

Юстина Иванова

Программист, data scientist

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

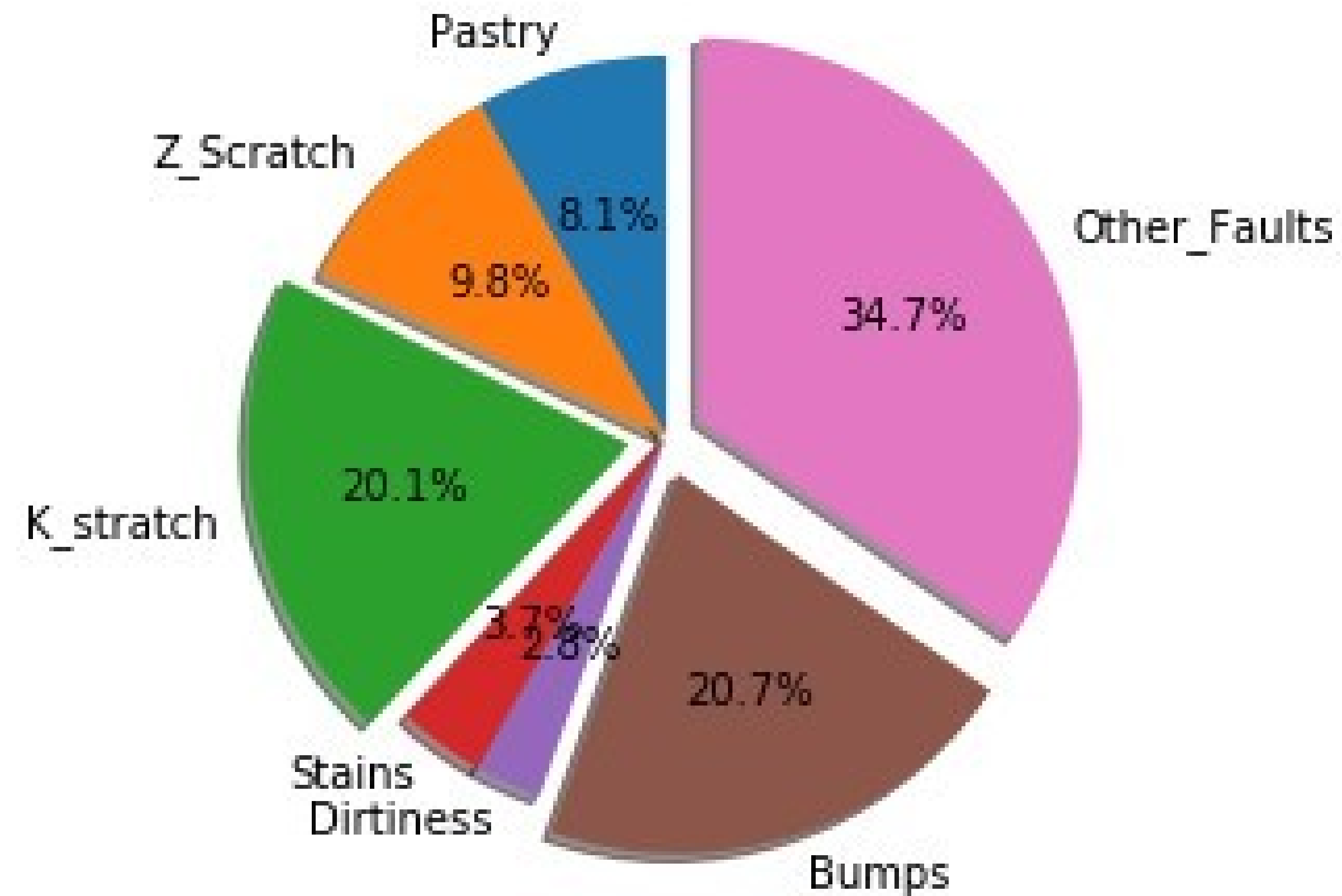
Спикер

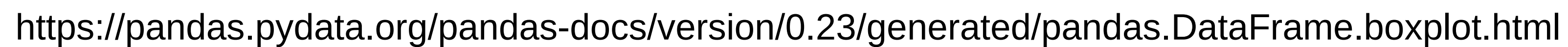


Юстина Иванова,

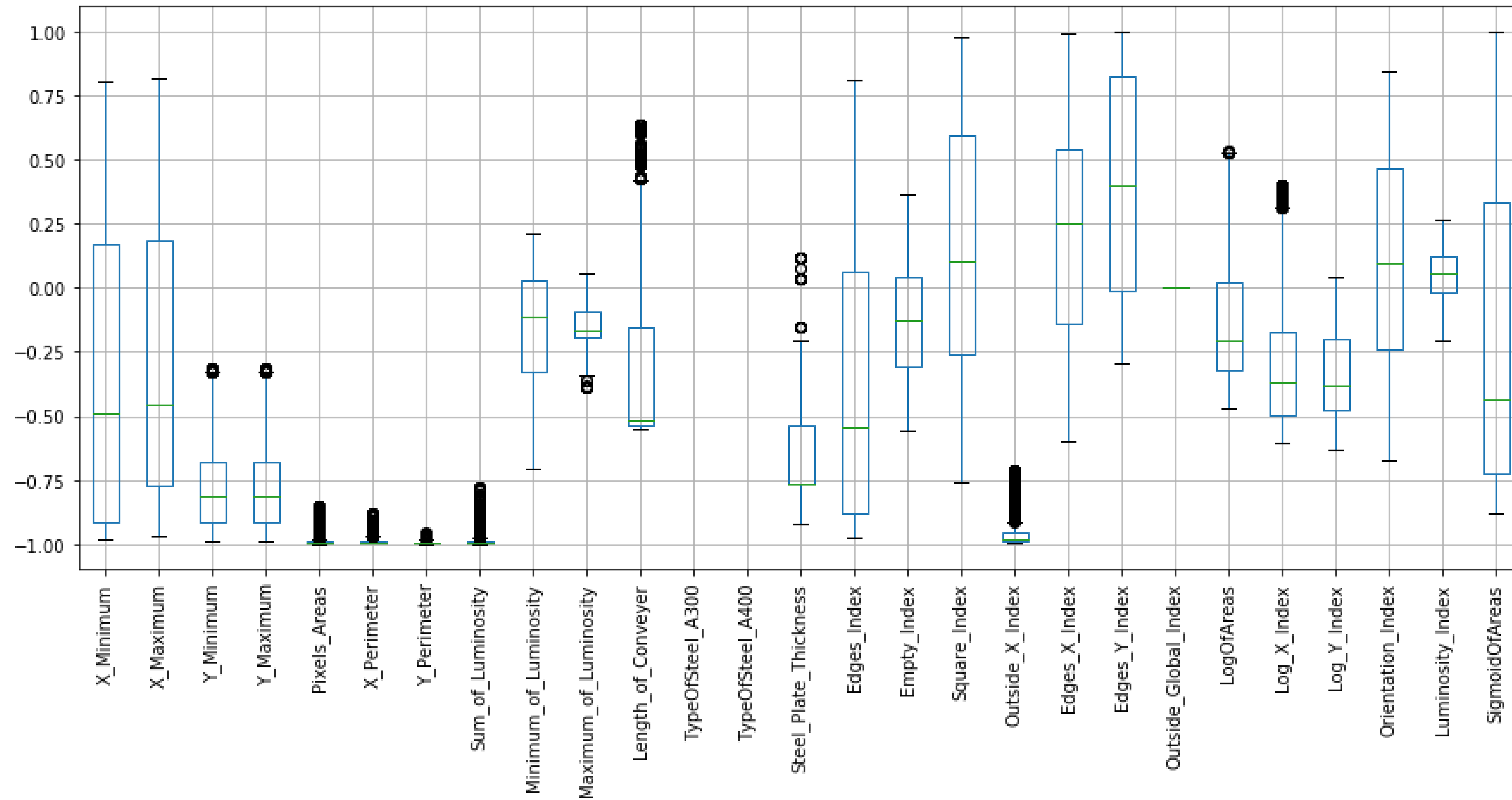
- PhD в Университете Больцано
- Выпускница МГТУ им. Баумана
- Магистр по Artificial Intelligence в University of Southampton

Датасет Faulty Steel Plates.

