

Юстина Иванова

Программист, data scientist

Статистика в python. Кейс-стади №1.
Датасеты: faulty steel plates,
Titanic.

Спикер

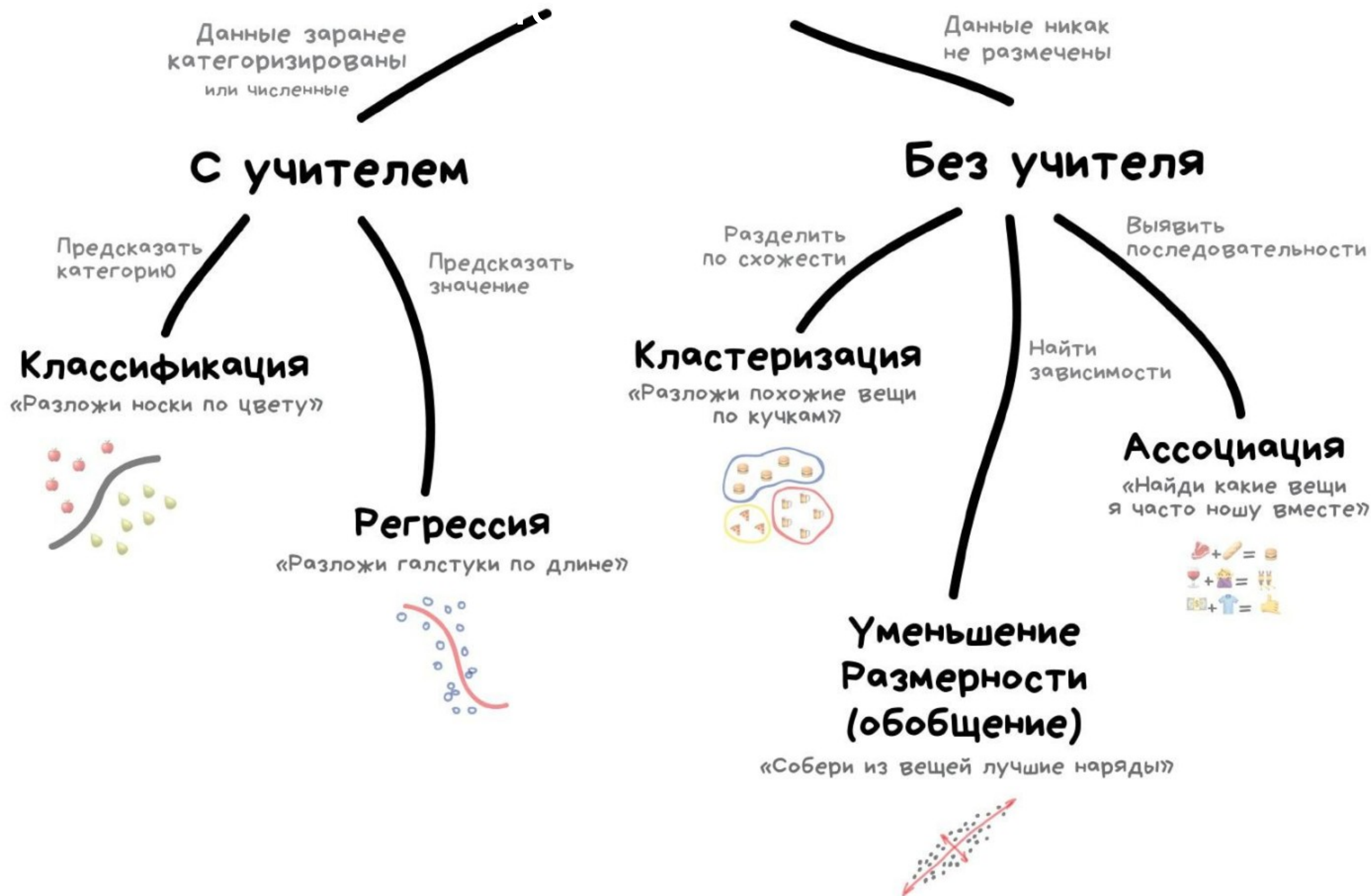


Юстина Иванова,

- PhD в Университете Больцано
- Data scientist по компьютерному зрению в компании ОЦРВ,
- Выпускница МГТУ им. Баумана
- Магистр по Artificial Intelligence в University of Southampton

