Предсказание магнитных свойств наночастиц для биомедицинских применений

Обработка данных

SCAMT

Что такое обработка данных в ML проекте?

Данные – таблица (DataFrame), колонками которого являются дексрипторы
Строка – вектор, содержащий информацию об одном эксперименте

Что с этим делать?

Понять, какие типы данных присутствуют в нашей таблице (строковый, чистовой, списки тд)

Удаление дубликатов

Feature engineering – использование собранных данных для создания новых дескрипторов, отбор независимых параметров

Missing data handling – некоторые алгоритмы машинного обучения не могут работать с пустыми строками: удаление или заполнение (какой алгоритм?)

Удаление выбросов – как распознать выброс (визуально, Z-score, квартили?). Особенность химических данных

Нормализация данных – привести мультимодальные данные к одному виду, сгладить разницу в значениях

Feature engineering

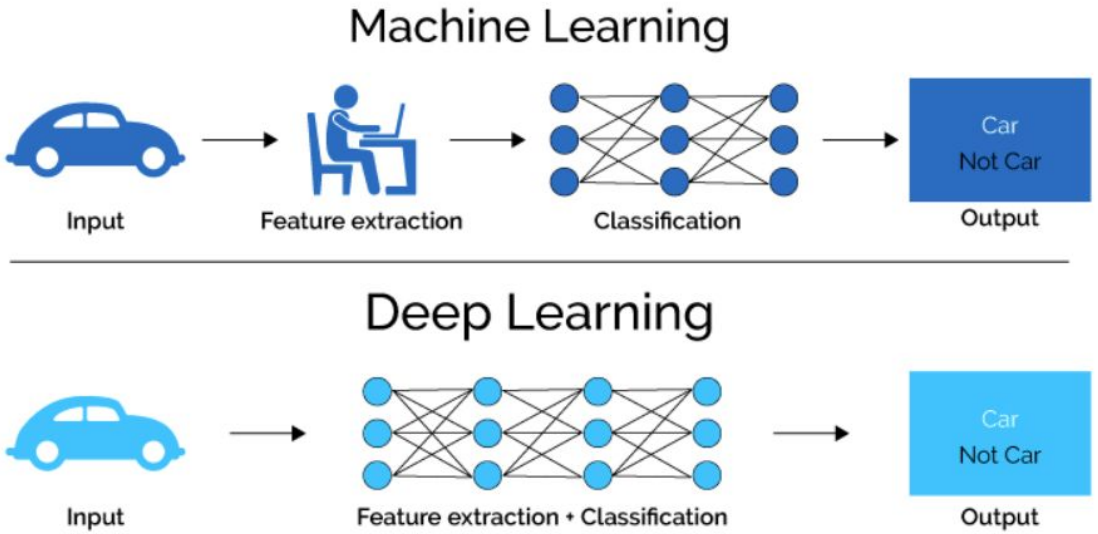
Алгоритм работает с числовыми векторами

Как компьютер поймет химическую формулу? А форму наночастицы?

Нужно использовать уникальное свойство:

- Элементный состав – электроотрицательности, число валентных электронов, порядковый номер таблицы Менделеева, магнитный момент, спин …
- Форма определяет то, каким образом из трех измерений частицы можно получить её площадь и объем?

Не забываем про физический смысл – у нас же НаУкА
И зачастую от качества фич зависит качество предсказаний моделей МО

