DATA STEWARDSHIP DMP EXERCISE

A Data Management Plan created using DMPonline

Creator: Moritz Leidinger

Affiliation: Other

Template: Digital Curation Centre

ORCID iD: https://orcid.org/0000-0002-4537-6648

Project abstract:
This experiment aims to apply two machine learning models to solve a classification task for two different datasets. Each dataset is evaluated with each of the machine learning models, using different parameter settings and preprocessing strategies to compare the respective results and analyze across datasets and/or machine learning models.

Last modified: 22-04-2019

DATA STEWARDSHIP DMP EXERCISE

DATA COLLECTION

What data will you collect or create?

First and formmost, source code will be created in this experiment. This source code will be written in the programming language Python and will be partly contained in so-called Jupyter notebooks (special JSON files) and partly in python files (standard text files with the .py extension). Additionally, some documentation will be produced and stored in both a Microsoft Word file and a markdown file. The outputs of the source code will be stored partly in .csv files consisting of two columns (an ID & the result of classification) and partly directly in said Jupyter notebooks). Additionally, images will be produced showing a visual representation of the decision tree used for classifcation.
The total size of the above-mentioned data will not exceed 10 MB in size, with the exception of the result files, which are generated in a cumulative way and could therefore exceed this 10 MB if the source code is run many times. In general, the file size should be easily handleable by consumer machines for both execution and storage. The chosen file formats have long been in use and are used by a large community, increasing the changes of long-time support immensely.
In addition to the data created in this experiment, data from two online repositories will be used. The "Arrhythmia" dataset was downloaded from https://archive.ics.uci.edu/ml/datasets/Arrhythmia and is provided in two .txt files, whereas the "Breast Cancer" dataset was sourced from https://www.kaggle.com/c/184702-tu-ml-ws-18-breast-cancer/data and in .csv format (both long-standing standards). Both datasets are useable free of charge.

How will the data be collected or created?

The source code will be created in an Integrated Development Environment (IDE) called PyCharm (Build PY-191.6655.12) which is provided by JetBrains s.r.o., the documentation will be created in Microsoft Word (v16.23) and the output csv and image files will be created by running the source code.
The source code will be under git version control during creation and centrally stored in a public github repository which can be found at https://github.com/buboh/data-stewardship-ex1. In this repository, pre-existing data is stored in the "data" directory with "Breast_Cancer" and "Arrhythmia" subdirectories, whereas the source code can be found in the "src" directory. Common functions will be defined in the "base.py" file, whereas the processing of data and execution of funtions is done in the "Breast_Cancer.ipynb" and the "Arrhythmia.ipynb" files. Documentation can be found in the README.md file and the license for reusing the source code in the "LICENCE" file. Output files from

running the source code are stored in a "results" directory which again contains "Breast_Cancer" and "Arrhythmia" subdirectories. Additionally the root directory contains a .gitignore file which contains the files to not commit into version control.
The source code is written in Python (v3.7.3) and will adhere to the Python coding conventions as specified by the Pyhton Documentation for version 3.7.3.

DOCUMENTATION AND METADATA

What documentation and metadata will accompany the data?

There will be a short report containing a description of the experimental setup, as well as an overview of what frameworks and datasets were used. It will also a short report of the findings of the experiment. This report will be available as a Microsoft Word Document as well as a PDF and it will be included in the README file of the repository. This README will also include a short manual on how to set up the environment to run the sourcecode. Additionally, some of the more complicated parts of the source code will have accompaning comments describing the functionality.
Metadata describing the Arrhythmia dataset is included in the downloaded dataset and can be found in the "arrhythmia.names.txt" file. There is no such description for the "Breast Cancer" dataset.
For each result file produced by running the source code, the files name will describe the algorithm and settings used to produce the result as well as the time of creation.

ETHICS AND LEGAL COMPLIANCE

How will you manage any ethical issues?

No ethical issues have been identified, all pre-existing data is fully anonymized and so are the result files of the classification procedure. The only identifying data is of the author of the experiment himself.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

A copy of the MIT Licence will be provided here, as well as in the repository which specifies how the data can be reused:

STORAGE AND BACKUP

How will the data be stored and backed up during the research?

The data will be stored on a local machine for manipulation and in a GitHub repository for Version Control. Additionally, a periodical backup of the local machine is stored on a local harddrive.

How will you manage access and security?

The local machine is secured by a password and the harddrive is locked inside an apartment, but the GitHub repository will be publically available for viewing and forking. Pushing changes can only be done by collaborators, whose accounts are secured by passwords as well.

SELECTION AND PRESERVATION

Which data are of long-term value and should be retained, shared, and/or preserved?

Since the data produced in this experiment is for the sake of learning to write a DMP only, none if it has any long-term value except maybe the DMP itself, which could be useful for future reference for the author himself and maybe other students writing their first DMP.

What is the long-term preservation plan for the dataset?

The pre-existing datasets are already stored on public repositories and will remain to do so. Additionally, the exact data used for the experiment will be stored in the GitHub repository alongside the source code and the result files.

DATA SHARING

How will you share the data?

The GitHub repository containing the data will be publicy available. The source code as well as the result data will be assigned a DOI, just like this DMP. The data will be made available during the experiment and will continue to be updated until the experiment is finished. The data of completion will be noted in the README file.
The source code has the DOI: 10.5281/zenodo.2648741
The result dataset has the DOI: 10.5281/zenodo.2648713
This DMP has the DOI: 10.5281/zenodo.2648751

Are any restrictions on data sharing required?

The data provided can be shared in accordance with the terms of the provided MIT license. This license gives virtually no restrictions for reusing the data.

RESPONSIBILITIES AND RESOURCES

Who will be responsible for data management?

Moritz Leidinger (moritz.leidinger@student.tuwien.ac.at)

What resources will you require to deliver your plan?

No resources will be needed to deliver the plan.