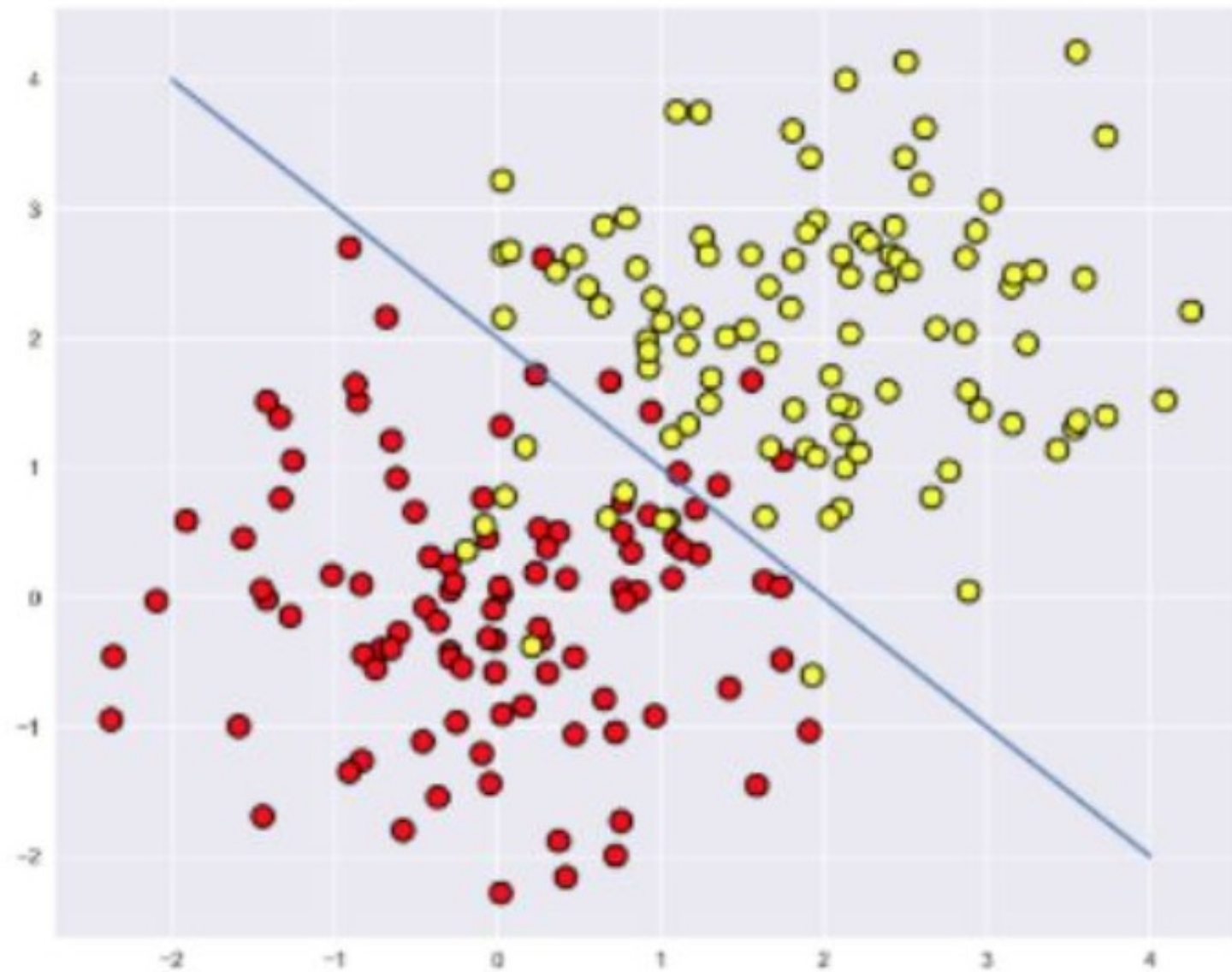
Классическое Обучение

ОЛОГИЯ

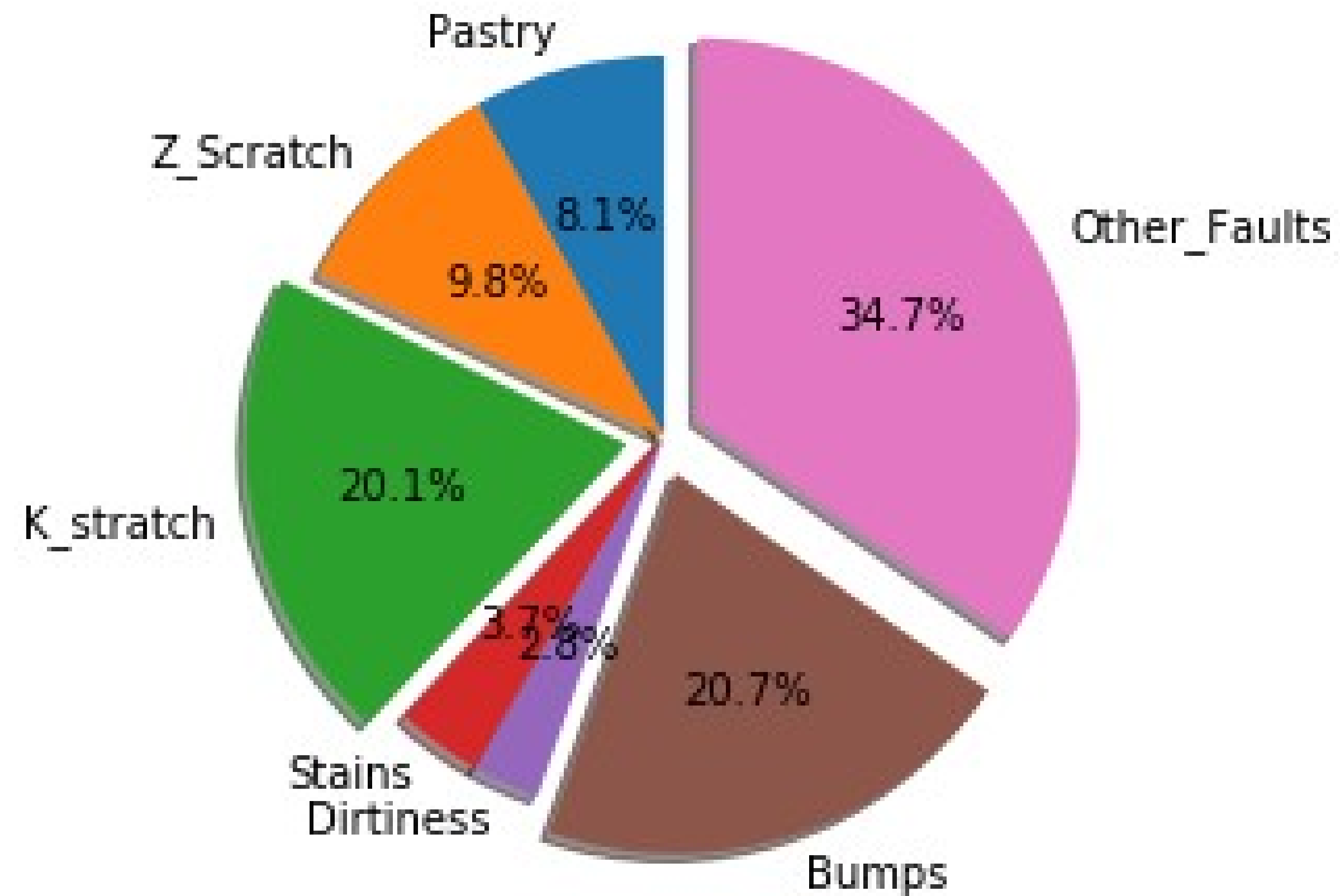


Классификация

Множество допустимых ответов конечно. Их называют метками классов (class label). Класс — это множество всех объектов с данным значением метки.



Проблема несбалансированности классов.

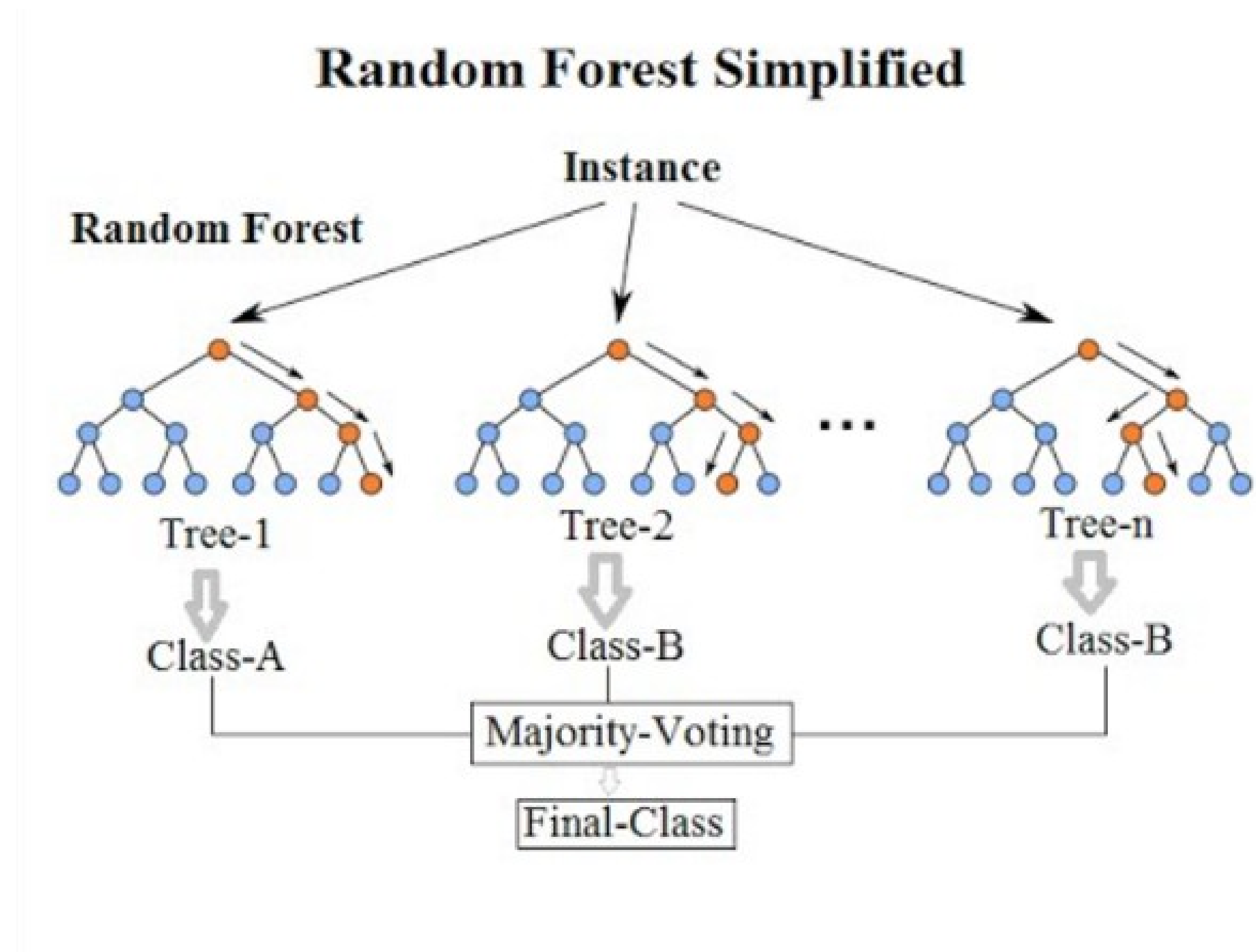


Деревья решений.