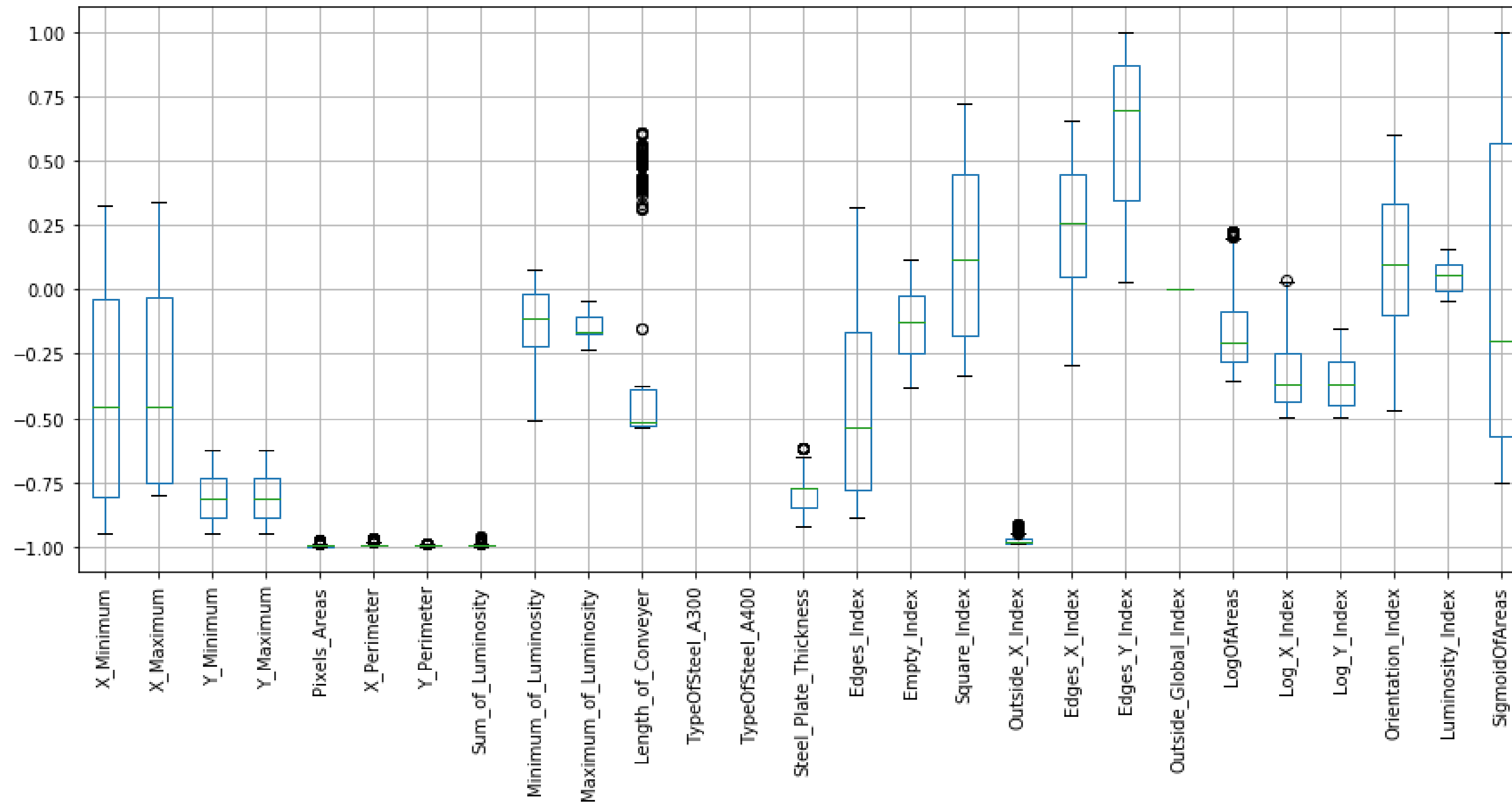




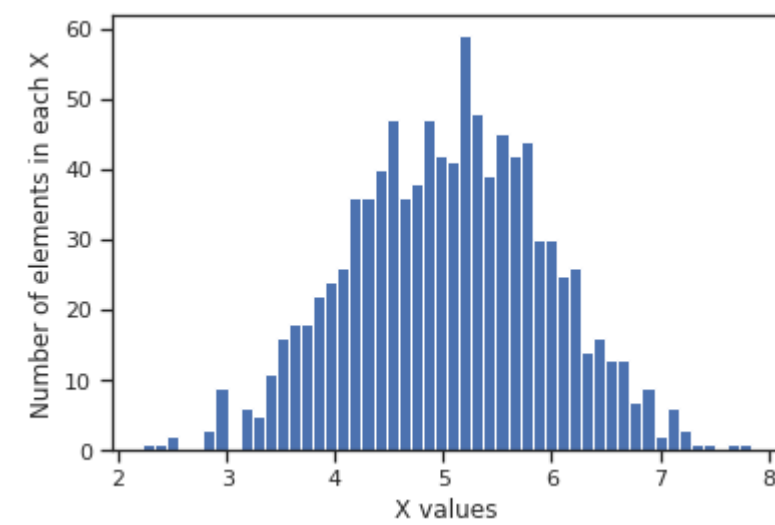
Удаление элементов вне интерквартильного интервала..



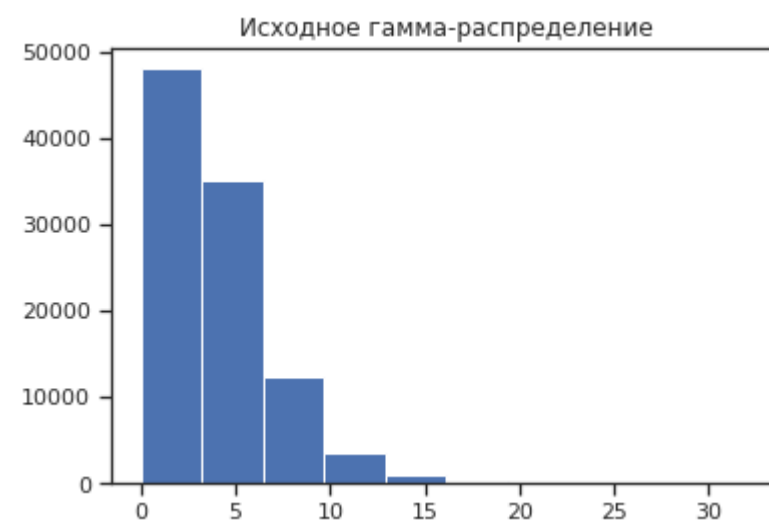
Удаление элементов вне квантилей 20% и 80%.



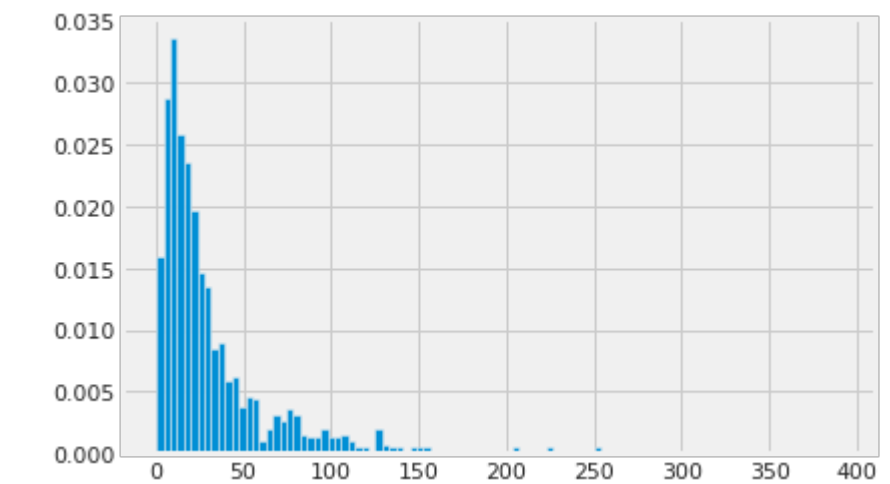
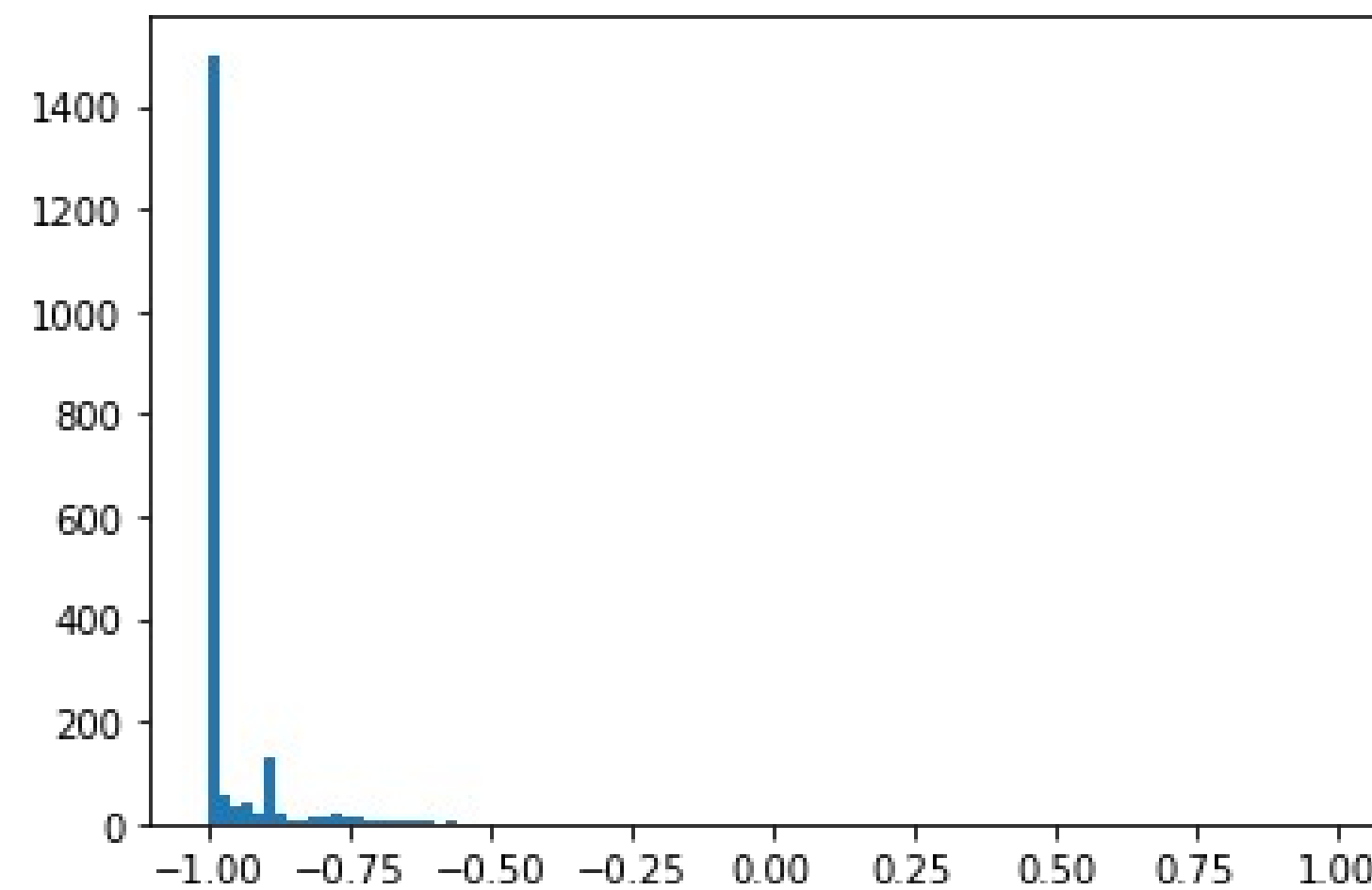
Тесты на согласие: какое это распределение?



