

# Юстина Иванова

Программист, data scientist

Статистика в python. Кейс-стади №1.  
Датасеты: faulty steel plates,  
heart disease record,  
Brent oil prices.



Спикер

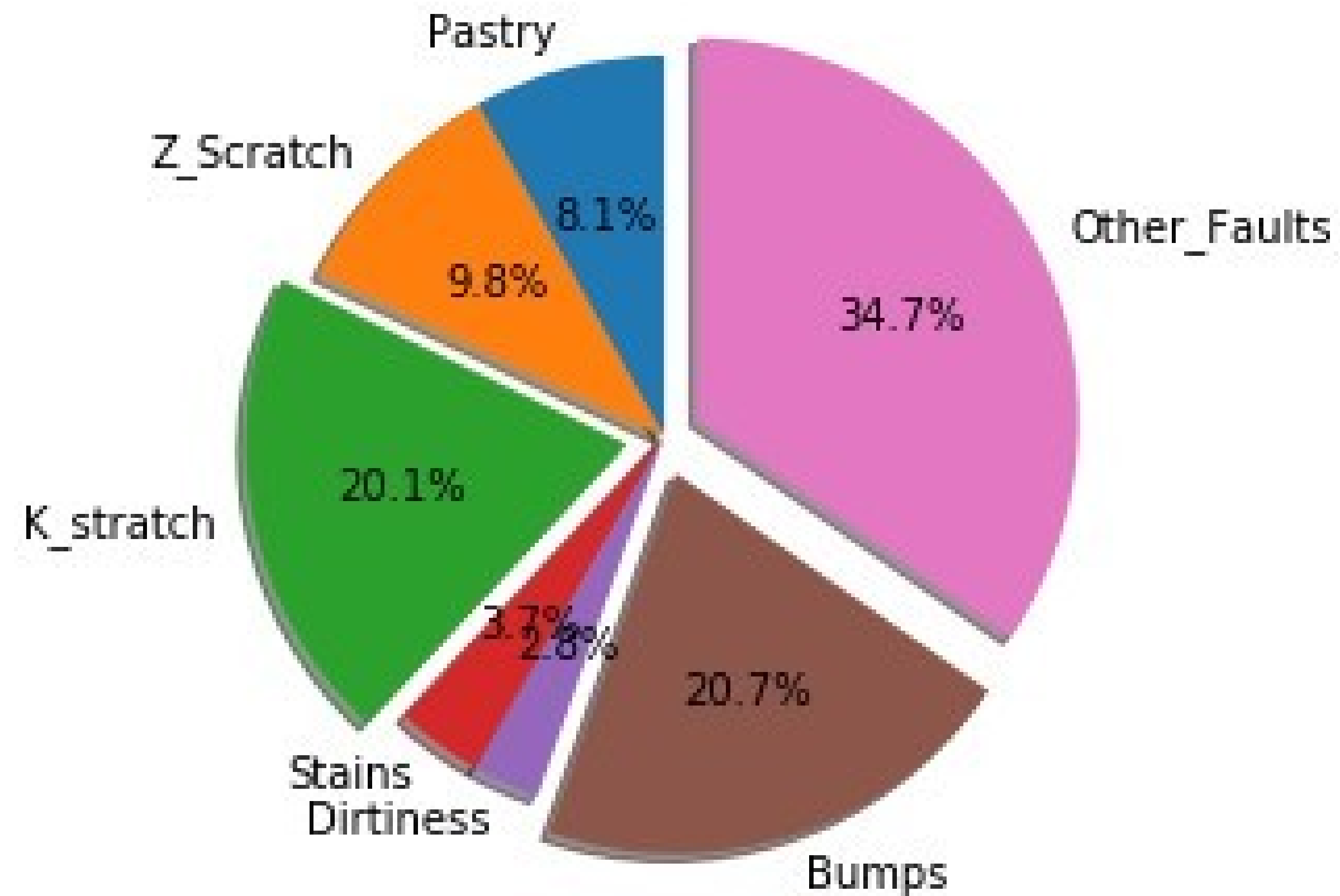


## Юстина Иванова,

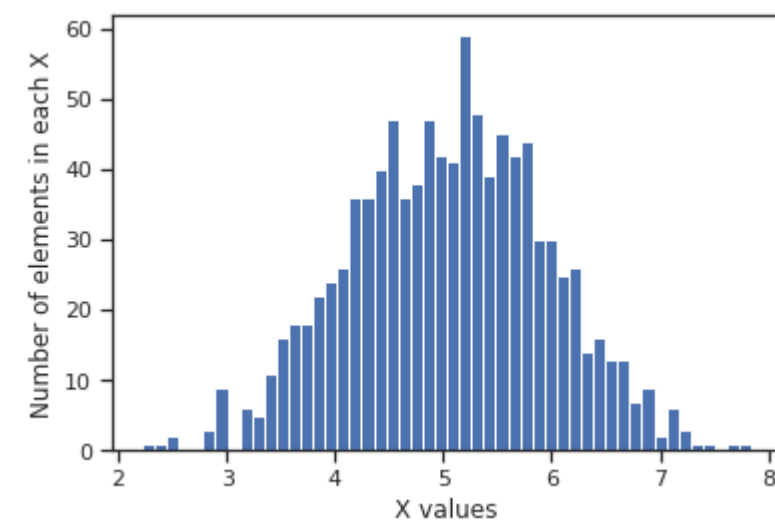
- PhD в Университете Больцано
- Выпускница МГТУ им. Баумана
- Магистр по Artificial Intelligence в University of Southampton



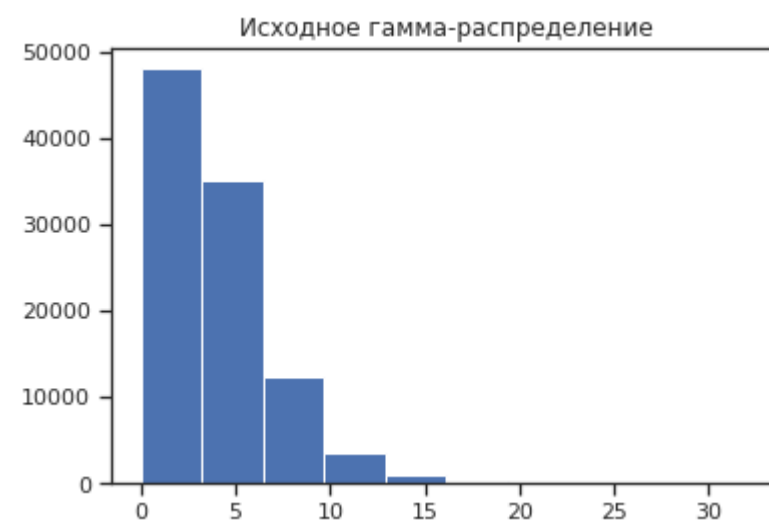
# Датасет Faulty Steel Plates.



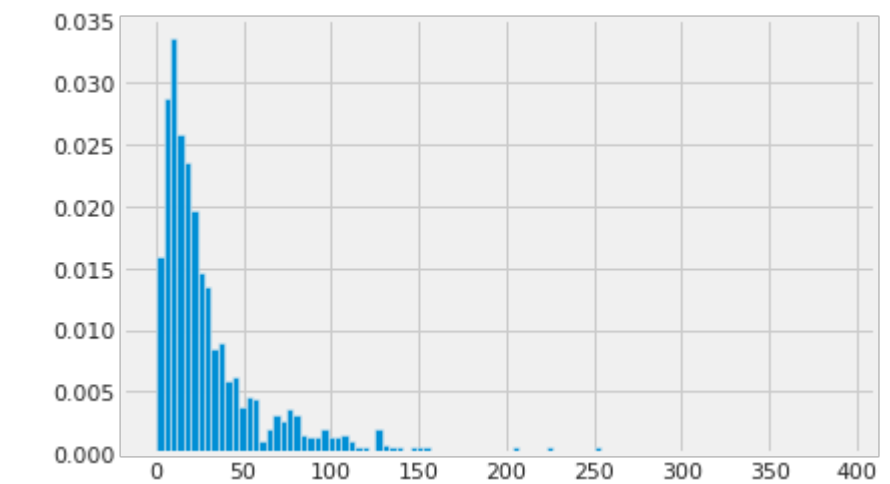
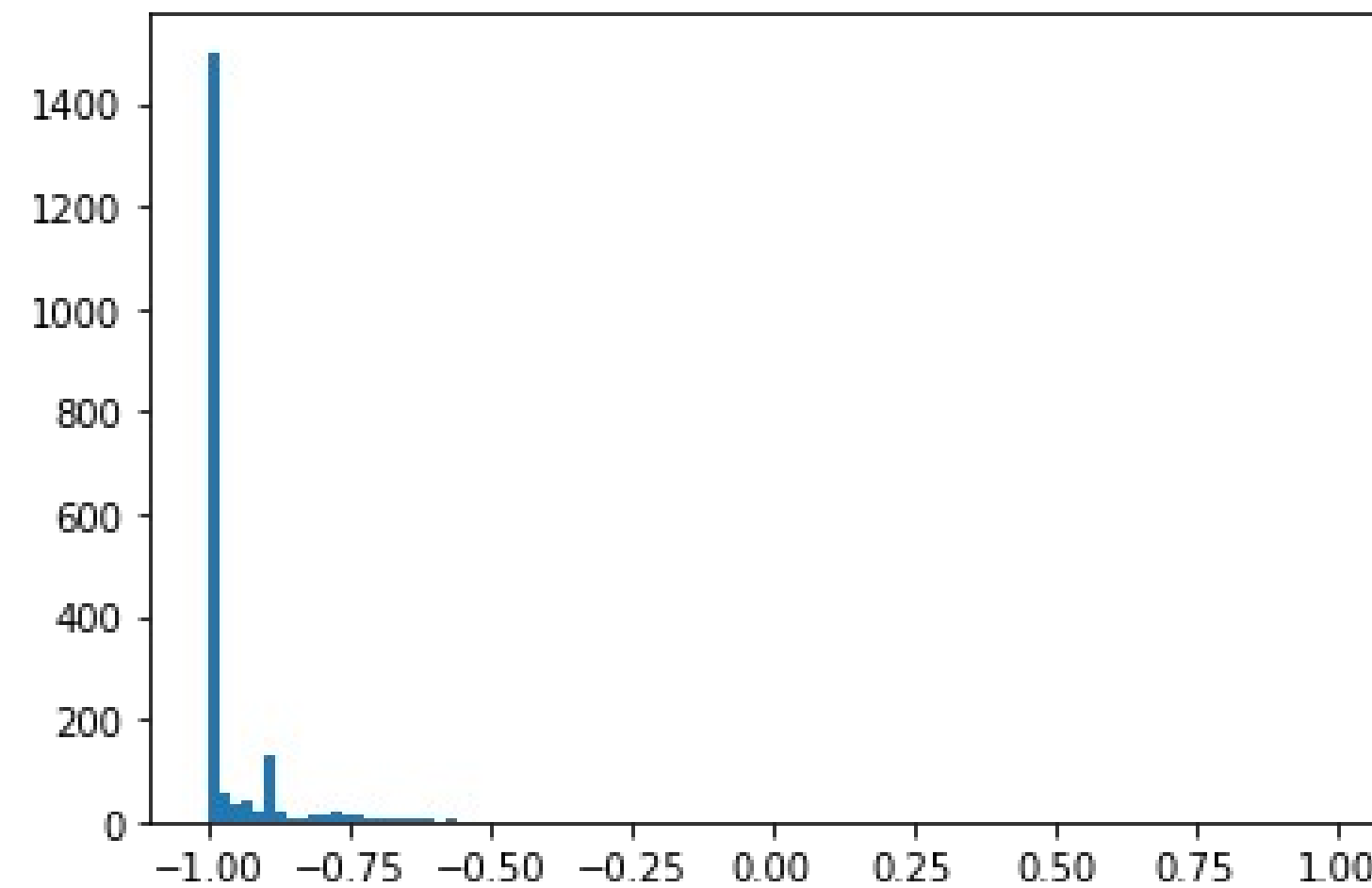
# Тесты на согласие: какое это распределение?



