Appendices for the Paper: 'Completeness of Datasets Documentation on ML/AI repositories: an Empirical Investigation'

Marco Rondina, Antonio Vetrò, and Juan Carlos De Martin

Politecnico di Torino, Torino, Italy
{marco.rondina,antonio.vetro,juancarlos.demartin}@polito.it

A Documentation Test Sheet

Table 1: Documentation Test Sheet

Dataset:		
Test Field ID	**Test Field Name**	**Presence Check**
1.01	*Purpose for the dataset creation*	
1.02	*Dataset creators*	
1.03	*Dataset funders*	
1 *Motivation Presence Average*		
2.01	*Description of the instances*	
2.02	*Number of the instances*	
2.03	*Information about missing values*	
2.04	*Recommended data splits*	
2.05	*Description of errors, noise or redundancies*	
2.06	*Information about data confidentiality*	
2.07	*Information about possible data dangerousness (offensive, insulting, threatening or cause anxiety) or biases*	
2.08	*Information about people involved in data production and their compensation (if people related)*	
2.09	*Description of identifiability for individuals or subpopulations (if people related)*	
2.10	*Description of data sensitivity (if people related)*	
2.11	*Statistics*	
2.12	*Pair plots*	
2.13	*Probabilistic model*	
2.14	*Ground truth correlations*	
2 *Composition Presence Average*		
Continue on next page		

Table 1 – continued from previous page

Dataset:		
Test Field ID	**Test Field Name**	**Presence Check**
3.01	*Description of instances acquisition and data collection processes*	
3.02	*Information about people involved in the data collection process and their compensation*	
3.03	*Time frame of data collection*	
3.04	*Information about ethical review processes*	
3.05	*Information on individuals' knowledge of data collection (if people related)*	
3.06	*Information on individuals' consent for data collection (if people related)*	
3.07	*Analysis of potential impacts of the dataset and its use on data subjects*	
3 *Collection processes Presence Average*		
4.01	*Description of sampling, preprocessing, cleaning, labelling procedures*	
4.02	*Information about people involved in the data sampling, preprocessing, cleaning, labelling procedures and their compensation*	
4.03	*Description of others possible sampling, preprocessing, cleaning, labelling procedures*	
4 *Data processing procedures Presence Average*		
5.01	*Description of the tasks in which the dataset has already been used and their results*	
5.02	*Description of recommended uses or tasks*	
5.03	*Description of not recommended uses*	
5.04	*Repository that links to papers or system that use the datasets*	
5.05	*Description of license and terms of use*	
5 *Uses Presence Average*		

Continue on next page

Table 1 – continued from previous page

Dataset:		
Test Field ID	**Test Field Name**	**Presence Check**
6.01	*Information about subject supporting, hosting, maintaining the dataset*	
6.02	*Contact of the owner*	
6.03	*DOI*	
6.04	*Erratum*	
6.05	*Information about dataset updates*	
6.06	*Information about management of older dataset versions*	
6.07	*Information about the mechanism to extend, augment, build on, contribute to the dataset*	
	6 *Maintenance Presence Average*	
	Dataset *Presence Average*	
Dataset	*Characteristics*	
c.01	*Data is people related*	
c.02	*Presence of explicit target variable*	
c.03	*Dataset is a sample or a reduction of a larger set*	
c.04	*Recently updated*	

B Fields of Information

Here we describe the choices of *Test Fields*: since the main source for constructing the schema is Datasheet for Datasets (DfD), their descriptions are outlined as a difference from the latter. Where not specifically indicated, a 1-to-1 mapping was carried out between the DfD's questions and the test fields of the *Documentation Test Sheet*.

In the section **2** *Composition*, questions on what the instances represent, what data each instance consists of, about relationships between individual instances and if the dataset relies on external resources were collapsed into field **2.01** *Description of the instances.* Questions about the fact that the dataset identifies any subpopulations and about the possibility to identify individuals were collapsed into field **2.09** *Description of identifiability for individuals or subpopulations (if people related).* Questions related to the presence of a target variable and to the fact that the dataset is a sample/reduction of a larger set were moved to the *Characteristics* section. In this section, we added some fields about the different statistical properties coming from [3]: **2.11** *Statistics*, **2.12** *Pair plots*, **2.13** *Probabilistic model*, **2.14** *Ground truth correlations.*

In the section **3** *Collection processes*, questions on how was data acquired, on the mechanism used to collect the data and on data sources were collapsed in field **3.01** *Description of instances acquisition and data collection processes.* Information about the sampling strategy was moved to the first field of *Data processing procedures*, i.e. **4.01** *Description of sampling, preprocessing, cleaning, labelling procedures.* Questions about the user's consent to data collection/use and on the presence of a mechanism to revoke their consent were collapsed in field **3.06** *Information on individuals' consent for data collection (if people related).*

In section **4** *Data processing procedures*, questions on preprocessing, cleaning, labelling description, presence of 'raw' data and on availability of used software were collapsed in field **4.01** *Description of sampling, preprocessing, cleaning, labelling procedures.* Two more fields were added in this section: the first one is the description of other possible (and perhaps recommended) procedures, and the second one is the description of the information about people involved in these procedures. The latter is inspired by the demographic information of contributors to the construction of NLP datasets, as exposed in [1].

