Life Expectancy Analysis and Prediction

A Data Management Plan created using DMP OPIDoR

DMP Creator: Maroua Jaoua

Principal Investigator Affiliation: Other Organisation

DMP Template: Horizon 2020 DMP

Last modified: 27-04-2020

Funder: European Comission (Horizon 2020)

Project abstract:

The puprose of this project is to investigate the topic of life expectancy. A couple of questions will be answered based on the available and selected data. It is intended to compute the average human life expectancy in the world and to understand the influence of the income status of the country on the life expectancy. The difference of life expectancy between male and female is also evaluated. Finally, the countries with the highest and lowest imporvement in terms of life expectancy at birth are determined.

Principal Investigator: Maroua Jaoua

ORCID iD: https://orcid.org/0000-0001-8109-9644

Data Contact Person: Maroua Jaoua

Research outputs :

1. Merged and Preprocessed data generated from the initial collected data (Dataset)
2. Jupyter notebook including the CRISP-DM workflow and the final results of prediction and analysis (Other: Jupyter Notebook)

Life Expectancy Analysis and Prediction - Initial DMP (within first 6 months)

For each data set specify the following:

Merged and Preprocessed data generated from the initial collected data

Data set reference and name

The data of interest is a collection of data from three different sources. These sources are the World Health Organization, World Bank data and Kaggle. This data will be then merged and preprocessed.

Data set description

The data is saved as csv files. The data volume is small (3.84 MB). The data mainly includes information about the health, the bank information and the status of the country. The sources of the data are World Health Organization, World Bank data and Kaggle. This data exists over platforms under the licenses CC0 1.0, CC-BY 4.0 and CC BY-NC-SA 3.0 IGO. All of these licenses allow users to copy, modify and distribute data. The different collected sources will be combined and merged into one data source which can be later on reused. The types of the generated files are csv files which have also small volume. (data/mergedData.csv [2346 kB] and data/preprocessedData.csv [615 kB])

Standards and metadata

Metadata about the project is included in documentation/metadata.xml . It includes experiment title, authors, date, tools, ... Metadata describing the filenames of the different csv files are also included under data/metadata.txt .

Data sharing

The data access is open under the license CC BY-NC-SA 3.0 IGO. Embargo periods are not necessary. The data will be saved in a Github repository under the folder "data" (https://github.com/MarouaJaoua/DataStewardship1). Additionally, metadata and links of the sources are included in txt files. The repository is public and has a license.

Archiving and preservation (including storage and backup)

The files that should preserved are :

- README.md
- The Jupyter notebook life_exp_notebook_group_30.ipynb
- requirements.txt
- documentation/metadata.xml

The preservation will be in the Github repository. This will require no additional cost and not a lot of resources. The contributers of the project will do it themselves.

Jupyter notebook including the CRISP-DM workflow and the final results of prediction and analysis

Data set reference and name

Question answered.

Data set description

Question answered.

Standards and metadata

Question answered.

Data sharing

Question answered.

Archiving and preservation (including storage and backup)

Question answered.

Life Expectancy Analysis and Prediction - Final review DMP

Scientific research data should be easily: 1. Discoverable

Merged and Preprocessed data generated from the initial collected data

Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?

The data collected has a Digital Object Identifier created by zenodo. Similarly, the data generated by the notebook has a DOI. A DOI for the repository for the latest release is also created. All DOIs are linked in the Github repository.

DOI for collected data: https://doi.org/10.5281/zenodo.3770486
DOI for merged and preprocessed data: https://doi.org/10.5281/zenodo.3770405
DOI for repository: https://doi.org/10.5281/zenodo.3770415
Github Repository: https://github.com/MarouaJaoua/DataStewardship1

Jupyter notebook including the CRISP-DM workflow and the final results of prediction and analysis

Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?

Question r nswered.

2. Accessible

Merged and Preprocessed data generated from the initial collected data

Are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses?

The data is accessible. The repository is public and the project and data are easily accessible. A license is also provided for the project and the data.

Jupyter notebook including the CRISP-DM workflow and the final results of prediction and analysis

Are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses?

Question not answered.

3. Assessable and intelligible

Merged and Preprocessed data generated from the initial collected data

Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review?

The project is intelligible. It has a clear folder structure. It includes a READ.me file including the instructions. It includes a requirements text file to ensure that the right versions are selecte. The code is well commented and follows the conventions. The notebook also includes a clear description of the intentions and gives a great idea about the workflow. A presentation is also added for extra clarity.

Jupyter notebook including the CRISP-DM workflow and the final results of prediction and analysis

Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review?

Question not answered.

4. Usable beyond the original purpose for which it was collected

Merged and Preprocessed data generated from the initial collected data

Are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data?

The data is reusable. A license is created for both the data and the project. The license chosen is CC BY-NC-SA 3.0 IGO. This is because one of the data sources requires this license. The commercial use is not possible in this case. Therefore, users are allowed to freely copy, reproduce, reprint, distribute, translate and adapt the work and data for non-commercial purposes.

Jupyter notebook including the CRISP-DM workflow and the final results of

prediction and analysis

Are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data?

Question not answered.

5. Interoperable to specific quality standards

Merged and Preprocessed data generated from the initial collected data

Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc?

The data is interoperable. It follows the standards and conventions of machine learning. It clearly includes and describe the data pipeline. The documentation is detailed and metadata is provided.

Jupyter notebook including the CRISP-DM workflow and the final results of prediction and analysis

Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc?

Question not answered.