Dataset Description

1. Gene expression cancer RNA-Seq

- This dataset is a subset of the RNA-Seq (HiSeq) PANCAN dataset, consisting of randomly extracted gene expression data from patients with different types of tumors: BRCA (Breast Cancer), KIRC (Kidney Renal Clear Cell Carcinoma), COAD (Colon Adenocarcinoma), LUAD (Lung Adenocarcinoma), and PRAD (Prostate Adenocarcinoma). It was donated on June 8, 2016.

Dataset Characteristics

- **Subject Area (Scope):** Biology
- **Feature Type:** Real-valued
- **Number of Instances:** 801
- **Number of Features:** 20,531
- **Missing values:** No
- **Associated tasks:** Classification, Clustering

Dataset Information

- Each row in the dataset represents a sample (patient), while each column corresponds to a gene expression level measured using the Illumina HiSeq sequencing platform. The attributes are labeled with dummy names (e.g., *gene_XX*), maintaining the order of the original submission. For the complete list of probe names, refer to the original dataset submission at Synapse.

2. Isolet

- The ISOLET dataset is designed for spoken letter recognition, where the goal is to predict which letter-name was spoken. It was donated on September 11, 1994 and is commonly used for classification tasks in speech recognition and machine learning.

Dataset Characteristics

- **Subject Area:** Computer Science
- **Feature Type:** Real-valued
- **Number of Instances:** 7,797
- **Number of Features:** 617
- **Missing Values:** No
- **Associated Tasks:** Classification

Dataset Information

- The dataset was collected from 150 speakers, each pronouncing the name of every letter in the alphabet twice, resulting in 52 samples per speaker. The speakers are divided into five groups: isolet1, isolet2, isolet3, isolet4, and isolet5.
    - The first four groups (isolet1-4) form the training set.
    - The fifth group (isolet5) is used as the test set, stored separately.
    - Three instances are missing due to recording issues.

- This dataset is valuable for studying noisy, perceptual classification tasks and evaluating the scalability of machine learning algorithms. Notably, decision tree-based methods like C4.5 have been observed to run slower than backpropagation-based models on this dataset.
- For convenience, the dataset has been formatted for C4.5, including a corresponding C4.5-style names file.

3. lfw people

- The Labeled Faces in the Wild (LFW) dataset is designed for face recognition and classification tasks. It consists of images of faces collected from the web, labeled with the identities of the individuals. The dataset is widely used in machine learning for evaluating face recognition models.

Dataset Characteristics

- **Subject Area:** Computer Science
- **Feature Type:** Real (pixel values ranging from 0 to 255)
- **Number of Instances (Total Samples):** 13,233
- **Number of Classes (Unique Individuals):** 5,749
- **Dimensionality:** 5,828
- **Associated Tasks:** Classification

Dataset Information

- Each instance in the dataset corresponds to a grayscale face image, preprocessed and resized to 62×47 pixels, with the option to extract a subset of the image using a slicing parameter. The dataset is available in a funneled variant for improved alignment and can be resized using the `resize` parameter.
    - The images are stored as a NumPy array of shape (13,233, 62, 47), where each row represents a face.
    - The dataset includes a target array that contains numerical labels for each image, ranging from 0 to 5,748, corresponding to different individuals.
    - The target_names array provides the actual names of the individuals.
    - The dataset is structured as a dictionary-like Bunch object in Scikit-learn, with attributes such as `data`, `images`, `target`, and `target_names`.
- This dataset is particularly useful for evaluating face recognition algorithms using techniques like eigenfaces and Support Vector Machines (SVMs).

4. Musk version 2

- The Musk (Version 2) dataset is designed for molecular classification, where the goal is to predict whether a given molecule is a musk or non-musk based on its conformations. It was donated on September 11, 1994, and is widely

used for studying classification tasks in computational chemistry and machine learning.

Dataset Characteristics

- **Subject Area:** Physics and Chemistry
- **Feature Type:** Integer
- **Number of Instances (Total Conformations):** 6,598
- **Number of Features:** 166
- **Missing Values:** No
- **Associated Tasks:** Classification

Dataset Information

- The dataset consists of 102 molecules, of which 39 are classified as musks and 63 as non-musks based on human expert judgment. However, because molecules can adopt multiple conformations due to bond rotation, the dataset contains a total of 6,598 conformations, each represented as a 166-dimensional feature vector.
- This dataset presents a "multiple instance learning" problem:
    - A molecule is classified as musk if at least one of its conformations is identified as musk.
    - A molecule is classified as non-musk if none of its conformations are musk.
- This characteristic makes the dataset particularly useful for developing and evaluating machine learning models that handle multiple-instance problems.

5. Olivetti faces
    - The Olivetti Faces dataset, originally from AT&T Laboratories Cambridge, is used for face recognition and classification tasks. It contains grayscale images of 40 individuals, each with multiple facial images under varying conditions. The dataset is widely used in machine learning and computer vision research.

Data Characteristics

- **Subject Area:** Computer Science
- **Feature Type:** Real (pixel values between 0 and 1)
- **Number of Instances (Total Samples):** 400
- **Number of Classes (Unique Individuals):** 40
- **Dimensionality:** 4,096 (64×64 pixels per image)
- **Missing Values:** No
- **Associated Tasks:** Classification

Dataset Information

- Each instance in the dataset is a 64×64 grayscale image of a face, which has been flattened into a 4,096-dimensional vector. The dataset includes:

    - A data matrix (`data`) of shape (400, 4096), where each row is a flattened face image.
    - An images array (`images`) of shape (400, 64, 64), where each row represents a face.
    - A target array (`target`) containing labels ranging from 0 to 39, corresponding to 40 different individuals.
- By default, the dataset is not shuffled, but an option is available to shuffle it for better generalization in model training. It is formatted as a Scikit-learn Bunch object, which can be accessed using `fetch_olivetti_faces()`.
- This dataset is commonly used for testing face recognition algorithms and dimensionality reduction techniques such as Principal Component Analysis (PCA) and Support Vector Machines (SVMs).

6. Taiwanese bankruptcy prediction
    - The Taiwanese Bankruptcy Prediction dataset contains financial indicators of companies collected from the Taiwan Economic Journal between 1999 and 2009. The dataset aims to predict company bankruptcy based on financial metrics, following the business regulations of the Taiwan Stock Exchange. It was donated on June 27, 2020 and is commonly used for financial risk assessment and corporate bankruptcy prediction.

Dataset Characteristics

- **Subject Area:** Business
- **Feature Type:** Integer
- **Number of Instances:** 6,819
- **Number of Features:** 95
- **Missing Values:** No
- **Associated Tasks:** Classification

Dataset Information

- Each instance represents a company's financial profile, with 95 features capturing key financial ratios and performance indicators. The first attribute is the class label, indicating whether a company went bankrupt or not.
- Some notable financial variables include:
    - Liquidity Ratios (e.g., Current Ratio, Acid Test, Quick Assets/Total Assets)
    - Profitability Metrics (e.g., Operating Income/Capital, Net Income to Stockholder's Equity)
    - Leverage & Debt Ratios (e.g., Total Liability/Equity Ratio, Interest-bearing Debt/Equity)
    - Growth Indicators (e.g., Total Asset Growth, Net Income Growth, Return on Total Assets Growth)
    - Stock Performance Metrics (e.g., Earnings Per Share (EPS), Book Value Per Share)

- The dataset is useful for corporate financial analysis, credit risk modeling, and bankruptcy forecasting. It enables machine learning models to identify patterns in financial stability and predict potential business failures.