In the section **5** *Uses*, questions about tasks for which the dataset should not be used and about the possible impact of some data characteristics or procedures were collapsed in field **5.03** *Description of not recommended uses.* Moreover, the question related to the dataset's terms of use, originally located in the Distribution section, was moved to this section.

The Distribution section of the DfD was discarded (except for DOI and terms of use) because it was found to be very inapplicable to testing the documentation of datasets already published in public repositories. In fact, questions about distribution of datasets to third parties, distribution methods, third party restrictions and export controls were not useful for this study as it focused on online public repositories. The 'repository' represents a third party with whom the dataset has already been shared, often in a way that is common to all datasets within a repository.

Finally, in the section **6** *Maintenance*, questions on planned updates and on people-related data retention were collapsed in field **6.05** *Information about dataset updates.*

In addition to dataset metadata, some characteristics of the data were tracked. These characteristics, aggregated in section **c** *Characteristics*, are four: **c.01** *Data is people related*, **c.02** *Presence of explicit target variable*, **c.03** *Dataset is a sample or a reduction of a larger set* and **c.04** *Recently updated.*

The field **c.01** *Data is people related* was checked as True if the dataset contained people-related data. This is derived from the presence of specific additional questions in the case of datasets with people-related data in DfD's scheme. To check this characteristic, an interpretation was required. In the academic literature, it is possible to find some recommendations to take a very broad interpretation of whether a dataset relates to people [2]. During the documentation analysis, that recommendation was not taken literally, so not all human artefacts were considered personal data: further details can be found in Appendix C.

The field **c.02** *Presence of explicit target variable* was checked as True if the dataset contained an explicit target variable (derived from the related question in the Composition section of DfD).

The field **c.03** *Dataset is a sample or a reduction of a larger set* was checked as True if it was clear that the dataset was the subset (in terms of rows or columns) of another dataset (derived from the related question in the Composition section of DfD).

Finally, the field **c.04** *Recently updated* was checked as True if the dataset was published or updated after 01/01/2021[1].

[1] The analysis took place between the last few months of the year 2021 and the first half of the year 2022.

C Selected Datasets

Table 2: Huggingface selected datasets (20/11/2021).

ID	Name	Download	Duplicate of	URL
hug01	glue	719706		https://huggingface.co/datasets/glue
hug02	super_glue	490789		https://huggingface.co/datasets/super_glue
hug03	anli	171100		https://huggingface.co/datasets/anli
hug04	wikitext	114761		https://huggingface.co/datasets/wikitext
hug05	wino_bias	102485		https://huggingface.co/datasets/wino_bias
hug06	squad	98446		https://huggingface.co/datasets/squad
hug07	imdb	93646		https://huggingface.co/datasets/imdb
hug08	trec	71906		https://huggingface.co/datasets/trec
hug09	adversarial_qa	70084		https://huggingface.co/datasets/adversarial_qa
hug10	race	68456		https://huggingface.co/datasets/race
hug11	duorc	67179		https://huggingface.co/datasets/duorc
hug12	squad_v2	66750		https://huggingface.co/datasets/squad_v2
hug13	winogrande	58213		https://huggingface.co/datasets/winogrande
hug14	hellaswag	54372		https://huggingface.co/datasets/hellaswag
hug15	common_voice	54179		https://huggingface.co/datasets/common_voice
hug16	cnn_dailymail	53207		https://huggingface.co/datasets/cnn_dailymail
hug17	piqa	53155		https://huggingface.co/datasets/piqa
hug18	xsum	50393		https://huggingface.co/datasets/xsum
hug19	cosmos_qa	50151		https://huggingface.co/datasets/cosmos_qa
hug20	mlqa	49740		https://huggingface.co/datasets/mlqa
hug21	quail	49413		https://huggingface.co/datasets/quail
hug22	paws	48998		https://huggingface.co/datasets/paws
hug23	wmt16	48694		https://huggingface.co/datasets/wmt16
hug24	ai2_arc	47424		https://huggingface.co/datasets/ai2_arc
hug25	rotten_tomatoes	46131		https://huggingface.co/datasets/rotten_tomatoes

Table 3: Kaggle selected datasets (18/11/2021).