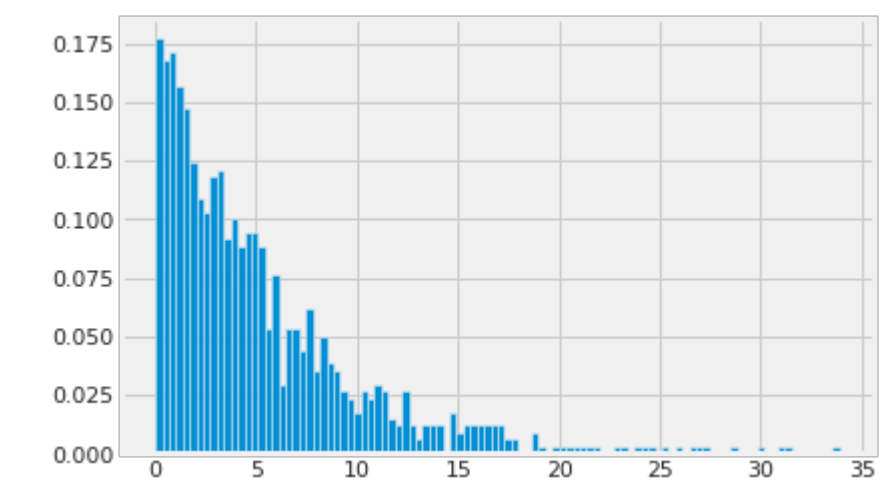
нормальное



гамма

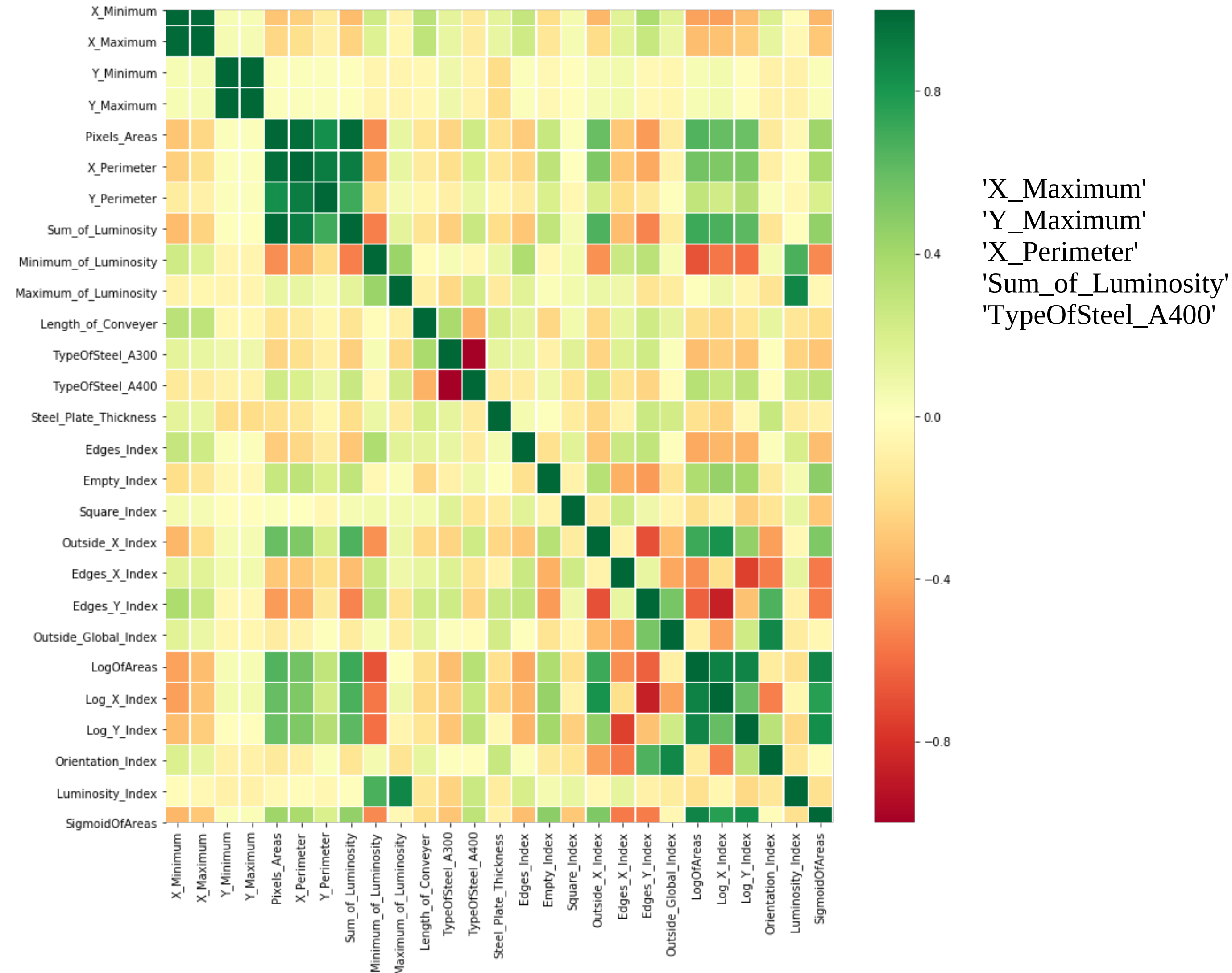


экспоненциальное



экспоненциальное

# Удаление мультиколлинеарности



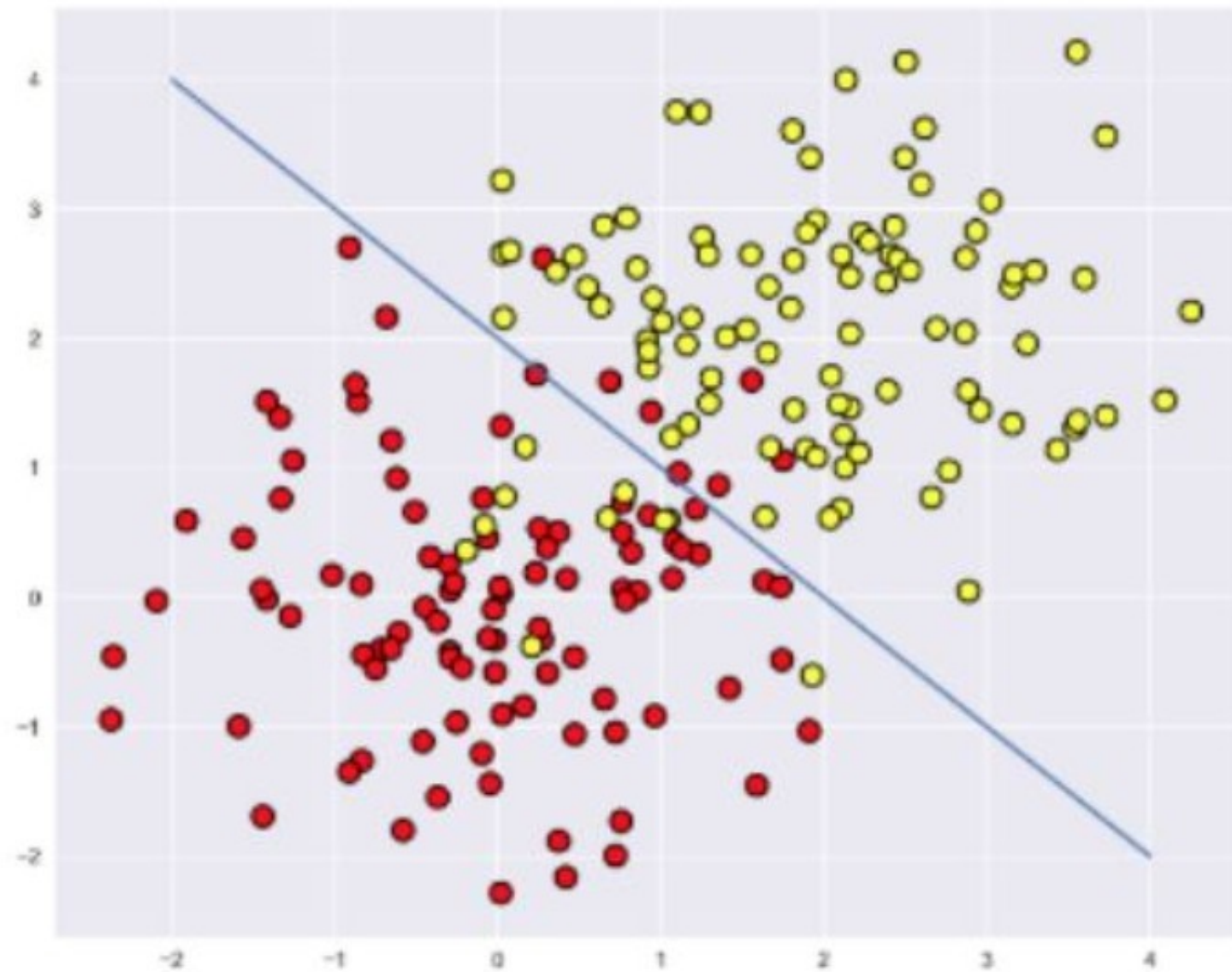
# Классическое Обучение





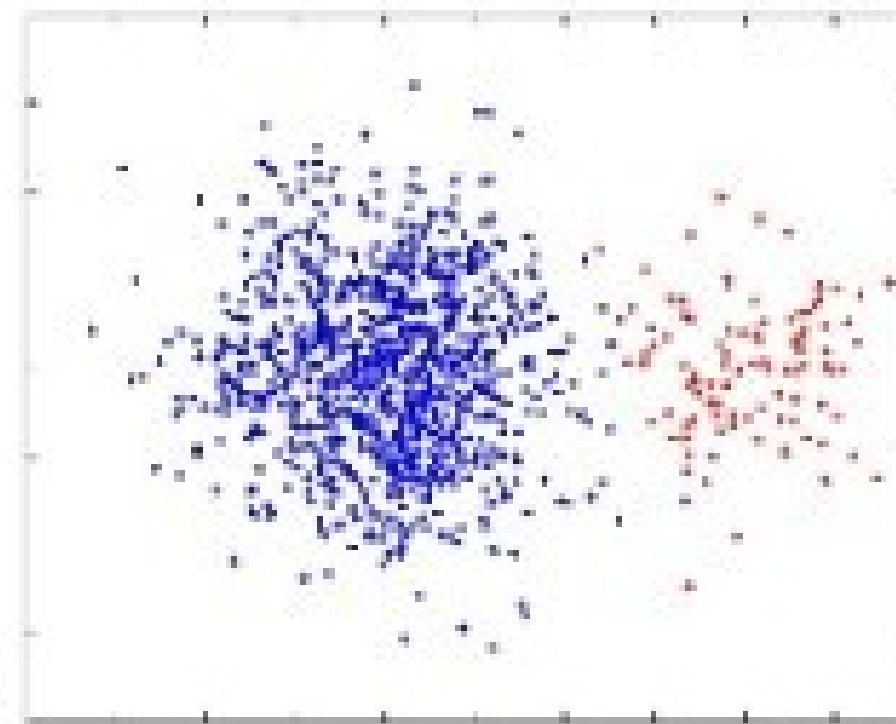
