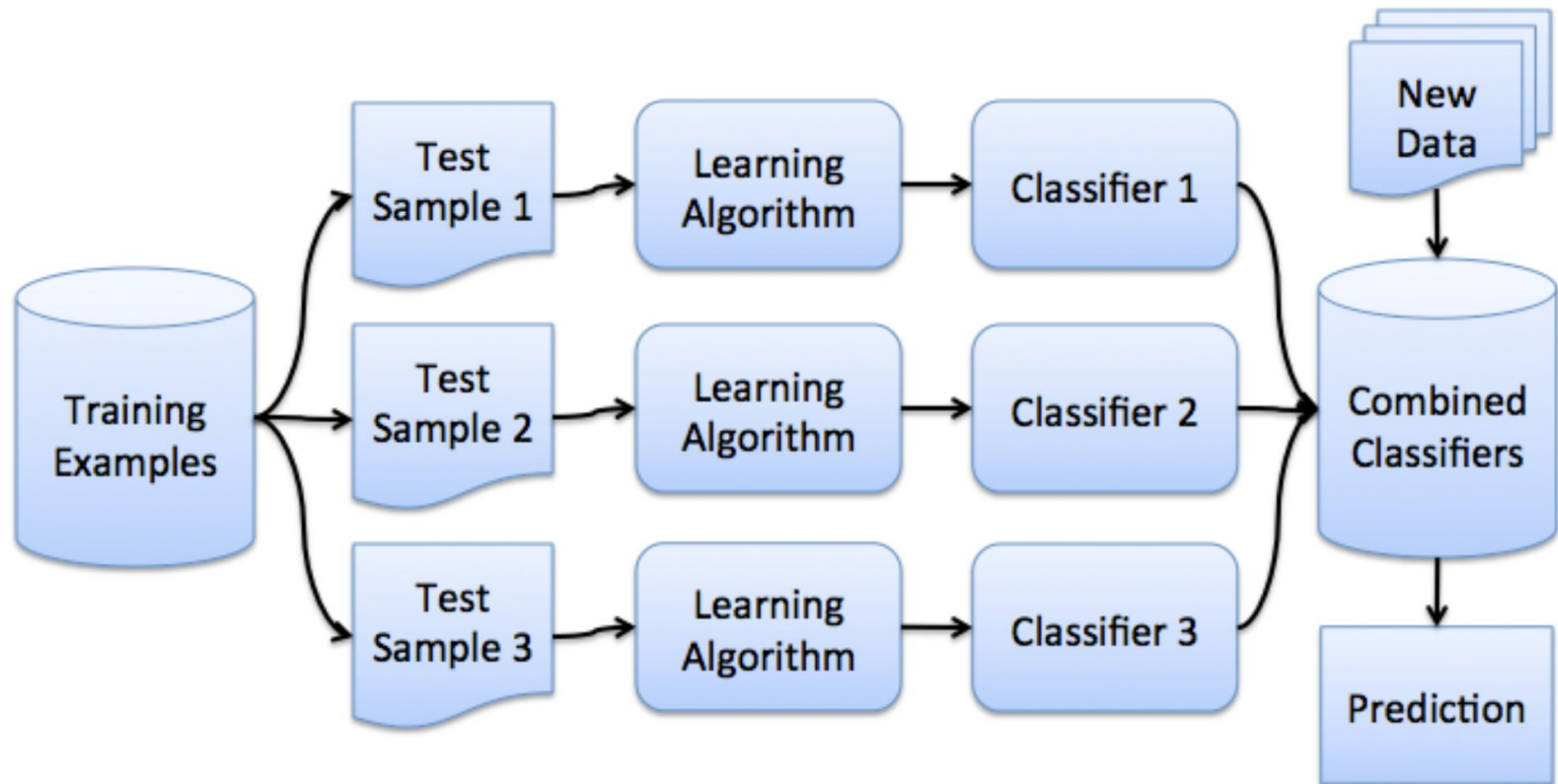
"Мудрость толпы»(1906).

Фрэнсис Гальтон посетил рынок, где проводилась некая лотерея для крестьян. Их собралось около 800 человек, и они пытались угадать вес быка, который стоял перед ними. Бык весил 1198 фунтов. Ни один крестьянин не угадал точный вес быка, но если посчитать среднее от их предсказаний, то получим 1197 фунтов.

