**Регресія**

**Мета роботи**

Ознайомитись з основами машинного навчання та аналізу даних для розв’язання задачі

регресії, реалізувати методи, що базуються на алгоритмі XGBoost. Застосувати регуляризації для (Lasso / Elastic Net)

**Рекомендована література**

**Про роботу з лінійними регресіями в Python**:

https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html

https://scikit-learn.org/stable/modules/linear\_model.html

https://realpython.com/linear-regression-in-python/

**XGBoost**:

https://drive.google.com/file/d/1kuyu5L\_1Hwb7QQFsyIpPPCTEwaO9Vj4X/view?usp=sharing

**PyTorch/Tensorflow/Keras/SciKit Learn**:

<https://drive.google.com/file/d/1faM-aL5Bsea5DcXpFVr6ApOL5qWNWved/view?usp=sharing>

<https://drive.google.com/file/d/1tAD_-4hZ1nqe5Fqfyic92hmyrQlZnzPe/view?usp=sharing>

<https://drive.google.com/file/d/1D5m901531N9-v-UZQC77YhtRCTdbhAG4/view?usp=sharing>

<https://drive.google.com/file/d/10-cRqjUrSp_qOquyjKLuzqO8KPA9yA3r/view?usp=sharing>

**Ансамблінг у SciKit learn:**

https://scikit-learn.org/stable/modules/ensemble.html#voting-regressor

**Хід роботи**

**Дані**

В роботі використовується публічний набор даних з https://www.kaggle.com/datasets (для тих, хто претендує на максимальний бал).

**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. 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 (70% на навчанання, 30% на тест).

**Реалізація моделі**

Написати код у Python / R, який реалізований з використанням двох підходів на вибір:

1. SciKit learn
2. XGBoost
3. Vanilla Python (Numpy/Pandas)

**Аналіз результатів**

1. Вибір оптимальних параметрів регресій, їх обгрунтування

2. Оцінка помилок на начальній та тестовій вибірках

3. Порівняння результатів різних підходів (співпали чи ні, причини чому могли не співпасти, runtime тощо)

4. Результати оформити протоколом

**Студенти, що розраховують на високий бал мають:**1. Відкоментувати код (що кожна строка робить, окрім import / library)

1. Якісно візуалізувати результати та проміжні етапи (за необхідністю)
2. Обрати порівняно складний датасет, який потребує додаткового очищення
3. Ансамблювання за прикладом у посиланні з розділу “Рекомендована література” - додатковий бал.

**Контрольні питання**

1. Що таке ансамблювання та для чого воно використовується?

2. Які моделі базуються на принципі лінійної регресії?

3. Що таке XGBoost