## Классификация

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

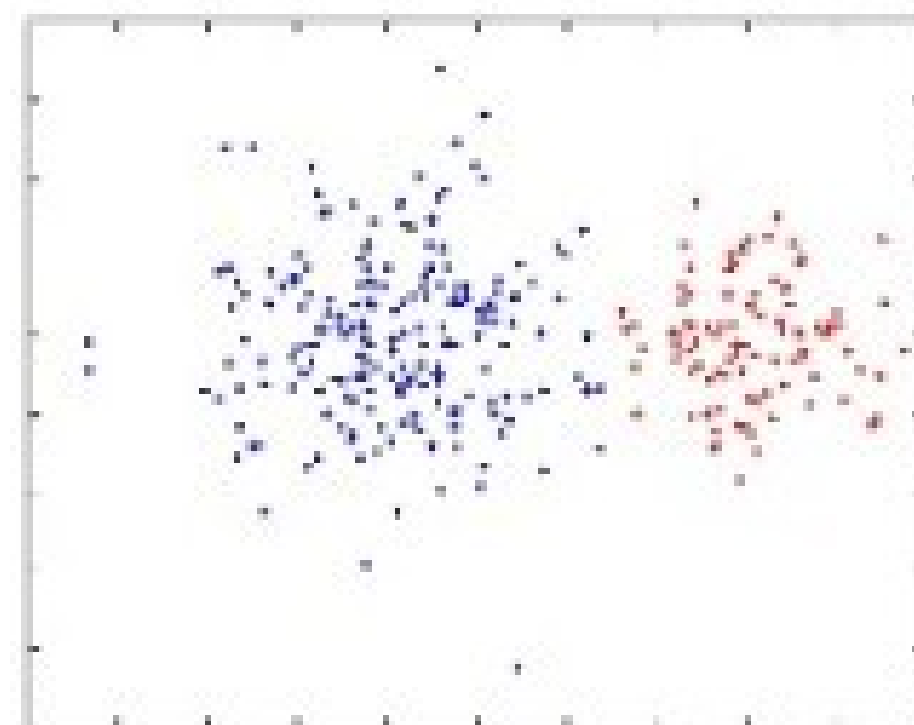


# Проблема несбалансированности классов.

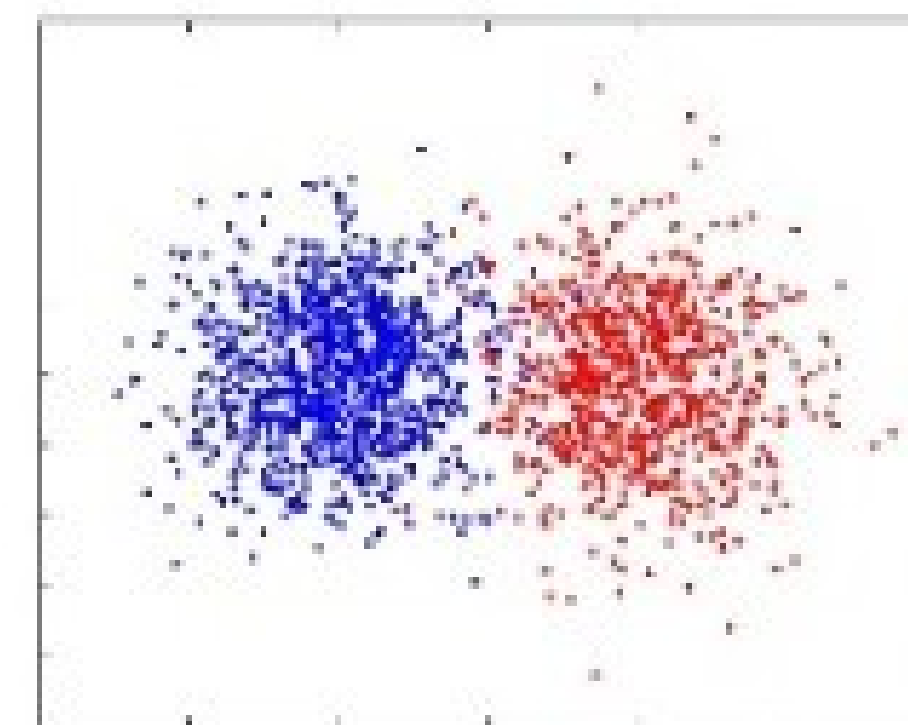
**Sampling:** Rebalancing the dataset



Under-sampling



Over-sampling



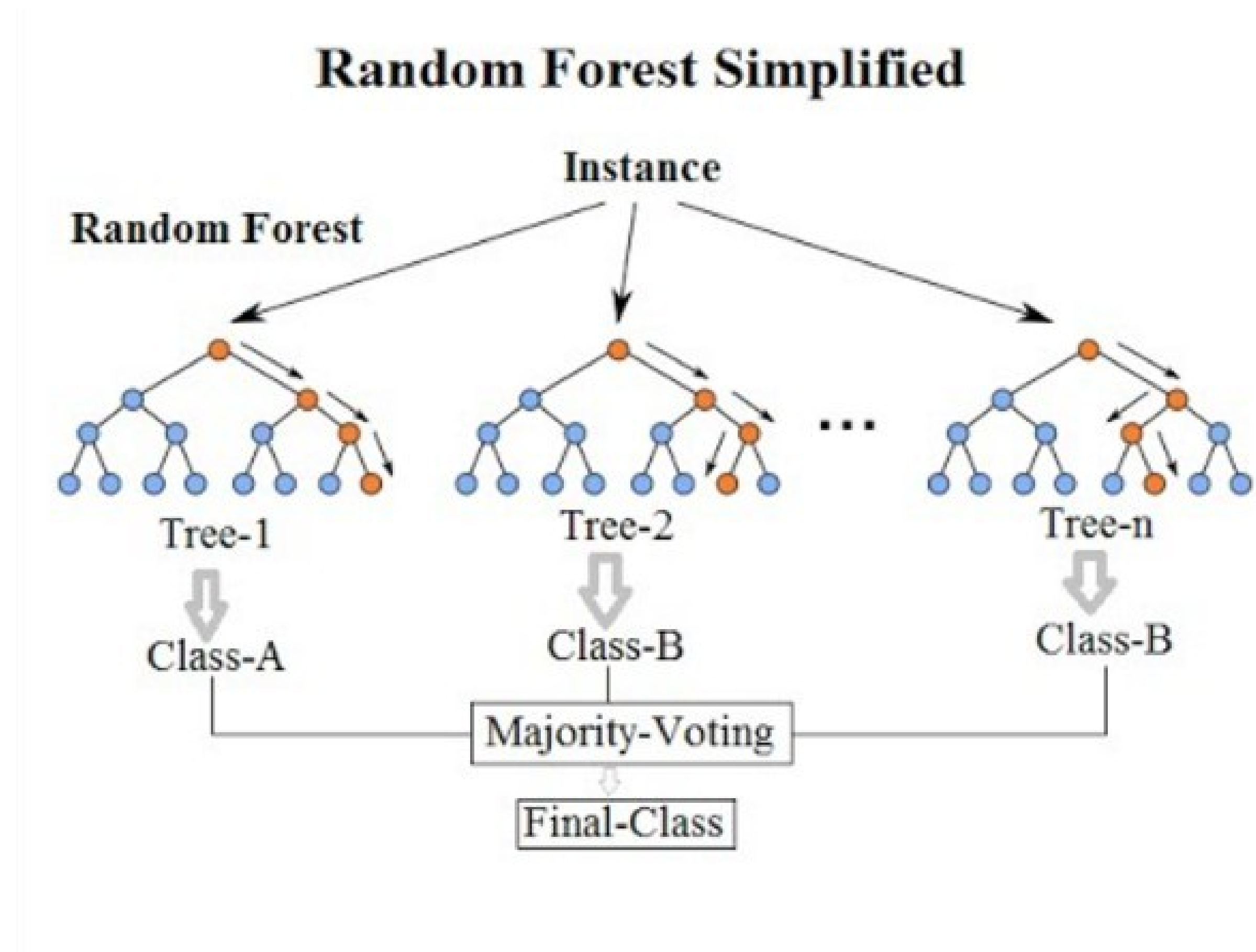