Bootstrap – метод генерации выборки с помощью выбора элементов исходной выборки с возвращением.

Bootstrap используется для корректировки смещения, тестирования гипотез, построения доверительных интервалов.

Bagging

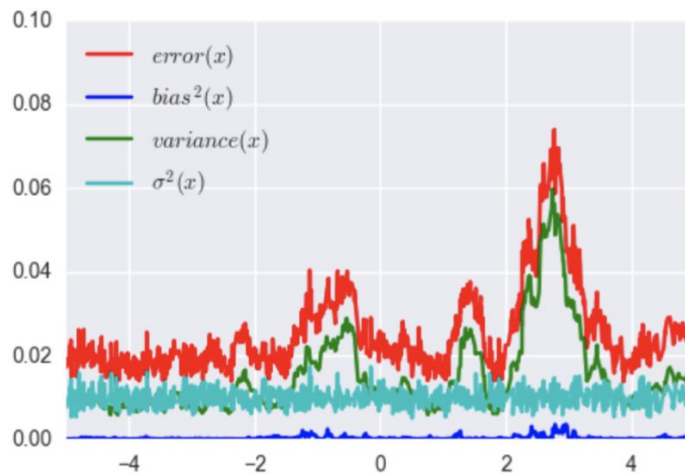
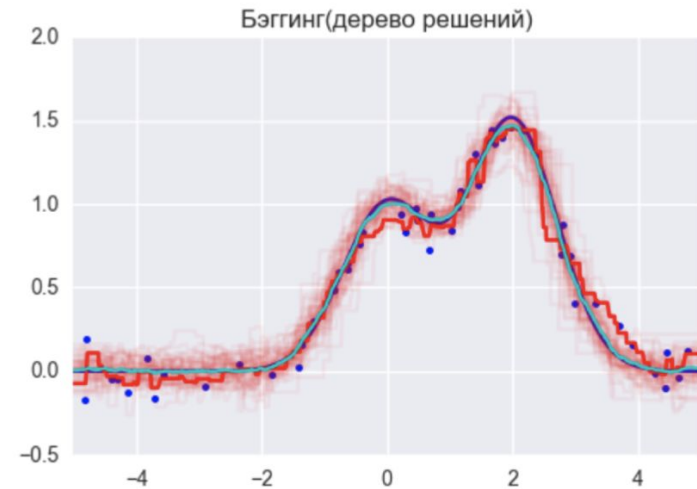
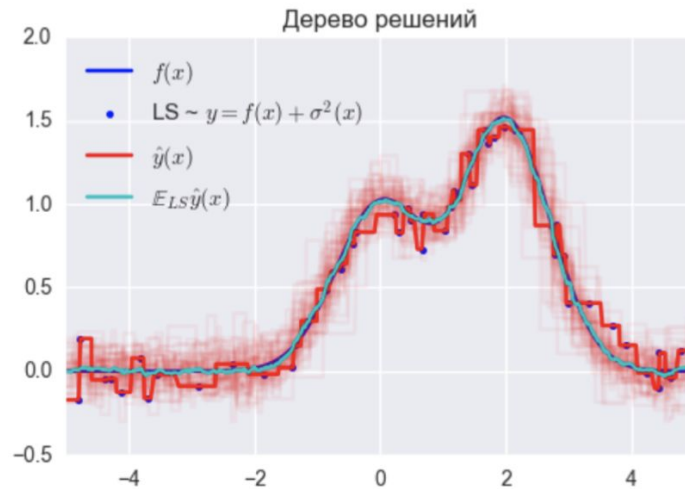


$$E_1 = \frac{1}{n} E_x \sum_{i=1}^n \varepsilon_i^2(x)$$

$$\begin{aligned} E_n &= E_x \left(\frac{1}{n} \sum_{i=1}^n b_i(x) - y(x) \right)^2 \\ &= E_x \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \right)^2 \\ &= \frac{1}{n^2} E_x \left(\sum_{i=1}^n \varepsilon_i^2(x) + \sum_{i \neq j} \varepsilon_i(x) \varepsilon_j(x) \right) \\ &= \frac{1}{n} E_1 \end{aligned}$$