ID	Name	Download	Duplicate of	URL
kag01	Credit Card Fraud Detection	360828		https://www.kaggle.com/mlg-ulb/creditcardfraud
kag02	Novel Corona Virus 2019 Dataset	347779		https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
kag03	Video Game Sales	264773		https://www.kaggle.com/gregorut/videogamesales
kag04	Heart Disease UCI	256102	uci04	https://www.kaggle.com/ronitf/heart-disease-uci
kag05	Pima Indians Diabetes Database	235317	oml04	https://www.kaggle.com/uciml/pima-indians-diabetes-database
kag06	Iris Species	228045	uci01	https://www.kaggle.com/uciml/iris
kag07	World Happiness Report	202882		https://www.kaggle.com/unsdsn/world-happiness
kag08	Netflix Movies and TV Shows	183020		https://www.kaggle.com/shivamb/netflix-shows
kag09	The Movies Dataset	178101		https://www.kaggle.com/rounakbanik/the-movies-dataset
kag10	Breast Cancer Wisconsin (Diagnostic) Data Set	177162	uci07	https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
kag11	TMDB 5000 Movie Dataset	174636		https://www.kaggle.com/tmdb/tmdb-movie-metadata
kag12	COVID-19 Dataset	168132	kag02	https://www.kaggle.com/imdevskp/corona-virus-report
kag13	Google Play Store Apps	166169		https://www.kaggle.com/lava18/google-play-store-apps
kag14	Trending YouTube Video Statistics	158082		https://www.kaggle.com/datasnaek/youtube-new
kag15	Wine Reviews	148561		https://www.kaggle.com/zynicide/wine-reviews
kag16	Chest X-Ray Images (Pneumonia)	143227		https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
kag17	European Soccer Database	140647		https://www.kaggle.com/hugomathien/soccer
kag18	COVID-19 in India	137213		https://www.kaggle.com/sudalairajkumar/covid19-in-india
kag19	COVID-19 Open Research Dataset Challenge (CORD-19)	134256		https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Continue on next page

Table 3 – continued from previous page

ID	Name	Download	Duplicate of	URL
kag20	Students Performance in Exams	134210		https://www.kaggle.com/spscientist/students-performance-in-exams
kag21	FIFA 19 complete player dataset	130521		https://www.kaggle.com/karangadiya/fifa19
kag22	Avocado Prices	126093		https://www.kaggle.com/neuromusic/avocado-prices
kag23	House Sales in King County, USA	111243		https://www.kaggle.com/harlfoxem/housesalesprediction
kag24	Suicide Rates Overview 1985 to 2016	111006		https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
kag25	New York City Airbnb Open Data	110090		https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data
kag26	Red Wine Quality	108089		https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
kag27	Amazon Fine Food Reviews	108069		https://www.kaggle.com/snap/amazon-fine-food-reviews
kag28	Fashion MNIST	102001		https://www.kaggle.com/zalando-research/fashionmnist
kag29	Telco Customer Churn	101166		https://www.kaggle.com/blastchar/telco-customer-churn
kag30	Bitcoin Historical Data	98823		https://www.kaggle.com/mczielinski/bitcoin-historical-data

Table 4: UC Irvine Machine Learning Repository selected datasets (27/01/2022).

ID	Name	Download	Duplicate of	URL
uci01	Iris	122721		https://archive-beta.ics.uci.edu/ml/datasets/iris
uci02	Diabetes	85583		https://archive-beta.ics.uci.edu/ml/datasets/diabetes
uci03	Adult	81206		https://archive-beta.ics.uci.edu/ml/datasets/adult
uci04	Heart Disease	77318		https://archive-beta.ics.uci.edu/ml/datasets/heart+disease
uci05	Wine	63145		https://archive-beta.ics.uci.edu/ml/datasets/wine
uci06	Car Evaluation	60395		https://archive-beta.ics.uci.edu/ml/datasets/car+evaluation
uci07	Breast Cancer Wisconsin (Diagnostic)	54593		https://archive-beta.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+diagnostic
uci08	Abalone	45356		https://archive-beta.ics.uci.edu/ml/datasets/abalone
uci09	Breast Cancer	44779		https://archive-beta.ics.uci.edu/ml/datasets/breast+cancer
uci10	Mushroom	44738		https://archive-beta.ics.uci.edu/ml/datasets/mushroom
uci11	Glass Identification	40148		https://archive-beta.ics.uci.edu/ml/datasets/glass+identification
uci12	Census Income	34569	uci03	https://archive-beta.ics.uci.edu/ml/datasets/census+income
uci13	Breast Cancer Wisconsin (Original)	33993		https://archive-beta.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+original
uci14	Statlog (German Credit Data)	33688	oml01	https://archive-beta.ics.uci.edu/ml/datasets/statlog+german+credit+data
uci15	Thyroid Disease	28521		https://archive-beta.ics.uci.edu/ml/datasets/thyroid+disease
uci16	Liver Disorders	28141		https://archive-beta.ics.uci.edu/ml/datasets/liver+disorders
uci17	Optical Recognition of Handwritten Digits	27391		https://archive-beta.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits
uci18	Ionosphere	26767		https://archive-beta.ics.uci.edu/ml/datasets/ionosphere

