Lily Yan
LIS 545
Term Project - Data and Metadata Profile

https://www.kaggle.com/datasets/adityakadiwal/water-potability

Data

For this assignment I have chosen 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. The metadata provided on the origin of this data on Kaggle are included under *Provenance*. However, there isn't much origin information for this dataset, simply marking it as synthetically generated. Although the methodology for this is not specified, synthetically generated data suggest that this dataset was computationally generated rather than collected from real-world observations.

There are various key stakeholders for this dataset, including users, data scientists, researchers interested in the environmental sector, those involved in environmental advocacy, and government regulatory members. As this is a synthetically generated dataset, key stakeholders would be users who are interested in using this data to train a model geared toward water quality predictive analytics. Those interested in the environmental sector and policy will value this dataset for the purposes of research, analytics, use-case and scenario testing but not direct decision making as this data does not reflect the complexities of the real world.

This dataset comes with one .csv file, containing data on variable columns such as pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity and lastly, our binary-labeled outcome variable, Potability. This data file comes in the format of a .csv file, which is a common data format that can be accessed with various software tools. Although not a data analytics tool, the MacOS software application Numbers allows users to open and view .csv files. Spreadsheet softwares and platforms, such as Microsoft Excel and Google Sheets can also be used to open, view, and interact with .csv files. For data analysis, Python along with Pandas library, and R can be used to read and manipulate .csv files, especially working with large datasets like this one which contains over 3,000 rows of data.

This dataset is designated as public domain, which lifts any copyright restrictions, allowing users to use, modify, and share this data.

Metadata

The webpage of the dataset includes metadata categories such as *Collaborators, Authors, Coverage, DOI Citation, Provenance, License,* and *Expected Update Frequency.* These metadata fields are structured, standardized, and managed by Kaggle themselves.

The structure of this metadata in practice provides users with basic information about the dataset. For example, the field Coverage expands to subcategories *Temporal Coverage Start*

Date, End Date, Geospatial Coverage, which cover temporal and spatial range our data encompass. This is important for understanding the scope and context in which our data covers. However, this information is sparse, as most of these are empty or vague. The Coverage section unfortunately was blank. As mentioned above, Provenance provides information on the origins of the dataset, but the only information I could get was that it was synthetically generated. This presents a challenge to the usability and reliability of the dataset, as shown in the discussion section where several users raise doubts about the authenticity of the data and its acquisition methodology. Upon looking at the community's comments, users also questioned how potability, the outcome, was calculated from all the variables provided. From a generative AI and machine learning use-case perspective, it would be important to provide clarity on the methodology in which potability was calculated to ensure interpretability of the data and the use of data for model training. This transparency is also essential for ethical usage of AI, in which this interpretability allows for fair assessment of results and avoids generalization due to unknown factors and data. Moreover, this dataset only came with one .csv file of the data itself, with no secondary metadata resources attached. The original creator did provide scope notes for the data columns, presented on the webpage along with a preview of the data, which provides users with contextual domain information about the intended use of each variable. There are also visible keyword tags on the webpage, which tags this specific dataset pertaining to the subject of the data such as tags Earth and Nature, Environment as well as use-cases such as tags Beginner, Binary Classification.

Overall, the metadata attached to this dataset is not comprehensive, as very little information about the acquisition, origin, and assurance processes is provided. To enrich the metadata to improve user discoverability and guide users who are unfamiliar with the data, a detailed description of the data creation process is needed. Within this, information on provenance, processes, methodology, and purposes of the dataset should be documented to provide users with the context and allow them to assess the potential usability of the dataset.

I would also suggest collocating metadata into one consistent file format or location. All the metadata listed were a mix of information written in the description, and inserted via Kaggle's structured metadata fields. It would be useful for this information on coverage, origin, scope notes, and acquisition of data to be provided in a file accompanying the data file to guide new users. Real-world water quality observations would benefit decision makers directly in understanding current water quality analytics, allowing them to make informed decisions on policy and organization levels. However, the nature of this synthetically generated dataset affects the possible use cases of the data, which should be more visibly communicated to users. Collocating metadata into consistent formatting and location can ensure this metadata is preserved across different platforms, as not all platforms like Kaggle provide the same structures for metadata. Moreover, proper documentation of the nature and background of the data not only communicates the purposes of the dataset, but also the limitations of it. Having and documenting this information not only enhances the usability and visibility of the data, but also facilitates responsible and ethical use of synthetic datasets across different domains.

Upon Google searching the dataset and creator, it returned a few published articles published by various open access publishers such as Atlantis Press, and Multidisciplinary Digital Publishing Institute (MPDI). Published articles citing this dataset often pertain to neural network models using the dataset (Rustam et al., 2022), or an analysis of the dataset itself (Yang, 2022).

References

DataOne Best Practices Primer, https://dataoneorg.github.io/Education/bestpractices/

Kadiwal, A. Water Quality – Drinking water potability. Kaggle. Accessed January 19th, 2024 from https://www.kaggle.com/datasets/adityakadiwal/water-potability

Rustam F, Ishaq A, Kokab ST, de la Torre Diez I, Mazón JLV, Rodríguez CL, Ashraf I. (2022). An Artificial Neural Network Model for Water Quality and Water Consumption Prediction. Water. 2022; 14(21):3359. https://doi.org/10.3390/w14213359

Yang, R. (2022). Analyses of Approaches to Deal with Missing Data in Water Quality Data Set. Atlantis Press. https://www.atlantis-press.com/article/125973965.pdf