Бэггинг позволяет снизить дисперсию (variance) обучаемого классификатора

Bagging



Ошибка дерева решений

$$0.0255(Err) = 0.0003(Bias^2) + 0.0152(Var) + 0.0098(\sigma^2)$$

Ошибка бэггинга на деревьях решений

$$0.0196(Err) = 0.0004(Bias^2) + 0.0092(Var) + 0.0098(\sigma^2)$$

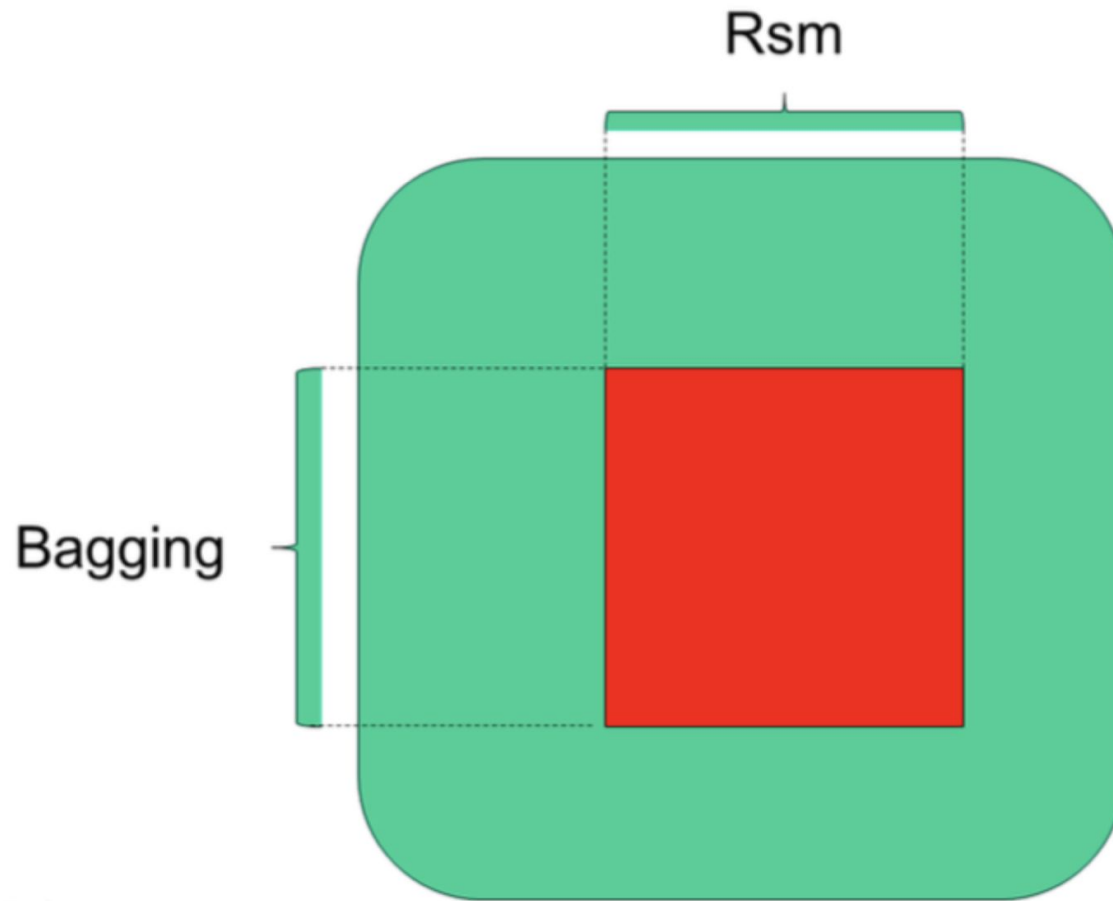
- **Bagging = Bootstrap aggregation**

Обучаем алгоритмы по случайным подвыборкам размера N , полученным с помощью выбора с возвращением.

- **RSM = Random Subspace Method**

Метод случайных подпространств. Обучаем алгоритмы по случайным подмножествам признаков.

Благодаря описанным стратегиям добиваемся максимального различия между базовыми алгоритмами.