# Дерево решений.

## Давать ли кредит?

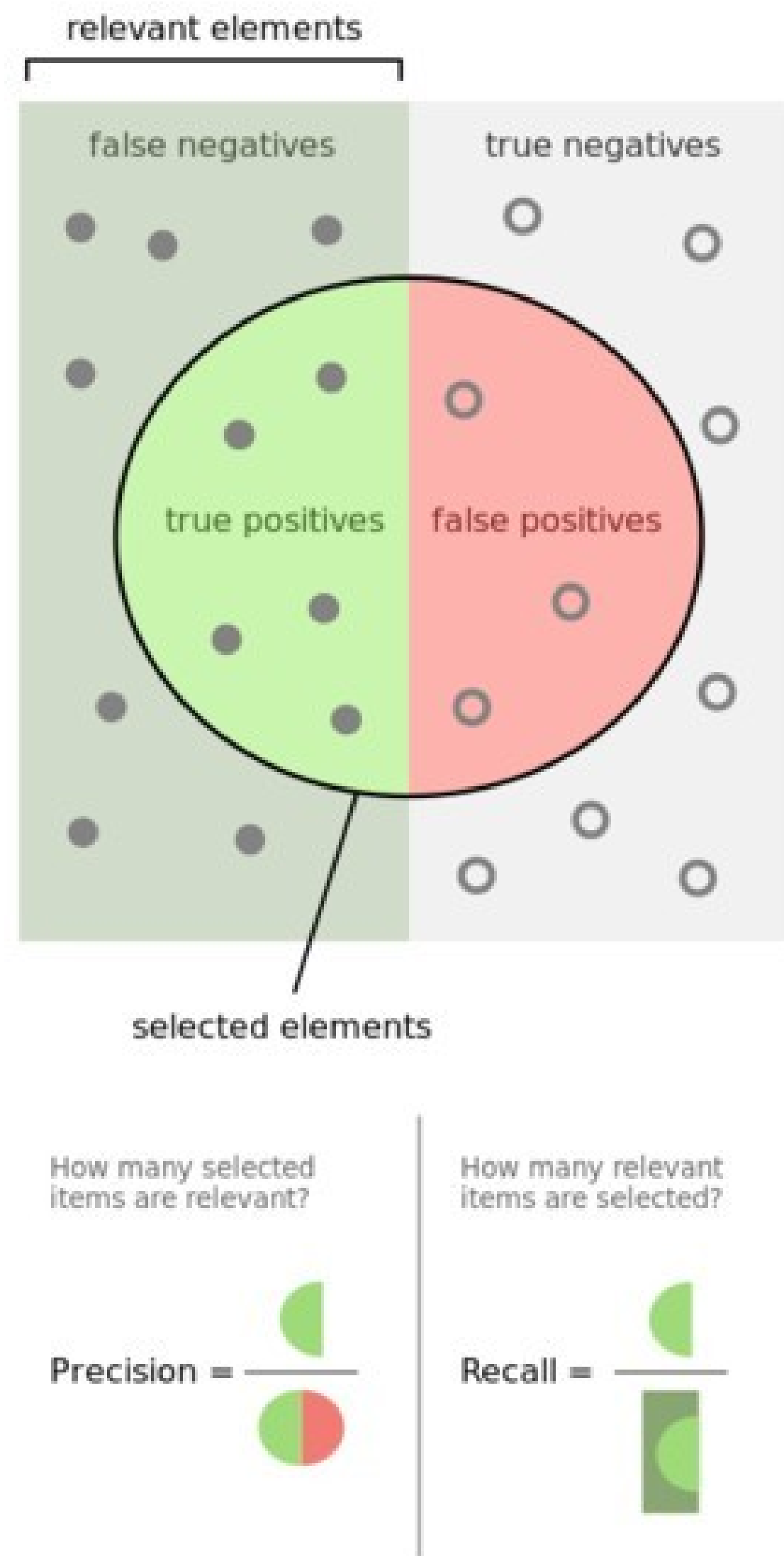


# Случайный лес.





# Метрики классификации



**Precision**

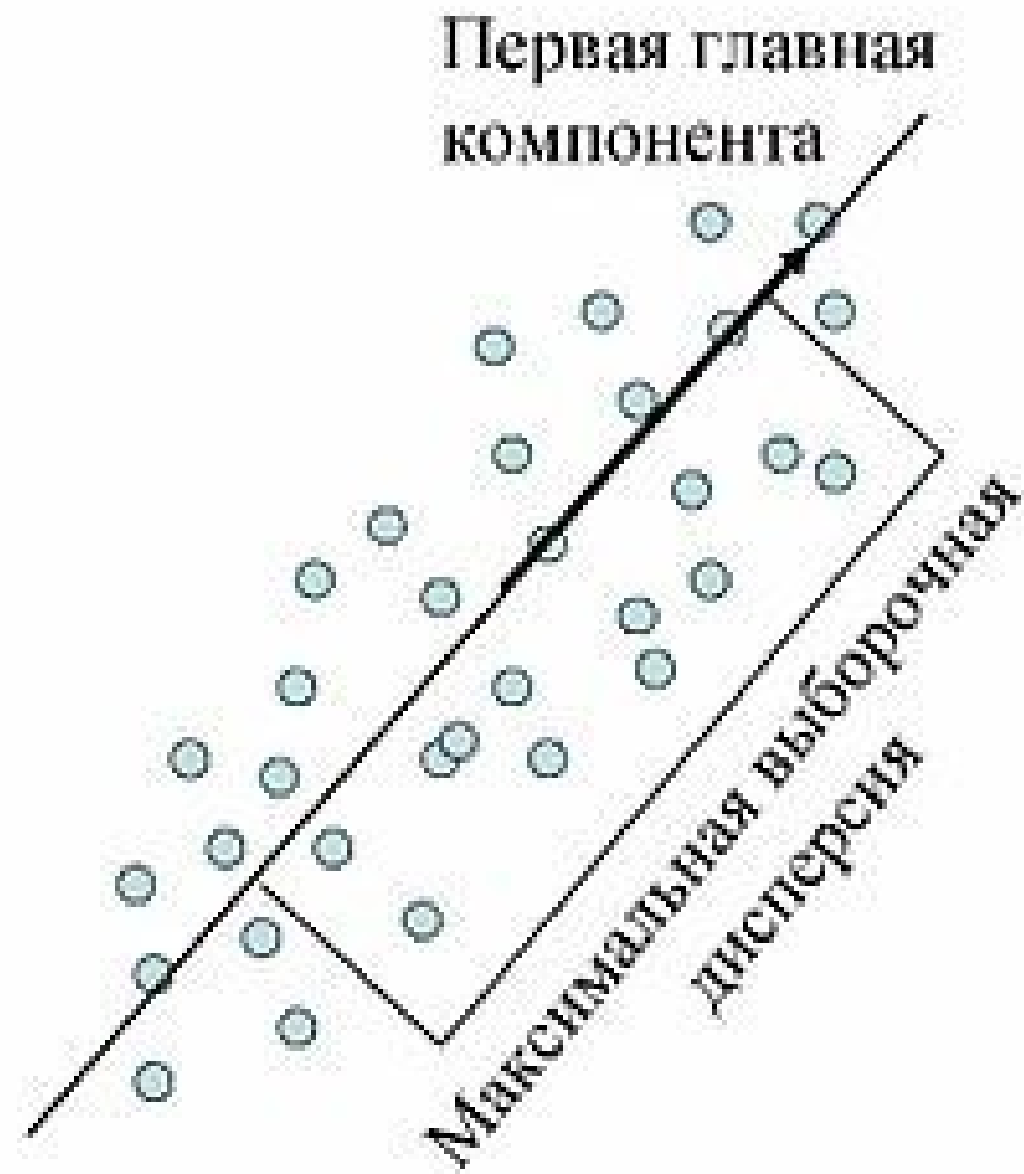
**Recall**

**F1-мера**

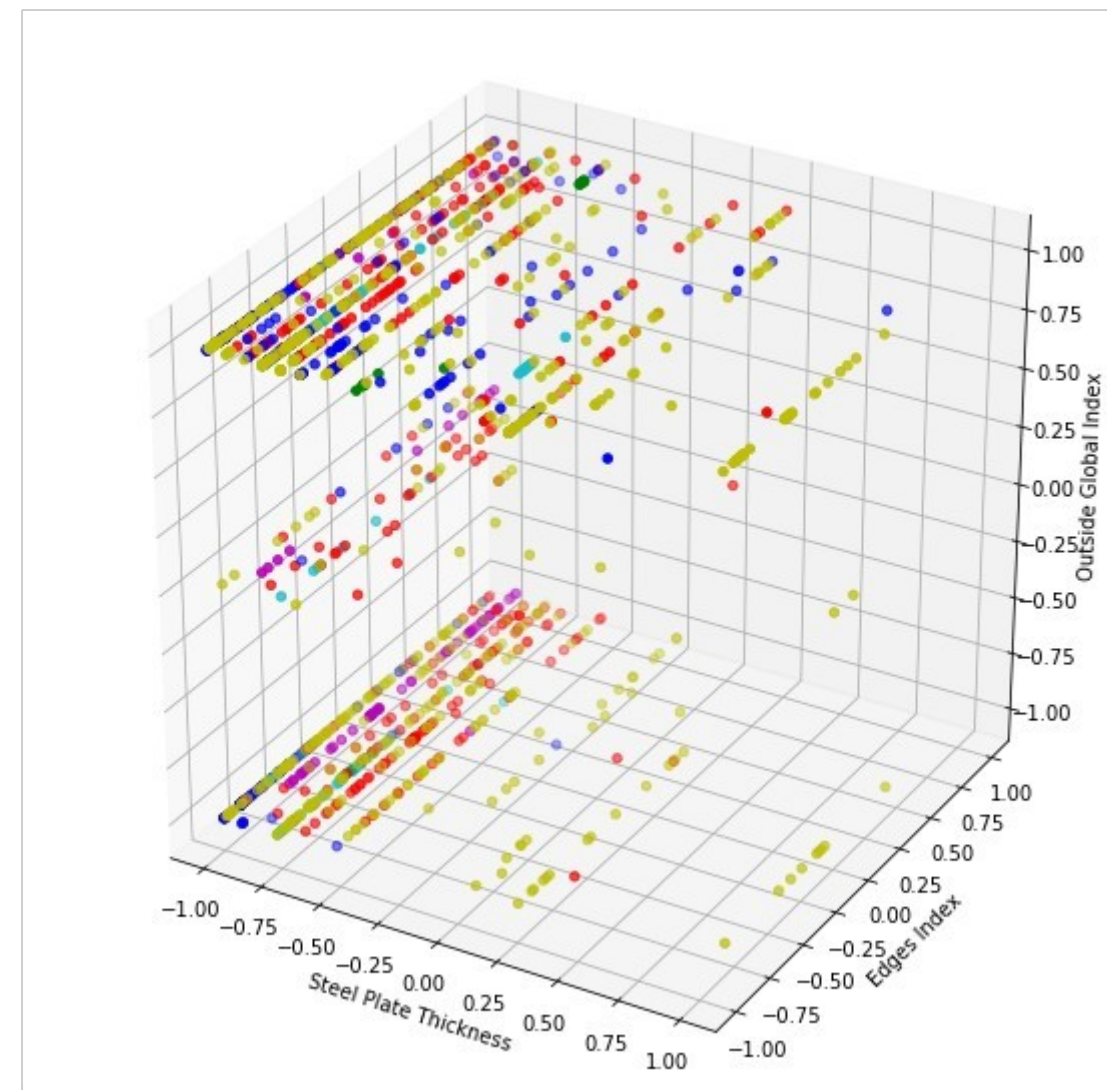