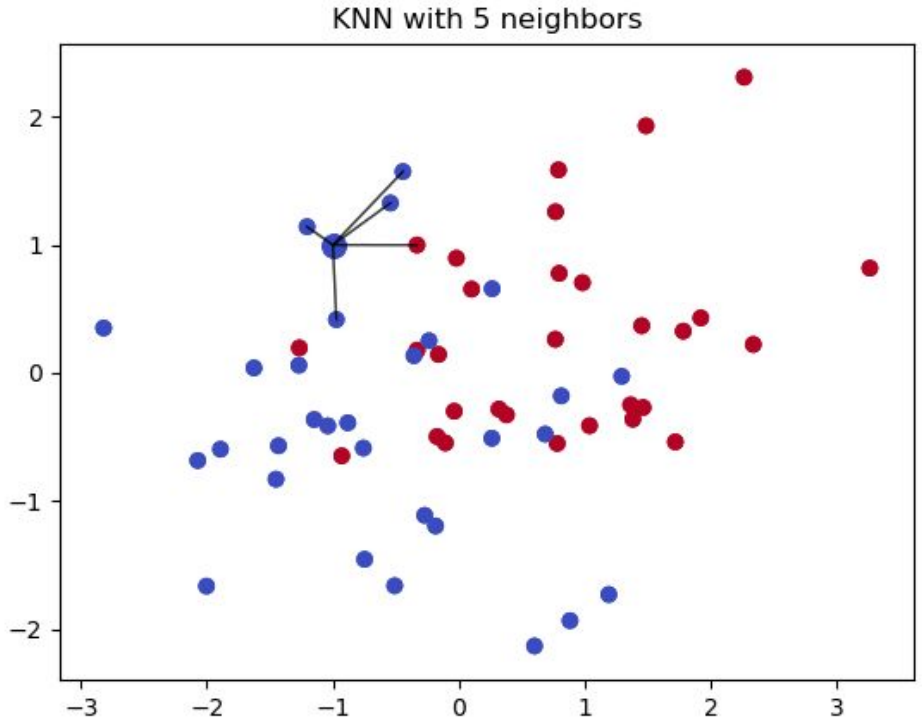


Missing data handling

Удаление строк (а тем более столбцов) с пропущенными значениями – непозволительная роскошь для нас, так как данных мало. Но иногда приходится делать ☹

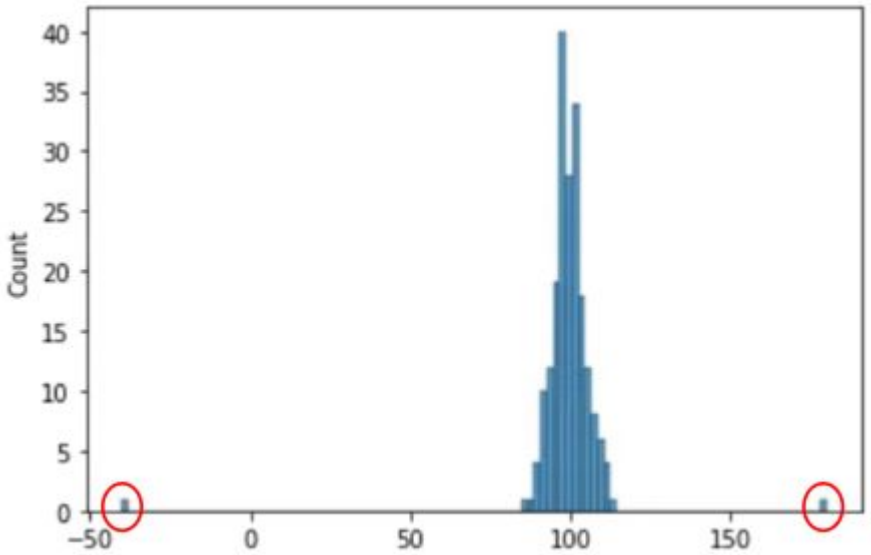
Нам остается заниматься заполнением пропущенных значений
Есть несколько стратегий (для числовых данных): использование среднего, медианы, моды (не очень, так как не учитывает возможные взаимосвязи между параметрами, если пропуски неслучайны). Также нам могут помочь модели МО (алгоритм k nearest neighbors (kNN) является одним из самых популярных и простых в использовании)

Сколько соседей?
Какая метрика?