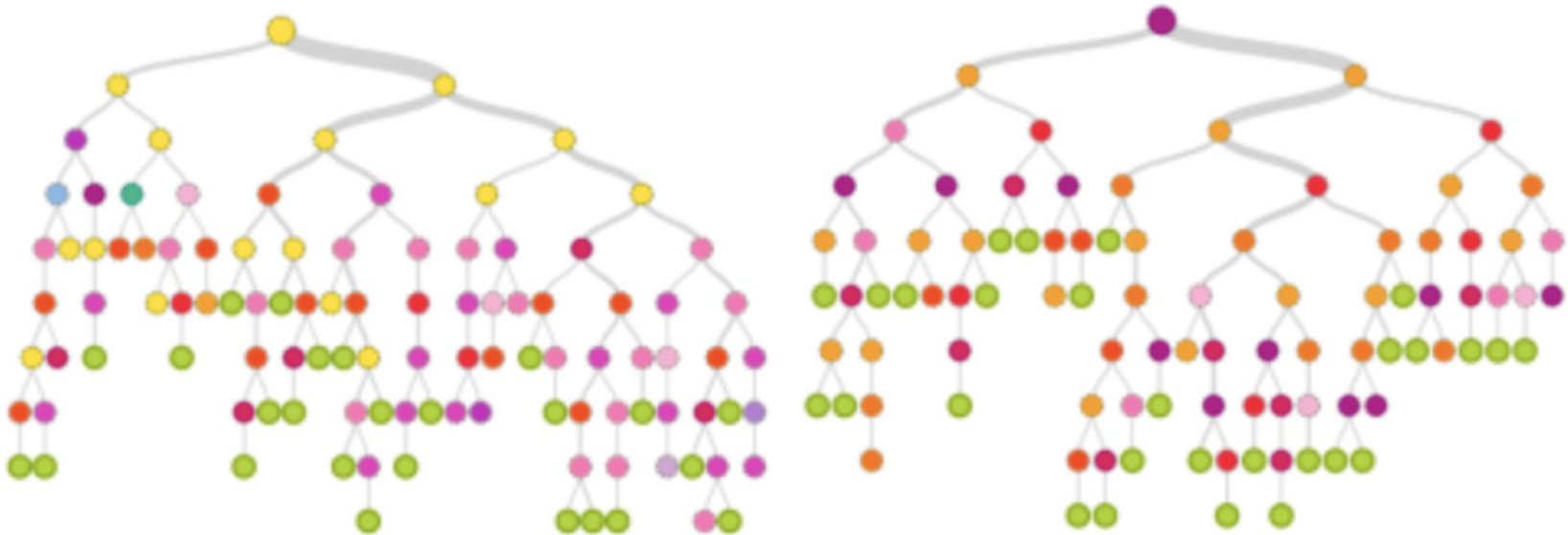


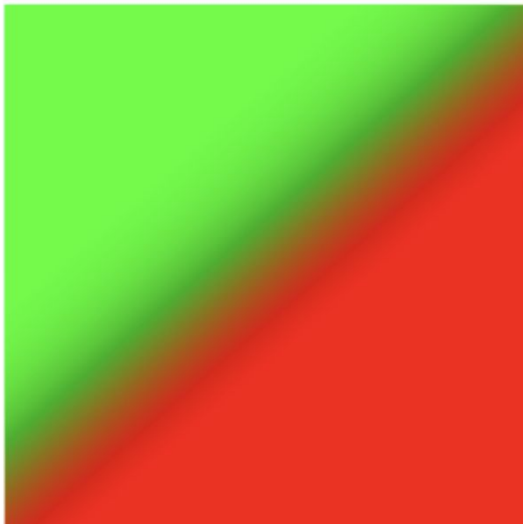
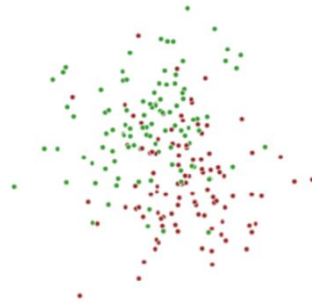
Случайный лес (random forest) - bagging & rsm над решающими деревьями.



