# Практична сесія Алгоритми дерева рішень і k-найближчих сусідів на реальних датасетах

**Мета:** засвоїти основні відомості про роботу з алгоритмом дерева прийняття рішень та робота з різними наборами даних.

**Частина 1. Підготовка даних**

* 1. Завантажити дані згідно варіанту.
  2. Розділити на тренувальну, валідаційну та тестові вибірки в пропорції 70/15/15
  3. Обробити NaN значення – або видаленням, або заповненням середнім, або якимось іншим чином. Обраний варіант обґрунтувати.
  4. Нормалізувати дані – для тренувальних даних обрахувати значення mean(середнього) та standard deviation(стандартного відхилення) для кожної з ознак і перетворити дані віднявши обраховане середнє та поділивши на стандартне відхилення. Зауважте, що на відміну від Z-score тут немає модуля.
  5. Нормалізувати валідаційні та тестові даних на основі обрахованих mean та std для тренувальних даних

**Частина 2. Побудова дерева рішень**

* 1. Побудувати 3 різні дерева рішень(Decision tree) з різною глибиною на основі даних згідно варіанту.
  2. Візуалізувати отримані дерева рішень. Порівняти точність отриманих дерев рішень на валідаційній та тренувальній вибірках.

**Частина 3. Побудова k-nn**

* 1. Побудувати 3 різні k-nn з різним k, типом алгоритму на основі даних згідно варіанту.
  2. Візуалізувати отримані результати

**Частина 4. Кінцева валідація**

* 1. Обрахувати метрики, які відповідають задачі, порівняти їх та запропонувати варіанти покращення

**🖇Приклад створення тренування Дерева прийняття рішень** <https://jakevdp.github.io/PythonDataScienceHandbook/05.08-random-forests.html>

Більшість датасетів з <https://seaborn.pydata.org/generated/seaborn.load_dataset.html>

| **Варіант #** | **Датасет** | **Назва колонки яку потрібно передбачити** |
| --- | --- | --- |
| 1 | [diabetes](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset) | ‘Diabetes\_012’ |
| 2 | [bankruptcy](https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction) | ‘Bankrupt’ |
| 3 | planets | 'mass', |
| 4 | tips | 'tip' |
| 5 | diamonds | 'clarity', |
| 6 | exercise | ‘pulse’ |
| 7 | titanic | 'survived', |
| 8 | iris | 'species' |

# 🌟Датасети з зірочкою

| **Dataset Name** | **Task Type** | **Data Format** | **Size** | **Difficulty** | **Primary Link (Kaggle/HF)** | **Description** | **Learning Objectives** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| California Housing Dataset | Regression | Tabular + Coordinates | 20640 samples | ⭐☆☆ | https://www.kaggle.com/datasets/camnugent/california-housing-prices | House prices with longitude/latitude coordinates | Spatial regression; coordinate mapping; basic GIS concepts |
| NYC Taxi Trip Dataset | Regression | Tabular + GPS | 100K-1M trips | ⭐⭐☆ | https://www.kaggle.com/datasets/neilclack/nyc-taxi-trip-data-google-public-data | Taxi trip data with pickup/dropoff coordinates | Distance calculation; spatial relationships; trip prediction |
| Weather Station Data | Regression | Tabular + Coordinates | 1K-10K stations | ⭐⭐☆ | https://www.kaggle.com/datasets/noaa/gsod | Global weather station data with GPS coordinates | Spatial interpolation; climate modeling |
| Real Estate Dataset | Regression | Tabular + Coordinates | 10K-50K properties | ⭐⭐☆ | https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset | Property prices with location coordinates | Neighborhood effects; spatial feature engineering |
| Twitter Geospatial | Classification/NLP | Text + GPS | Variable | ⭐⭐⭐ | https://www.kaggle.com/datasets/yashdogra/ibhf-jhjbjh | Geo-tagged tweets for spatial analysis | Social media analytics; spatial-temporal patterns |
| Crime Spatial Dataset | Classification | Tabular + Coordinates | 1K-50K incidents | ⭐⭐☆ | https://www.kaggle.com/datasets/paultimothymooney/denver-crime-data | Crime incidents with location and temporal data | Spatial clustering; hotspot analysis; crime prediction |
| Air Quality Dataset | Regression | Time Series + Coordinates | 100-1K stations | ⭐⭐⭐ | https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates | Air pollution monitoring with station coordinates | Spatial interpolation; environmental modeling; time series |
| Crop Yield Dataset | Regression | Tabular + Coordinates | County-level data | ⭐⭐⭐ | https://huggingface.co/datasets/CropNet/CropNet | Agricultural yield data with county coordinates | Agricultural prediction; yield forecasting; rural analysis |

# Питання для самоконтролю

* Що таке дерева прийняття рішень?
* Яким чином відбувається розбиття на піддерева/листки в деревах прийняття рішень?
* Що таке Ентропія та Information gain? Яку роль вони грають в деревах прийняття рішень?
* Що таке k-nn?
* Поясніть чому на вашу думку ви отримали кращі/гірші результати на різних алгоритмах