Continue on next page

Table 4 – continued from previous page

ID	Name	Download	Duplicate of	URL
uci19	Auto MPG	26543		https://archive-beta.ics.uci.edu/ml/datasets/auto+mpg
uci20	Pen-Based Recognition of Handwritten Digits	26233		https://archive-beta.ics.uci.edu/ml/datasets/pen+based+recognition+of+handwritten+digits
uci21	Image Segmentation	24985		https://archive-beta.ics.uci.edu/ml/datasets/image+segmentation
uci22	Congressional Voting Records	24684		https://archive-beta.ics.uci.edu/ml/datasets/congressional+voting+records
uci23	Zoo	24438		https://archive-beta.ics.uci.edu/ml/datasets/zoo
uci24	Letter Recognition	23194	oml10	https://archive-beta.ics.uci.edu/ml/datasets/letter+recognition
uci25	Lung Cancer	23075		https://archive-beta.ics.uci.edu/ml/datasets/lung+cancer
uci26	Spambase	21980	oml08	https://archive-beta.ics.uci.edu/ml/datasets/spambase
uci27	Yeast	21867		https://archive-beta.ics.uci.edu/ml/datasets/yeast
uci28	Hepatitis	21415		https://archive-beta.ics.uci.edu/ml/datasets/hepatitis
uci29	Internet Advertisements	20802		https://archive-beta.ics.uci.edu/ml/datasets/internet+advertisements

Table 5: OpenML selected datasets (03/02/2022).

ID	Name	Download	Duplicate of	URL
oml01	credit-g (1)	302		https://www.openml.org/d/31
oml02	SpeedDating (1)	168		https://www.openml.org/d/40536
oml03	iris (1)	157	uci01	https://www.openml.org/d/61
oml04	diabetes (1)	104		https://www.openml.org/d/37
oml05	blood-transfusion-service-center (1)	100		https://www.openml.org/d/1464
oml06	tic-tac-toe (1)	96		https://www.openml.org/d/50
oml07	eeg-eye-state (1)	95		https://www.openml.org/d/1471
oml08	spambase (1)	93		https://www.openml.org/d/44
oml09	mnist_784 (1)	81		https://www.openml.org/d/554
oml10	letter (1)	75		https://www.openml.org/d/6
oml11	isolet (1)	71		https://www.openml.org/d/300
oml12	Satellite (1)	70		https://www.openml.org/d/40900
oml13	one-hundred-plants-texture (1)	67		https://www.openml.org/d/1493
oml14	creditcard (1)	59	kag01	https://www.openml.org/d/1597
oml15	soybean (1)	56		https://www.openml.org/d/42
oml16	waveform-5000 (1)	54		https://www.openml.org/d/60
oml17	gisette (2)	53		https://www.openml.org/d/41026
oml18	glass (1)	52	uci11	https://www.openml.org/d/41
oml19	steel-plates-fault (1)	50		https://www.openml.org/d/1504
oml20	arrhythmia (1)	50		https://www.openml.org/d/5
oml21	mammography (1)	49		https://www.openml.org/d/310
oml22	amazon-commerce-reviews (1)	48		https://www.openml.org/d/1457
oml23	electricity (1)	45		https://www.openml.org/d/151
oml24	kr-vs-kp (1)	44		https://www.openml.org/d/3
oml25	spectrometer (1)	44		https://www.openml.org/d/313
oml26	mushroom (1)	42	uci10	https://www.openml.org/d/24
oml27	Titanic (1)	42		https://www.openml.org/d/40945
oml28	bank-marketing (1)	41		https://www.openml.org/d/1461
oml29	phoneme (1)	40		https://www.openml.org/d/1489

D Dataset Documentation Reading Principles

The website of the selected datasets was analysed by manual inspection. Potential ambiguities arise from the fact that much of the information is in general text fields. For this reason, interpretation of the documentation reading is necessary and the results may be influenced by the bias introduced by the author. This section aims to provide a general overview of the interpretative choices made for each *Documentation Test Sheet* field in order to facilitate the reproducibility of this study. The readings were carried out during the month of February 2022.

D.1 Motivation

The *Motivation* section aims to provide general background information about who created the datasets, who founded them, and for what purposes. The field **1.01** *Purpose for dataset creation* has been checked as present if a clear purpose for the dataset emerged (e.g. benchmarking) or if it was possible to derive the main task for which the dataset was designed (e.g. credit scoring). The field **1.02** *Dataset creators* represents the check for the presence of information about the persons or institutions that created the dataset. There is some ambiguity on this point on platforms that allow users to upload datasets in an unmediated way, such as Kaggle. On such platforms, it is not uncommon to find datasets that provide an easier way to access data created elsewhere (e.g. datasets with information about YouTube videos) or even other datasets. In such situations it is difficult to distinguish clearly between the creator and the subject that supports, hosts or maintains the dataset (discussed in field **6.01** *Information about the subject supporting, hosting, maintaining the dataset*). During the analysis, an attempt was made to interpret ambiguous situations by trying to identify the subjects or institutions that created the dataset. This was done by defining the creation of the dataset as a process that requires some kind of collection and/or transformation in order to be used in some way. Field **1.03** *Dataset funders* suffers from the same problems as above. In fact, this field required an interpretation of institutions and companies that in some way funded the creation of the dataset. One of the elements used to determine the value field is related to acknowledgements from companies or institutions.

D.2 Composition

The *Composition* section aims to provide information about data characteristics. In order to find a value for the field **2.01** *Description of instances*, information about the value contained in a single instance was sought.