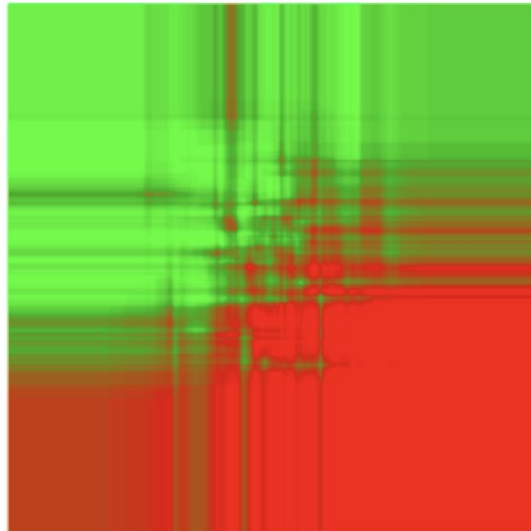
Неустойчивость.

Незначительные изменения в данных приводят к значительным изменениям в топологии дерева.

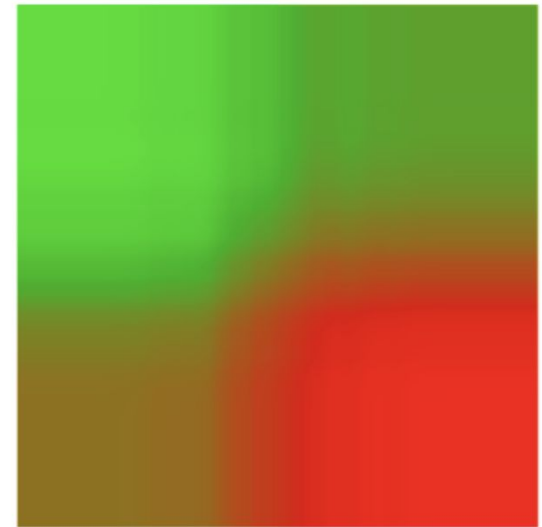




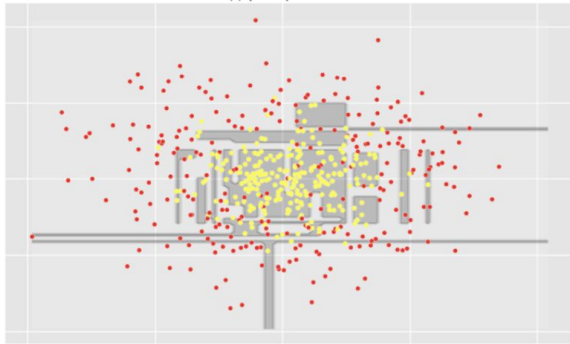
(a) Original data



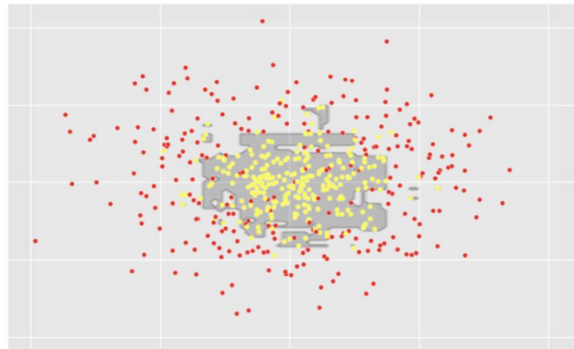
(b) RF (50 Trees)



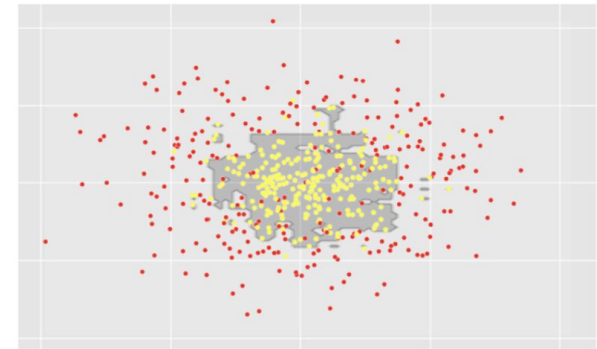
(c) RF (2000 Trees)



