How to get datasets for Machine Learning

The key to success in the field of machine learning or to become a great data scientist is to practice with different types of datasets. But discovering a suitable dataset for each kind of machine learning project is a difficult task. So, in this topic, we will provide the detail of the sources from where you can easily get the dataset according to your project.

Before knowing the sources of the machine learning dataset, let's discuss datasets.

What is a dataset?

**A dataset** is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table. Below table shows an example of the dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | **Age** | **Salary** | **Purchased** |
| India | 38 | 48000 | No |
| France | 43 | 45000 | Yes |
| Germany | 30 | 54000 | No |
| France | 48 | 65000 | No |
| Germany | 40 |  | Yes |
| India | 35 | 58000 | Yes |

A tabular dataset can be understood as a database table or matrix, where each column corresponds to a **particular variable,** and each row corresponds to the **fields of the dataset.** The most supported file type for a tabular dataset is **"Comma Separated File,"** or **CSV.** But to store a "tree-like data," we can use the JSON file more efficiently.

Types of data in datasets

* **Numerical data:**Such as house price, temperature, etc.
* **Categorical data:**Such as Yes/No, True/False, Blue/green, etc.
* **Ordinal data:**These data are similar to categorical data but can be measured on the basis of comparison.

Note: A real-world dataset is of huge size, which is difficult to manage and process at the initial level. Therefore, to practice machine learning algorithms, we can use any dummy dataset.

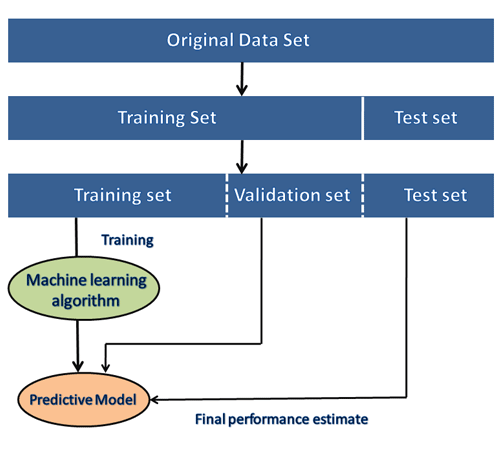
Need of Dataset

To work with machine learning projects, we need a huge amount of data, because, without the data, one cannot train ML/AI models. Collecting and preparing the dataset is one of the most crucial parts while creating an ML/AI project.

The technology applied behind any ML projects cannot work properly if the dataset is not well prepared and pre-processed.

During the development of the ML project, the developers completely rely on the datasets. In building ML applications, datasets are divided into two parts:

* **Training dataset:**
* **Test Dataset**

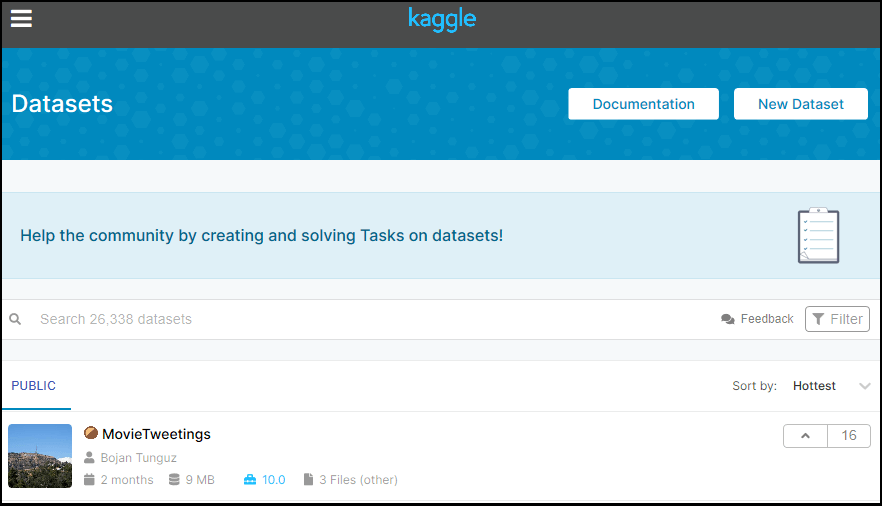


Note: The datasets are of large size, so to download these datasets, you must have fast internet on your computer.