Field **2.02** *Number of instances* refers to the number of rows in the dataset.

Field **2.03** *Information about missing values* refers to the presence of some explicit references to the absence or presence of missing values and any details about symbols representing these missing values. One element to bear in mind is that if the repository does not provide an explicit metadata field which requires this information, authors may be inclined to discuss missing values only if the submitted datasets contain them. This warning should also be considered for any of the following fields that represent data issues, such as **2.05** *Description of errors, noise or redundancies*, **2.06** *Information about data confidentiality*, **2.07** *Information about possible data dangerousness (offensive, insulting, threatening or cause anxiety) or biases* and **2.09** *Description of identifiably for individuals or subpopulations (if people related)*.

Field **2.04** *Recommended data splits* represent clear suggestions about specific data division in different subgroups, such as training set, dev set and training set.

2.08 *Information about people involved in data production and their compensation (if people related)*, **2.09** *Description of identifiability for individuals or subpopulations (if people related)*, **2.10** *Description of data sensitivity (if people related)* are only applicable to those datasets that contain people-related data.

2.08 *Information about people involved in data production and their compensation (if people related)* refers to information about the people from whom the data were produced, such as demographics and whether they were remunerated in any way.

In order to check the fields **2.11** *Statistics*, **2.12** *Pair plots*, **2.13** *Probabilistic model* and **2.14** *Ground truth correlations*, graphical elements useful for clarifying the statistical properties of the data were sought.

D.3 Collection Processes

The fields related to *Collection processes* section, aims to check the presence of information related to how data has been collected.

In order to check the field **3.01** *Description of instances acquisition and data collection processes*, a description of how each data instance was collected (e.g. questionnaires, automatic scraping tool, etc.) was sought.

The field **3.02** *Information about people involved in the data collection process and their compensation* is similar to field **2.08**. *Information about people involved in data production and their compensation (if people related)*, except that in this case the focus is on those involved in data collection rather than those involved in data production.

The field **3.03** *Timeframe of data collection* was checked as present if it was clear when the data collection took place.

For field **3.04** *Information about ethical review processes* was searched for details of any discussion and outcome of any ethical review relating to the dataset.

The fields **3.05**, **3.06** are only applicable to those datasets containing people-related data. In particular, the field **3.05** *Information on individuals' knowledge of data collection (if people related)* was checked as present if the documentation discussed in some way the fact that the data producers knew about the data collection, while the field **3.06** *Information on individuals' consent for data collection (if people related)* was checked as present if it was clear that the individuals not only knew about the data collection, but also gave their explicit consent. In order to check these two fields, in the case of data collected by questionnaire, it was assumed that the subjects were aware of the data collection and had given their consent. These fields suffer from the problem that this information tends to be made explicit only in the most virtuous cases.

3.07 *Analysis of potential impacts of the dataset and its use on data subjects* reflects the presence of discussions about how the datasets may affect on the subjects who produced the data.

D.4 Preprocess, Cleaning, Labelling

Fields related to the section **4** *Preprocess, Cleaning, Labelling processes*, aims to check the presence of information on how data has been transformed.

Field **4.01** *Description of sampling, preprocessing, cleaning, labelling procedures* was checked as present if there was some reference in the documentation material to how the data had been sampled, preprocessed, cleaned or labelled.

Field **4.02** *Information about people involved in the data sampling, preprocessing, cleaning, labelling procedures* concerns instead information about the persons carrying out the procedures described in field **4.1**, following the same principles as for fields **2.7** and **3.2**.

The field **4.03** *Description of others possible sampling, preprocessing, cleaning, labelling procedures* represents the presence of details of how the data could be further sampled, pre-processed, cleaned or labelled.

D.5 Uses

Fields related to *Uses* section, aims to check the presence of information on how data has been transformed.

In order to check the presence of the information related to the first field of this section, the field **5.01** *Description of the tasks in which the dataset has already been used and their results*, references were sought to some kind of model trained on the data in the dataset, and possibly to the results obtained. In the case of platforms that allow users to share their models, such as Kaggle, OpenML and Huggingface, user-uploaded models related to the datasets were considered as examples of tasks using the dataset. In the case of the UCI Machine Learning Repository, the sometimes present "Evals" section, which shows the accuracy and precision of some classification algorithms (such as Support Vector Classification, Random Forest Classification, Logistic Regression, etc.), was considered as useful to check this field.

Fields **5.02** *Description of recommended uses or tasks* and **5.03** *Description of not recommended uses* were checked as present in the presence of some reference to the proposed tasks for using or not using the dataset.

Field **5.04** *Repository that links to papers or system that use the datasets* refers to the presence of a way to access papers that use the dataset. For the datasets included in the Huggingface repository, the presence of the paperswithcode unique id was considered useful in this regard. For the datasets included in the UC Irvine Machine Learning Repository, the webpage section "Papers citing this datasets" was considered useful in this regard, although it is not verifiable how up-to-date these lists are.