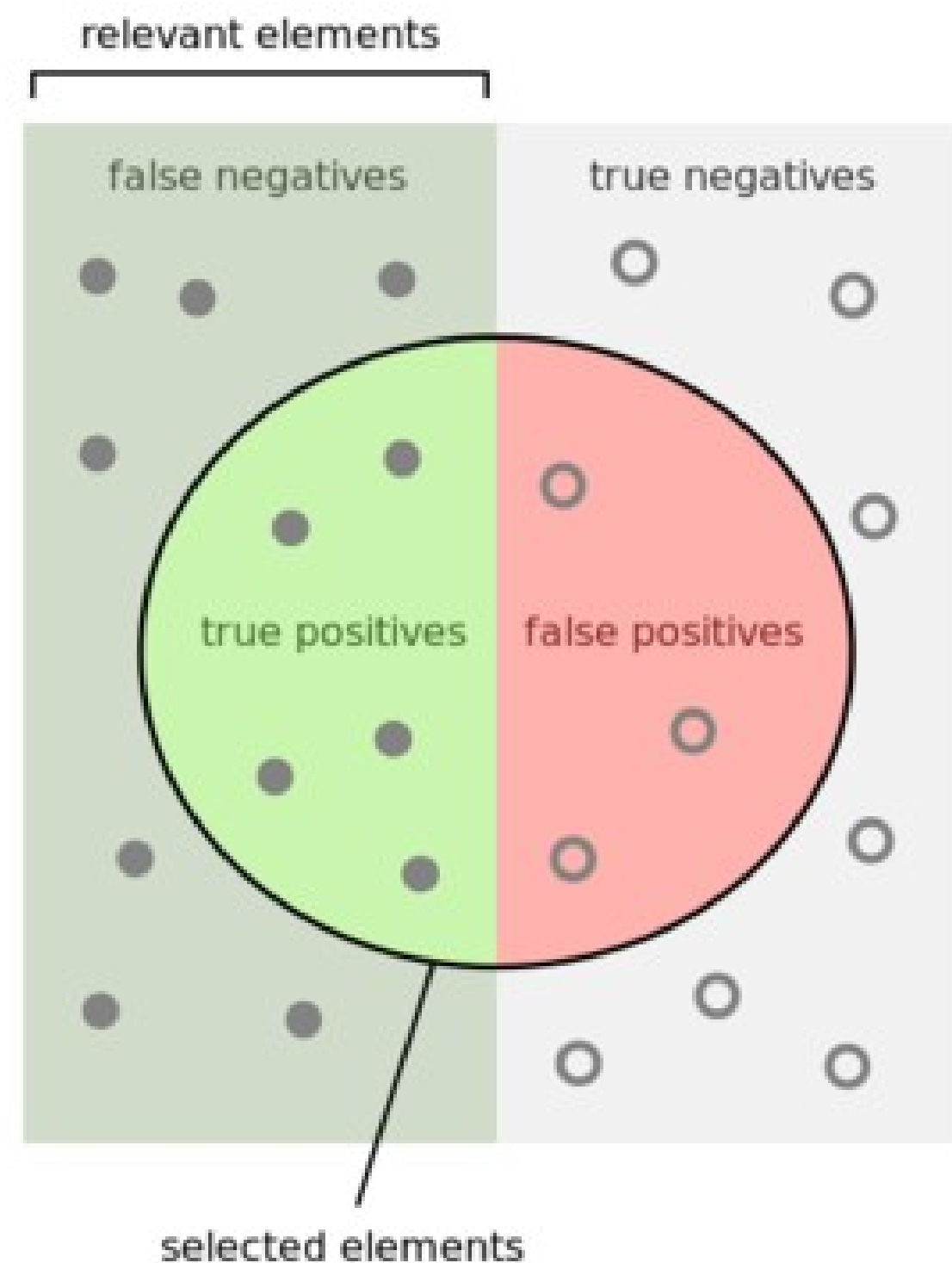


Дерево Решений

Случайный лес.



Метрики классификации



How many selected items are relevant?

Precision = $\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$

How many relevant items are selected?

