Suggested datasets

The list of datasets and tasks are intended as suggestions. Students can personalize their projects based on their interests, considering new datasets or different tasks (just be reasonable). You can find a list of repositories to take inspiration from in case you want to explore more datasets.

Basic

2. Breast Cancer Wisconsin (Diagnostic)

Domain: Bioinformatics/Medical

- **Details**: 569 samples with 30 cell nucleus features (e.g., radius, texture) from FNA biopsies.
- **Learning Task**: Binary classification (malignant vs. benign).
- **Challenges**:
    - Feature engineering with correlated metrics (mean, SE, extreme values).
    - Handling class imbalance (357 benign vs. 212 malignant).
- **Difficulty**: Entry-level (structured tabular data).
- **Link**: UCI WDBC Dataset

5. Wine Quality Dataset

Domain: Food Science

- **Details**: 1,599 red wine samples with 11 physicochemical features (e.g., pH, alcohol).
- **Learning Task**: Regression to predict quality scores (0–10).
- **Challenges**:
    - Highly skewed quality ratings (most scores 5–6).
    - Feature correlation analysis (e.g., sulfur dioxide vs. density).
- **Difficulty**: Entry-level (tabular regression).
- **Link**: UCI Wine Quality

7. Heart Disease (4 Databases)

Domain: Healthcare

- **Details**: 303 patient records across 4 hospitals (Cleveland, Hungary, Switzerland, VA Long Beach) with 13 attributes (these are pre-filtered from a total of 76 potential attributes).
- **Learning Task**: Binary classification (heart disease presence).
- **Challenges**:
    - Missing values and categorical feature encoding.
    - Small sample size per hospital.

- **Difficulty**: Entry-level (classification with missing data).
- **Link**: UCI Heart Disease

8. Yeast Dataset

Domain: Bioinformatics

- **Details**: 1,484 yeast protein samples with 8 features for localization site prediction.
- **Learning Task**: Multi-class classification (10 cellular localization sites).
- **Challenges**:
  - Class imbalance (e.g., MIT site dominates).
  - Feature relevance analysis (e.g., sequence homology).
- **Difficulty**: Intermediate (biological feature interpretation).
- **Link**: UCI Yeast Dataset

Advanced

4. Skin Cancer MNIST: HAM10000

Domain: Medical Imaging

- **Details**: 10,015 dermatoscopic images across 7 skin lesion classes (e.g., melanoma, benign keratosis).
- **Learning Task**: Multi-class classification with imbalanced data.
- **Challenges**:
  - Class imbalance (e.g., melanocytic nevi dominate).
  - Requires data augmentation (rotation, flipping).
- **Difficulty**: Intermediate (image preprocessing/CNN tuning).
- **Link**:
  HAM10000 Paper
  Kaggle dataset

1. Chest X-Ray Images for Classification

Domain: Medical Imaging

- **Details**: 377,110 chest X-ray images (frontal/lateral views) with 14 diagnostic labels (e.g., pneumonia, edema).
- **Learning Task**: Multi-label classification of pathologies from high-resolution images.
- **Challenges**:
  - Variable image resolutions (e.g., 2500×3056 pixels).
  - Requires preprocessing (CLAHE contrast enhancement, DICOM-to-JPG conversion).
- **Difficulty**: Intermediate to Advanced (handling large-scale imaging data).

- **Link**: MIMIC-CXR-JPG Dataset

3. Gas Sensor Array Drift Dataset

Domain: Environmental Sensing

- **Details**: Measurements 13910 measurements from 16 chemical sensors utilized in simulations for drift compensation over months, for 6 different gases.
- **Learning Task**: Gas discrimination (classification) and time-series analysis (regression).
- **Challenges**:
  - Temporal sensor drift affecting accuracy.
  - Multi-gas concentration variability.
- **Difficulty**: Intermediate to Advanced (time-dependent data preprocessing).
- **Link**: Gas Sensor Drift Dataset

6. Fluorescent Neuronal Cells v2

Domain: Fluorescent Microscopy

- **Details**: Hundreds of high-resolution images of neuronal cells
- **Learning Task**: Cell counting, detection or segmentation
- **Challenges**:
  - Class-imbalance, artifacts, noisy labels
  - Requires domain-specific curation
- **Difficulty**: Advanced (image analysis with several challenges)
- **Link**: AMS Acta Repository

More Open Data Archives

- UCI Machine Learning: https://archive.ics.uci.edu/
- OpenML: https://www.openml.org/search?type=data&sort=runs&status=active
- Kaggle: https://www.kaggle.com/datasets
- Awesome-public-datasets: https://github.com/awesomedata/awesome-public-datasets

Extra (physics-related)

1. JetClass

Domain: High-Energy Physics

- **Details**: 125M jets simulated with MadGraph + Pythia + Delphes, containing 10 classes of different jet types
- **Learning Task**: Multi-class classification (10 jet classes).

- **Challenges**:
  - Class imbalance (e.g., MIT site dominates).
  - Feature relevance analysis (e.g., sequence homology).
- **Difficulty**: Advanced (suitable for research projects)
- **Link**: https://zenodo.org/records/6619768

2. JetNet

Domain: High-Energy Physics

- **Details**: Particle cloud dataset containing gluon, top quark, and light quark jets saved in CSV format. Each jet is represented by its constituent particles with properties such as momentum and energy
- **Learning Task**: Classification task for jet tagging (identifying the origin of particle jets)
- **Challenges**:
  - Dealing with variable-length sets of particles per jet
  - Handling complex physical relationships between particles
  - Extracting meaningful features from particle-level information
- **Difficulty**: Advanced (requires understanding of particle physics concepts; suitable for research projects)
- **Links**:
  https://zenodo.org/records/6975118
  https://joss.theoj.org/papers/10.21105/joss.05789

3. LHC Olympics

Domain: High-Energy Physics

- **Details**: Multiple "black box" datasets containing 1M events each, formatted as particle clouds. Events are stored as pandas dataframes in compressed h5 format with each event containing up to 700 particles represented by their detector coordinates (pT, eta, phi). The array format is (Nevents=1M, 2100)
- **Learning Task**: Anomaly detection - identifying potential signal events within vast amounts of background data that might indicate new physics
- **Challenges**:
  - Extremely rare signals within large background data
  - Mismatch between simulation and "data" in black boxes
  - Unknown signal characteristics (model-agnostic search)
  - Dealing with high-dimensional, sparse particle data
- **Difficulty**: Advanced (designed as a research challenge for physics community; suitable for research projects)

- **Link**:
  https://zenodo.org/records/4536624
  https://lhco2020.github.io/homepage/