Field **5.05** *Description of license and terms of use* was considered present if at least the name of the licence was clearly displayed in the documentation.

D.6 Maintenance

Fields related to *Maintenance* section, aims to check the presence of information concerning how data has been maintained.

Field **6.01** *Information about subject supporting, hosting, maintaining the dataset* refers to the presence of information not about the creators of the dataset, but about the subject hosting and maintaining it. Within Kaggle and OpenML repositories, the subject that uploaded the dataset was considered to be the subject that is maintains it. Within the UC Irvine Machine Learning Repository, since there is a mechanism for donating datasets and the repository management is similar to that of a library, the subjects responsible for the site itself were considered to be the subjects maintaining the datasets.

To verify the existence of a **6.02** *Contact of the owner*, contacts of any kind were sought with the creators of the dataset (UC Irvine Machine Learning Repository) or with the subject who made it public (Kaggle, OpenML, Huggingface). Any means of sending a message to these individuals, such as email, in-platform messages, was considered a contact.

Field **6.03** *DOI* was checked as present if a DOI number was given.

Field **6.04** *Erratum* refers to the presence of an erratum: this field suffers from the same problem presented for data issues presented in section **2** *Composition.*

A relevant aspect of **6.05** *Information about dataset updates* field checking concerns the fact that, for the Huggingface repository, the commit information present in the Files and Versions tab has been taken into account.

Field **6.06** *Information about management of older dataset versions* refers to the presence of information about older versions of the dataset that are still available or retired.

Field **6.07** *Information about the mechanism for extending, augmenting, building upon, contributing to the dataset* has been checked as present where the presence of a mechanism for these purposes is conspicuous, such as the GitHub link present in datasets within the Huggingface repository.

D.7 Characteristics

Field **c.01** *Data is people related* was checked as True, if the dataset contains people-related data. In order to check this characteristic, an interpretation was required. For this purpose, has been taken into account the recommendation, provided by Gebru et al. [2], to take a broad interpretation of whether a dataset relates to people. To give an example, the authors suggest that any dataset that contains text written by people relates to people.

During the documentation analysis, this recommendation was not taken literally, especially for NLP specific datasets (typical Huggingface). In order to clarify the interpretation method, some examples will follow. Datasets containing any kind of user text production such as reviews, search engine queries, etc.

were considered as people-related data. Datasets containing Wikipedia texts, newspaper articles, book corpus, school examination texts or ad-hoc text production were not. The provenance is not always clear, but an attempt was made to infer the specificity of the data from the context. Image recognition of handwritten letters or numbers was not considered as people-related data. Datasets containing medical information about patients were considered as personal data.

Field **c.02** *Presence of explicit target variable* was checked as True, if the dataset contains an explicit target variable.

Field **c.03** *Dataset is a sample or a reduction of a larger set* was checked as True if the dataset is a subset of another dataset.

Finally, field **c.04** *Recently updated* was checked as True if the dataset was published or updated after 01/01/2021. The Huggingface update date is the date given in the *Files and versions* tab of the dataset web page. The Kaggle update date is the date obtained from the *lastUpdated* field via the APIs. The OpenML update date refers to the publication date given on the website. The UC Irvine Machine Learning Repository update date refers to the *Donated on* date given on the website.

E Raw Data

In the following tables, each column represents a dataset, while each row represents a *Test Field*. The cell referring to dataset *d* and *Test Field f* contains one of the three possible *Presence Check Values*.

E.1 Huggingface

Recommended information				Datasets																									
Section	**Sec %**	**Field ID**	**Field %**	**hug01**	**hug02**	**hug03**	**hug04**	**hug05**	**hug06**	**hug07**	**hug08**	**hug09**	**hug10**	**hug11**	**hug12**	**hug13**	**hug14**	**hug15**	**hug16**	**hug17**	**hug18**	**hug19**	**hug20**	**hug21**	**hug22**	**hug23**	**hug24**	**hug25**	**Repo AVG**
Motivation	0.56	1.01	0.64	1	1	1	0	1	0	1	0	1	1	1	0	1	0	0	1	1	0	1	1	1	1	0	1	0	0.36
		1.02	0.88	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		1.03	0.16	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	
Composition	0.28	2.01	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		2.02	0.92	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	
		2.03	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.04	0.92	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	
		2.05	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.06	0.08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	
		2.07	0.12	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	
		2.08	0.25	NA	0	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	
		2.09	0.50	NA	1	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	
		2.10	0.25	NA	0	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	
		2.11	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.12	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.13	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.14	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Collection processes	0.19	3.01	0.52	0	0	0	1	0	1	0	1	1	1	1	1	0	0	1	1	1	0	0	0	0	1	1	0	1	
		3.02	0.16	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	
		3.03	0.04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
		3.04	0.04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
		3.05	0.25	NA	0	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	
		3.06	0.25	NA	0	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	
		3.07	0.00	NA	0	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	
Preprocess / cleaning / labelling	0.25	4.01	0.32	0	0	0	1	0	0	0	0	1	0	1	1	0	0	1	1	1	0	0	0	0	1	0	0	0	
		4.02	0.44	0	0	0	1	1	1	0	0	1	0	1	1	0	0	1	1	1	0	0	0	0	1	0	1	0	
		4.03	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Uses	0.59	5.01	0.92	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	
		5.02	0.56	1	0	0	1	1	1	0	0	1	0	1	1	0	0	1	1	1	1	0	0	0	1	1	1	0	
		5.03	0.08	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
		5.04	0.92	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	
		5.05	0.48	1	0	0	1	1	1	0	0	1	0	1	1	0	0	1	1	1	0	0	0	0	1	0	1	0	
Maintenance	0.40	6.01	0.36	0	0	0	0	0	0	1	0	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1	1	
		6.02	0.20	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	
		6.03	0.24	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	
		6.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.05	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		6.06	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.07	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Datasets AVG				0.30	0.23	0.24	0.42	0.27	0.39	0.26	0.27	0.61	0.33	0.48	0.45	0.30	0.27	0.56	0.73	0.42	0.27	0.27	0.27	0.27	0.45	0.30	0.39	0.23	
Charateristics		c.01	0.16	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	
		c.02	0.84	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	
		c.03	0.04	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		c.04	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	