Recall = $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

Precision

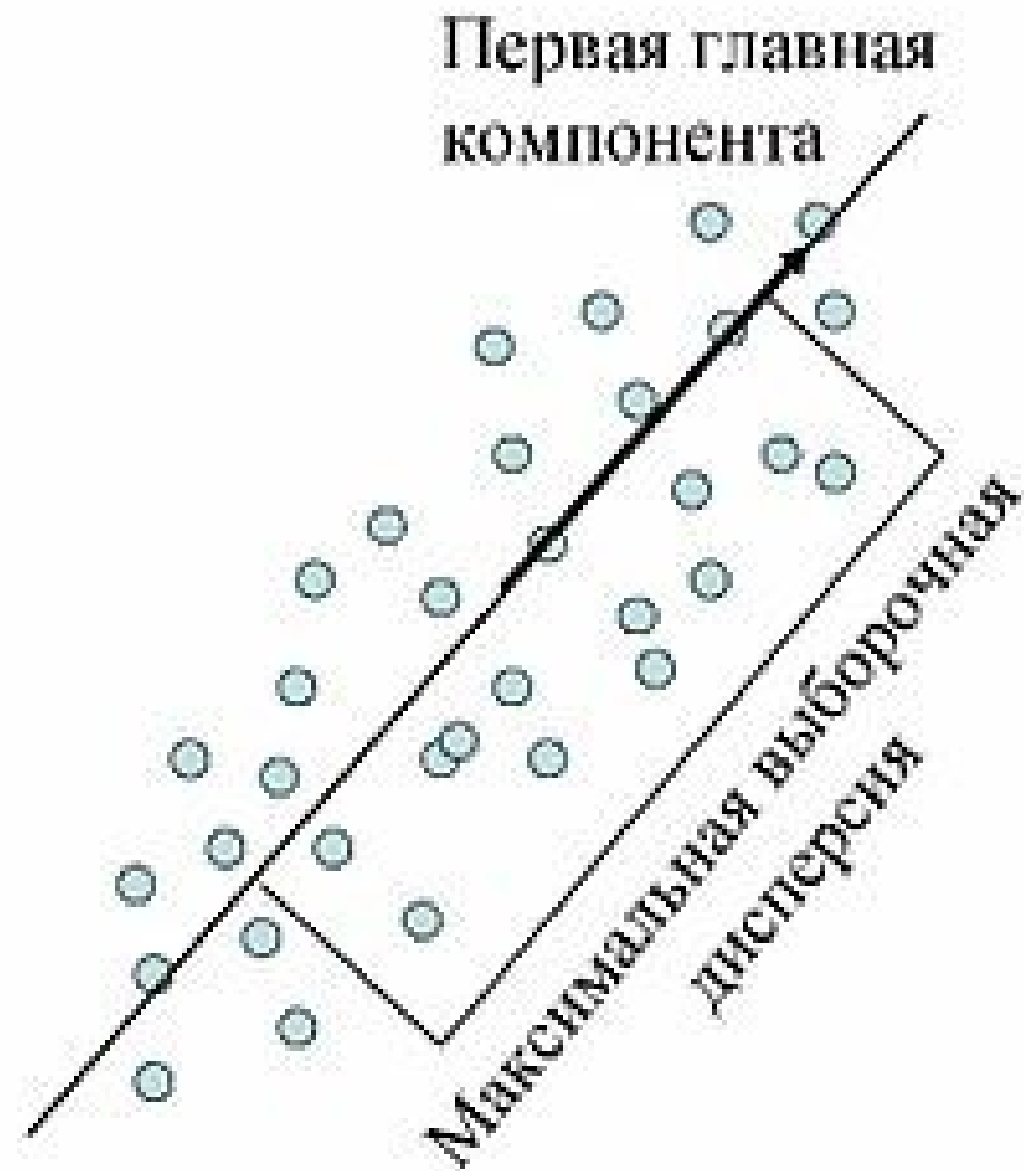
Recall

F1-мера

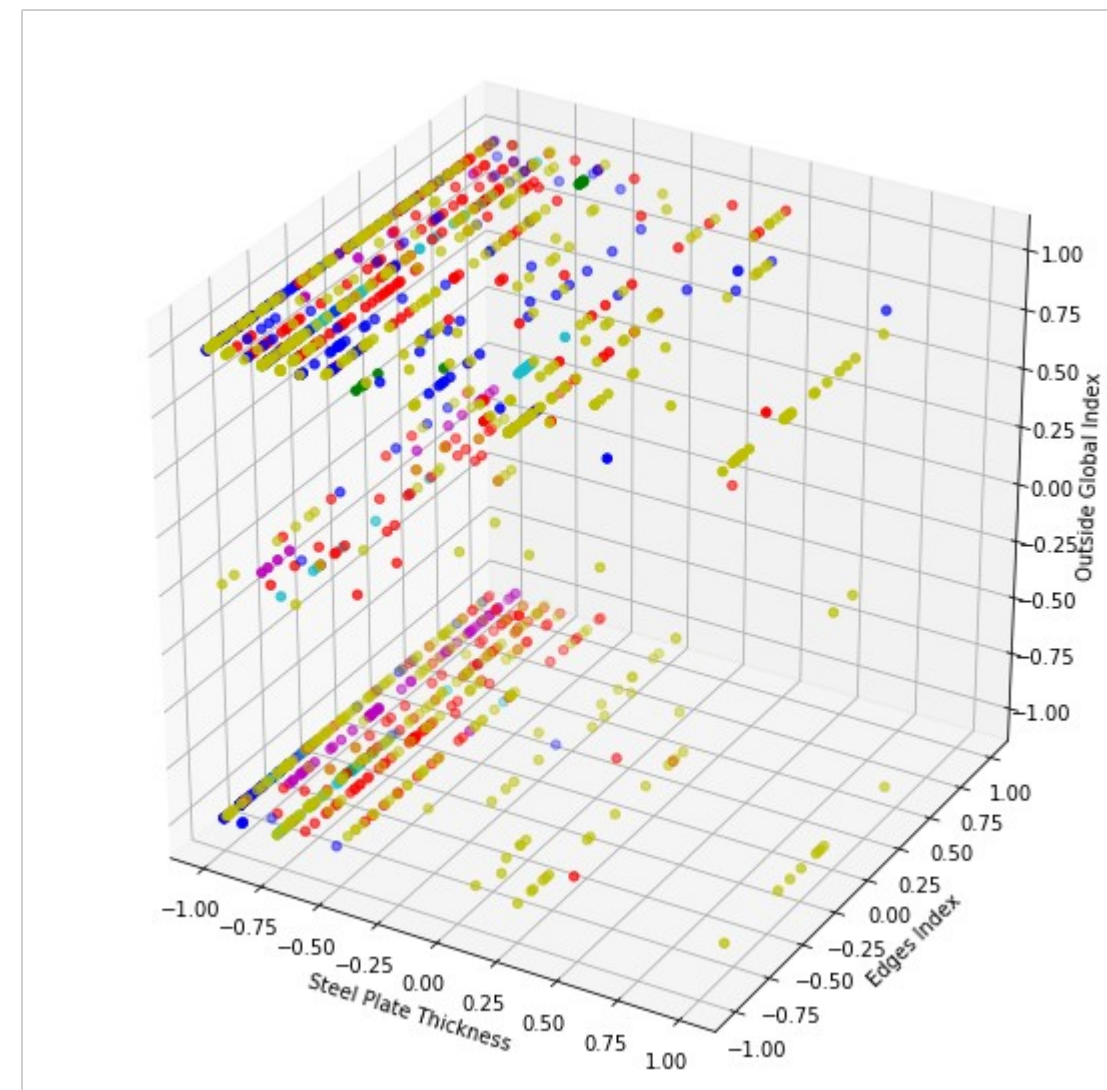
$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Принцип минимальных компонент.

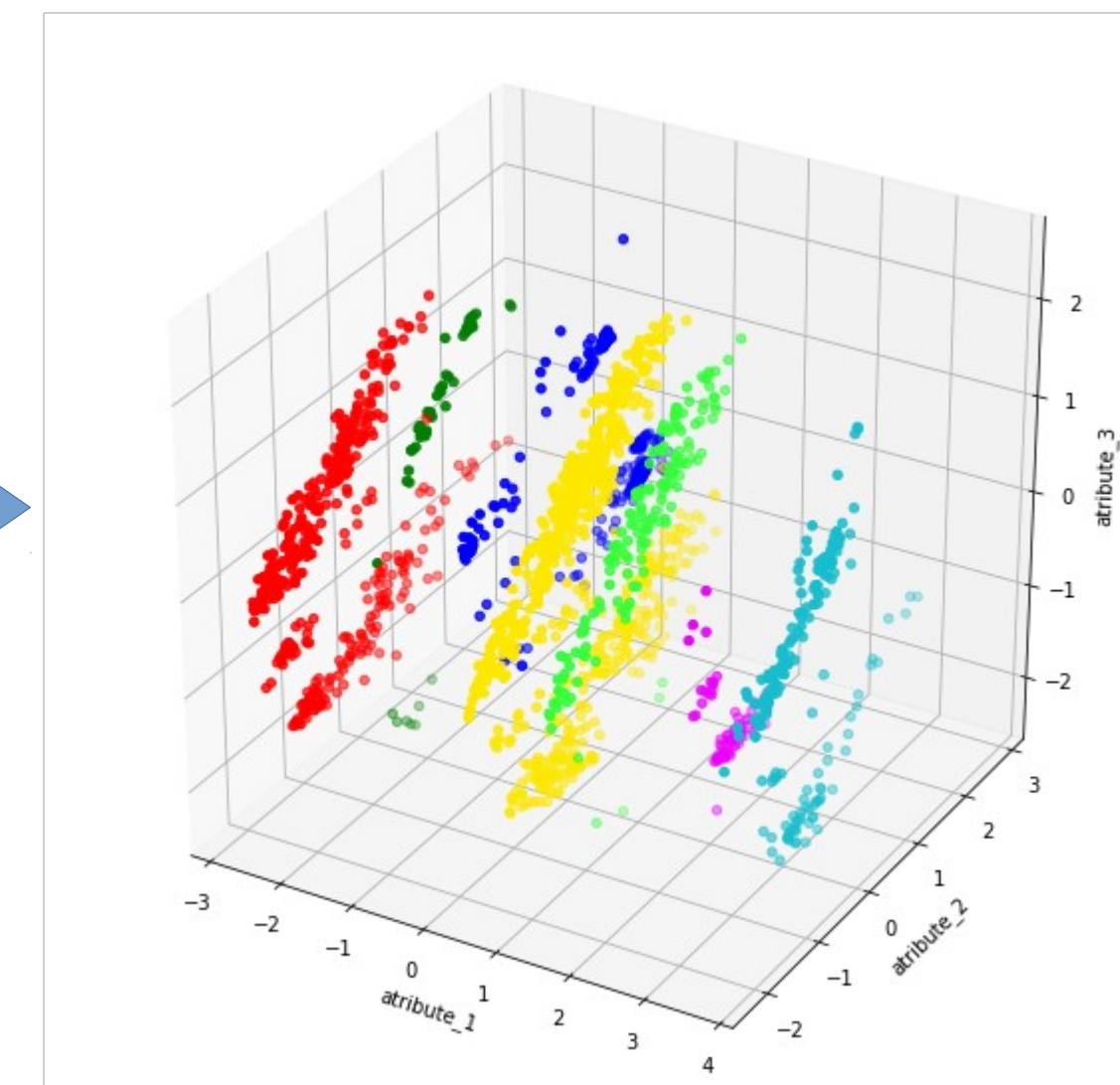
Поиск ортогональных проекций с наибольшим рассеянием



