**ДОМАШНЄ ЗАВДАННЯ 3**

**Мета**: навчитися кластеризувати дані, зменшувати розмірність даних, налаштовувати гіперпараметри моделей.

**Кроки виконання завдання:**

**Підготовка даних:**

1. Отримати датасет для даної задачі – [посилання на датасет](https://raw.githubusercontent.com/yehorchankov/udata-school/master/hw_3_data.csv).
2. Використовуючи бібліотеки pandas або numpy, завантажити дані в jupyter notebook.
3. Дослідити дані. Проаналізуйте типи даних, їх можливу відсутність, збалансованість, наявність викидів або колонок, що сильно корелюють. Візуалізуйте їх за допомогою пакетів matplotlib, seaborn або будь-якого іншого, за потреби зменште розмірність даних.
4. Якщо виявилося, що дані мають викиди, видаліть такі точки. Якщо деякі дані відсутні – використайте відомі вам підходи, щоб виправити це.
5. Якщо дані незбалансовані, використайте або один з підходів до ребалансування даних, або оберіть відповідну для таких випадків метрику.
6. Оберіть метрику, яка найбільше підходить для вашої задачі та даних.

**Задача класифікації:**

1. Оберіть дві моделі для задачі класифікації, одна з яких має бути SVM, і порівняйте їх метрики якості, використовуючи параметрами за замовчуванням.
2. Налаштуйте гіперпараметри обох моделей так, щоб отримати найкращий результат метрики якості. Для налаштування параметрів використовуйте один з підходів: баєсівська оптимізація, оптимізація на основі градієнтів або оптимізація на основі еволюційних алгоритмів. Для SVM одним з параметрів має бути тип ядра. Виведіть значення метрики.

**Задача кластеризації:**

1. Обрати два алгоритми кластеризації даних, серед яких обов’язково має бути алгоритм на основі центроїд і/або DBSCAN.
2. Налаштувати гіперпараметри алгоритмів, якщо це потрібно. Кількість кластерів *k* вважати за *n \* 2,* де *n* – число класів в датасеті.
3. Провести кластеризацію даних та візуалізувати кластери.
4. Обрати одну внутрішню та одну зовнішню метрики оцінки якості моделі. Проаналізувати результати кластеризації – отримати значення метрик і візуалізувати (наприклад, за допомогою silhouette plot) по можливості.

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) – сторінка HW\_3.

Дані представляють собою csv файл, де перша строчка це назва колонок. Колонки f0, f1, f2, f3, f4, f5, f6, f7, f8, f9 – це колонки значень, в колонці lbl – назва класу.

Дедлайн здачі роботи – 9 квітня до кінця дня.

**Зверніть увагу**: робота оцінюється! За невчасну здачу роботи зніматиметься 30 балів за кожні 12 години запізнення.