

Publicly Available Datasets for Machine Learning

UCI Machine Learning Repository

The UCI Machine Learning Repository is one of the most widely used sources for machine learning datasets. It hosts over 500 datasets across domains such as classification, regression, and clustering, including famous datasets like Iris, Adult Income, and Wine Quality.

Access: <https://archive.ics.uci.edu/ml/index.php>

Kaggle Datasets

Kaggle offers a vast collection of datasets uploaded by its global data science community. These datasets span a wide range of topics such as finance, healthcare, NLP, and computer vision, and are often used in competitions and public notebooks.

Access: <https://www.kaggle.com/datasets>

Google Dataset Search

Google Dataset Search is a specialized search engine that helps users locate publicly available datasets across the internet. It aggregates metadata from repositories and data providers to offer access to structured data across disciplines.

Access: <https://datasetsearch.research.google.com/>

AWS Open Data Registry

Amazon Web Services (AWS) Open Data Registry provides a curated list of publicly available datasets hosted on AWS infrastructure. These datasets are commonly used in research and industry, covering domains such as genomics, satellite imagery, and COVID-19.

Access: <https://registry.opendata.aws/>

OpenML

OpenML is an online platform that allows sharing, organizing, and analyzing machine learning datasets, tasks, and experiments. It supports reproducibility and benchmarking, and includes built-in APIs for integration with ML frameworks.

Access: <https://www.openml.org/>

CMU StatLib Dataset Archive

Maintained by Carnegie Mellon University, the StatLib archive offers a diverse set of statistical datasets, often used in academic and pedagogical contexts. These datasets are

particularly valuable for classical statistical learning problems.

Access: <http://lib.stat.cmu.edu/datasets/>

Microsoft Research Open Data

Microsoft Research Open Data provides free access to datasets from research across various Microsoft projects. Topics include computer vision, NLP, time series, and biomedical signals. It promotes reproducibility and academic collaboration.

Access: <https://msropendata.com/>

Data.gov

Data.gov is the U.S. government's open data portal, offering access to over 250,000 datasets from federal agencies. These datasets span topics like climate, agriculture, finance, education, and public health.

Access: <https://www.data.gov/>

The McGraw Hill logo is displayed in white text on a light red rectangular background. The text is arranged in three lines: "Mc" on the top line, "Graw" on the middle line, and "Hill" on the bottom line. The font is a clean, sans-serif typeface.