Удаление выбросов

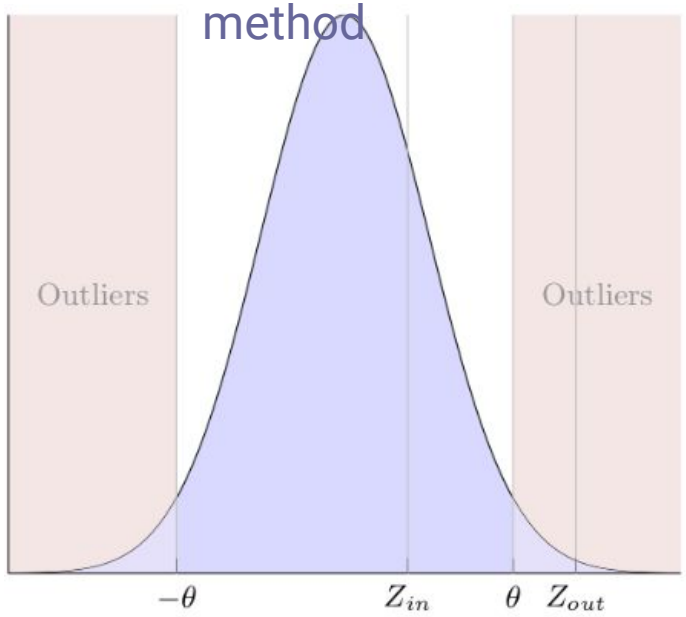
Визуально



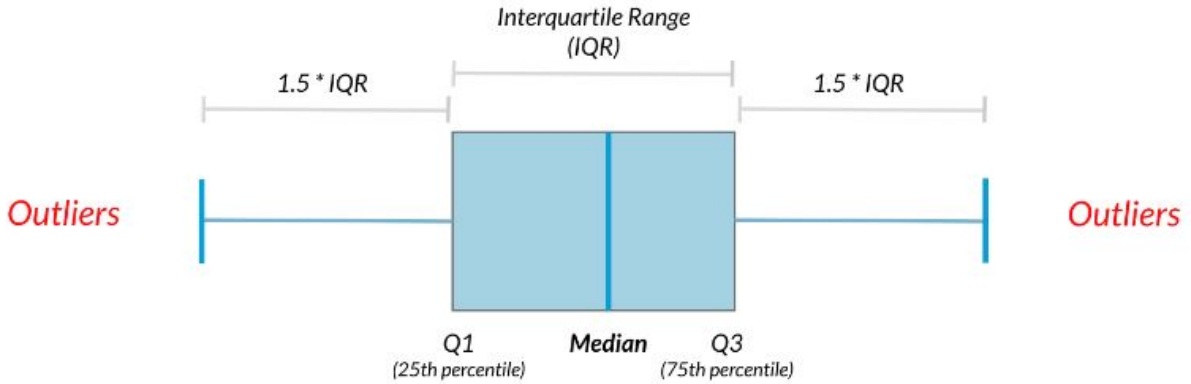
$$z\_score = \frac{y - y_{mean}}{y_{std}}$$