$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

# Принцип минимальных компонент.

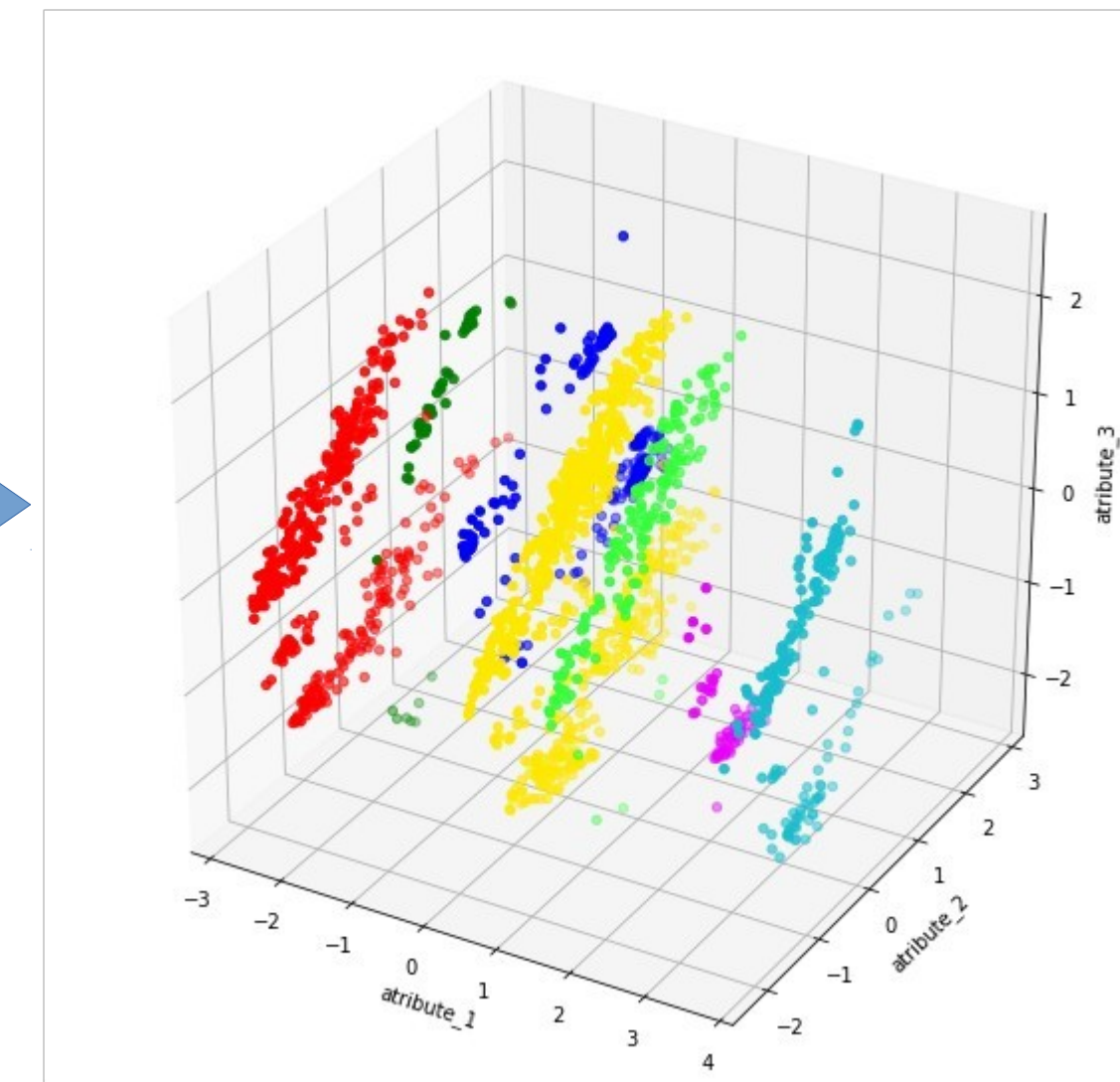
Поиск ортогональных проекций с наибольшим рассеянием



Было



Стало





# Логистическая регрессия.

Задача логистической регрессии – определить вероятность принадлежности к классу.