E.2 Kaggle

Recommended information				Datasets																									
Section	Sec %	Field ID	Field %	kag01	kag02	kag03	kag07	kag08	kag09	kag11	kag13	kag14	kag15	kag16	kag17	kag18	kag19	kag20	kag21	kag22	kag23	kag24	kag25	kag26	kag27	kag28	kag29	kag30	Repo AVG
Motivation	0.52	1.01	0.52	1	1	0	1	0	0	0	0	0	1	1	1	1	1	0	0	0	0	1	0	1	0	1	1	1	0.31
		1.02	0.96	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	
		1.03	0.08	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	
Composition	0.28	2.01	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		2.02	0.72	1	0	1	1	0	1	1	1	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	
		2.03	0.12	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	
		2.04	0.08	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
		2.05	0.16	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	
		2.06	0.08	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
		2.07	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.08	0.42	0	0	NA	0	NA	0	NA	NA	NA	1	1	NA	0	0	1	NA	NA	NA	1	NA	NA	0	NA	1	NA	
		2.09	0.17	1	0	NA	0	NA	0	NA	NA	NA	1	0	NA	0	0	0	NA	NA	NA	0	NA	NA	0	NA	0	NA	
		2.10	0.00	0	0	NA	0	NA	0	NA	NA	NA	0	0	NA	0	0	0	NA	NA	NA	0	NA	NA	0	NA	0	NA	
		2.11	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		2.12	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.13	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.14	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Collection processes	0.22	3.01	0.60	0	0	1	1	0	1	1	1	1	1	1	1	0	1	0	1	1	0	1	1	0	0	1	0	0	
		3.02	0.12	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	
		3.03	0.48	1	1	0	1	0	0	0	0	0	1	0	0	1	1	0	1	1	1	0	1	0	1	0	0	1	
		3.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		3.05	0.00	0	0	NA	0	NA	0	NA	NA	NA	0	0	NA	0	0	0	NA	NA	NA	0	NA	NA	0	NA	0	NA	
		3.06	0.00	0	0	NA	0	NA	0	NA	NA	NA	0	0	NA	0	0	0	NA	NA	NA	0	NA	NA	0	NA	0	NA	
		3.07	0.00	0	0	NA	0	NA	0	NA	NA	NA	0	0	NA	0	0	0	NA	NA	NA	0	NA	NA	0	NA	0	NA	
Preprocess / cleaning / labelling	0.09	4.01	0.24	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	1	0	1	
		4.02	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		4.03	0.04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
Uses	0.48	5.01	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		5.02	0.72	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0	1	0	0	
		5.03	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		5.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		5.05	0.68	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	
Maintenance	0.34	6.01	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		6.02	0.80	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	
		6.03	0.04	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.05	0.52	1	1	0	0	1	0	0	0	0	1	1	1	0	1	0	1	1	0	0	0	1	0	1	1	1	
		6.06	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.07	0.04	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
Datasets AVG				0.36	0.26	0.27	0.28	0.27	0.26	0.30	0.27	0.27	0.41	0.38	0.45	0.26	0.31	0.23	0.36	0.33	0.30	0.33	0.33	0.39	0.18	0.39	0.26	0.42	
Charateristics		c.01	0.48	1	1	0	1	0	1	0	0	0	1	1	0	1	1	1	0	0	0	1	0	0	1	0	1	0	
		c.02	0.32	1	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	1	0	1	1	0	
		c.03	0.08	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		c.04	0.20	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	

