**ДОМАШНЄ ЗАВДАННЯ 1**

**Мета**: навчитися передоброблювати «брудні» дані

**Кроки виконання завдання:**

1. Знайти «брудний» датасет.
2. Використовуючи бібліотеки pandas або numpy, завантажити дані в jupyter notebook.
3. Використайте базові функції pandas, щоб охарактеризувати дані (наприклад, дізнатися кількість відсутніх значень, дослідити типи даних в колонках).
4. Без використання сторонніх бібліотек (використання базових функцій numpy дозволяється) написати функції обробки даних, що відсутні.
5. Реалізувати наступні функції для передоброки даних у вигляді окремого python модулю:

* Функція для видалення колонок або строчок з відсутньою інформацією;
* Функція для заміни відсутніх даних на середнє значення, медіанне значення та моду;
* Функція для заміни відсутніх даних числового типу на значення, встановлене за допомогою лінійної регресії (можна використовувати реалізацію з бібліотеки sklearn - linear\_model.LinearRegression);
* Функція для заміни відсутніх даних на значення, встановлене за допомогою методу KNN (створити власну реалізацію методу К найближчих сусідів);
* Функція для стандартизації даних ();
* Функція для масштабування значень ().

1. Імпортувати реалізовані функції у раніше створений jupyter notebook.
2. Обрати декілька реалізованих раніше функцій та обробити датасет за їх допомогою.
3. Розбити дані на підвибірку для тренування та для тестування у пропорції 80% на 20% (використайте функцію model\_selection.train\_test\_split з бібліотеки sklearn) – для задачі навчання з учителем
4. Натренувати модель машинного навчання ваш вибір та підрахувати метрику якості.
5. Спробувати інші способи передобробки даних на кроці 7 та подивитися, як зміниться результат на кроці 9.

Виконане завдання - jupyter notebook та модуль з функціями обробки даних закомітити на ваш github. Відмітитись в наступному файлі – записати ім'я/прізвище та посилання на виконане завдання github: [гугл док](https://docs.google.com/spreadsheets/d/1Uw7prdKOwOd1qMdjHHUM_cWVzwZNjfxzDY-FP4_IDaY/edit?usp=sharing). В описі до репозиторію залишити посилання на дані, які ви використали в домашній роботі.

Дедлайн здачі роботи – 15 січня 20:00.

**Зверніть увагу**: робота оцінюється!

**Корисні посилання**

Де шукати дані:

* <https://github.com/awesomedata/awesome-public-datasets>
* <https://www.kaggle.com/datasets>
* <https://toolbox.google.com/datasetsearch>