Popular sources for Machine Learning datasets

Below is the list of datasets which are freely available for the public to work on it:

1. Kaggle Datasets

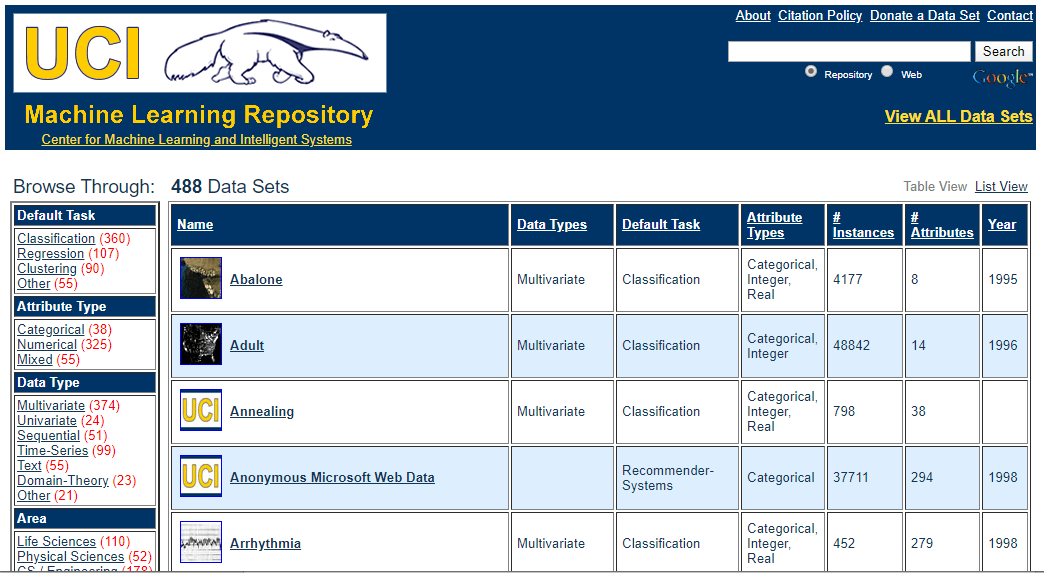


Kaggle is one of the best sources for providing datasets for Data Scientists and Machine Learners. It allows users to find, download, and publish datasets in an easy way. It also provides the opportunity to work with other machine learning engineers and solve difficult Data Science related tasks.

Kaggle provides a high-quality dataset in different formats that we can easily find and download.

The link for the Kaggle dataset is <https://www.kaggle.com/datasets>.

2. UCI Machine Learning Repository



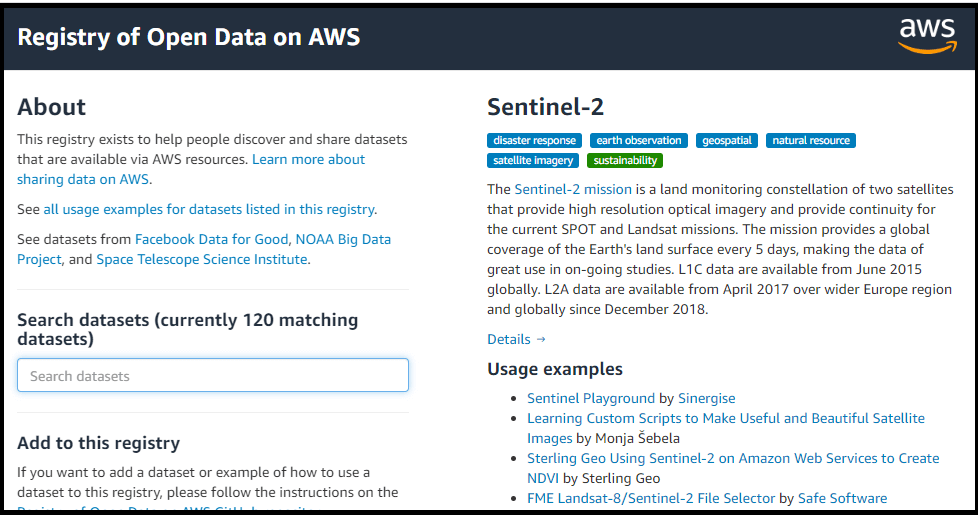
UCI Machine learning repository is one of the great sources of machine learning datasets. This repository contains databases, domain theories, and data generators that are widely used by the machine learning community for the analysis of ML algorithms.

