# Datasheet for the datasets used in analysis of education level in Philippines in 1993

Zihao Liu

Extract of the questions from Gebru et al. (2021)

The datasheets is written using R programming language and a R Markdown file (R Core Team 2020).

**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 analyze the education level of male in Philippines in 1993. The dataset was generated from a table in a demographic report of Philippines in 1993 ((NSO) and (MI) 1994).
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?* -The original table is created by National Statistics Office in Philippines and Macro International Inc. in USA.

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.*
    - Macro International Inc. provided the funding for the project that includes the original table ((NSO) and (MI) 1994).

**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.*
    - The Dataset is regarding the education level of male Filipinos in 1993, by age groups, region, and residence.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are mainly two instances, the male Filipinos and Regions in Philippines.
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 contain a sample of the whole population of Philippines. It is a representative of the larger set, in terms of age groups, and geographic coverage. The number of samples for each region and age group is planned according to the demographic distribution.
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.*
    - The data is just regarding the education level of male Filippinos and the level in different age groups and regions.
5. *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 generated from a table in NDS of Philippines in 1993. The report is published and will exist and remain constant.

6. *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.*
   - No
7. *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
8. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
   - The dataset is regarding the education level of male in Philippines in 1993.
9. *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.*
   - Not possible

### 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 collected by doing interviews and survey with selected individuals ((NSO) and (MI) 1994).
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 data was collected by surveys and interviews.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
   - Probabilistic and proportion to the size of the regions ((NSO) and (MI) 1994).
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)?*
   - Field staff were trained in Baguio City, Manila, Cebu City and Davao City, some of them were interviewers and some are field editors ((NSO) and (MI) 1994).
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 data was collected in 1993.
6. *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 comes from the NDS of Philippines in 1993 report.

### 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.*
   - Some of the text are misidentified in the process of OCR, and I had to correct them mannually.
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.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R programming language was used to preprocess the data (R Core Team 2020).

**Uses**

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset is used in the analysis of male Filipinos education level.
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.*
    - National Demographic Survey of Philippines in 1993 ((NSO) and (MI) 1994).

**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.*
    - The dataset will be opensource on GitHub.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - GitHub

**Maintenance**

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Zihao Liu
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - zihaohans.liu@mail.utoronto.ca

# References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

(NSO), National Statistics Office, and Macro International Inc. (MI). 1994. "National Demographic Survey of Philippines 1993."

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.