Lily Yan
LIS 545
Term Project - Repository Profile

UC Irvine Machine Learning Repository
https://www.re3data.org/repository/r3d100010960
https://archive.ics.uci.edu/

For the first part of our term project, I chose the *Water Quality* dataset from Kaggle owned by Aditya Kadiwal. This is a synthetically generated dataset that consists of water potability as a binary outcome variable, along with data on various water attributes such as water pH, hardness, total dissolved solids, chloramine, sulfate, and trihalomethane levels. Additionally, I have chosen the UC Irvine Machine Learning Repository for my dataset. As mentioned, this dataset is computationally generated rather than collected from real-world observations, suggesting that the purposes of this dataset is geared towards training and testing machine learning models rather than for contributing to domain-specific research or making direct decisions within the environmental sector. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators, serving the machine learning community and providing them with datasets across various domains.

The UCI Machine Learning Repository refers to data submissions as "donations," encouraging submissions from anyone across all subjects of knowledge. Although there is no defined scope, they note important criterias that have to be met before submission. The first criteria notes that those submitting data should have explicit permission to make the dataset available, meaning they either have been the original data collector, or have obtained consent from the original data collector. The second criteria calls for Personal Identifiable Information (PII) within datasets to be removed prior to submission to ensure that privacy of individuals are protected. The third and fourth points informs us that data that the repository accepts will be assigned a Digital Object Identifier (DOI), as well as licensed under a Creative Commons Attribution 4.0 International license (CC BY 4.0) which allows for data sharing, citation, and adaptation. The repository also allows linking externally for data on an external website as a form of submission, with the information stating that the data must be widely known and high-quality, and the external links must have a visible download button. This is ambiguous, as there are no other specifications as to how the repository determined what was well-known and high-quality.

There does not appear to be any limits to the data and domain types of content submitted. Upon exploring through the datasets in the repository, the repository covers a wide range of domains including but not limited to biology, physics, chemistry, finance, social science, games, and computer science. There are also various data types present such as multivariate, image, sequential, tabular, and other data types available. However, their filters also allow users to filter by *Task*, which contains keywords such as classification, regression, and clustering, suggesting that although there is no limit to domain or data types for the data accepted into the repository, there is a primary centering on the relevance of data to facilitate machine learning purposes.

The UCI Machine Learning Repository provides multiple access mechanisms for users to obtain datasets. Datasets are downloadable without having to login, and often download as a compressed file format such as .zip or .tar file. With that being said, logins are only needed for data submission, in which users can login with a newly created account, their Google account, or their Github account. When downloading datasets and extracting them from their compressed formats, the data file types range from .csv, .names, .data files. The distribution of datasets in compressed formats makes data easier to access, and facilitates organization to ensure files are not lost or corrupted during the downloading process. Furthermore, the repository also allows users to to import datasets straight into Python, allowing for the direct integration of datasets into users' Python-based applications, as well as facilitates data access especially with larger datasets.

It is worth noting that the information provided does not mention a specific structure or standard that attached metadata must be submitted in. As far as Dissemination Information Packages (DIP) go, this repository doesn't seem to create any in a formal sense. Datasets provided by the repository when downloaded, comes with metadata submitters choose to include. The case seems that users are also given the agency to organize and package the data with any relevant metadata for dissemination within their own use whether it be for research, analysis, or model training unless specifically disclosed in the original creator's DIP. Although no direct resources seem to be provided, the metadata profile provided for each dataset is structured in a way that suggests some guidance provided to users on what metadata should be included. For example, in line with the search and filtering functions of the repository, metadata on data characteristics, type, task, subject area, number of instances and number of features seem to be standard procedure for labeling datasets. There is also an area that displays additional information of the data, such as a summary, stating missing values, and purposes of the data. However, this information is inconsistent across different datasets, which probably means not every question is required to be answered. One key observation from exploring multiple datasets, was that most of them had an *Introductory Paper* attached. These papers were articles written by the original data collectors and creators, often published by open-source publishers and accessed through digital libraries, and research databases. These papers serve as comprehensive documentation, providing insight on the purpose, origins, and collection methodology of the dataset. However, not every single dataset provides this type of documentation so it seems that this information is not required during the submission process.

References

Markelle K., Rachel L., Kolby N. (n.d.).The UCI Machine Learning Repository. Accessed February 4th, 2024 from https://archive.ics.uci.edu