z имеет нормальное распределение

z-score method



Использование квартилей

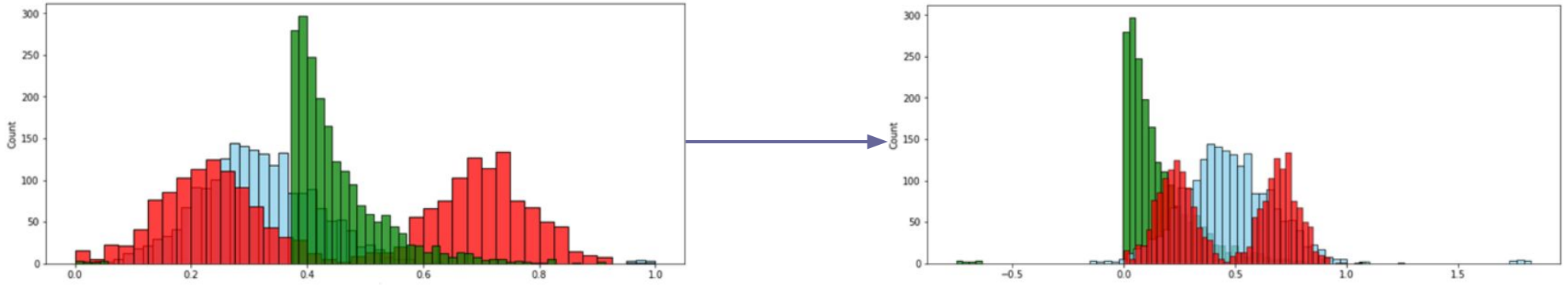


Нормализация данных

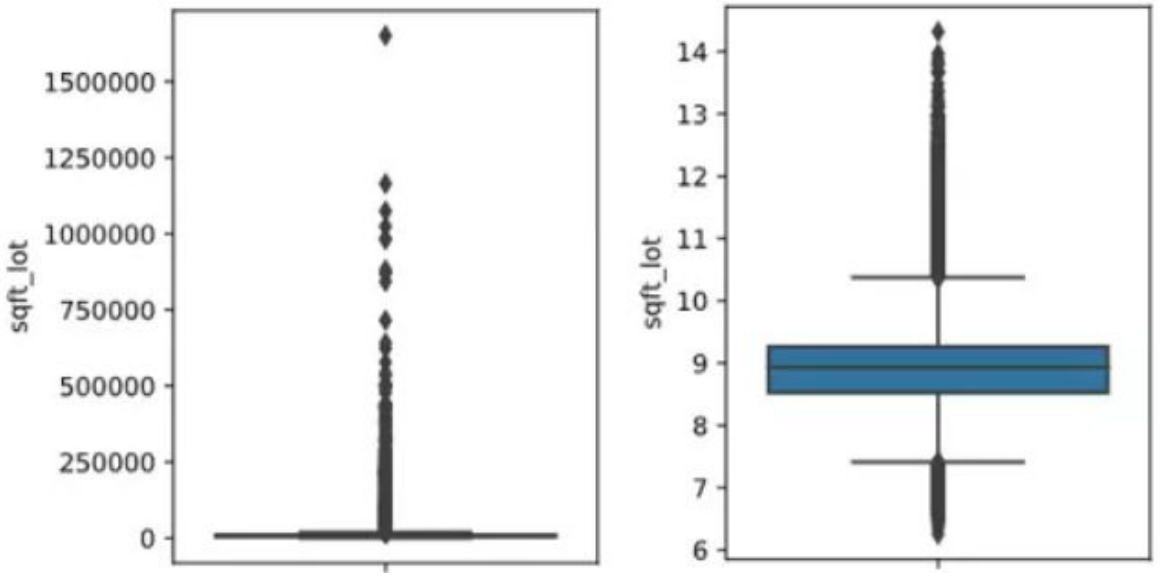
Сохраняем распределение

MinMaxScaler

$$x_{i,norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

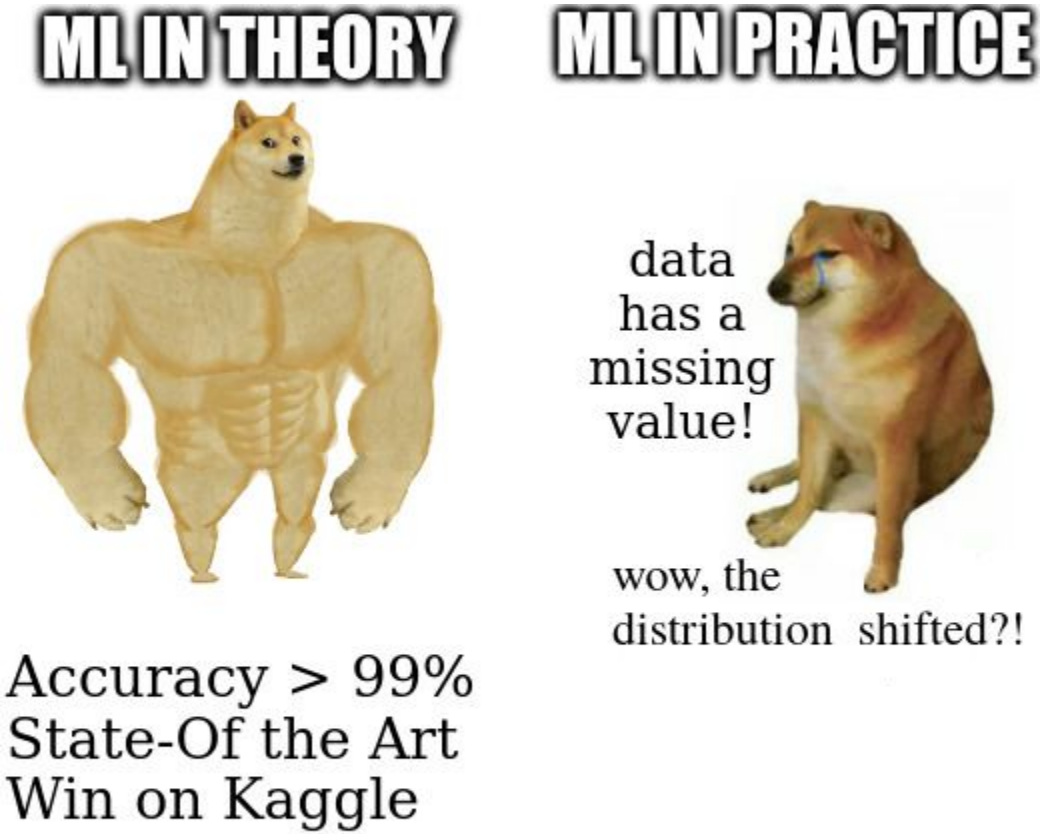


Логарифмирование



Позволяет сгладить датасет, особенно если данные различаются на несколько порядков

Practice


ML IN THEORY
Accuracy > 99%
State-Of the Art
Win on Kaggle
ML IN PRACTICE
data
has a
missing
value!
wow, the
distribution shifted?!