Было

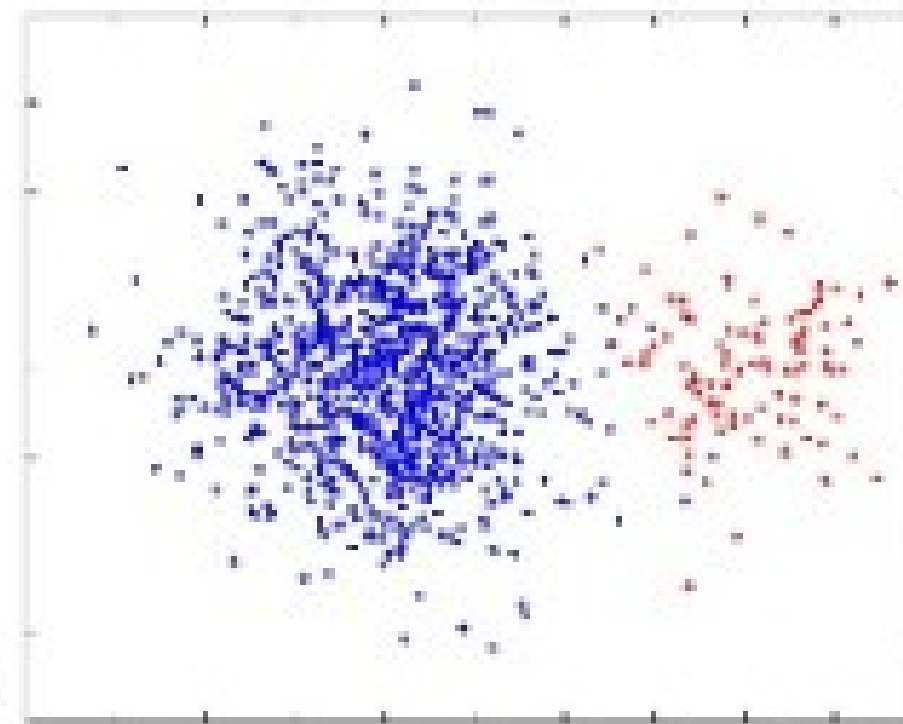


Стало

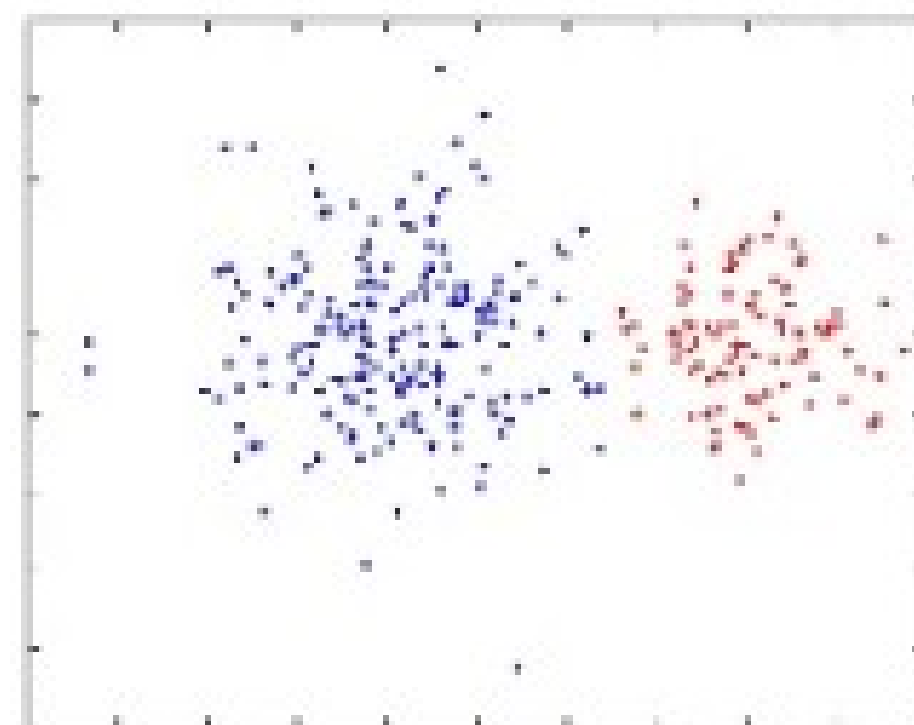


Проблема несбалансированности классов.

