**Data Sheet**

**Motivation.**

1. **For what purpose was the dataset created?**

The dataset was created to capture information regarding the effect of socioeconomic factors on infant mortality in Nepal.

1. **Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?**

It was created by the Ministry of Health (Nepal).

1. **Who funded the creation of the dataset?**

It was funded by the Nepali Government.

**Composition.**

1. **How many instances are there in total (of each type, if appropriate)?**

There are 5 instances of Mortality rates for all 16 socio-economic categories (i.e. 80 instances total).

1. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** The dataset is a summary of a sample of the population (all instances). The sample size is 8429.
2. **What data does each instance consist of?** Mortality Rate (Deaths per 1000 individuals).
3. **Is there a label or target associated with each instance?** Yes, for example the mortality rate of
4. **Is any information missing from individual instances?** Yes this specific data set does not contain information about individuals, rather it is a group summary based on a socio-economic demography.
5. **Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)?** Yes, the individual instances are linked based on their type of grouping. For example geographical location, Education, etc.
6. **Are there recommended data splits (for example, training, development/validation, testing)?** No
7. **Are there any errors, sources of noise, or redundancies in the dataset?** Yes, some of the instances are Not available or have an asterisk or an in a parenthesis. Figures in parentheses are based on 250-499 births; an asterisk indicates that the figure is based on fewer than 250 exposed persons and has been suppressed.
8. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?** It is self-contained.
9. **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)?** No, it is summary data (not individual).
10. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** No

*People*

1. **Does the dataset identify any sub-populations (for example, by age, gender)?** Yes the subpopulations are divided by socioeconomic group.
2. **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?** No
3. **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)?** No
4. **Any other comments?**

**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)?** The data was reported by subjects.
2. **What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** It was collected through a direct interview with the subjects.
3. **If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?**

The data was collected through two-step stratified sampling. The sampling strategy was probabilistic. The first step was the allocation 253 PSU (primary sampling unit). 34 were allocated to urban areas, 219 were allocated to rural areas. Then in these areas, each subdivision was given a probability proportional to their population size (number of households). 25 Interviews were completed from each PSU probabilistically.

1. **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)?**

Workers from the Nepal Ministry of Health.

1. **Over what timeframe was the data collected?**

This data was collected for the year 1996.

1. **Were any ethical review processes conducted (for example, by an institutional review board)?** No.
2. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?**

Obtained it through DHS.

1. **Were the individuals in question notified about the data collection?** No
2. **Did the individuals in question consent to the collection and use of their data?** The individuals are kept anonymous, so they do not need to consent to each data collection (just the initial).
3. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** No.
4. **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?** No

**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)?** Yes, the initial dataset was transposed to make the rows columns as they represented a specific grouping.
2. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)?** Yes: raw\_data.csv
3. **Is the software that was used to preprocess/clean/label the data available?** Script/clean\_data.r

**Uses.**

1. **Has the dataset been used for any tasks already?** It was used to analyse the difference in mortality rates between different socioeconomic groups.
2. **Is there a repository that links to any or all papers or systems that use the dataset?** No
3. **What (other) tasks could the dataset be used for?**

None

1. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** No
2. **Are there tasks for which the dataset should not be used?** No

**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?** Yes, it can be accessed by anyone.
2. **How will the dataset be distributed (for example, tarball on website, API, GitHub)?** It is available online at DHS.
3. **When will the dataset be distributed?**

It has already been distributed.

1. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** No
2. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** No
3. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No

**Maintenance.**

1. **Who will be supporting/hosting/maintaining the dataset?**

No one

1. **How can the owner/curator/manager of the dataset be contacted (for example, email address)?**

You must contact the ministry of health (Nepal).

1. **Is there an erratum?** No
2. **Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)?** No
3. **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)?** No
4. **Will older versions of the dataset continue to be supported/hosted/maintained?** N/A
5. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** No