Датасет

Dataset

Выборка

- Объем выборки - Количество данных
  - Данные
  - Информация
    - Информативные данные
- Репрезентативность - Качество данных
  - Представительная выборка
    - Representative sample
  - Сбалансированная выборка
    - Balanced sample

Признаки

- Статистические признаки
- X Y
- Y = f(X)
- Независимая и зависимая переменная
- Факторный и результативный признак
- Вход и выход машинной модели
- Independent / dependent / target
- Features

ЭВМ

- Электронная вычислительная машина
- Вычислительная техника
- Компьютер
- Машина
- Человеко-машинный интерфейс
- HMI
- Human machine interface

Табличные данные

- Строки – объекты (статистические единицы)
- Столбцы – признаки
- Факторные и результативные (целевые)

Метаданные

- Данные о данных
- README
  - Общее описание датасета
  - Что – где – когда – как
- Описание признаков / переменных
  - Рост – Height
    - Рост человека в сантиметрах
    - Тип: целое

Формат

- CSV
  - Comma Separated Values
  - Точка – десятичный разделитель
  - Запятая – разделитель полей
- Заголовки столбцов
  - Названия переменных
  - Height,Weight
  - 175,69
- Ограничения

Формат

- Электронные таблицы
- Excel
  - XLSX
  - XML

- Демонстрация:
  - XLSX - 7Zip – XML - Sheet

Размещение датасета

- Открытые наборы данных
- Платформы соревнований по МО
  - Kaggle
- Библиотеки МО
  - SKLearn - SciKit Learn

Iris

- Датасет Ирис
  - iris dataset

- from sklearn import datasets
- iris = dataset(…
- Print(iris)
- Target_names
- DESCR

Модели данных

- Логическая модель
- Русские названия

- Физическая модель
- Английские названия+подчеркивания