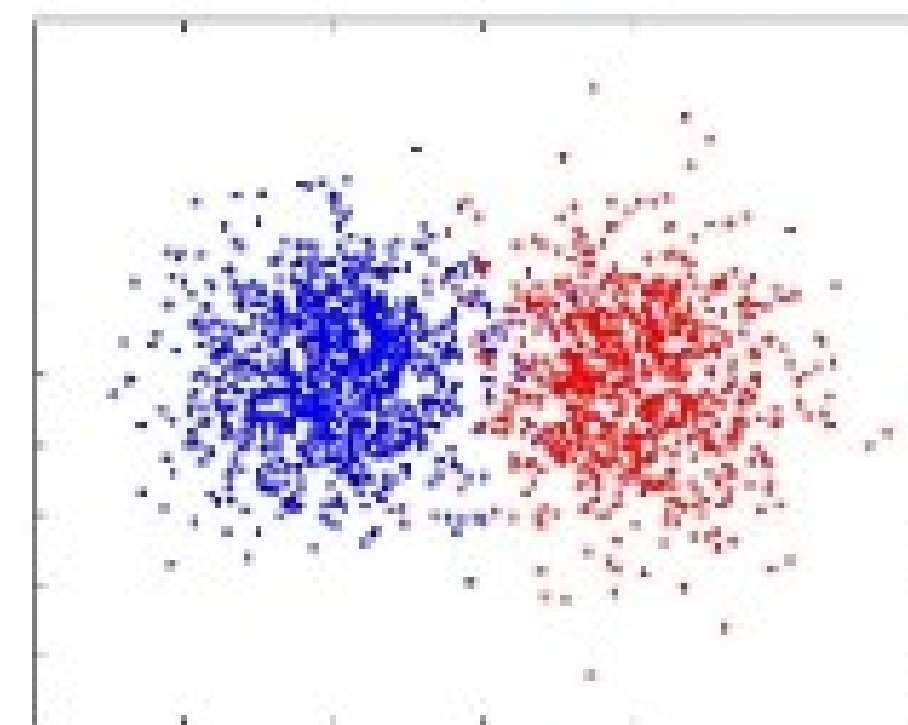
Sampling: Rebalancing the dataset



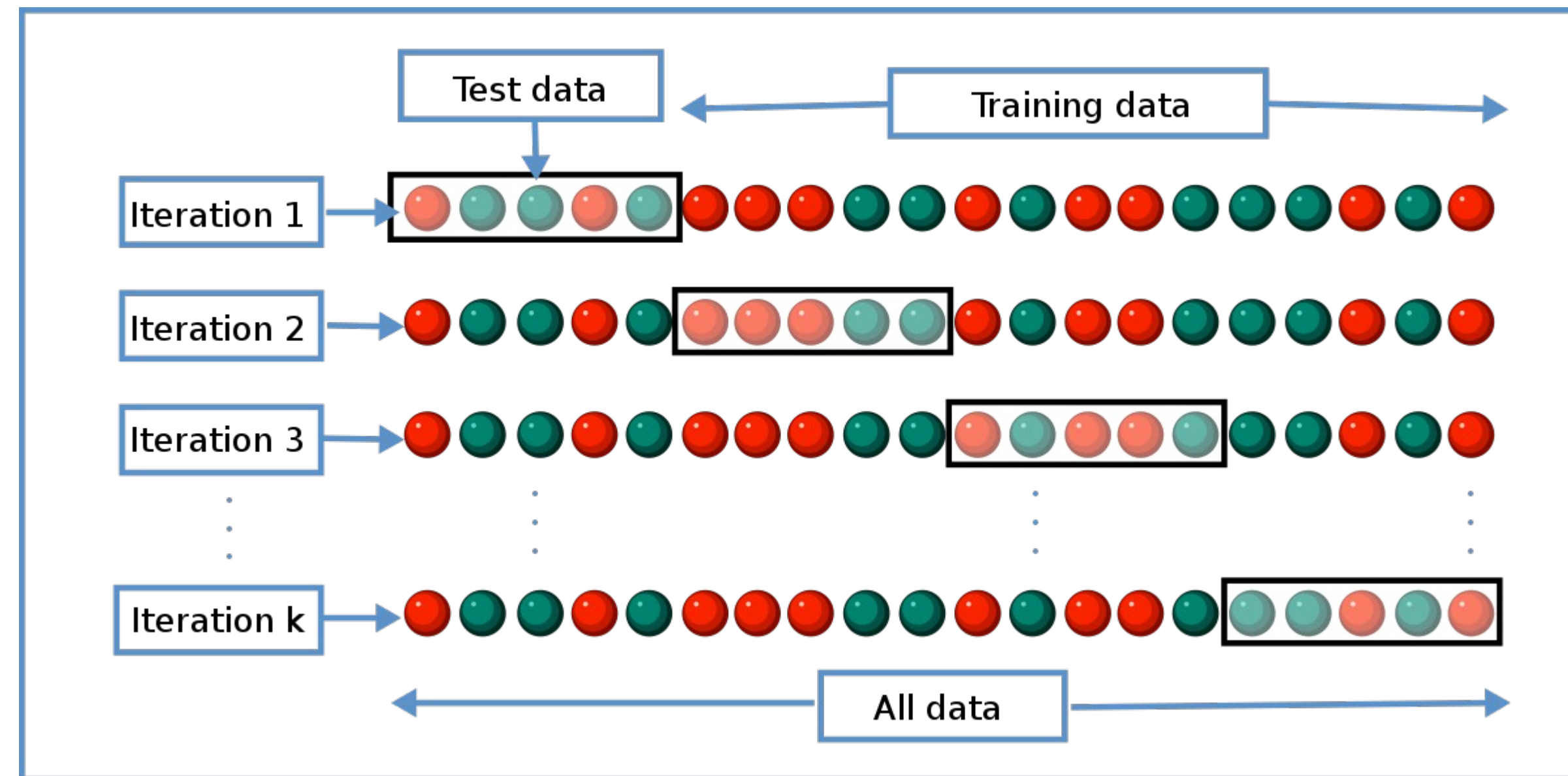
Under-sampling



Over-sampling

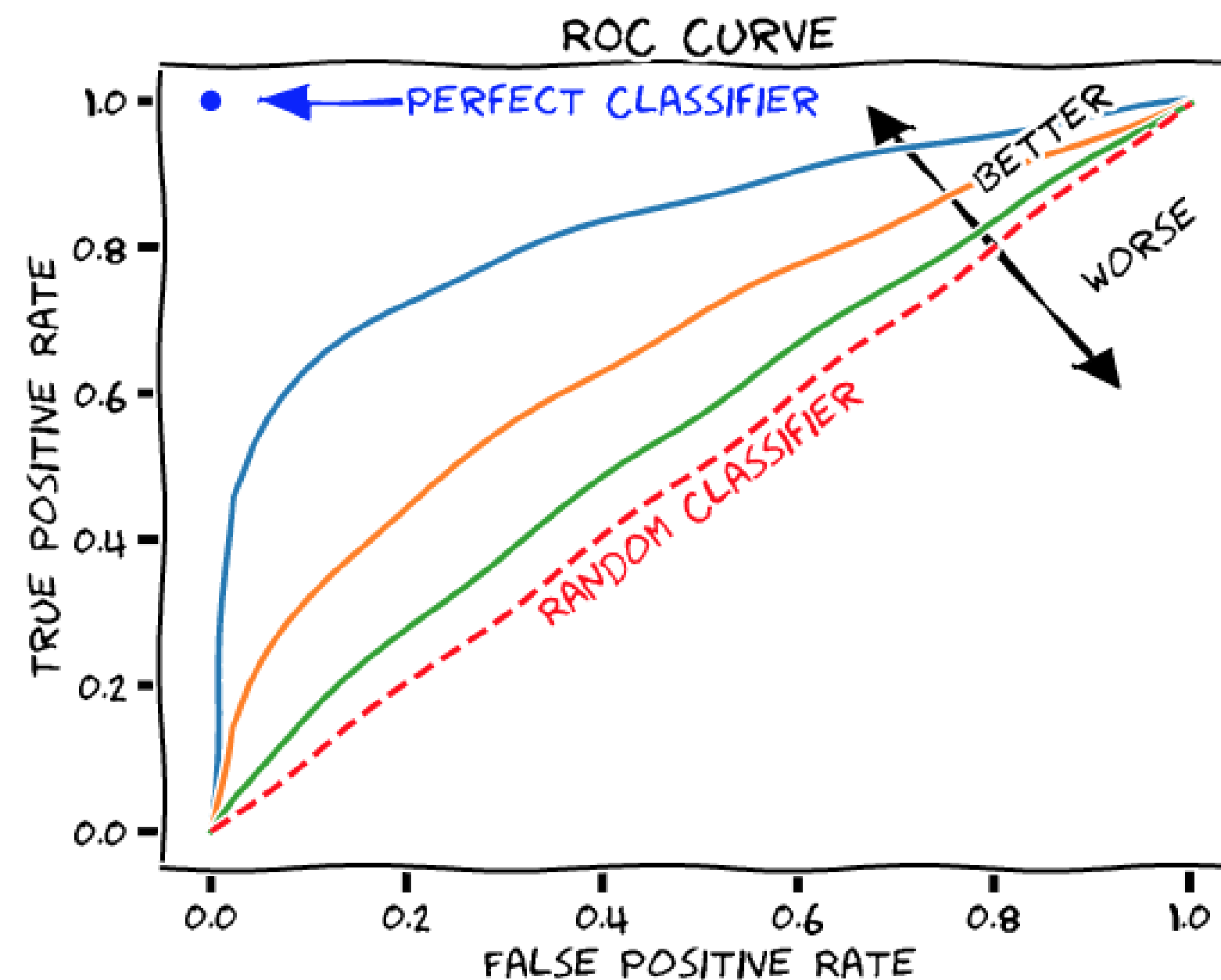


Метрики классификации: кросс-валидация



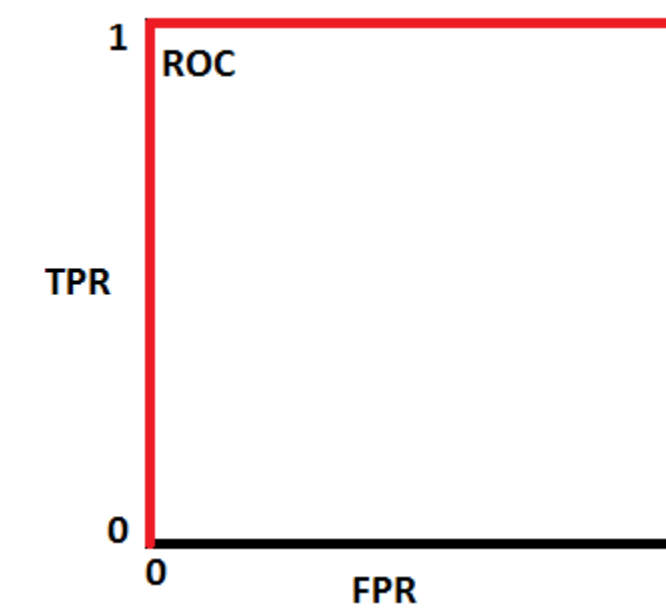
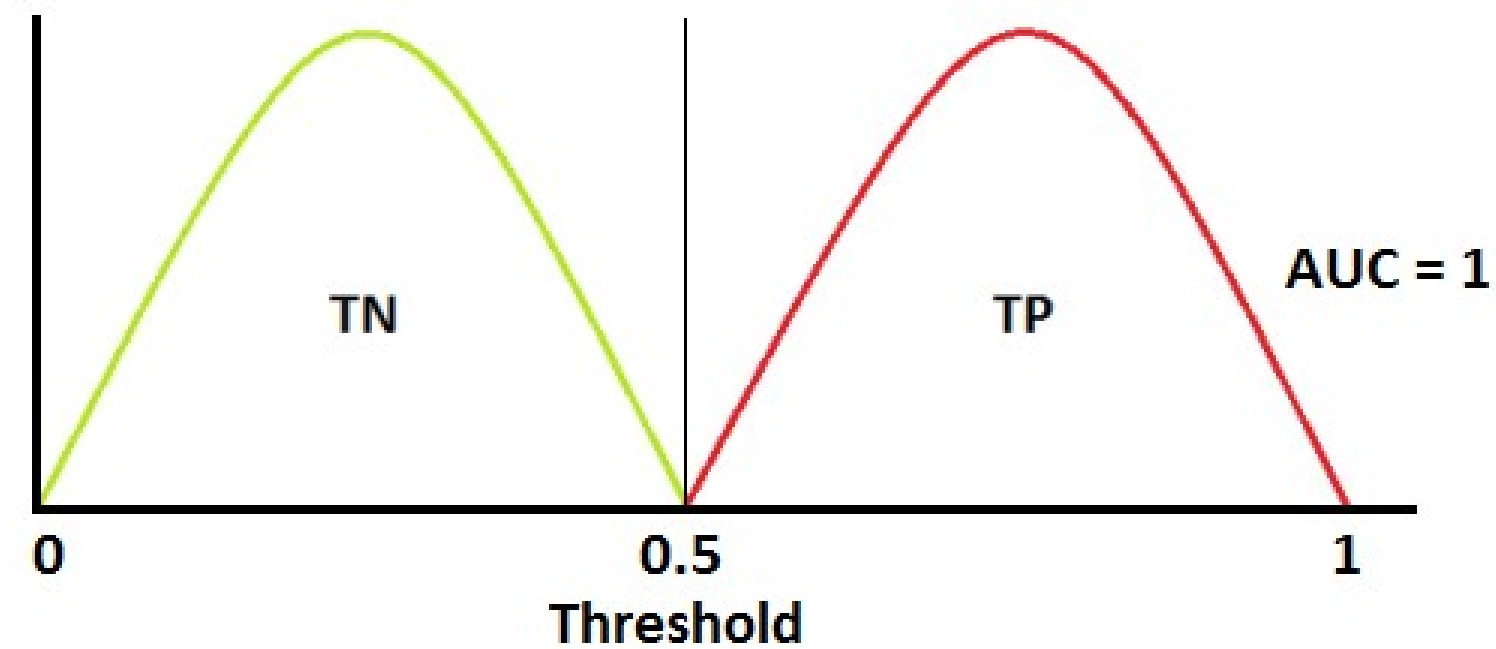
Оцениваем модель на нескольких тестовых данных