Since the year 1987, it has been widely used by students, professors, researchers as a primary source of machine learning dataset.

It classifies the datasets as per the problems and tasks of machine learning such as **Regression, Classification, Clustering, etc.** It also contains some of the popular datasets such as the **Iris dataset, Car Evaluation dataset, Poker Hand dataset, etc.**

The link for the UCI machine learning repository is <https://archive.ics.uci.edu/ml/index.php>.

3. Datasets via AWS



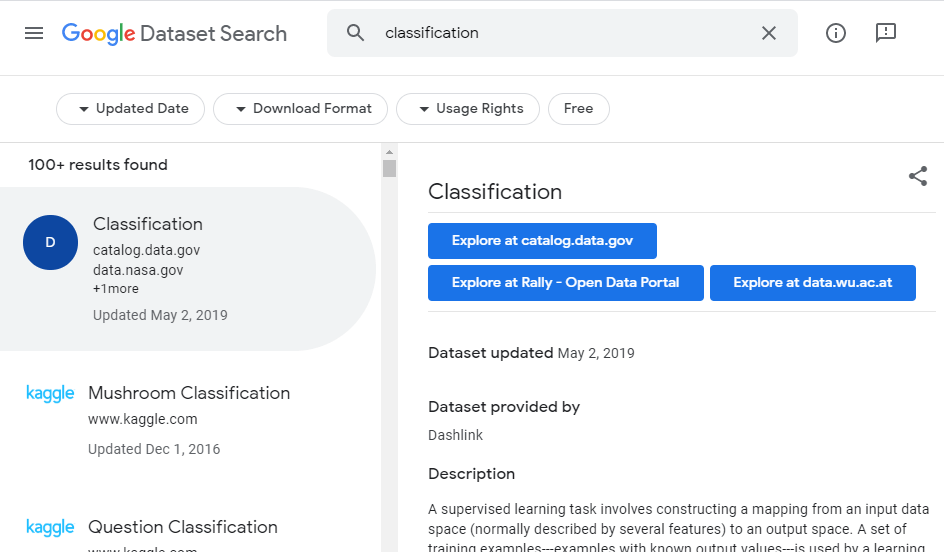
We can search, download, access, and share the datasets that are publicly available via AWS resources. These datasets can be accessed through AWS resources but provided and maintained by different government organizations, researches, businesses, or individuals.

Anyone can analyze and build various services using shared data via AWS resources. The shared dataset on cloud helps users to spend more time on data analysis rather than on acquisitions of data.

This source provides the various types of datasets with examples and ways to use the dataset. It also provides the search box using which we can search for the required dataset. Anyone can add any dataset or example to the **Registry of Open Data on AWS.**

The link for the resource is <https://registry.opendata.aws/>.

