Datasheet for Climate Change Dataset*

Lexi knight

December 14, 2024

This datasheet presents the results of a study examining the relationship between age and education and individuals' likelihood to adopt climate-friendly behaviors. The data reveals patterns showing that younger people and those with higher levels of education are more inclined to engage in actions like reducing car usage and conserving energy. In contrast, older individuals are more likely to focus on long-term sustainability efforts such as energy efficiency. Financial barriers and skepticism about individual impact remain significant obstacles to broader climate action. The findings highlight the need for targeted interventions to make sustainable behaviors more accessible across demographic groups.

Extract of the questions from @gebru2021datasheets.

Motivation

1. **For what 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 was created to explore how demographic factors such as age and education level influence the likelihood of individuals adopting climate-friendly behaviors. The data fills a gap in understanding the relationship between socio-demographic variables and sustainable behavior adoption.
2. **Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?**
    - Created by OpenDataToronto, under the research project on climate change perspectives, in collaboration with environmental studies groups and local climate organizations.

*Code and data are available at: https://github.com/LexiKnight/toronto_climate/tree/main.

3. **Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.**
    - Funded by personal resources and a minor grant from the City of Toronto's environmental studies program.
4. **Any other comments?**
    - The dataset is part of a broader effort to collect and analyze public opinions on climate change and sustainability in Toronto.

Composition

1. **What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.**
    - Each instance in the dataset represents an individual participant in the survey. The dataset includes demographic information and responses to questions about climate-related behaviors.
2. **How many instances are there in total (of each type, if appropriate)?**
    - The 2018 dataset contains 404 instances whereas the 2021 dataset contains 1400 instances, each corresponding to a single participant's responses.
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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).**
    - The dataset is a sample of Toronto residents aged 18-65. The sample is intended to be representative of the general population in terms of age, education level, and climate-related behavior, though certain demographic groups (e.g., low-income residents) are slightly underrepresented due to survey limitations.
4. **What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.**
    - Each instance consists of processed survey data, including categorical features (age group, education level, etc.) and numeric responses (e.g., likelihood of adopting behaviors, number of sustainable actions taken).

5. **Is there a label or target associated with each instance? If so, please provide a description.**

    - The target variable is the "likelihood of adopting climate-friendly behaviors," which is a scale ranging from 1 to 5 (1 being least likely, 5 being most likely).

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

    - There is some missing data for specific questions related to education, as some respondents chose not to provide this information.

7. **Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.**

    - There are no explicit relationships between individual instances; the data is cross-sectional and independent.

8. **Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.**

    - It not known based on the dataset itsef howevery we split the data into 80% training and 20% testing for machine learning tasks (to generate the models).

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

    - There are minor errors in the self-reported data, such as inconsistent responses. There is also instances where individuals did not respond such as stating their education level.

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.**

    - The dataset is self-contained and does not rely on external resources.

11. **Does the dataset contain data that might be considered confidential (for example, 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.**

    - The dataset does not contain any confidential information, though demographic data could be considered personal.

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 offensive or harmful data.

13. **Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these sub-populations are identified and provide a description of their respective distributions within the dataset.**

    - Yes, the dataset identifies sub-populations by age and education level. Age is split into 18-35, 36-50, 51-65 and 65+. Education levels range from high school to postgraduate education however the specific categories vary between the 2018 and 2021 survey.

14. **Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.**

    - No, the dataset is anonymized and does not allow for the identification of individual persons.

15. **Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race 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.**

    - The dataset contains demographic information, including education level, which may be sensitive, but no other sensitive data such as race, health, or financial status.

16. **Any other comments?**

    - The dataset is intended to support climate behavior research and advocacy efforts to improve public sustainability initiatives.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was acquired through a survey in which individuals self-reported their age, education level, and responses to questions about climate-related behaviors. The responses were validated through careful survey design and piloting with a small sample before the full data collection.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The 2018 survey data was collected through online interviews that lasted approximately ten minues. Quotas were set by region, age and gender ensuring sampling only those 18 years and older. The 2021 survey data was collected online and over the phone.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - N/A

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The data collection was managed by the researcher, the specifics are unknown.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The 2018 survey data was collected between October 11 and October 18, 2018. The 2021 survey data does not specify the timeframe of collection.