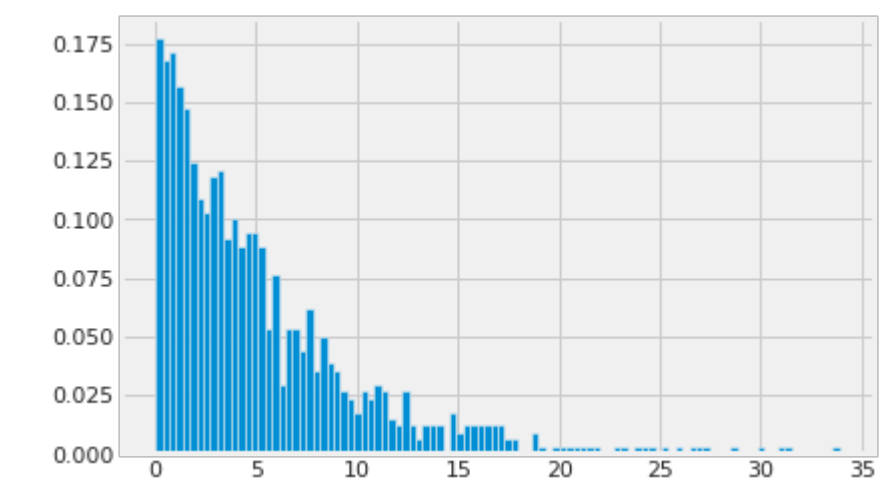
нормальное



гамма

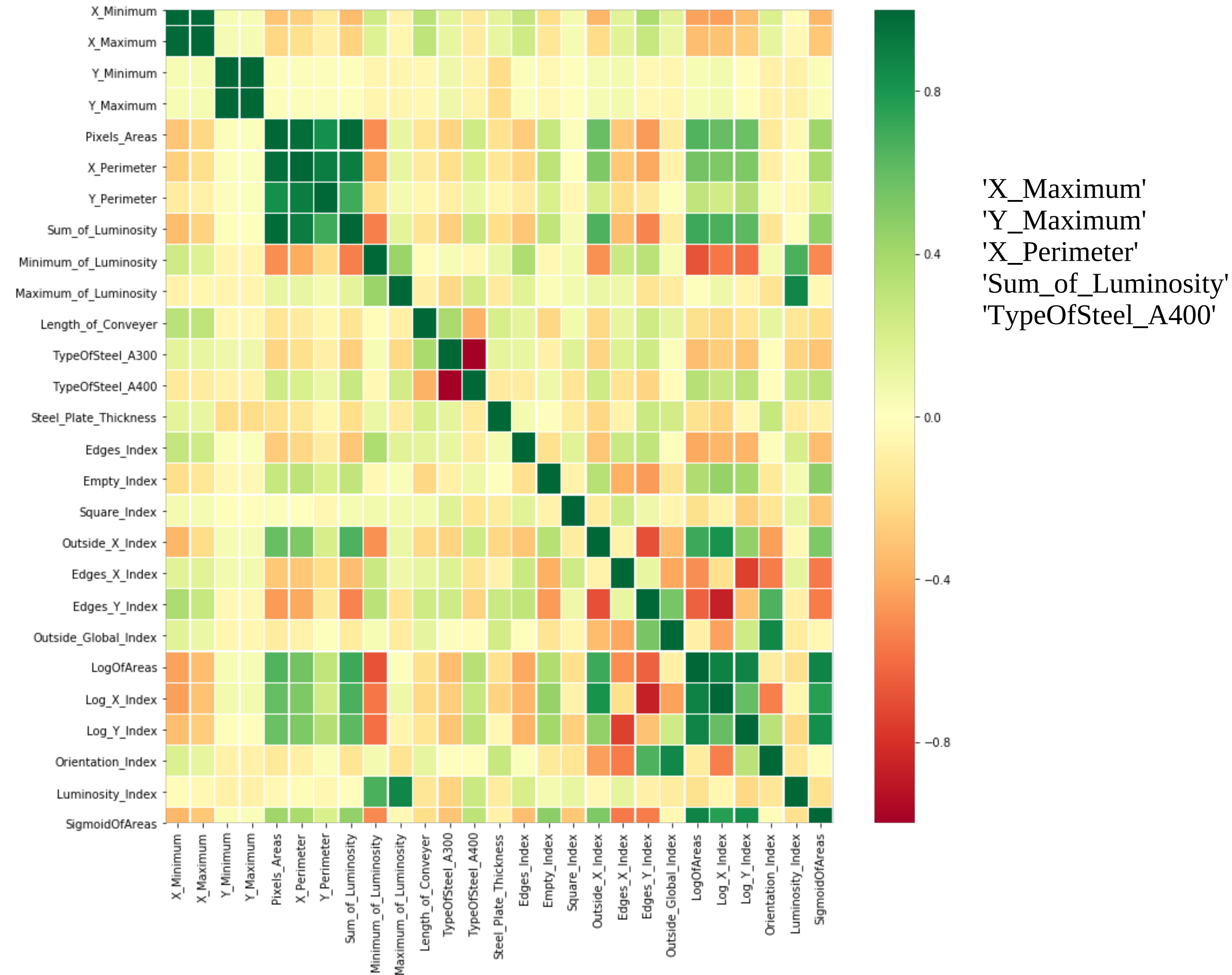


экспоненциальное



экспоненциальное

Удаление мультиколлинеарности

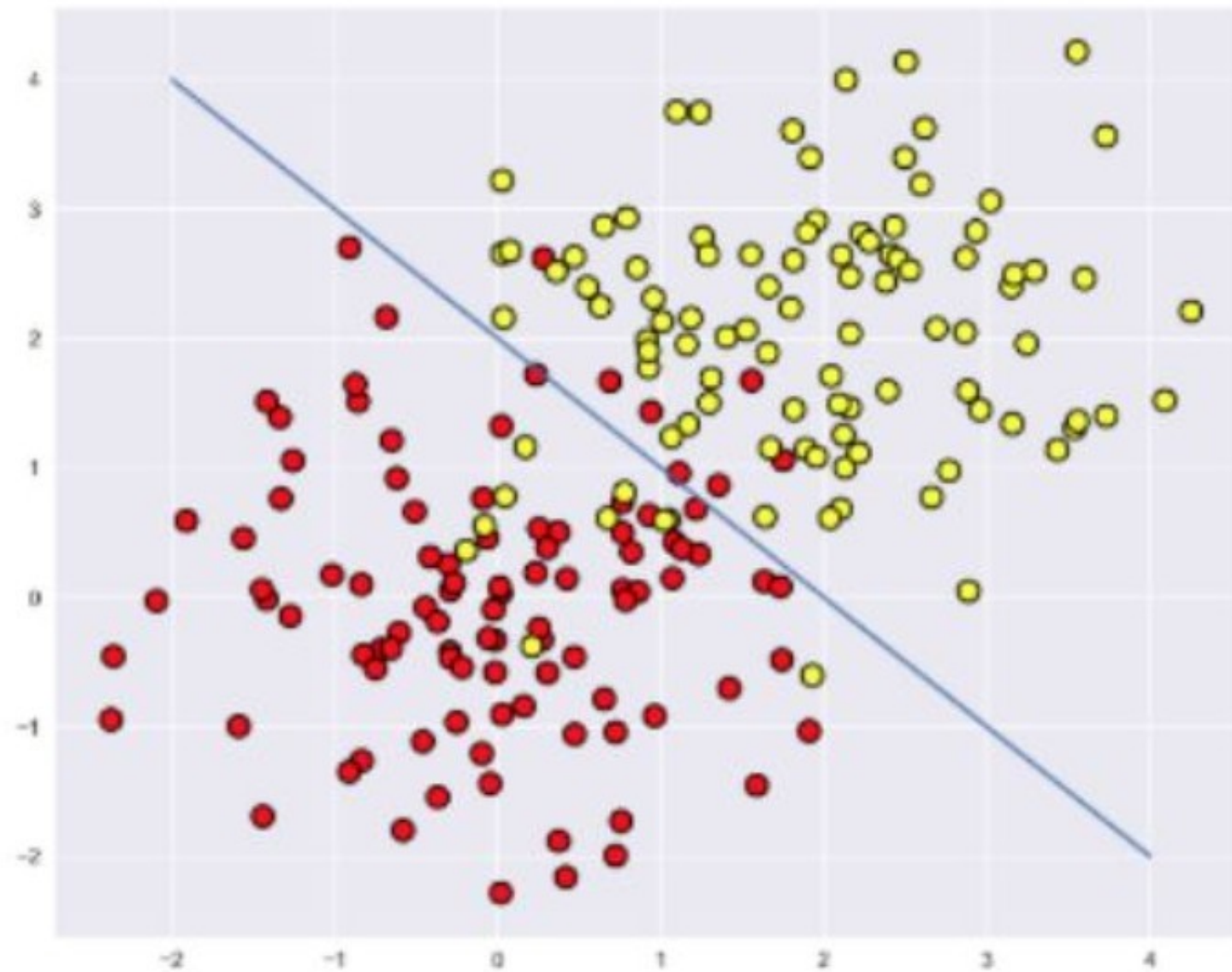