Построена на основе линейной функции.

$$h(x) = \theta^T x$$

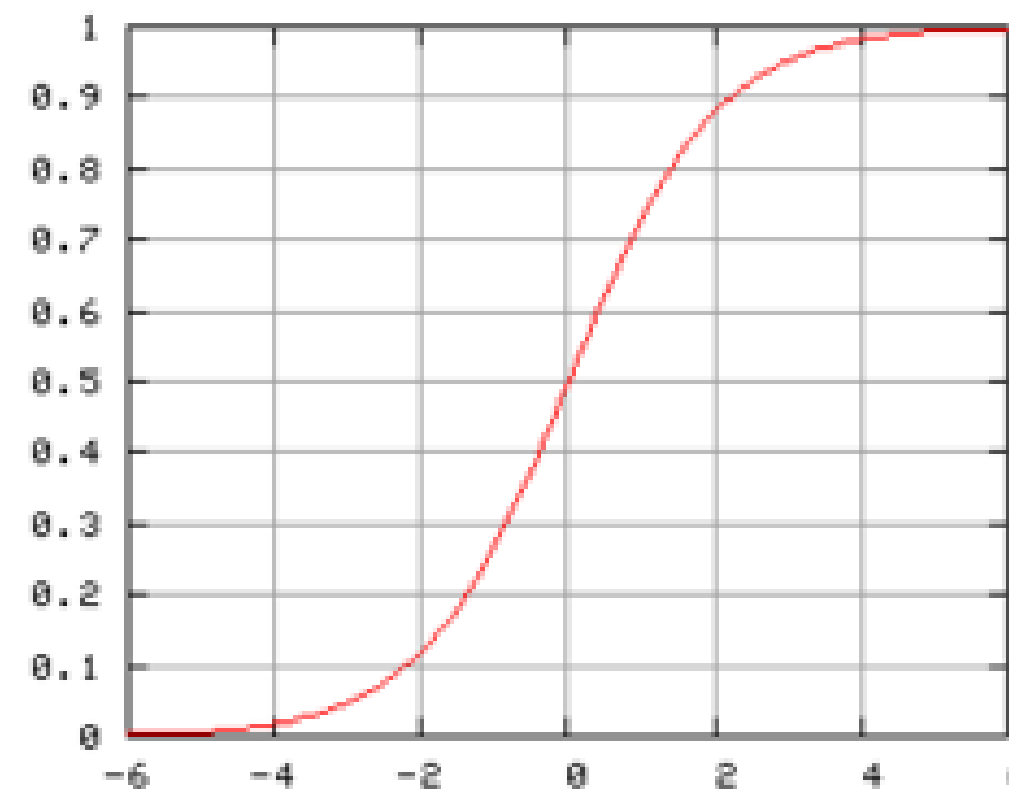
К линейной функции применяется функция активации:

$$h(x) = \sigma(\theta^T x)$$

Функция активации:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

# Сигмоида.



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Производная сигмоиды:

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$$

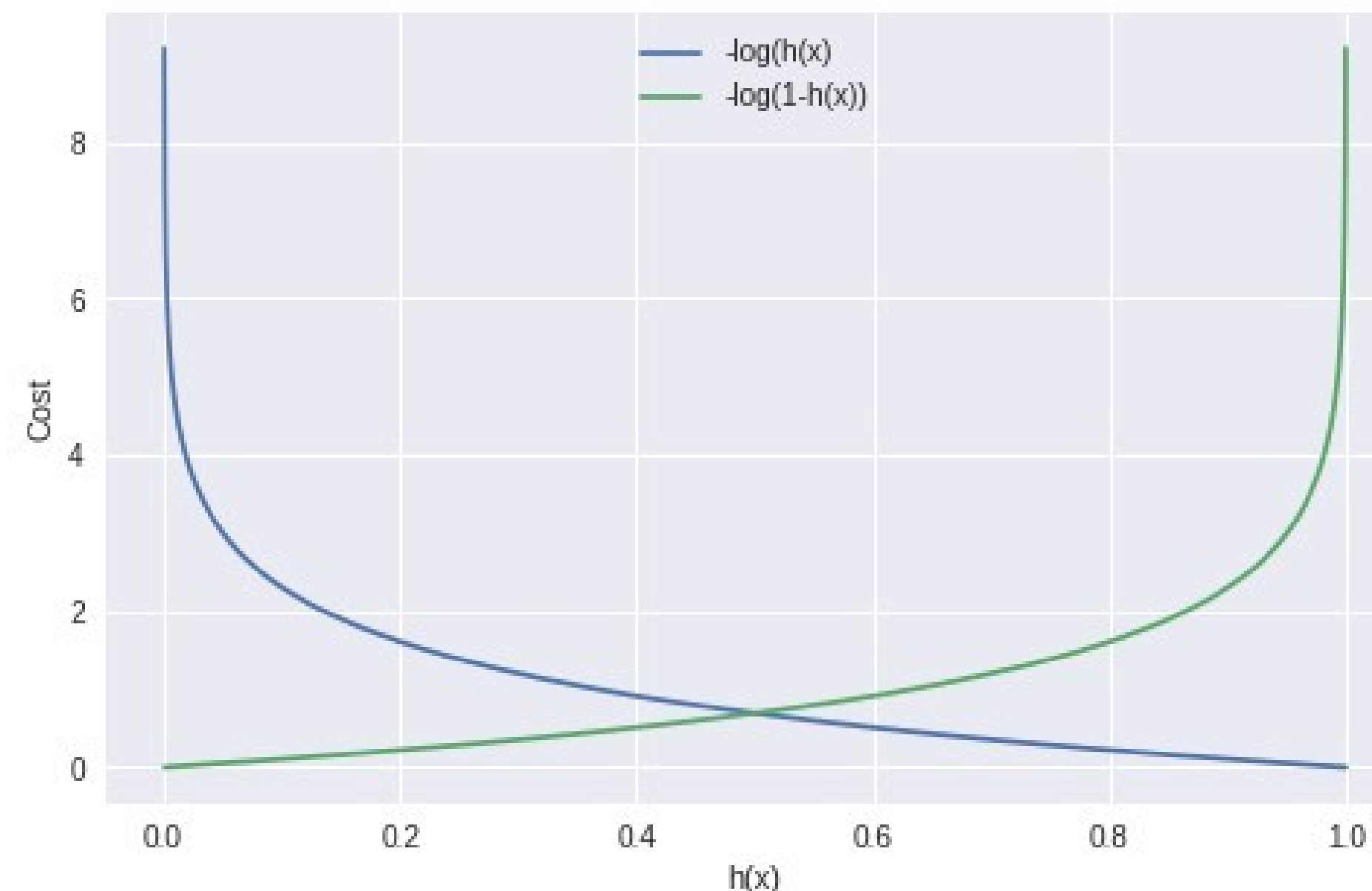


# Функция ошибки в логистической регрессии.

Модель ищет параметры, которые минимизируют функцию ошибки:

$$cost = \begin{cases} -\log(h(x)), & \text{if } y = 1 \\ -\log(1 - h(x)), & \text{if } y = 0 \end{cases}$$

Чем выше вероятность определения класса 1 при верном классе 0, тем выше стоимость ошибки.



# Функция ошибки в логистической регрессии.

Общий вид функции ошибки для модели:

$$\text{cost}(h(x), y) = -y \cdot \log(h(x)) - (1 - y) \log(1 - h(x))$$

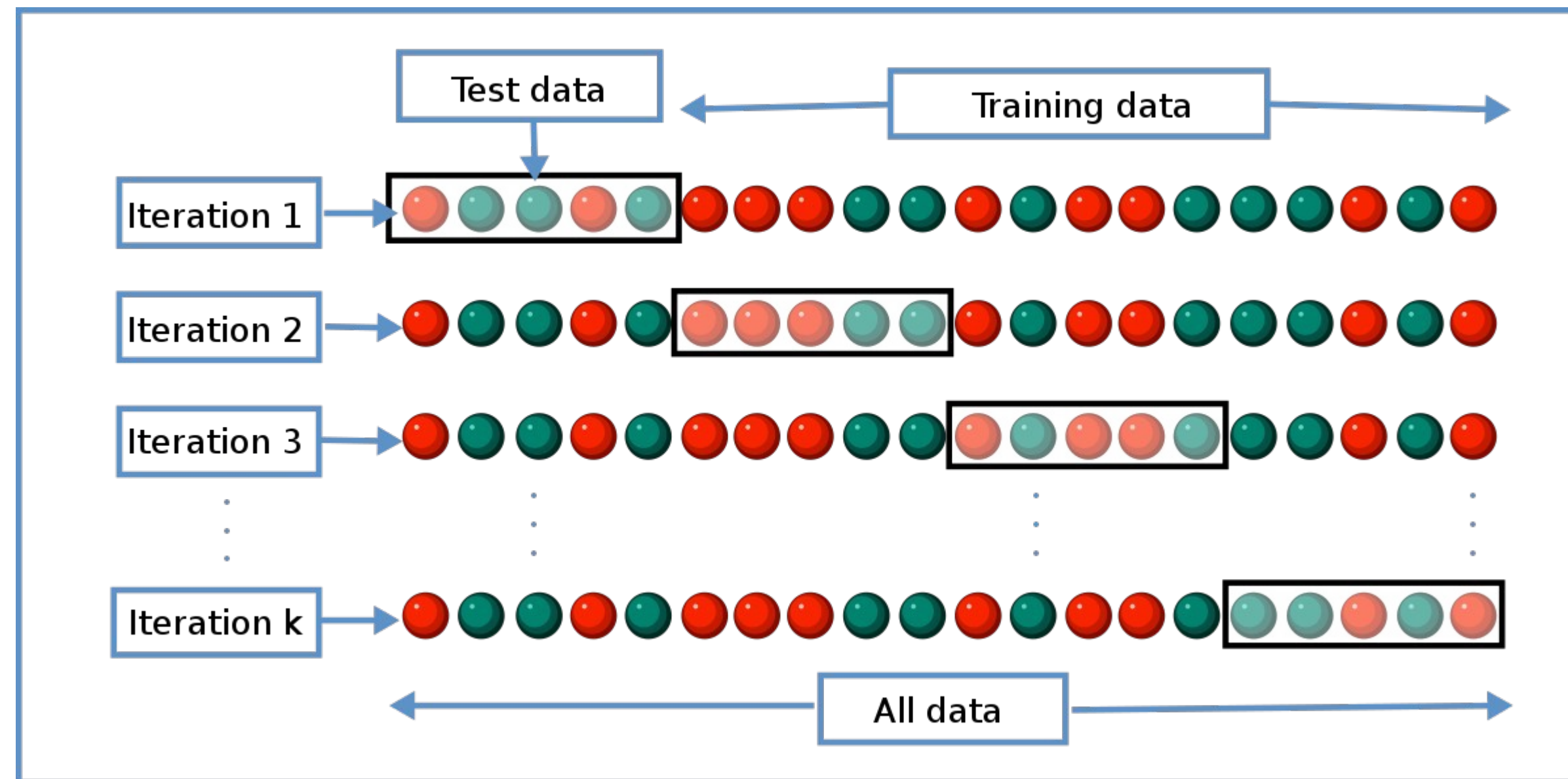
Ошибка для всех данных датасета:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^i \log(h(x^i)) + (1 - y^i) \log(1 - h(x^i))]$$

