

## Лабораторна робота #3

# Decision Tree. Random Forest

### Мета роботи

Ознайомитись з основами машинного навчання та аналізу даних, реалізувати процедуру побудови дерева прийняття рішень (*Decision Tree*) та ансамблю дерев прийняття рішень (*Random Forest*) із використанням бібліотеки *scikit-learn*.

### Рекомендована література

Російськомовна версія книги **Сабастьяна Рашки** “**Python и машинное обучение**” у pdf-форматі доступна для завантаження за посиланням:

<https://drive.google.com/file/d/1NOBI5mhAhlTxoxxl8aMWFQMU-MLuZ8ya/view?usp=sharing>

### Хід роботи

#### Дані

Рекомендоване джерело даних - публічні набори даних на <https://www.kaggle.com/datasets> (для тих, хто претендує на максимальний бал)

**Примітка:** на нижчу оцінку можна використовувати типові ідеальні набори даних (типу ірисів Фішера і т.п.)

#### Підготовчий етап

Провести аналіз вибраного набору даних, визначити вхідні та вихідні параметри, візуалізувати залежності входів на виходу, спробувати виявити основні залежності, детектувати аномалії, неповні зразки тощо у даних.

**Дорожня карта** для підготовки даних - стр. 34-35 рекомендованої книги.

Зокрема виконати кореляційний аналіз входів та виходів набору даних, виявити взаємозалежні фактори.

Провести підготовку даних до подальшого використання.

Обраний та відфільтрований набір даних розбити на навчальну та тестову частину (50% на навчання, 50% на тест) з рівними пропорціями представників всіх класів у навчальній та тестовій частині.

### Реалізація класифікатору

Реалізувати навчання дерева прийняття рішень (*Decision Tree*) та лісу дерев (*Random Forest*) відповідно до алгоритмів, описаних у рекомендованій книзі на стор. 93-103.

Приклад побудови дерева прийняття рішень (*Decision Tree*) для класичного набору даних (ірис Фішера) засобами бібліотеки *scikit-learn* на стор. 99 рекомендованої книги.

Приклад побудови ансамблю дерев прийняття рішень (*Random Forest*) для класичного набору даних (ірис Фішера) засобами бібліотеки *scikit-learn* на стор. 102 рекомендованої книги.

**!** Для отримання максимального балу навчання дерева має бути імплементовано самостійно, з обчисленням ентропії/gini impurity, визначення оптимального поділу на кожному етапі навчання

### Аналіз результатів

1. Результати кореляційного аналізу ознак та класів, ідентифікація взаємопов'язаних факторів
2. Результати пошуку проблем у датасеті (викиди, неповні дані тощо) та способи боротьби з проблемами, що були використані
3. Оцінка помилок на початковій та тестовій вибірках:
  - а. Для дерева прийняття рішень (*Decision Tree*)
  - б. Для ансамблю дерев прийняття рішень (*Random Forest*)
4. Результати оформити протоколом.

### Контрольні питання

1. Поняття дерева прийняття рішень (*Decision Tree*)
2. Поняття ансамблю дерев прийняття рішень (*Random Forest*)
3. Поняття чистих даних
4. Проблеми у реальних датасетах та способи боротьби з ними