Классическое Обучение



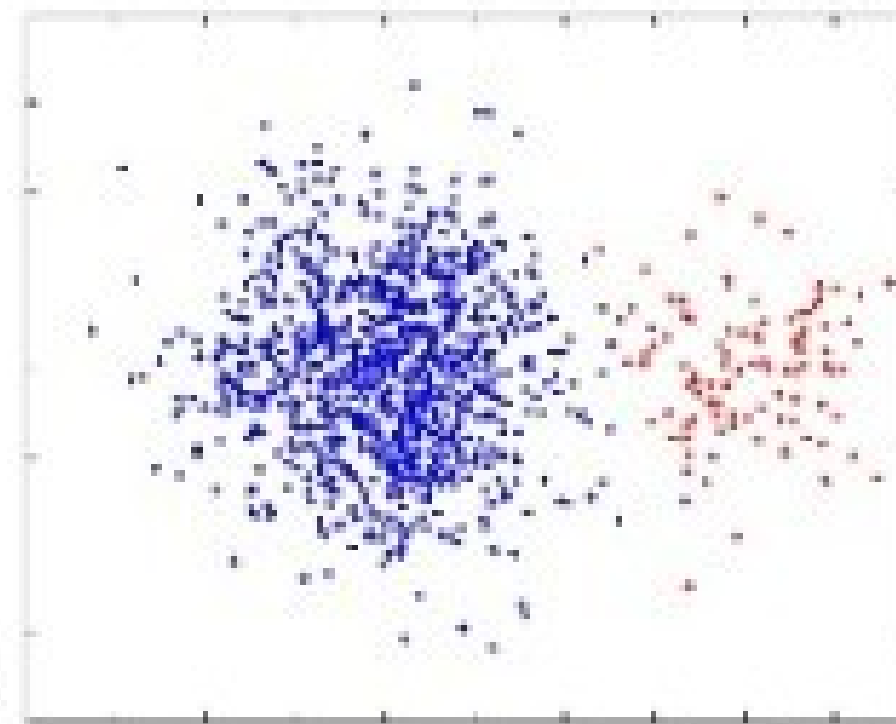
Классификация

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

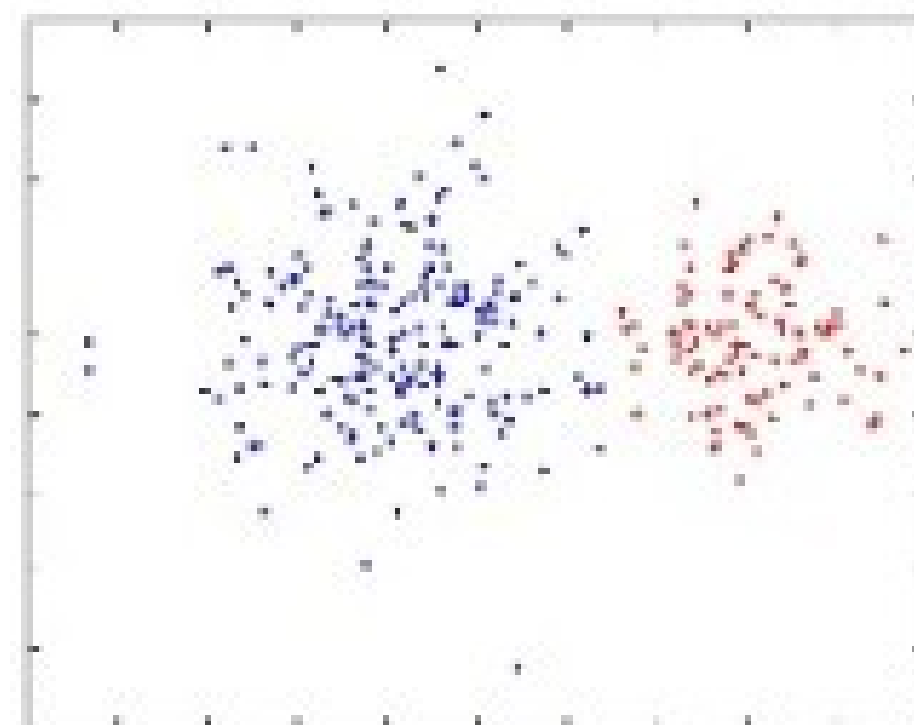


Проблема несбалансированности классов.

