DS Exercise DMP

ID (HTTP-ORCID)	Created	Modified	Language	Description
https://doi.org/10.5281/zenodo.2648751	22-04-2019	28.06.2019	en	First and formost, 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.

	Data Officer	Name: Moritz Leidinger Email: moritz.leidinger@student.tuwien.ac.at ID: https://orcid.org/0000-0002-4537-6648 (HTTP-ORCID)
I	**Data Characteristics**	
I.1	**Description of the data**	The data in this project will be created in collaboration with: Bernhard Kurz - Researcher - bernhard.kurz@student.tuwien.ac.at - ID: 1 (custom) Paul Lang - Librarian - paul.lang@student.tuwien.ac.at - ID: 2 (custom) One dataset will be created. This dataset is titled "ML data for project" and described as follows: First and formost, 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 datasets resource type genre is "dataset" and it's language is specified as "en" . It was issued on 28.06.2019. Keywords for the dataset are: Breast Cancer, Arrhythmia The data will be created over the course of a single project: Project "Projyproj Project": 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. The project starts at 28.06.2019 and is due to end at 31.08.2019. It will be funded by 1 (custom), with the Grant ID 1-1 (custom). The funding status is curently "granted".
II	**Documentation and Metadata**	
II.1	**Metadata standards**	Just some metadata Language: en ID: 2 ((custom))
II.2	**Documentation of data**	One technical resource is needed for this project: • ID: 0 (custom): A fridge to keep the cold brew coming.
II.3	**Data quality control**	• No assurance
III	**Data availability and storage**	

III.1	**Data sharing strategy**	ID-1: https://doi.org/10.5281/zenodo.202648713 (HTTP-ORCID) Data Stewardship DMP Exercise: GitHub repository containing the data. Access to the data is open. It is in text/csv format with a size of <10MiB bytes and can be accessed at https://github.com/buboh/data-stewardship-ex1 where is is available until unknown .
III.2	**Data storage strategy**	**Cost(s)** • Project costs: 110 EUR Total costs needed to complete the project. Dataset: "ML data for project" Data Stewardship DMP Exercise This distribution is hosted by "GitHub". The data will be stored in a GitHub repository for Version Control. This host is located in US. It is certified with "unknown". It supports versioning. The backup type used by the host is unknown. The data is backed with unknown frequency. The host provides persistent identifiers from the ['doi'] system.
IV	**Legal and ethical aspects**	
IV.1	**Legal aspects**	License: • MIT License Valid from: 2019-04-22
IV.2	**Ethical aspects**	Personal data: no Preservation statement: Must be preserved to keep research & development between stakeholders intact. Security & privacy: Access control 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. Sensitive data: no Ethical issues: no 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. https://very.ethical.reports.com