Где  $m$  – количество элементов.

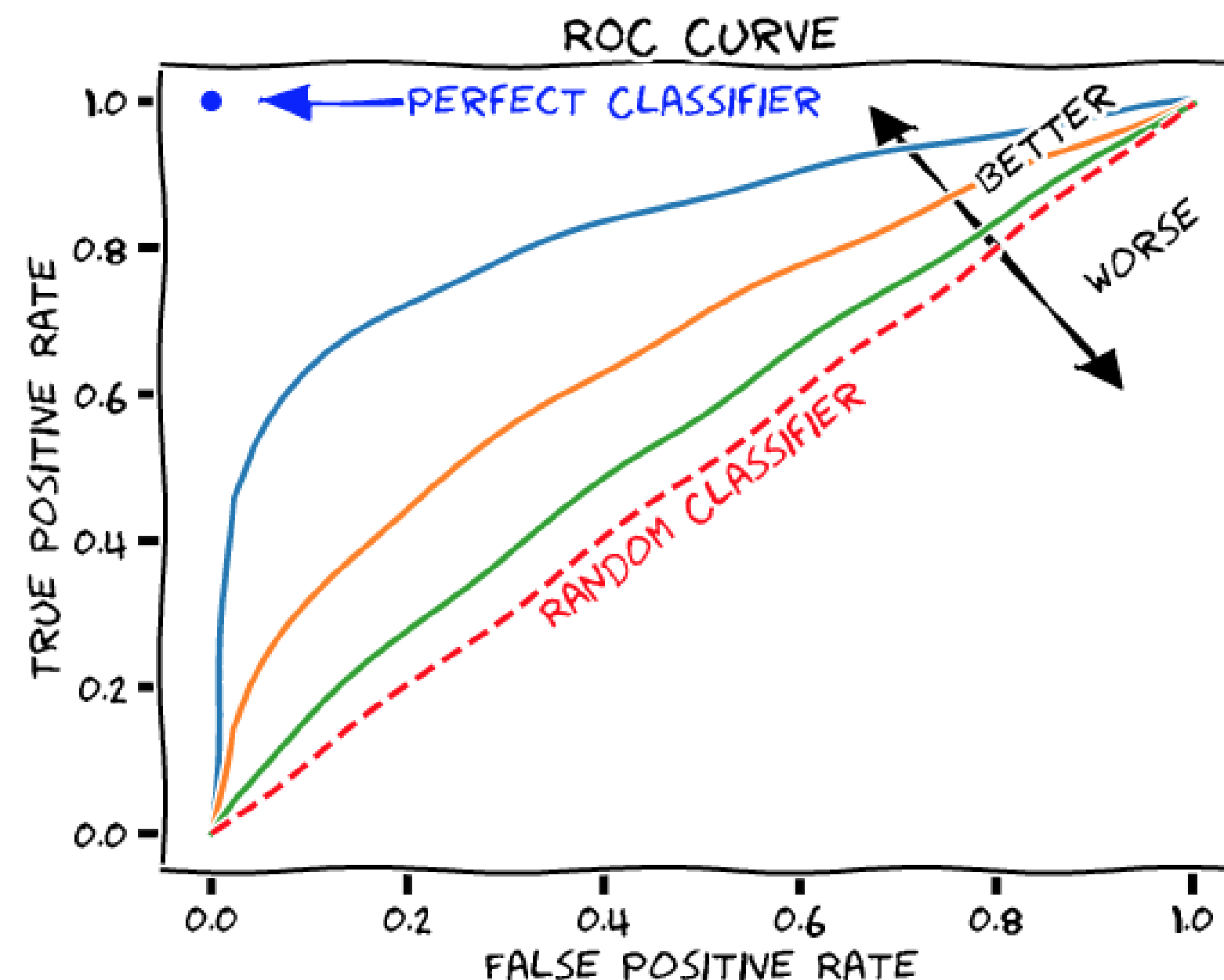


# Кросс-валидация



Оцениваем модель на нескольких тестовых данных

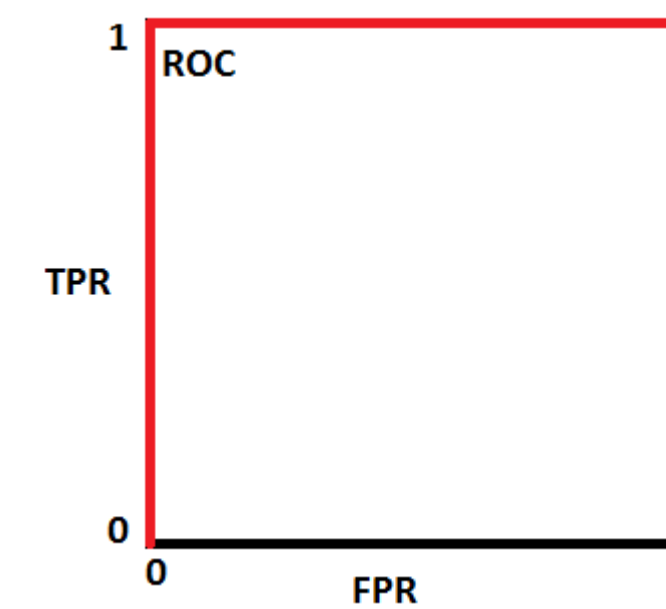
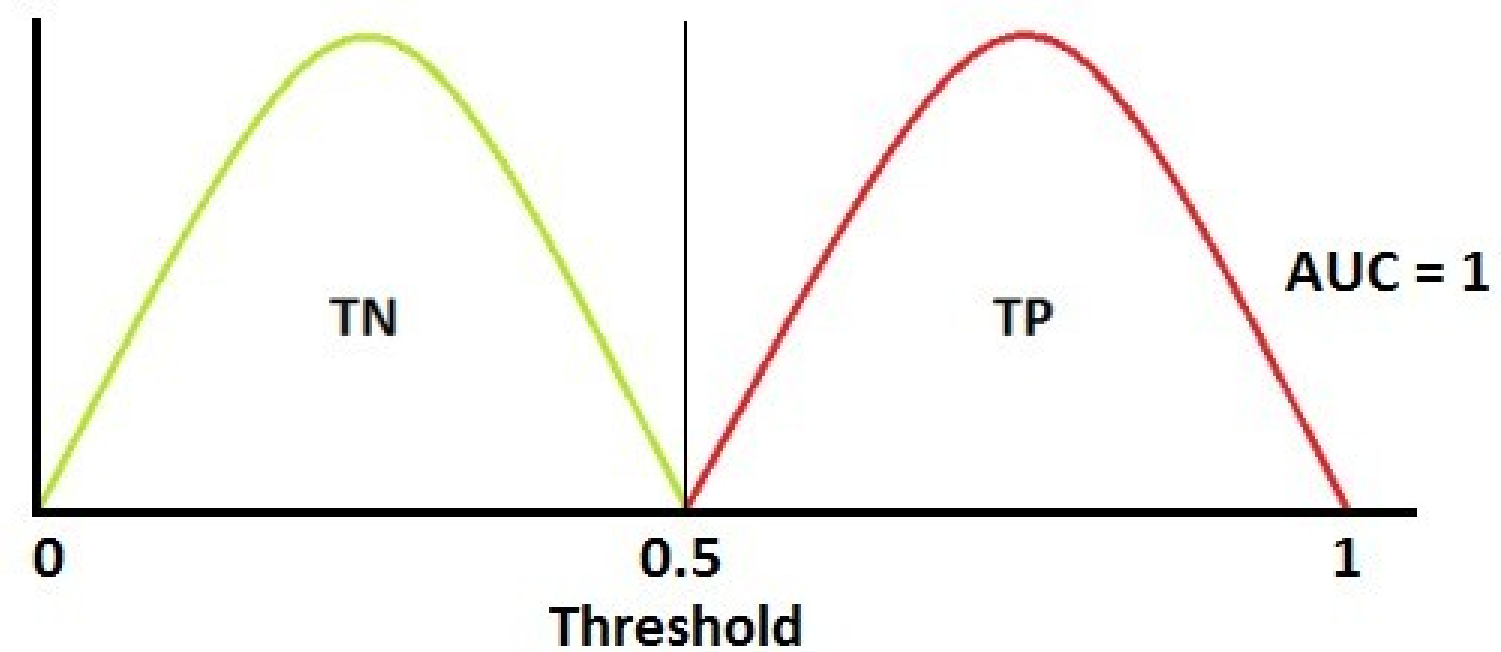
# Метрики классификации: ROC-кривая