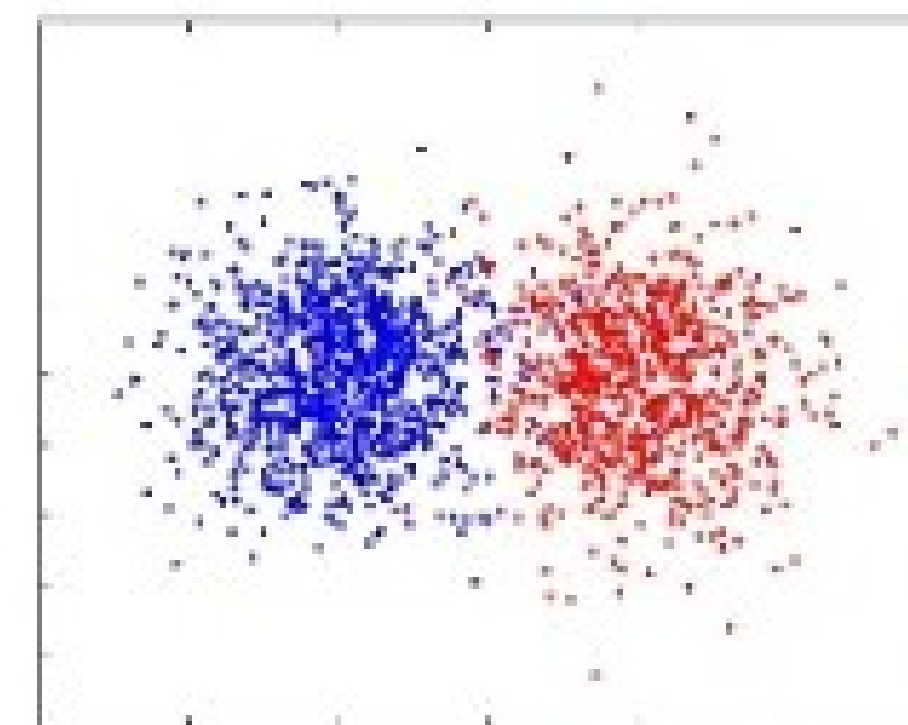
Sampling: Rebalancing the dataset



Under-sampling



Over-sampling

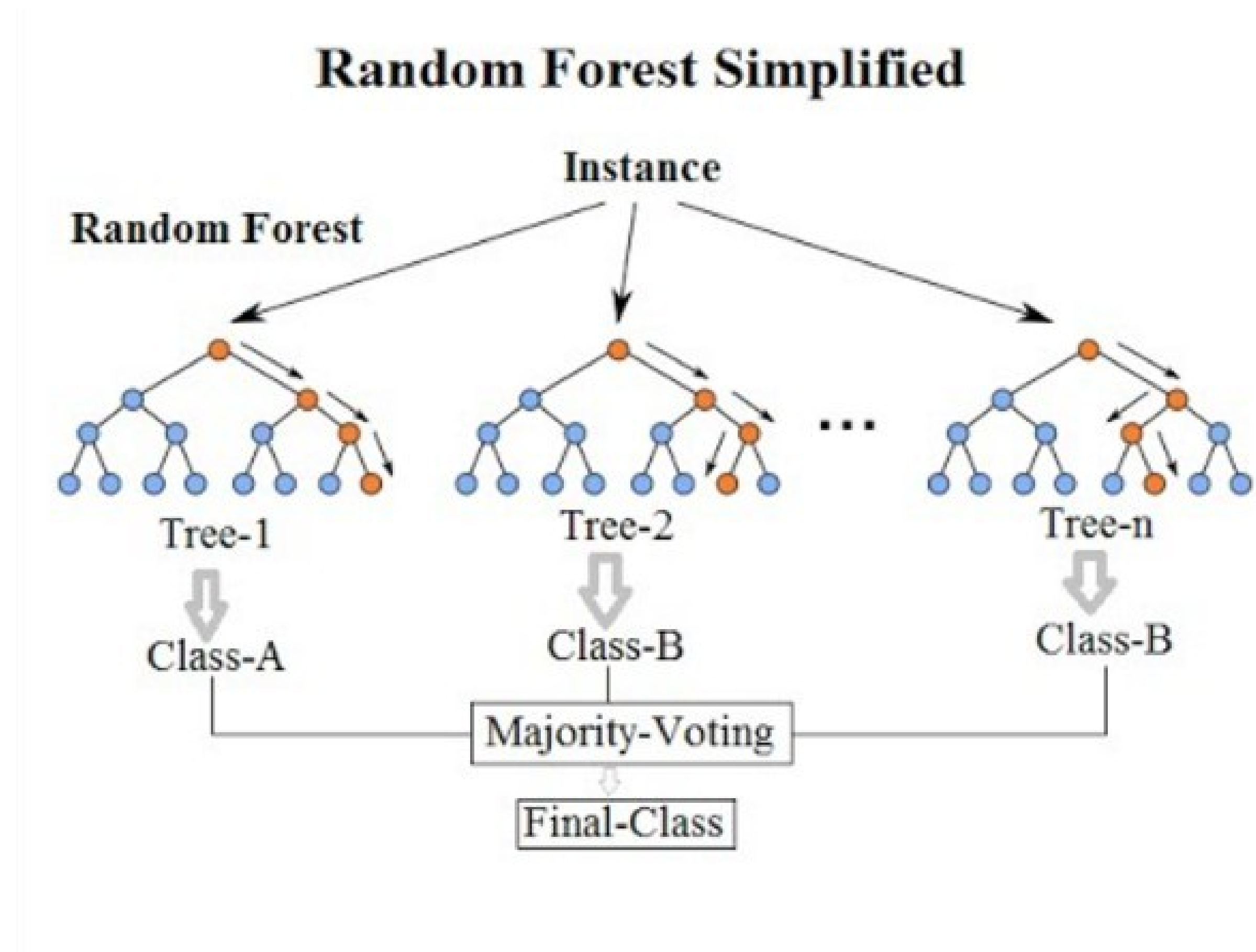


Дерево решений.

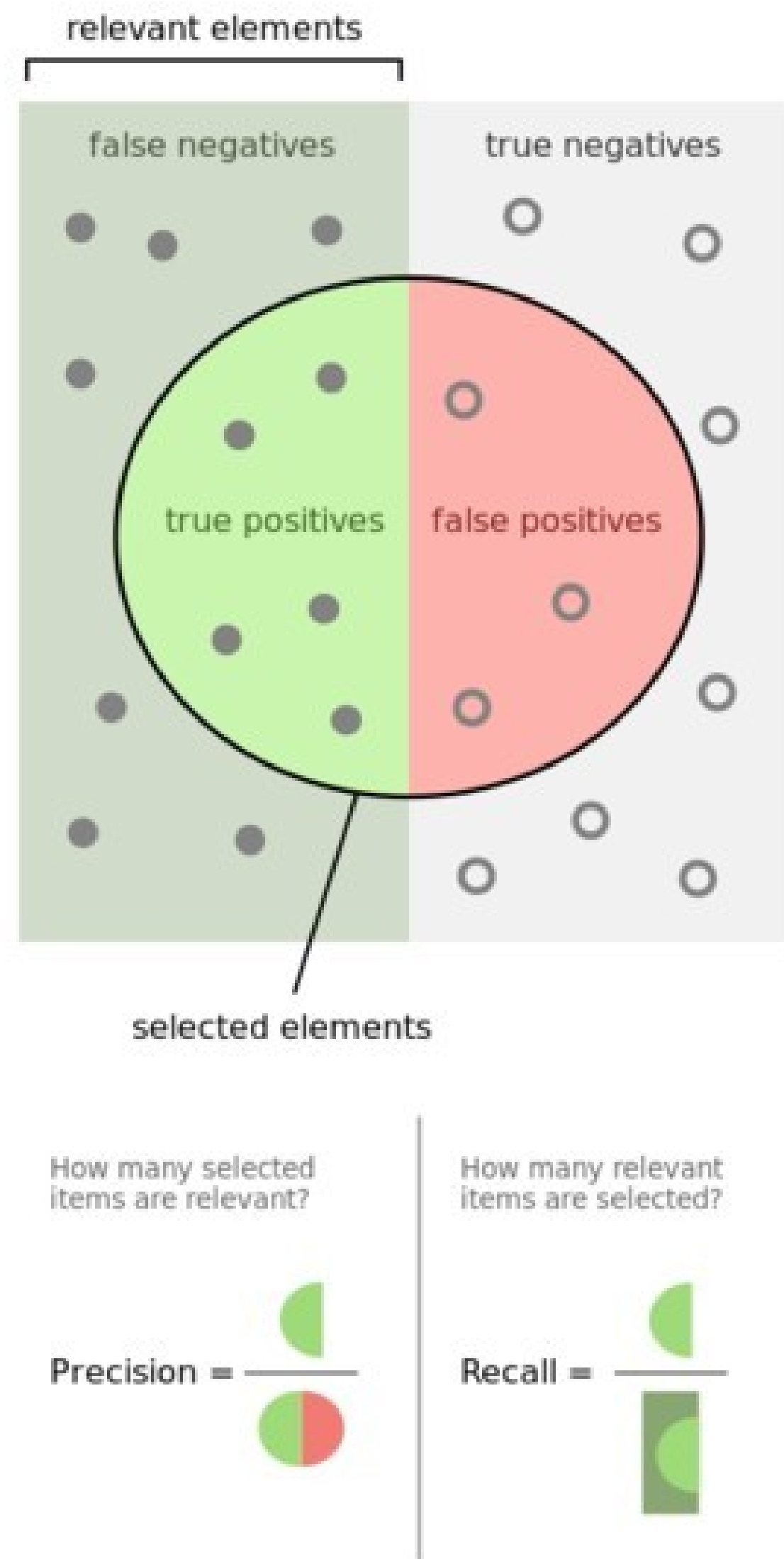
Давать ли кредит?



Случайный лес.