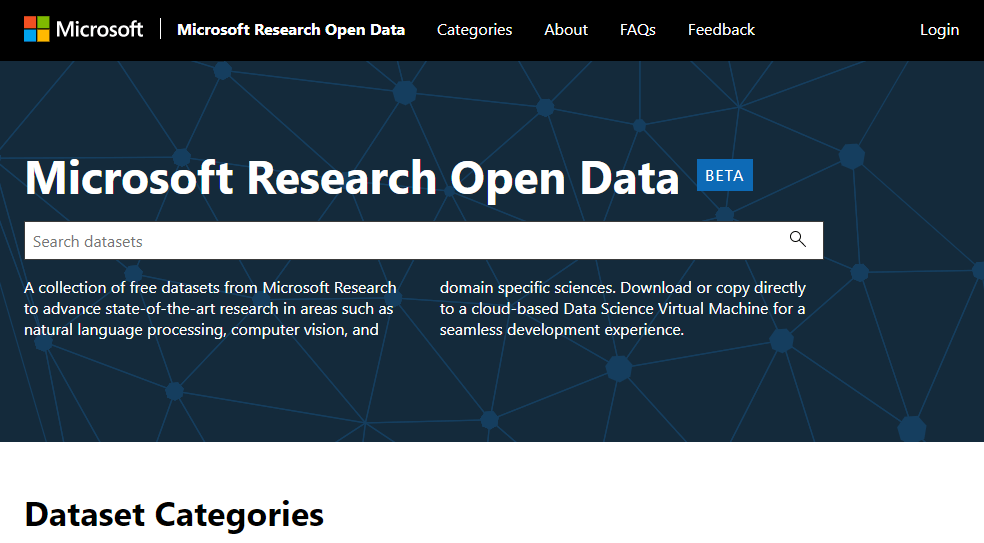
4. Google's Dataset Search Engine



**Google dataset search engine** is a search engine launched by **Google** on **September 5, 2018.** This source helps researchers to get online datasets that are freely available for use.

The link for the Google dataset search engine is <https://toolbox.google.com/datasetsearch>.

5. Microsoft Datasets

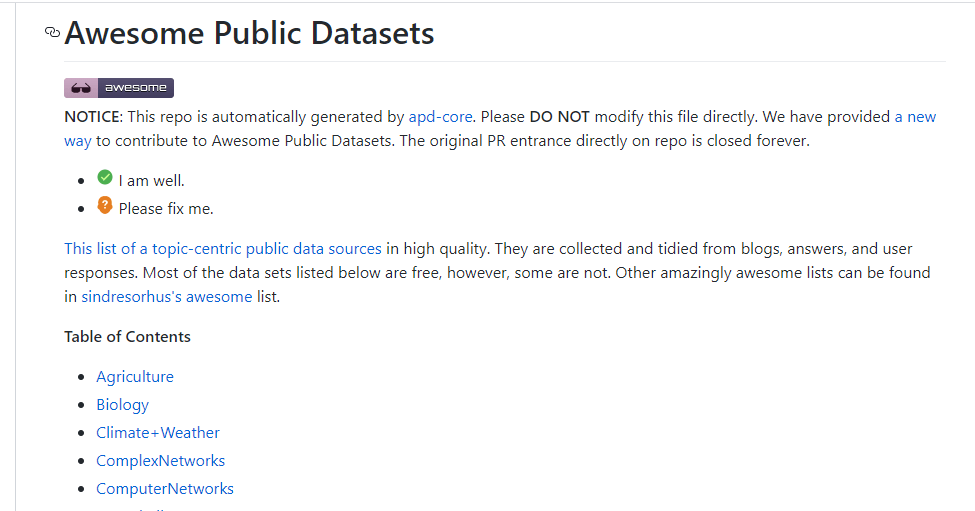


The Microsoft has launched the **"Microsoft Research Open data"** repository with the collection of free datasets in various areas such as **natural language processing, computer vision, and domain-specific sciences.**

Using this resource, we can download the datasets to use on the current device, or we can also directly use it on the cloud infrastructure.

The link to download or use the dataset from this resource is <https://msropendata.com/>.

6. Awesome Public Dataset Collection



Awesome public dataset collection provides high-quality datasets that are arranged in a well-organized manner within a list according to topics such as Agriculture, Biology, Climate, Complex networks, etc. Most of the datasets are available free, but some may not, so it is better to check the license before downloading the dataset.

The link to download the dataset from Awesome public dataset collection is <https://github.com/awesomedata/awesome-public-datasets>.