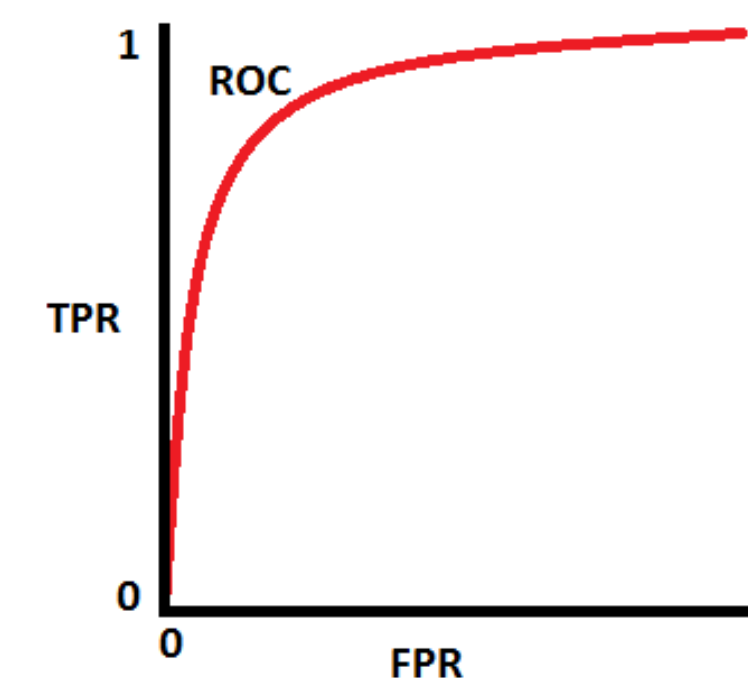
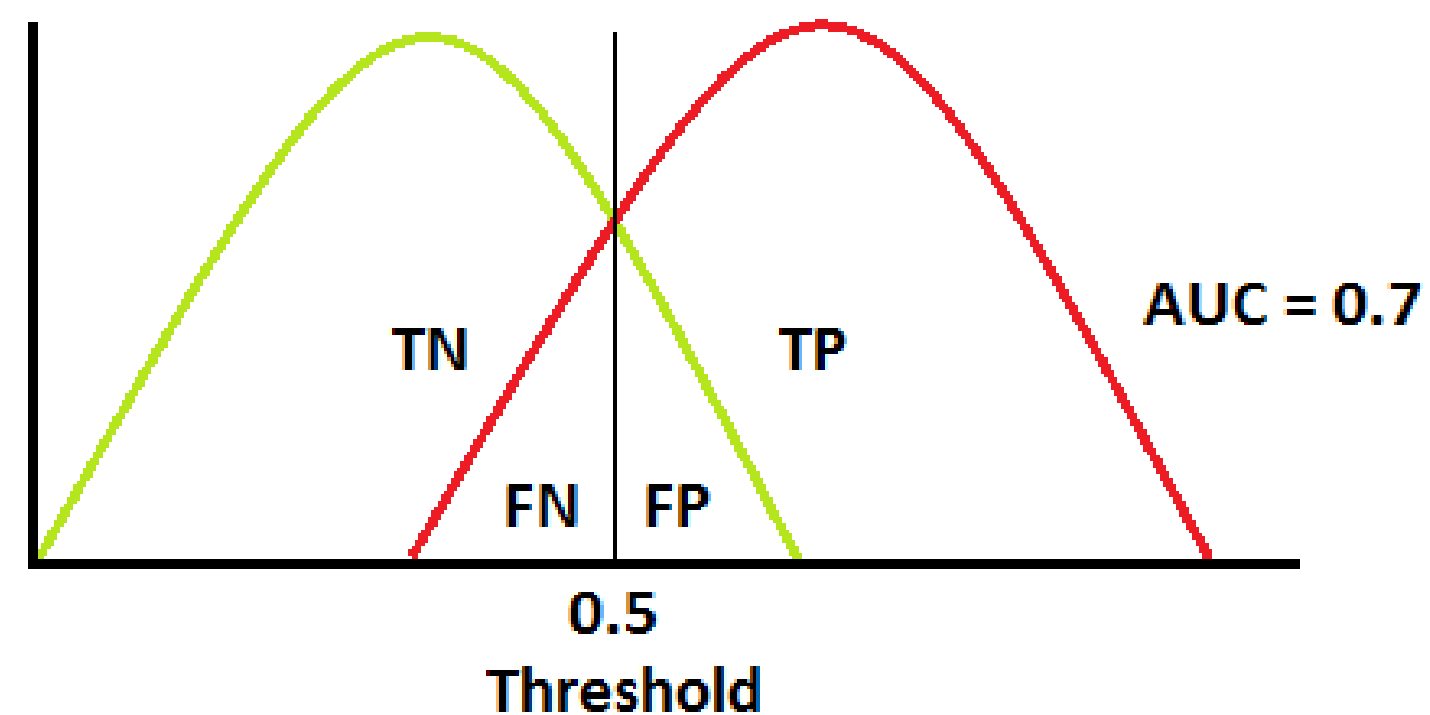
Позволяет определить порог,  
при котором мы будем отделять один класс от другого



# ROC-кривая



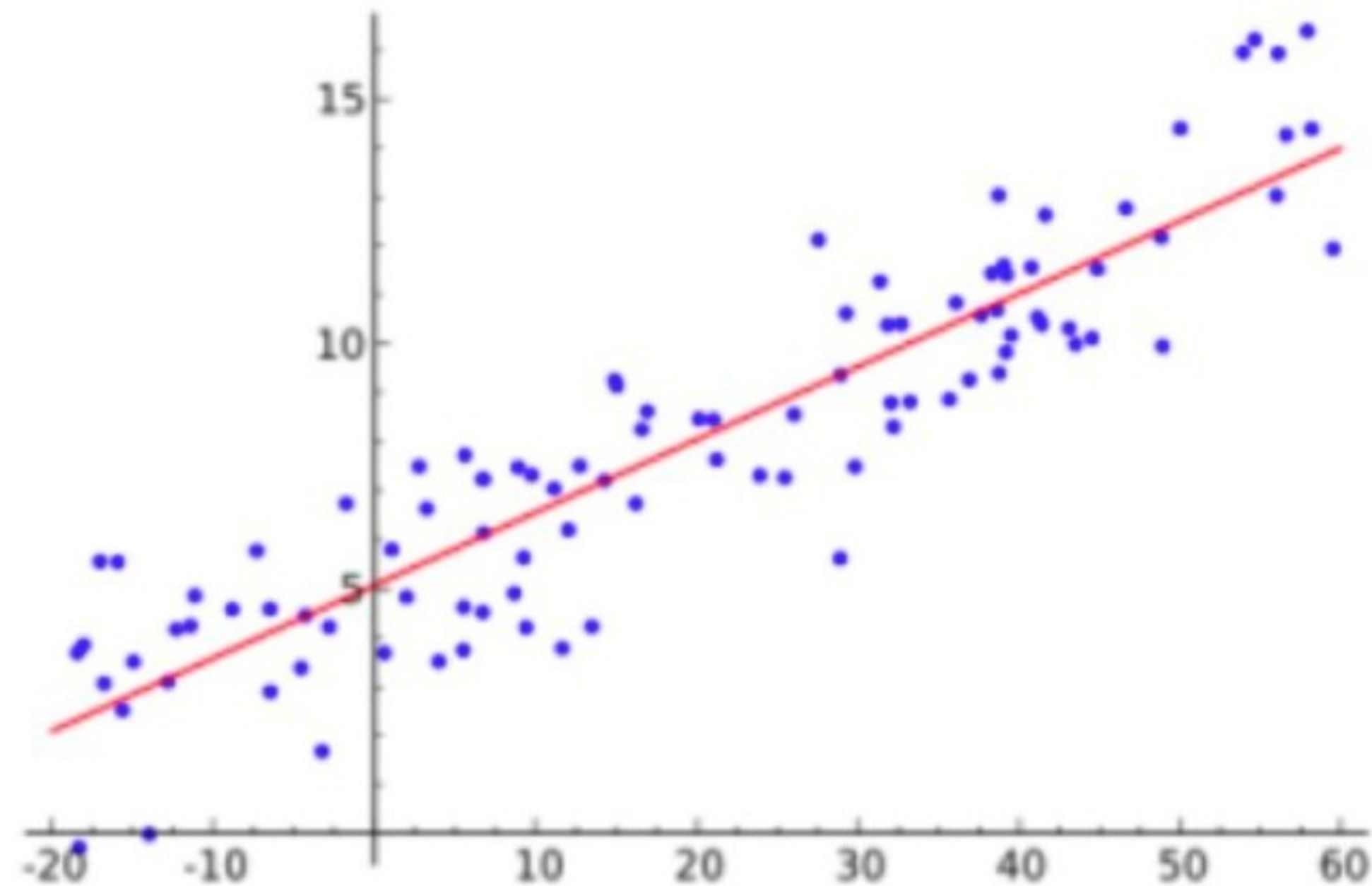
Идеальная модель — порог 50%



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

## Регрессия

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



# Вопросы?

Контакты спикера:  
[yustiks@gmail.com](mailto:yustiks@gmail.com)