Datasheet for the Dataset of Life Expectancy*

Pu Yuan

April 10, 2024

This dataset contains information related to life expectancy and various factors that might affect it in different countries. The data spans multiple years and includes health-related metrics along with economic indicators.

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 investigate the determinants of life expectancy globally, addressing a gap in comprehensive, multi-faceted health data analysis.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The data was collected from WHO and United Nations website.

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.*
    - As the data-sets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model data-set. The final merged file(final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors.

*Code and data are available at: https://github.com/scottyuan6/life_expectancy.git.

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

    - The final merged file(final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories:Immunization related factors, Mortality factors, Economical factors and Social factors.

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 dataset consists of aggregated data from various health and economic sources, validated through official health statistics procedures by the collecting entity.

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.*

    - Modifing column names, handling of missing values, and categorization.

Uses

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

    - Our analysis shows that both higher GDP and increased health expenditure are significantly associated with better life expectancy. More importantly, we discover that health spending has a more substantial impact in countries with lower GDP. This finding suggests that targeted health investments in poorer countries can lead to significant improvements in life expectancy, offering a powerful strategy for public health interventions..

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 project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website.

Maintenance

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

    - The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The datasets are made available to public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. .

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

    - You can find the author KUMARRAJARSHI from Kaggle.