Дерево решений



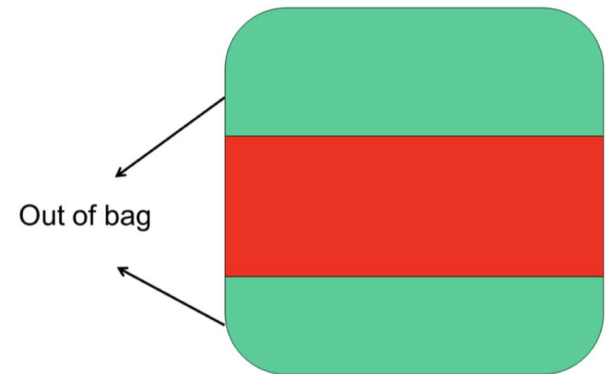
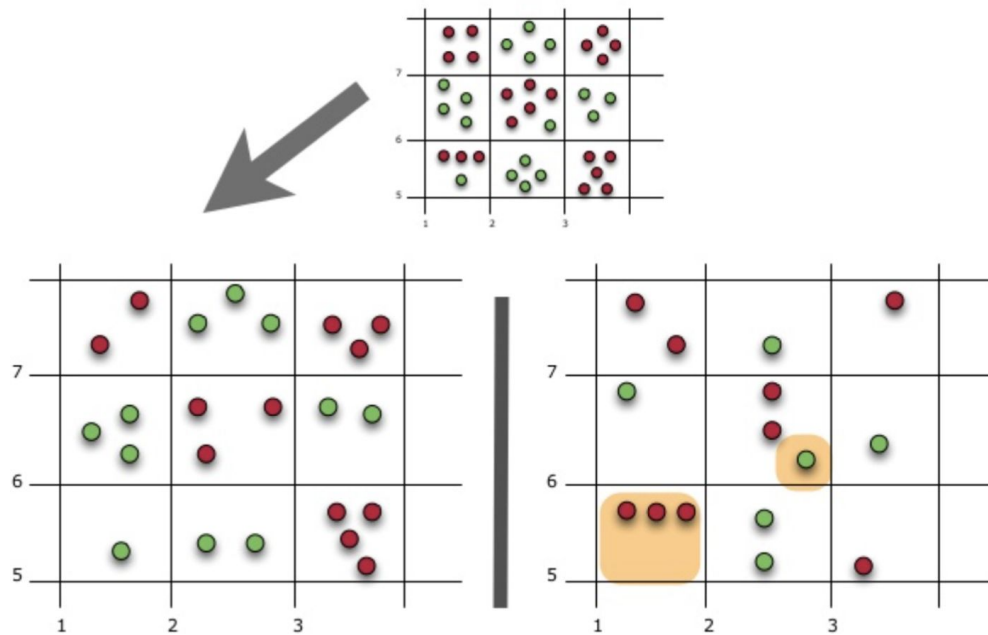
Бэггинг на деревьях
решений

