Finding predictive Features for countries living in extreme poverty. - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

The Data will be collected from UNESCO's Institute for Statistics. To be exact we will use the "demographic and socio-economic" dataset from the UIS website (http://data.uis.unesco.org/)
By collecting this data set from the UIS, we can create predictive linear models, that are trained to predict a countries poverty status.

Explain the relation to the objectives of the project

From the predictive linear models we can easily see, which features of countries where used to make the predictions and how important the linear model thought the feature was to its prediction. Finding these is this projects main objective.

Specify the types and formats of data generated/collected

All data is handled using pythons standard data types, scipy vectors and pandas dataframes.
Both the collected and generated data will be stored in .csv format.

Specify if existing data is being re-used (if any)

See above: The UIS' data is being used.

Specify the origin of the data

The data is from the UIS Site, labeled as "Demographic and socio-economic"
http://data.uis.unesco.org/RestSDMX/sdmx.ashx/GetData/DEMO_DS/SP_DYN_TFRT_IN+SP_DYN_LE00_IN+SP_DYN_IMRT_IN+200343+200144+200345+200151+SP_POP_GROW+SH_DYN_AIDS_ZS+SP_RUR_TOTL_ZS+200101+DT_TDS_DECT_GN_
startTime=2013&endTime=2019

State the expected size of the data (if known)

30MB

Outline the data utility: to whom will it be useful

The generated data will be useful for decision makers trying to predict countries that will be living in extreme poverty, and may also provide means to find axes on which counteraction might be taken to combat upcoming poverty.
It may also help countries living in extreme poverty to find guidance on how to overcome their status.

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

Metadata will manually be stored in an xml file called metadata.xsd
Unfortunately metadata from the UIS' side is very sparse, the metadata can only be accessed as pdf. It is nonetheless stored in the project's data folder.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

The project and all its resources (report, code, data) will be available on zenodo accessable via DOI.
https://doi.org/10.5281/zenodo.3757193

Outline naming conventions used

The data, program and metadata will be stored in the "project" folder.
The original data is in the data folder: continents.csv and unesco_poverty_dataset.csv
The generated data is in the data folder: transformed.csv and feature_descriptions.csv
The program files can be found at src/
The metadata is in the metadata.xml file in the root folder.
Metadata provided by the UIS is in the data folder.
A similar description to this can be found in the github repository's README.md
https://github.com/BrennerG/DS1

Outline the approach towards search keyword

The metadata file is written according to the Dublin Core XML standard, which should facilitate search- and findability.

Outline the approach for clear versioning

The program will be versioned using git (more precisely github) link: https://github.com/BrennerG/DS1
The intermediate versions of the used and generated data can be accessed by looking at previous github versions.
All the releases on github are linked to zenodo and will automatically be updated, so that the most updated version is visible at zenodo. https://doi.org/10.5281/zenodo.3757193

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

The metadata will be created using the Dublin Core standard using the following web-tool:
https://nsteffel.github.io/dublin_core_generator/generator_nq.html

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

All data will be made openly available with an MIT License.

Specify how the data will be made available

The data will be uploaded to zenodo. DOI: https://doi.org/10.5281/zenodo.3757193
and github: https://github.com/BrennerG/DS1

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

The generated intermediate data and conclusions will be stored in .csv and according visualizations can be found in .png format.
The source code of the relevant software can be found at github and will be included in the project folder.
The source code is written using python using streamlit notebooks. All necessary requirements for python are in the src/requirements.txt file. Instructions on how to install the requirements and run the code can be found in the README.md file.

Specify where the data and associated metadata, documentation and code are deposited

The data and associated metadata, documentation and code are deposited at github: https://github.com/BrennerG/DS1. Finished releases containing all of the above are automatically archived at zenodo: https://doi.org/10.5281/zenodo.3757193

Specify how access will be provided in case there are any restrictions

The github respository will be open for everyone to see, as will the zenodo uploads.

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

In order to make the data interoperable it is stored in .csv files.
The metadata is stored as .xsd according to the Dublin Core format.
These two file formats are fairly interoperable, as they are standard file formats, well known and well used.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Standard data types will be used in the .csv files - actually only Strings and Integers.
In the python source code scipy vectors, pandas dataframes are used in addition to the regular data types.

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

The data will be licensed with the MIT License.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

As no data embargo is needed, the data will be made available as soon as the project reached its conclusion.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

As the MIT License states, the reuse of data is not restricted to any kind of usage.

Describe data quality assurance processes

The data quality will be assured primarily by self-observation. In addition to this, the replicability of this experiment will be tested by the lecture team of the Data Stewardship course.

Specify the length of time for which the data will remain re-usable

The data will be available at zenodo.org.
A physical backup of all the data and source code is stored on my personal hard drive.

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

I am not paid to execute this experiment, so no monetary cost in making the data fair arises.
The main cost in doing this was time - all in all I spent 2 hours setting up github, zenodo, linking them and documenting the software.

Clearly identify responsibilities for data management in your project

This project is/was a group effort the responsibilities were assigned as following:
Gabriel (me)

- Data Acquisition & Quality
- Data Transformation
- Data Storage / Backup
- Metadata Production

Sanja Prijselac

- Data Imputation
- Machine Learning

Carlos Vargas

- Machine Learning

Aaron Abebe

- Data Visualization

Describe costs and potential value of long term preservation

The data will be on zenodo.org for an unforseeable amount of time - probably long.
The source code will be on github.com for practically ever.
A Backup of the source code and all data will be on my physical hard drive.
None of these will cost money.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

The data is on github, zenodo and my personal harddrive.
https://github.com/BrennerG/DS1
https://doi.org/10.5281/zenodo.3757193
The data is not sensitive as no personal data is contained in the data files.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

No protection of data is needed, as the data is publicly available and does not contain personal information.
As long as no ethical concerns arise with github, zenodo or UNESCO no ethical issues will arise concerning data acquisition, storage or transfer.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Question not answered.