Метрики классификации: ROC-кривая

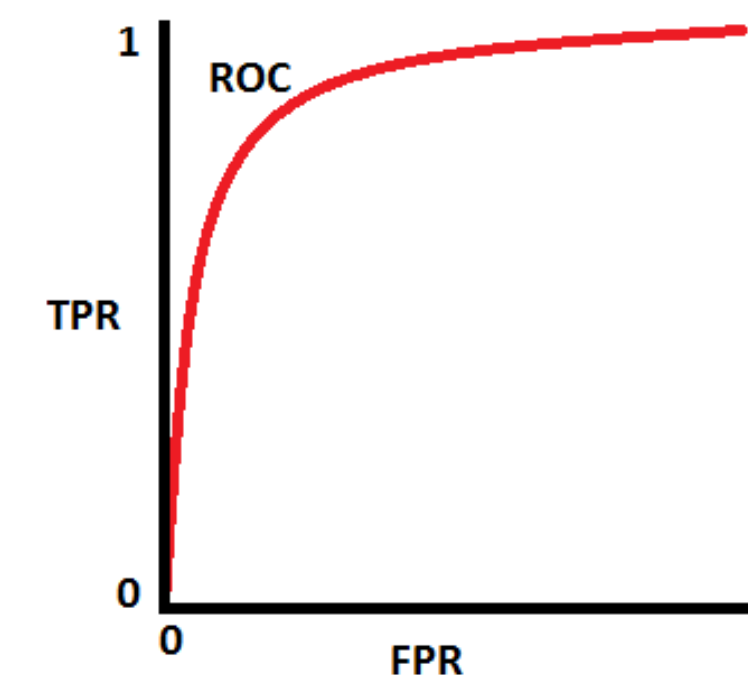
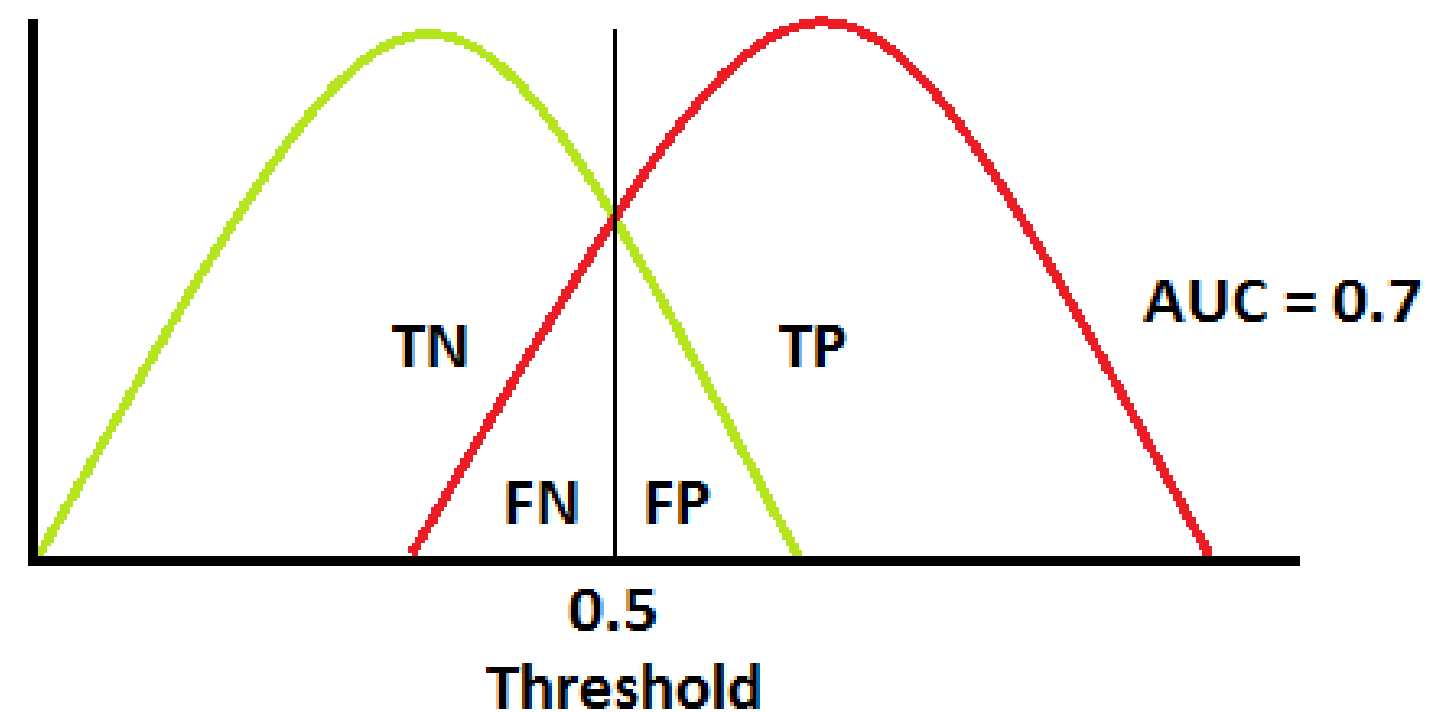