Метрики классификации



Precision

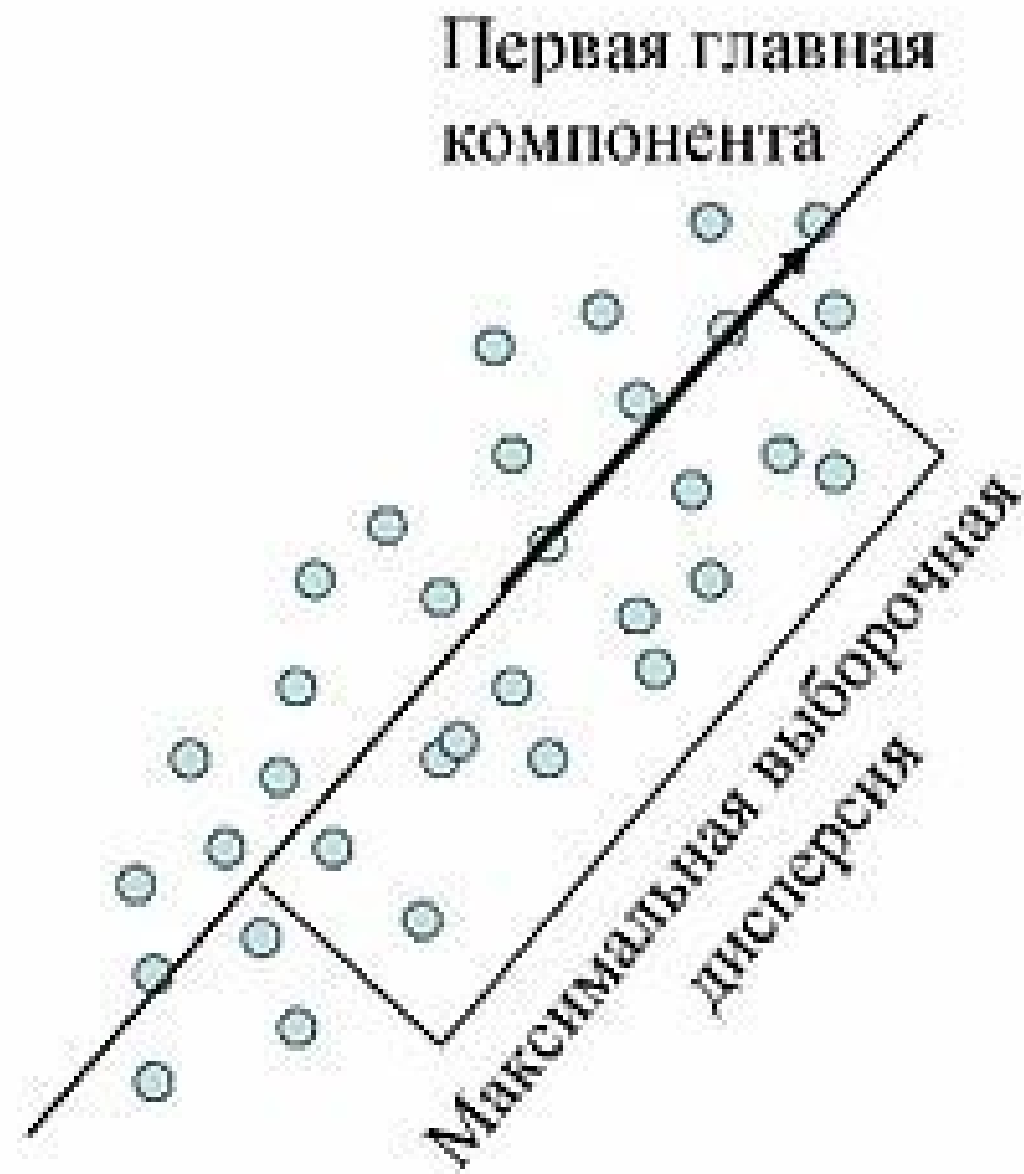
Recall

F1-мера

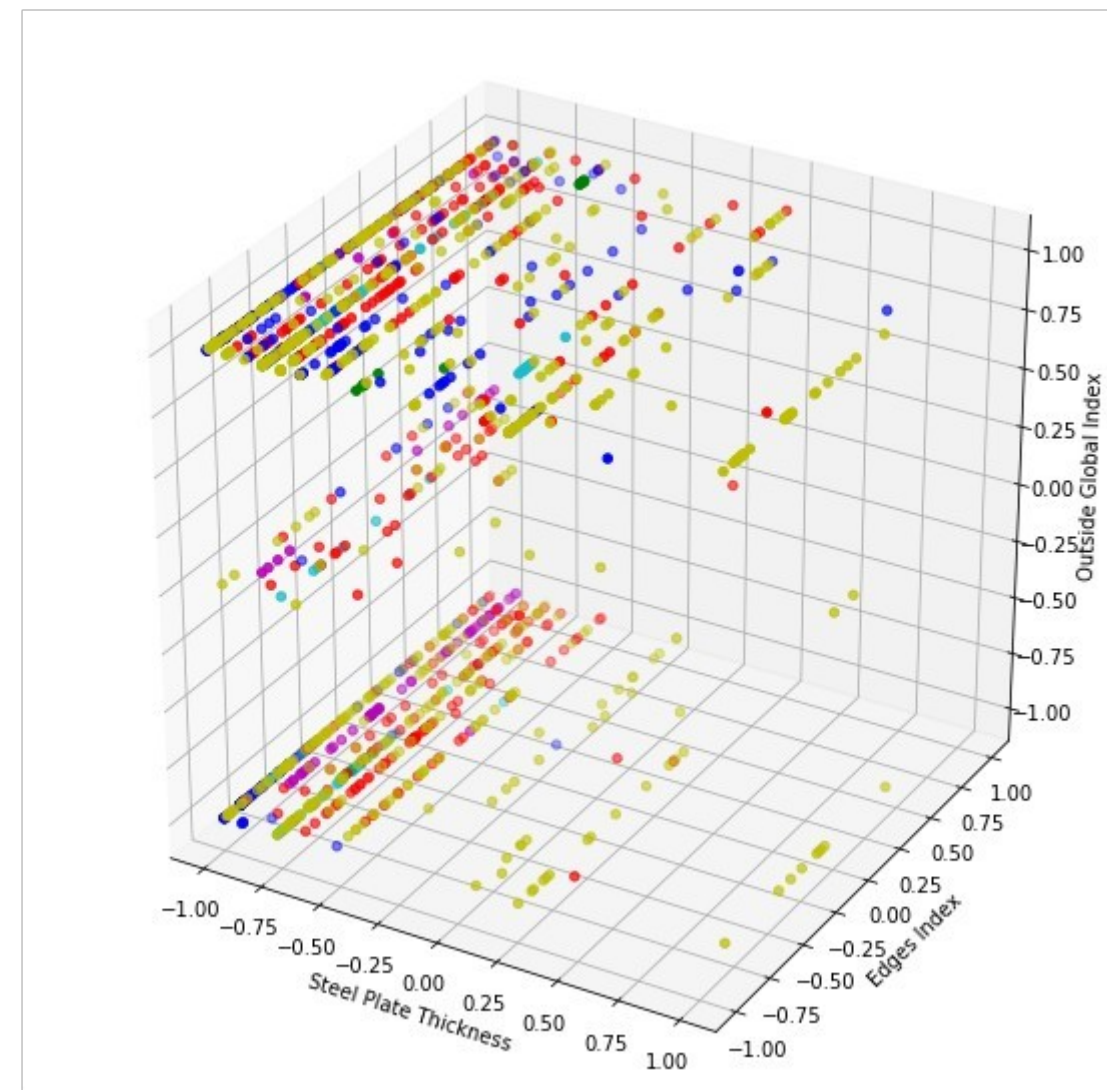
$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Принцип минимальных компонент.

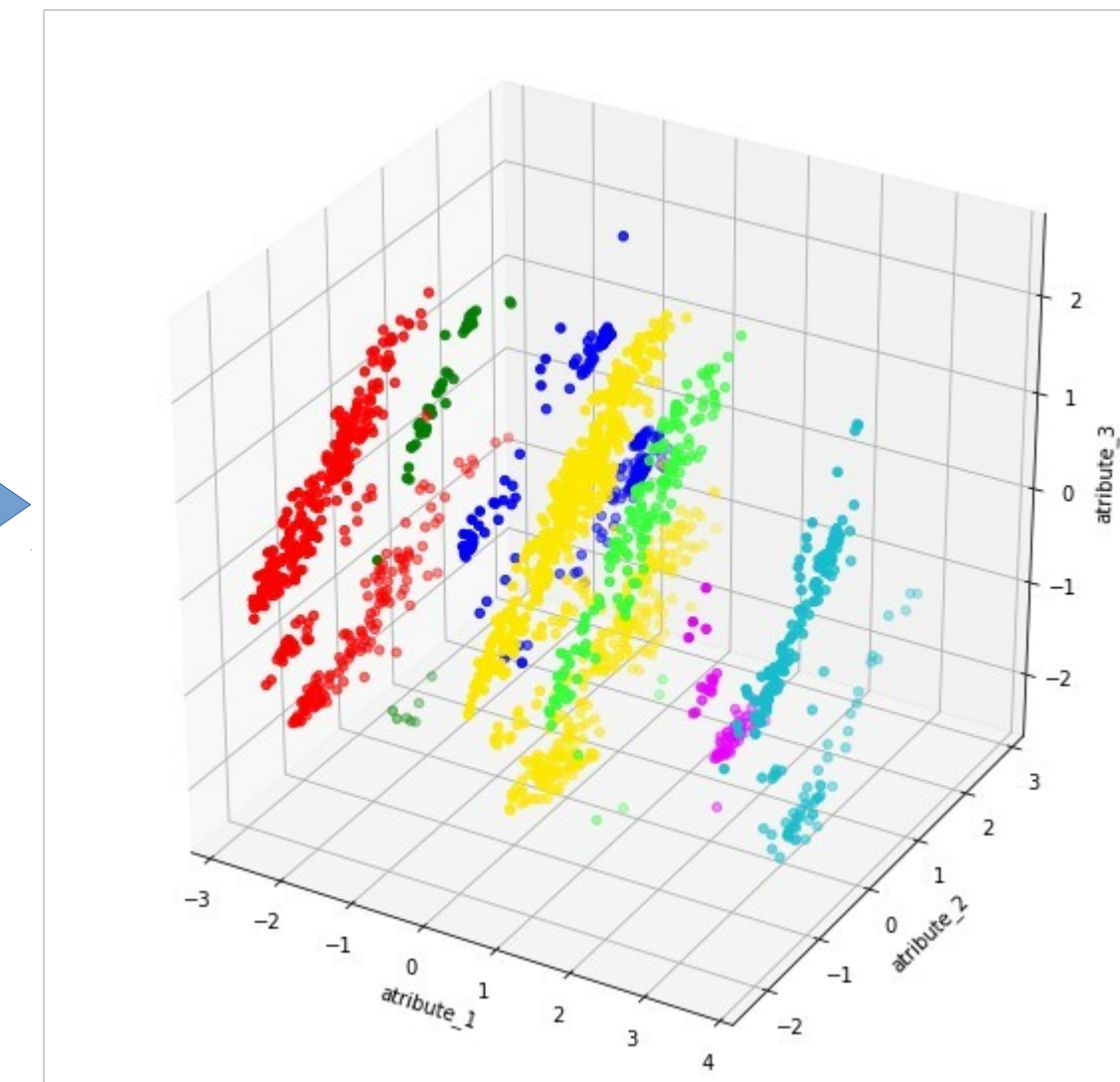
Поиск ортогональных проекций с наибольшим рассеянием