6. *Were any ethical review processes conducted (for example, 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.*

    - No, there is no mention of ethical review.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

- The data was collected directly from individuals via the online survey.

8. *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.*

    - Yes, participants were notified about the data collection via a consent form that was presented before they completed the survey. The notification provided information about the study's purpose, confidentiality, and data usage.

9. *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.*

    - Yes, participants consented to the data collection by agreeing to the terms provided in the online consent form before they were allowed to complete the survey.

10. *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).*

    - Participants were informed that they could withdraw their consent at any time before completing the survey. Contact details for data withdrawal requests were provided in the consent form.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.*

    - No, there is no mention of conduction of a data protection impact analysis.

12. *Any other comments?*

    - N/A

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, 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 remaining questions in this section.*

    - The data was cleaned to keep missing values and rename variables more meaningful descriptions.

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

    - Yes, the raw data has been saved in a separate file and can be accessed in the download script.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - Yes, the R scripts used for data preprocessing are available on the GitHub repository: https://github.com/LexiKnight/toronto_climate.

4. *Any other comments?*

    - N/A

Uses

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

    - The dataset has been used in an analysis of the impact of age and education level on climate-friendly behaviors, as part of a larger study on demographic influences on sustainability actions.

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.*

    - Yes, a link to the project repository is available here: https://github.com/LexiKnight/toronto_climate.

3. *What (other) tasks could the dataset be used for?*

    - The dataset could be used for further studies on demographic factors influencing sustainability behaviors, as well as the effectiveness of interventions designed to encourage climate-friendly actions.

4. *Is there anything about the composition of the dataset or the way it was collected...*

    - The dataset offers a rich resource for studying the intersection of demographics and climate behavior, with potential for comparative studies in other urban areas.

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

    - N/A

6. *Any other comments?*

    - N/A

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - No, the dataset will not be distributed to third parties unless requested for academic research, policy analysis, or public service projects related to climate change behavior. The dataset is however accessibly via GitHub.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset would be distributed via GitHub and is accessible to anyone with the link.
3. *When will the dataset be distributed?*
    - The dataset is available for distribution immediately.
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.*
    - No, the dataset will not be distributed under a copyright or other IP license nor others similar.
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.*
    - No third-party restrictions have been imposed on the dataset, but users are reminded to respect confidentiality and ethical guidelines regarding sensitive demographic data.
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.*
    - No export controls or regulatory restrictions apply to this dataset at the moment.
7. *Any other comments?*
    - The dataset is being made available to encourage academic, governmental, and non-governmental use, and contributors are encouraged to respect privacy and data security standards when using or sharing the dataset.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be hosted on GitHub and maintained by the Open Data Toronto project team, which includes data scientists and urban studies researchers.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - The dataset owner can be contacted at Opendatatoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
    - There is no erratum currently. If errors are identified, they will be documented and corrected in subsequent versions of the dataset.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - The dataset will be updated annually to reflect new data or to correct any errors. Updates will be communicated via Open Data Toronto.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the 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.*
    - The dataset will adhere to the privacy policy outlined in the informed consent process, with all data retention guidelines strictly followed. Data will be retained for a maximum of five years, after which it will be deleted unless otherwise required by law or for ongoing research purposes.
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 dataset consumers.*
    - Older versions of the dataset will be archived for historical reference. Users will be notified via GitHub whenever a new version is released, with clear indication of any major changes or updates.
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 dataset consumers? If so, please provide a description.*
    - Others are welcome to extend or augment the dataset through contributions via GitHub, following the guidelines provided. Contributions will be reviewed by the dataset maintainers for accuracy and relevance before being merged into the main repository.

8. *Any other comments?*
    - N/A