Позволяет определить порог,
при котором мы будем отделять один класс от другого

ROC-кривая



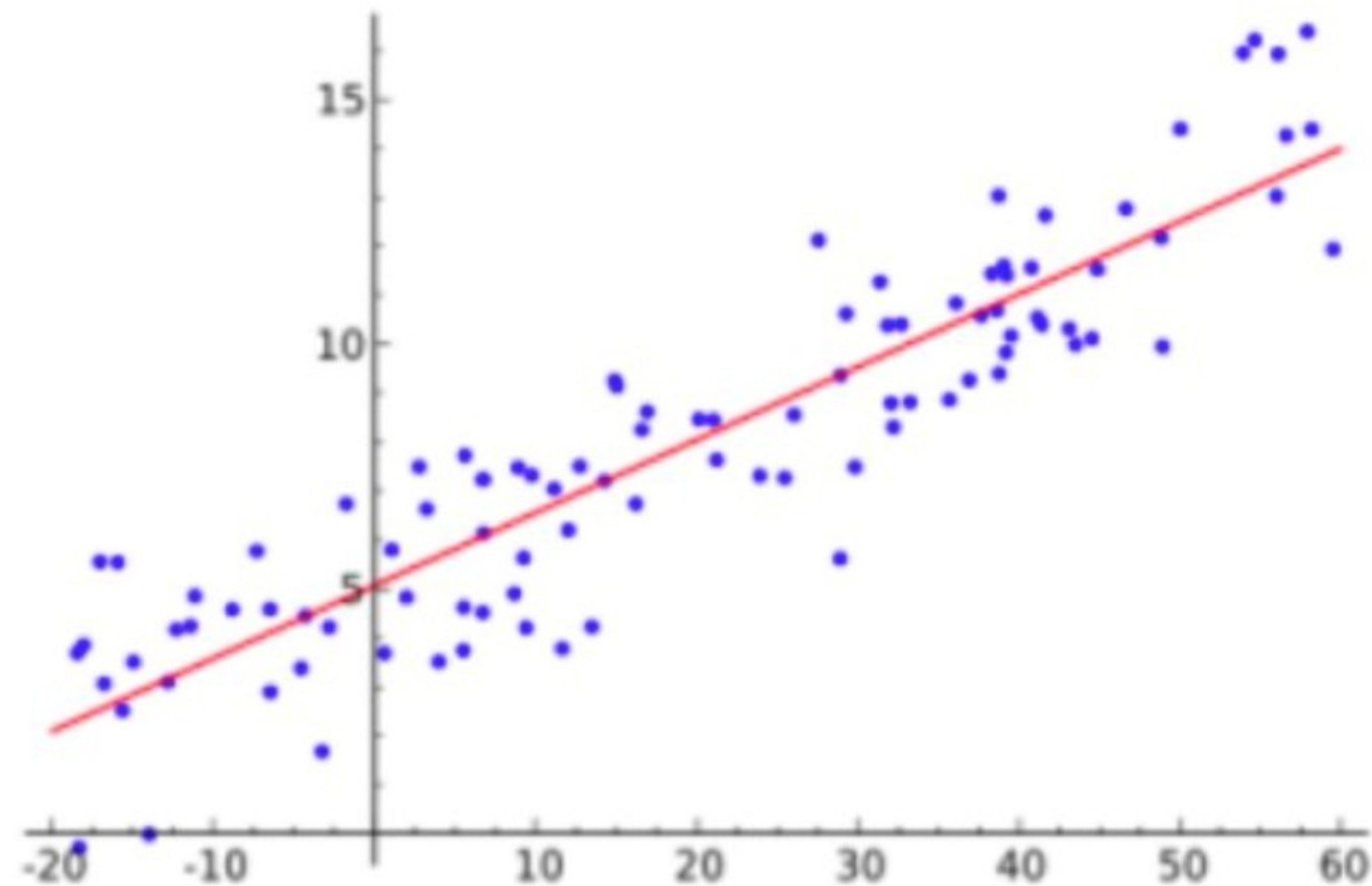
Идеальная модель — порог 50%



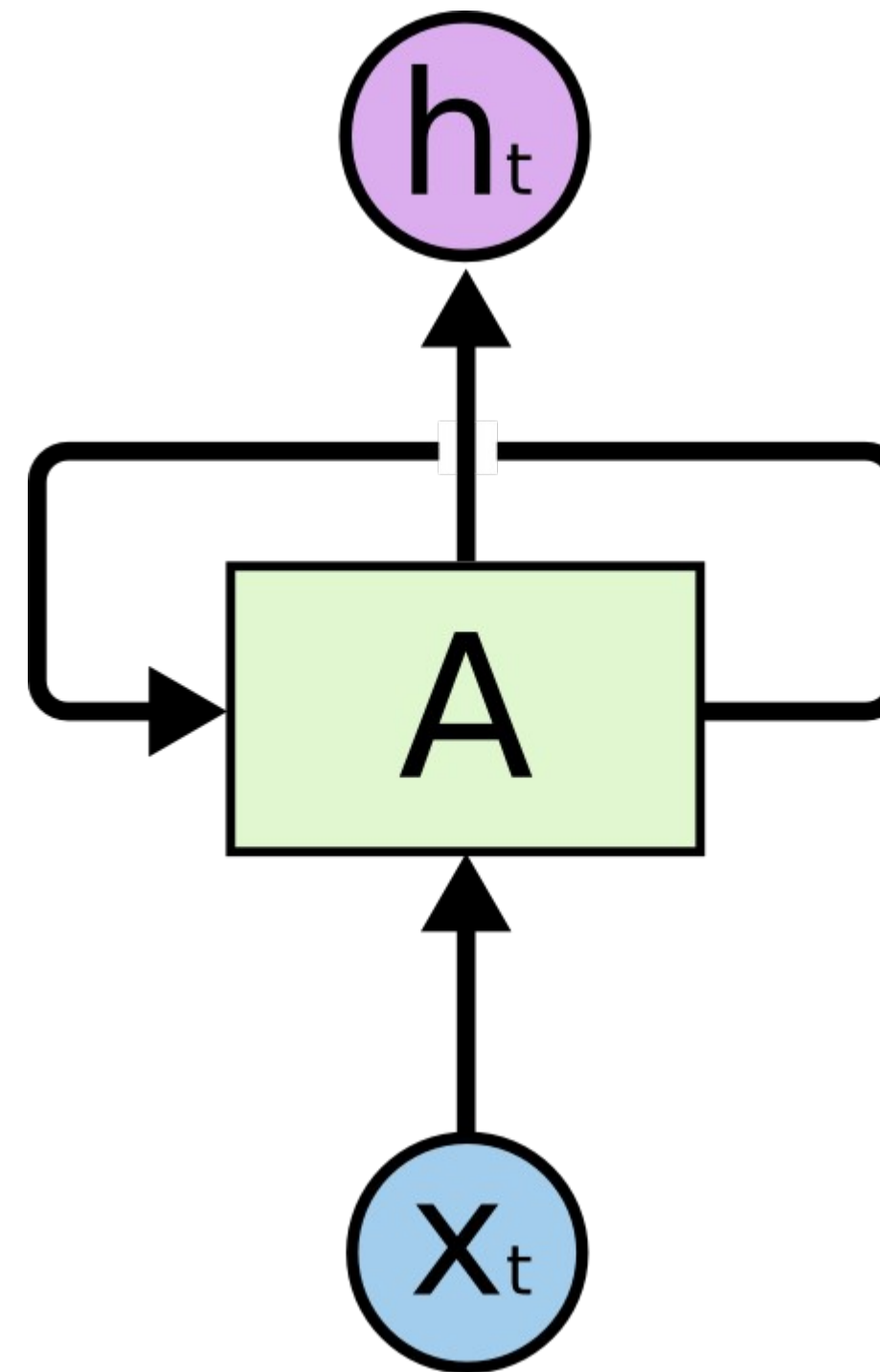
Модель с некоторыми ошибками — порог выбирается в зависимости от допускаемых ошибок

Регрессия

Отличается тем, что допустимым ответом является действительное число или числовой вектор.



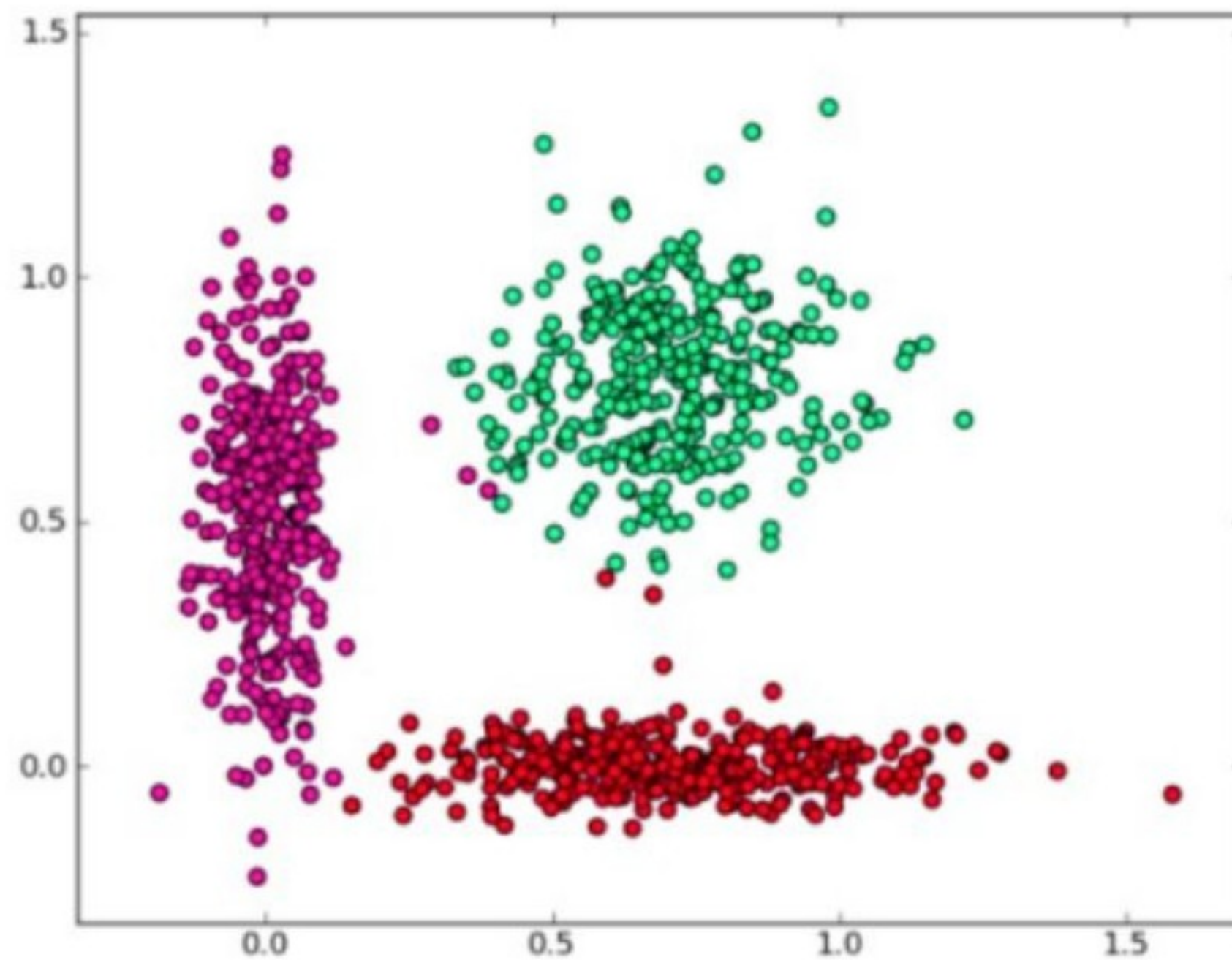
Модель LSTM.



Рекуррентные нейронные сети содержат обратные связи.

Кластеризация

Заключается в том, чтобы сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов. Функционалы качества могут определяться по-разному, например, как отношение средних межкластерных и внутрикластерных расстояний.



Вопросы?

Контакты спикера:
yustiks@gmail.com