Было



Стало

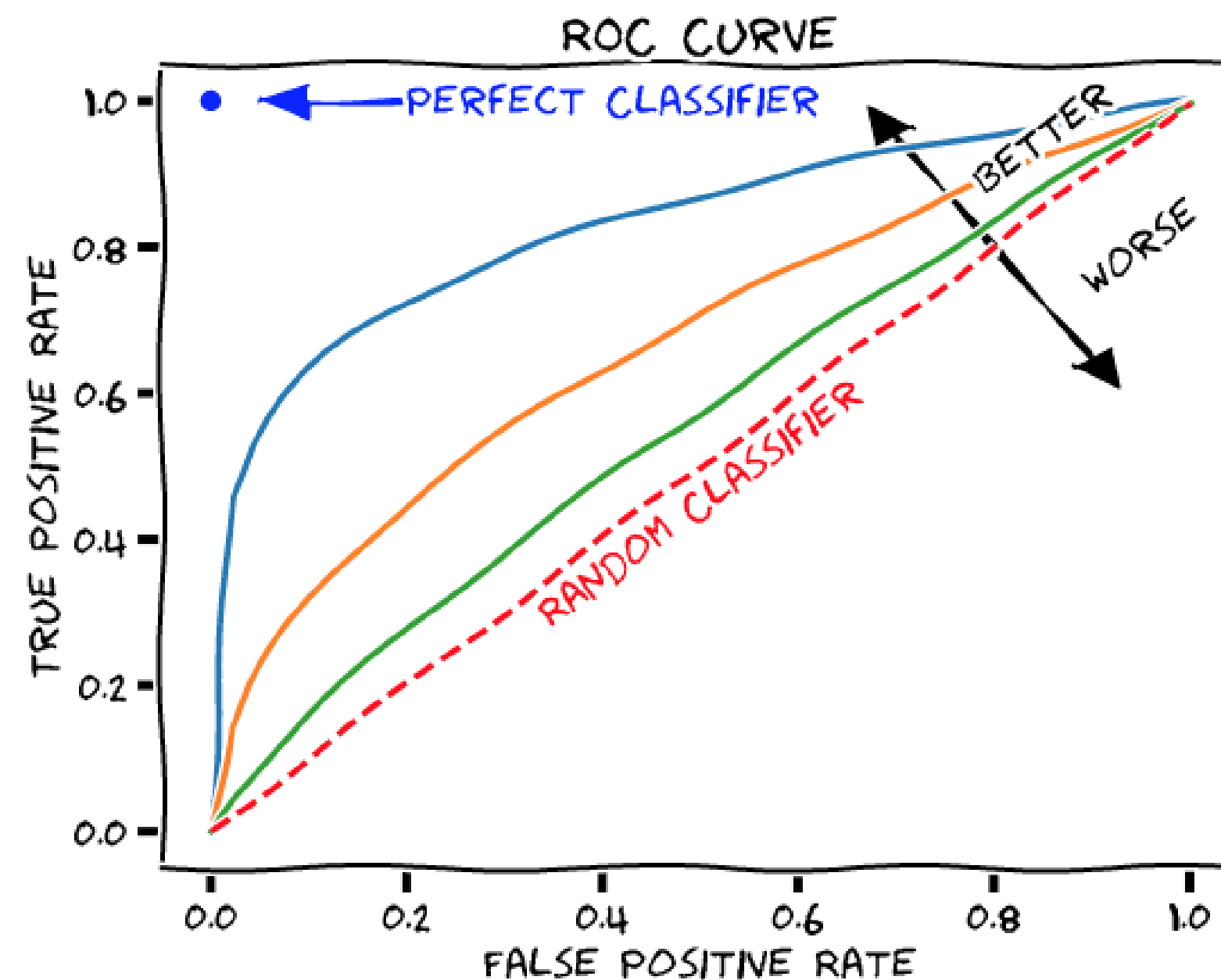


Кросс-валидация



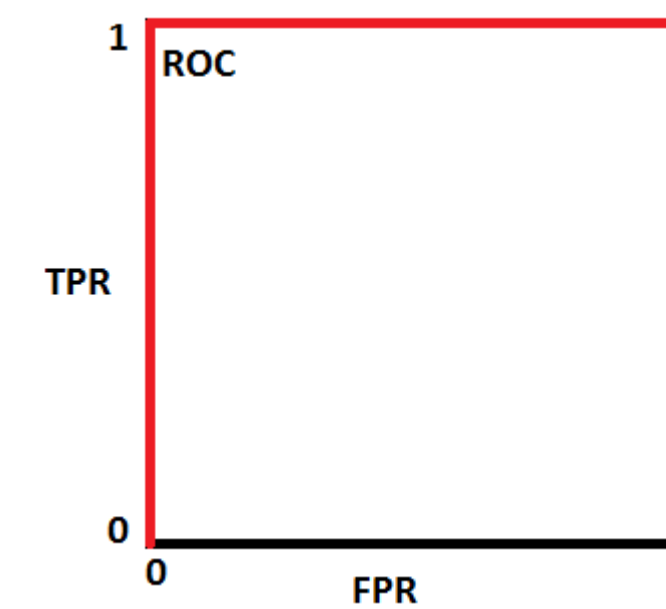
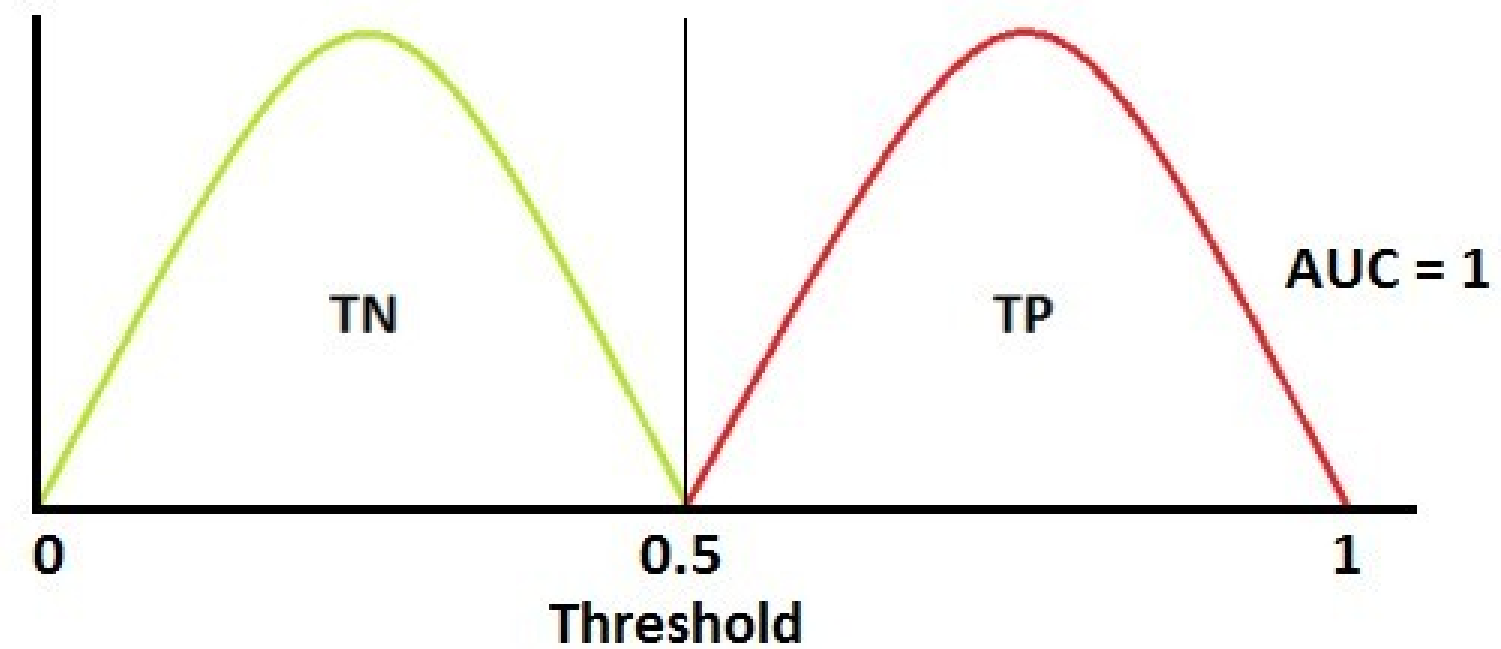
Оцениваем модель на нескольких тестовых данных

