Datasheet: Climate-Sensitive Waterborne Diseases Dataset for Predictive Machine Learning in Tanzania

Neema N. Lyimo, Kadeghe G. Fue, Silvia F. Materu, Ndimile C. Kilatu, Joseph P. Telemala
Sokoine University of Agriculture

1. Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

1. Purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
   The dataset's primary purpose is to support predictive machine learning modeling of climate-sensitive waterborne diseases, inform public health planning and interventions, and advance research on the interplay between climate, environmental conditions, and health outcomes in Tanzania.
   The specific gap that needed to be filled is the lack of centralized and open data for machine learning on climate-sensitive waterborne diseases in low resourced countries like Tanzania.

2. Who created this dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)?
   The dataset was created by a research team led by Neema Lyimo, Joseph Telemala Silvia Materu and Kadeghe Fue at Sokoine University of Agriculture in partnership with Ndimile Kilatu, a collaborator, from Morogoro Municipal.

3. What support was needed to make this dataset? (e.g. who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
   The dataset creation project was solely supported by Meridian Institute, through Lacuna Fund on Climate and Health

4. Any other comments?
   Your Answer Here

2. Composition

Dataset creators should read through the questions in this section prior to any data collection and then provide answers once collection is complete. Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

1. What do the instances that comprise the dataset represent (e.g. documents, photos, people, countries)? Are there multiple types of instances (e.g. movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
   The dataset consists of water sources, toilet quality and sanitation, and waste management facilities data points which were collected cross-sectionally between 2023 and 2024, and subsequently linked with five years (2019–2023) of monthly weather data. The dataset also contains health facilities locations data. Furthermore, there are disease incidence data points which have been aggregated monthly, spans the same five years and includes cases of Dysentery/Amoebiasis, Diarrhea (categorized by severity), Typhoid fever, Schistosomiasis, and Intestinal worms.

2. How many instances are there in total (of each type, if appropriate)?
   Water sources - 68,000 data points matched with weather data
   Toilet quality and sanitation - 13,560 data points associated with weather data.
   Waste management facilities – 1,440 data points associated with weather data.
   Health facilities – 319 data points
   Disease cases – 1,029,225 data points

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g. geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g. to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all the instances from the five selected municipal and city councils

4. What data does each instance consist of? "Raw" data (e.g. unprocessed text or images) or features? In either case, please provide a description.
   *The **water source** data includes detailed information on various attributes such as the type of water source, the approximate number of households using it, the water treatment methods applied, the water availability status, existing water source protection measures, and the geographical location of the source.*

   *The **toilet quality and sanitatio**n data provides information on various attributes, such as the type of toilet, the status of toilet waste leakage, and the geographical location of the facility (longitude, latitude, and altitude). It also includes details about the type of handwashing facility, the availability of water and soap, the number of people served by the toilet, and its usage status. Additionally, the dataset captures the availability and capacity (in liters) of water storage facilities, the water treatment methods used for drinking water, and the presence of water sources within a 50-meter radius of the toilet.*

   Waste management facilities –

   Health facilities –

   Disease cases –

5. Is there a label or target associated with each instance? If so, please provide a description.
   ***Water sources** – Water_Source_Type, Households_Using_Source, Water_Treatment_Method, Water_Availability, Source_Protection, and Source_Location.*

   ***Toilet quality and sanitation** – Type_of_toilet, Toilet_waste_leaking_status, Longitude, Latitude, Altitude, Type_of_handwashing_facility, Water_and_soap_availability, No_of_people_using_the_toilet, Toilet_use, Availability_of_water_storage_facility, Water_storage_capacity, Water_treatment_method, Water_sources_around_the_toilet*

   Waste management facilities –

   Health facilities –

Disease cases –

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). This does not include intentionally removed information, but might include, e.g. redacted text.
   Your Answer Here

7. Are relationships between individual instances made explicit (e.g. users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
   Your Answer Here

8. Are there recommended data splits (e.g. training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
   Your Answer Here

9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
   Your Answer Here

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g. websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g. licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
    Your Answer Here

11. Does the dataset contain data that might be considered confidential (e.g. data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
    Your Answer Here

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, *The dataset does not contain any data that might be offensive, insulting, threatening or might cause anxiety*

13. Does the dataset relate to people? If not, you may skip the remaining questions in this section.
    Yes, the dataset relates to people

14. Does the dataset identify any subpopulations (e.g. by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
    Your Answer Here

15. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
    No, it is not possible to identify individuals in the dataset as the data does not contain any information related to people's identity.

16. Does the dataset contain data that might be considered sensitive in any way (e.g. data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
    Yes, the data contains locations of water sources, waste management facilities, toilets, and health facilities. However, they were anonymised by shifting their location my a certain degree to hide their actual locations.

17. Any other comments?
    Your Answer Here

3. Collection

As with the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals of the prior section, the answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

1. How was the data associated with each instance acquired? Was the data directly observable (e.g. raw text, movie ratings), reported by subjects (e.g. survey

responses), or indirectly inferred/derived from other data (e.g. part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
Your Answer Here

2. What mechanisms or procedures were used to collect the data (e.g. hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
Your Answer Here

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g. deterministic, probabilistic with specific sampling probabilities)?
Your Answer Here

4. Who was involved in the data collection process (e.g. students, crowdworkers, contractors) and how were they compensated (e.g. how much were crowdworkers paid)?
Your Answer Here

5. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g. recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
Your Answer Here

6. Were any ethical review processes conducted (e.g. by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
Your Answer Here

7. Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
Your Answer Here

8. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g. websites)?
Your Answer Here

9. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
   Your Answer Here

10. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
    Your Answer Here

11. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
    Your Answer Here

12. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g. a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
    Your Answer Here

13. Any other comments?
    Your Answer Here

4. Preprocessing / Cleaning / Labeling

Dataset creators should read through these questions prior to any pre-processing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the “raw” data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a “bag-of-words” is not suitable for tasks involving word order.

1. Was any preprocessing/cleaning/labeling of the data done (e.g. discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not,

you may skip the remainder of the questions in this section.
Your Answer Here

2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g. to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
Your Answer Here

3. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Your Answer Here

4. Any other comments?
Your Answer Here

5. Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

1. Has the dataset been used for any tasks already? If so, please provide a description.
Your Answer Here

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
Your Answer Here

3. What (other) tasks could the dataset be used for?
Your Answer Here

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g. stereotyping, quality of service issues) or other undesirable harms (e.g. financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
Your Answer Here

5. Are there tasks for which the dataset should not be used? If so, please provide a description.
   Your Answer Here

6. Any other comments?
   Your Answer Here

6. Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

1. Will the dataset be distributed to third parties outside of the entity (e.g. company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
   Your Answer Here

2. How will the dataset will be distributed (e.g. tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
   Your Answer Here

3. When will the dataset be distributed?
   Your Answer Here

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
   Your Answer Here

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
   Your Answer Here

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or

other access point to, or otherwise reproduce, any supporting documentation.
Your Answer Here

7. Any other comments?
Your Answer Here

7. Maintenance

As with the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

1. Who is supporting/hosting/maintaining the dataset?
Your Answer Here

2. How can the owner/curator/manager of the dataset be contacted (e.g. email address)?
Your Answer Here

3. Is there an erratum? If so, please provide a link or other access point.
Your Answer Here

4. Will the dataset be updated (e.g. to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g. mailing list, GitHub)?
Your Answer Here

5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g. were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
Your Answer Here

6. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
Your Answer Here

7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
   Your Answer Here

8. Any other comments?
   Your Answer Here