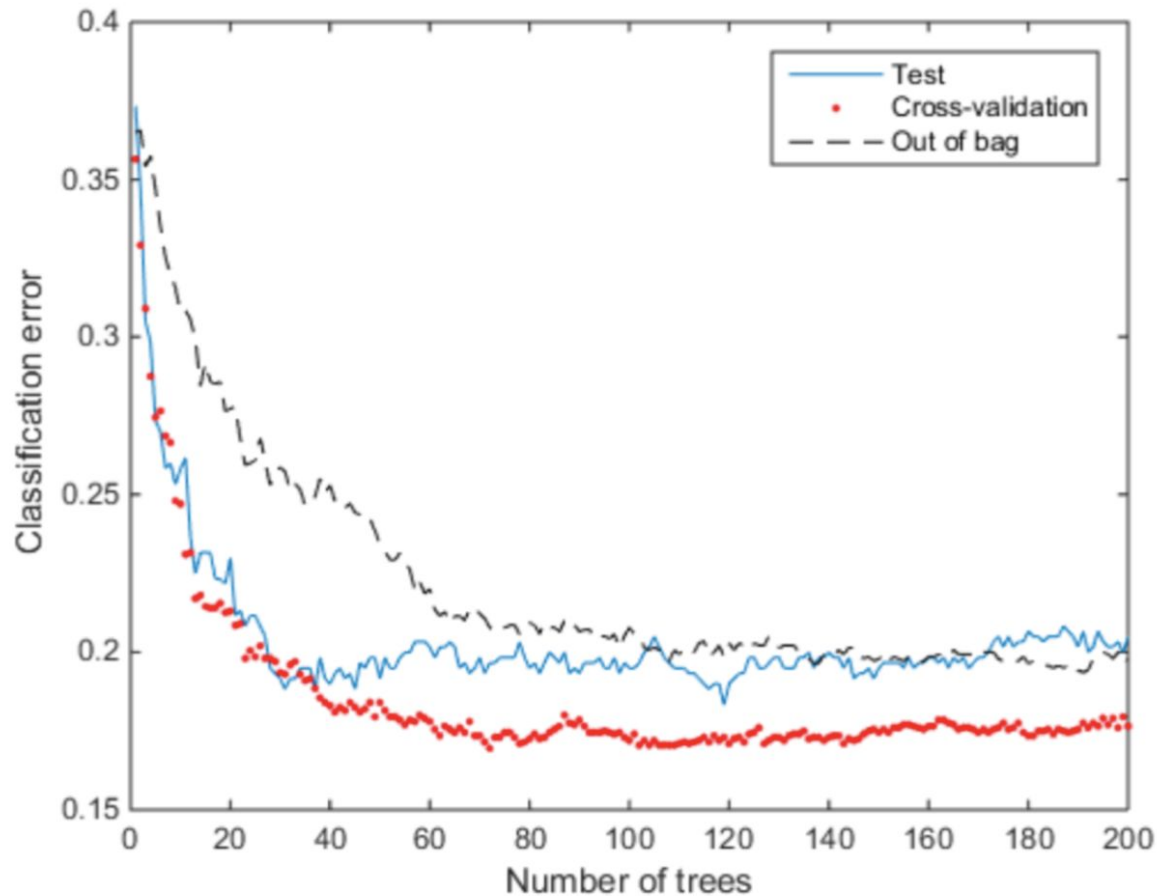


Случайный лес

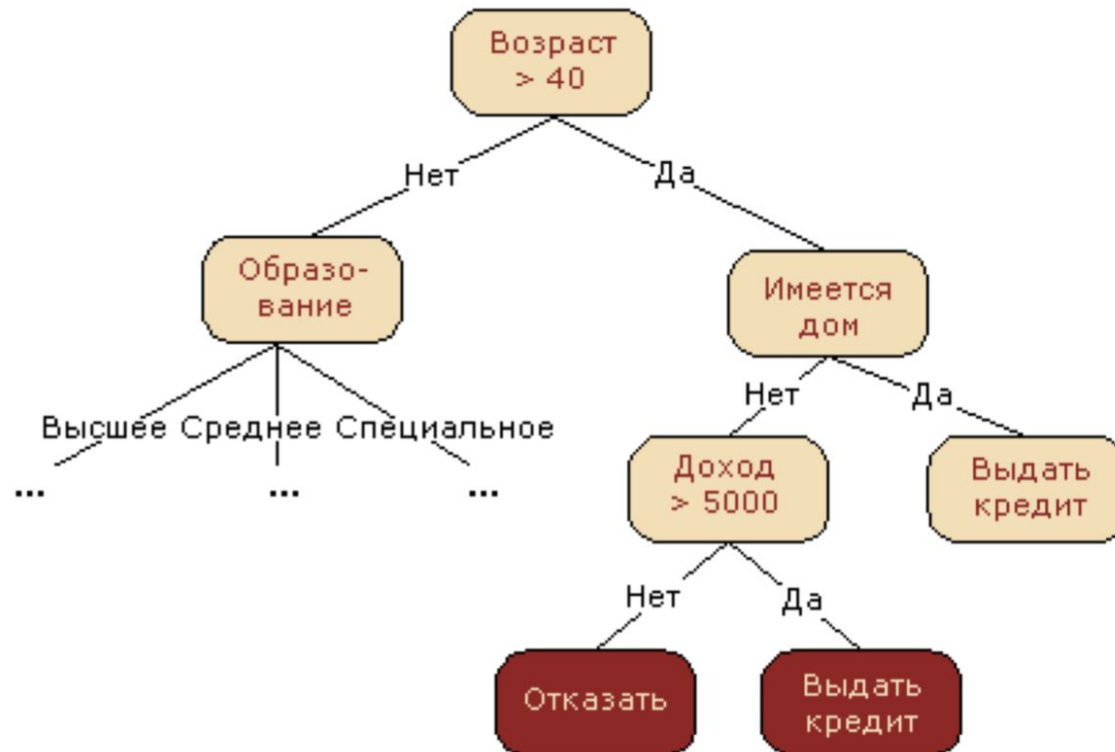
Out Of Bag

OOBE



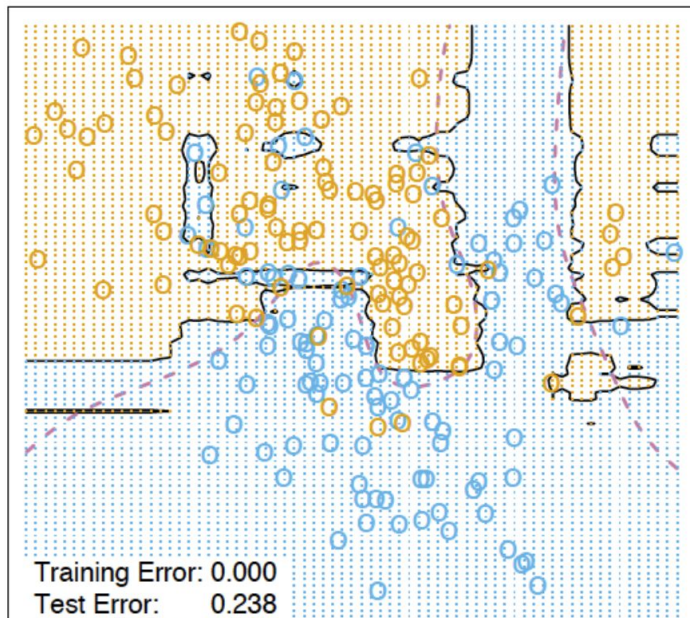


Не переобучается с ростом числа алгоритмов*

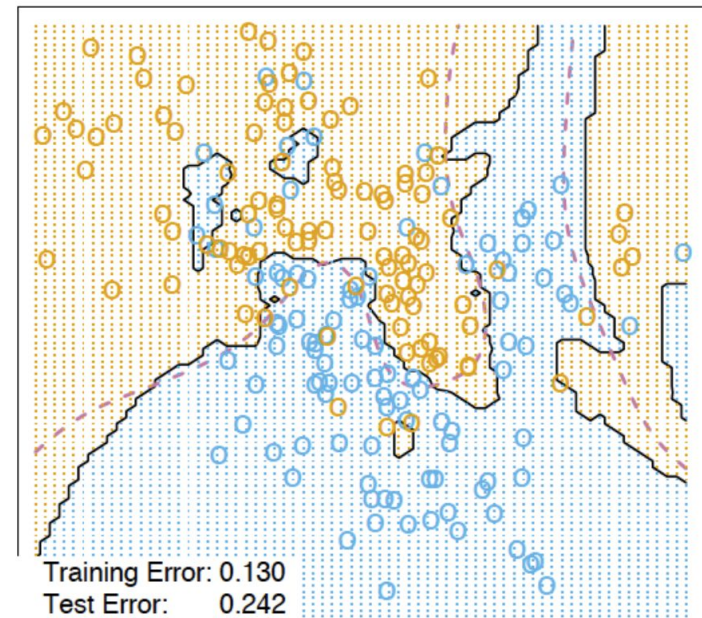


Признак	Важность признака
Возраст	1
Образование	0.67
Наличие дома	0.5
Доход > 5000	0.33

Random Forest Classifier



3-Nearest Neighbors



Плюсы

- Алгоритм прост
- Не переобучается*
- Хорошо параллелится
- Не требует сложной настройки параметров
- Не требует нормализации данных

Минусы

- Модели получаются большие и не интерпретируемыми
- Плохо работает с полиномиальными зависимостями
- Достаточно медленно работает для большого объема данных