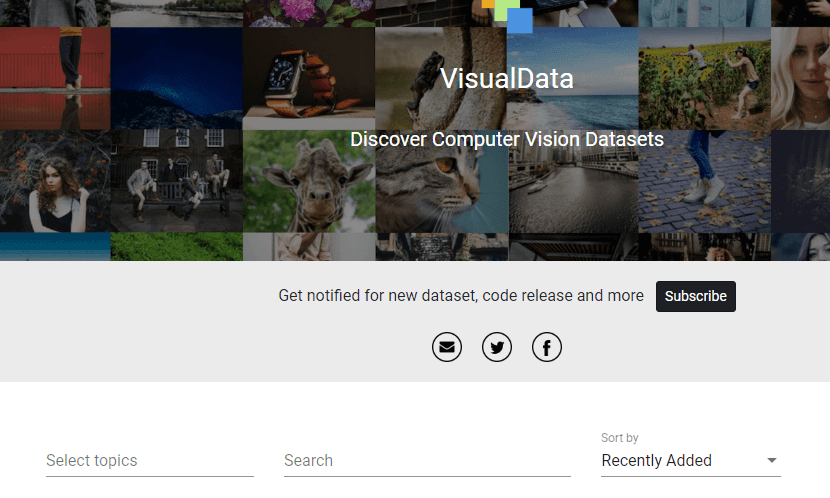
7. Government Datasets

There are different sources to get government-related data. Various countries publish government data for public use collected by them from different departments.

The goal of providing these datasets is to increase transparency of government work among the people and to use the data in an innovative approach. Below are some links of government datasets:

* [Indian Government dataset](https://data.gov.in/)
* [US Government Dataset](https://www.data.gov/)
* [Northern Ireland Public Sector Datasets](https://www.opendatani.gov.uk/)
* [European Union Open Data Portal](https://data.europa.eu/euodp/data/dataset)

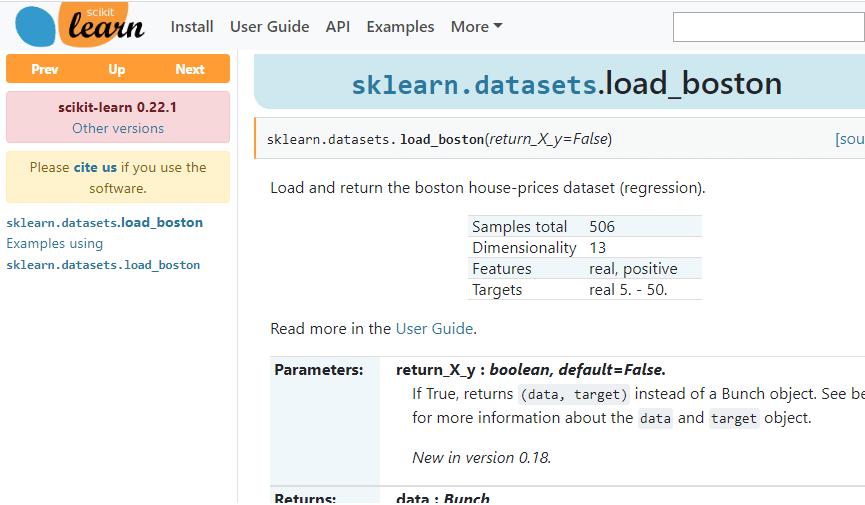
8. Computer Vision Datasets



Visual data provides multiple numbers of the great dataset that are specific to computer visions such as Image Classification, Video classification, Image Segmentation, etc. Therefore, if you want to build a project on deep learning or image processing, then you can refer to this source.

The link for downloading the dataset from this source is <https://www.visualdata.io/>.

9. Scikit-learn dataset



Scikit-learn is a great source for machine learning enthusiasts. This source provides both toy and real-world datasets. These datasets can be obtained from sklearn.datasets package and using general dataset API.

The toy dataset available on scikit-learn can be loaded using some predefined functions such as, **load\_boston([return\_X\_y]), load\_iris([return\_X\_y]),** etc, rather than importing any file from external sources. But these datasets are not suitable for real-world projects.

The link to download datasets from this source is <https://scikit-learn.org/stable/datasets/index.html>.