Метрики классификации: ROC-кривая

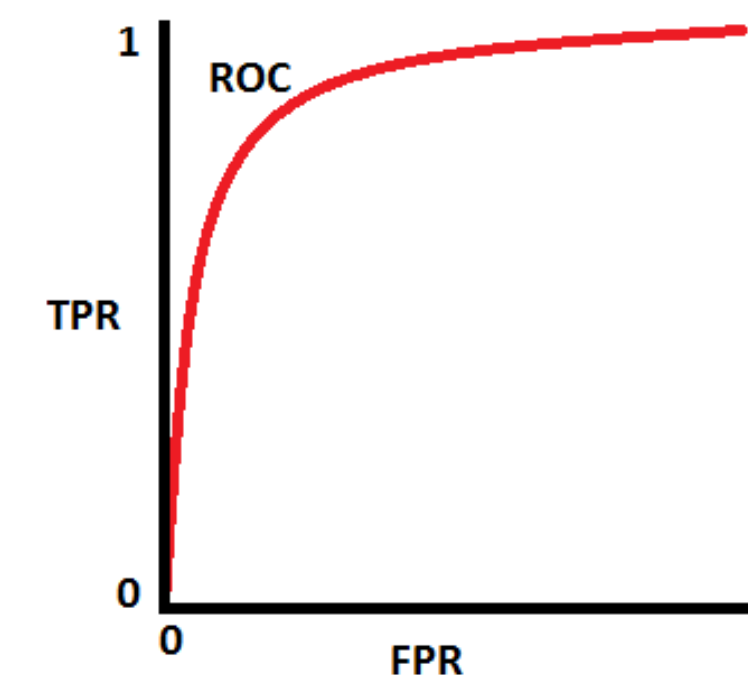
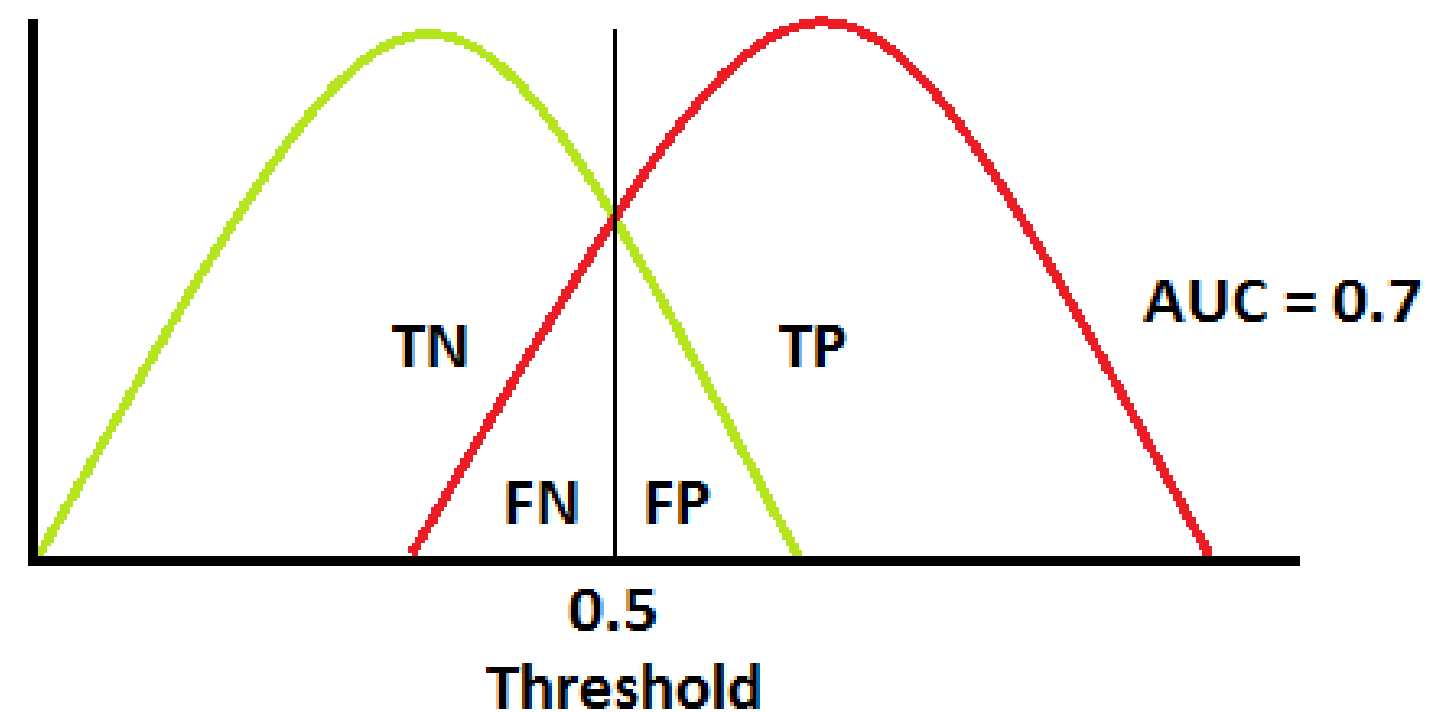


Позволяет определить порог,
при котором мы будем отделять один класс от другого

ROC-кривая



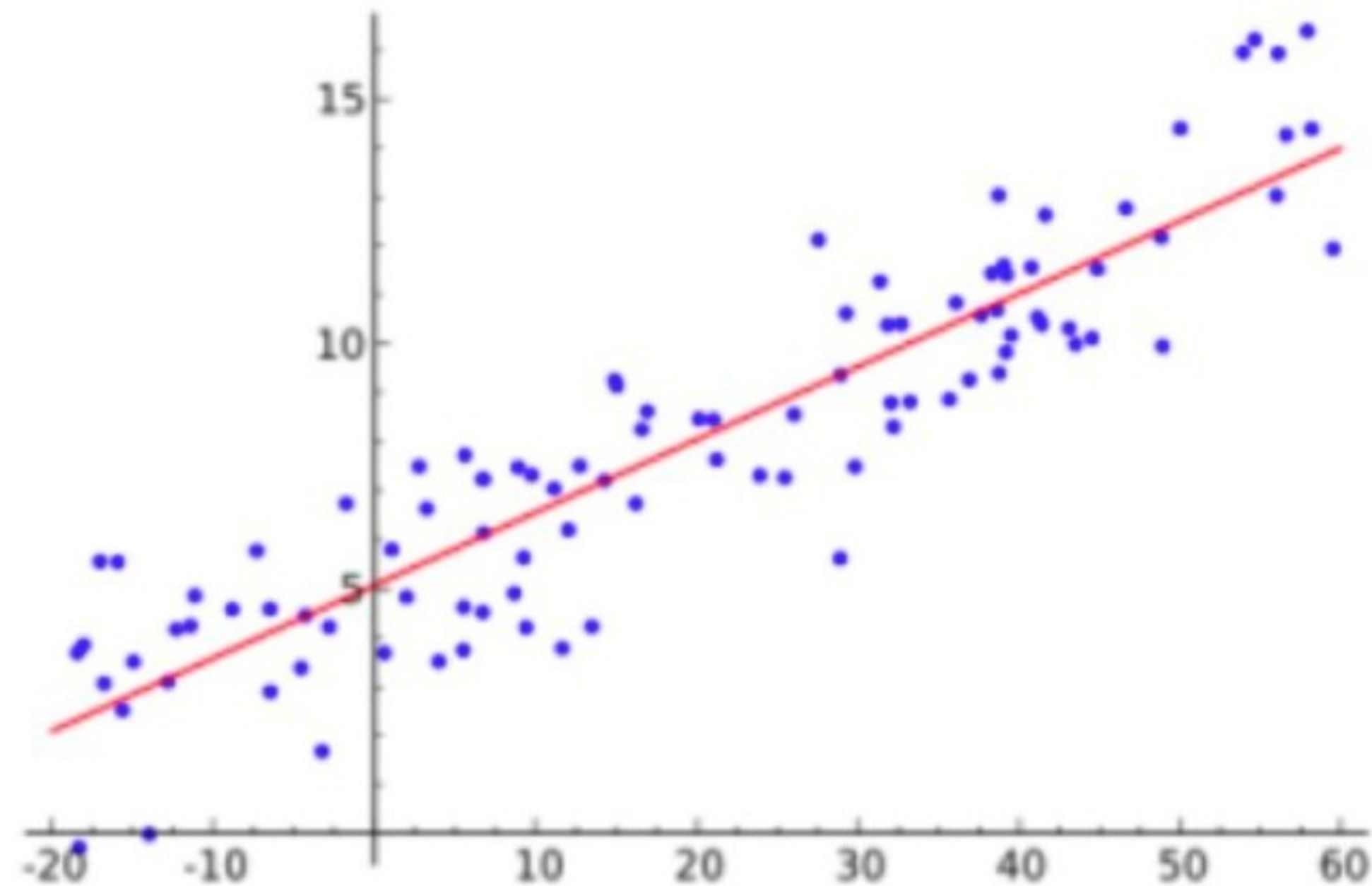
Идеальная модель — порог 50%



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