E.3 OpenML

Recommended information				Datasets																									
Section	**Sec %**	**Field ID**	**Field %**	**oml01**	**oml02**	**oml03**	**oml04**	**oml05**	**oml06**	**oml07**	**oml08**	**oml09**	**oml10**	**oml11**	**oml12**	**oml13**	**oml14**	**oml15**	**oml17**	**oml18**	**oml19**	**oml20**	**oml21**	**oml22**	**oml23**	**oml24**	**oml25**	**oml29**	**Repo AVG**
Motivation	0.56	1.01	0.68	1	0	1	1	0	0	1	1	1	1	1	0	0	0	1	1	1	0	1	1	1	1	0	1	1	0.32
		1.02	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		1.03	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Composition	0.35	2.01	0.80	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	0	0	1	1	1	1	1	1	
		2.02	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		2.03	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		2.04	0.12	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
		2.05	0.08	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	
		2.06	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.07	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.08	0.64	0	1	1	1	NA	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	0	1	NA	NA	NA	1	1	NA	
		2.09	0.09	0	0	0	0	NA	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	0	0	NA	NA	NA	0	0	NA	
		2.10	0.00	0	0	0	0	NA	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	0	NA	NA	NA	0	0	NA	
		2.11	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		2.12	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.13	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.14	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Collection processes	0.16	3.01	0.64	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	
		3.02	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		3.03	0.12	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
		3.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		3.05	0.09	0	1	0	0	NA	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	0	NA	NA	NA	0	0	NA	
		3.06	0.09	0	1	0	0	NA	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	0	NA	NA	NA	0	0	NA	
		3.07	0.00	0	0	0	0	NA	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	0	NA	NA	NA	0	0	NA	
Preprocess / cleaning / labelling	0.20	4.01	0.56	0	0	1	1	0	0	1	1	1	0	1	1	1	0	1	0	1	0	1	1	0	1	0	0	1	
		4.02	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		4.03	0.04	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Uses	0.53	5.01	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		5.02	0.64	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1	1	0	1	0	0	1	0	1	1	
		5.03	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		5.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		5.05	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Maintenance	0.18	6.01	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		6.02	0.16	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	
		6.03	0.08	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.05	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.06	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.07	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Datasets AVG				0.26	0.38	0.33	0.36	0.30	0.23	0.28	0.42	0.39	0.33	0.36	0.30	0.30	0.21	0.36	0.30	0.33	0.18	0.31	0.36	0.33	0.36	0.26	0.31	0.39	
Charateristics		c.01	0.44	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	0	
		c.02	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		c.03	0.28	0	0	1	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0	0	0	
		c.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

E.4 UC Irvine Machine Learning repository

Recommended information				Datasets																									
Section	**Sec %**	**Field ID**	**Field %**	**uci01**	**uci02**	**uci03**	**uci04**	**uci05**	**uci06**	**uci07**	**uci08**	**uci09**	**uci10**	**uci11**	**uci13**	**uci15**	**uci16**	**uci17**	**uci18**	**uci19**	**uci20**	**uci21**	**uci22**	**uci23**	**uci25**	**uci27**	**uci28**	**uci29**	**Repo AVG**
Motivation	0.35	1.01	0.44	1	0	0	1	1	1	0	1	0	1	1	0	0	0	1	0	0	1	0	0	0	1	1	0	0	0.27
		1.02	0.60	1	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	0	1	0	0	0	1	0	0	1	
		1.03	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Composition	0.26	2.01	0.88	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	
		2.02	0.96	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		2.03	0.88	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		2.04	0.12	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	
		2.05	0.28	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0	1	0	
		2.06	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.07	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.08	0.31	NA	0	1	1	NA	0	0	0	0	NA	NA	0	0	0	NA	NA	NA	NA	NA	1	NA	0	NA	1	NA	
		2.09	0.08	NA	0	0	1	NA	0	0	0	0	NA	NA	0	0	0	NA	NA	NA	NA	NA	0	NA	0	NA	0	NA	
		2.10	0.00	NA	0	0	0	NA	0	0	0	0	NA	NA	0	0	0	NA	NA	NA	NA	NA	0	NA	0	NA	0	NA	
		2.11	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.12	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.13	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		2.14	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Collection processes	0.09	3.01	0.36	0	1	0	0	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	1	0	0	0	0	0	
		3.02	0.04	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
		3.03	0.12	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
		3.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		3.05	0.00	NA	0	0	0	NA	0	0	0	0	NA	NA	0	0	0	NA	NA	NA	NA	NA	0	NA	0	NA	0	NA	
		3.06	0.00	NA	0	0	0	NA	0	0	0	0	NA	NA	0	0	0	NA	NA	NA	NA	NA	0	NA	0	NA	0	NA	
		3.07	0.00	NA	0	0	0	NA	0	0	0	0	NA	NA	0	0	0	NA	NA	NA	NA	NA	0	NA	0	NA	0	NA	
Preprocess / cleaning / labelling	0.15	4.01	0.44	0	0	1	1	0	0	1	1	0	0	0	1	0	0	1	1	1	1	1	0	0	1	0	0	0	
		4.02	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		4.03	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Uses	0.69	5.01	0.88	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	
		5.02	0.56	0	0	1	0	1	1	0	1	1	1	1	0	0	1	0	1	0	1	0	1	0	1	1	0	1	
		5.03	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		5.04	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		5.05	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Maintenance	0.15	6.01	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		6.02	0.04	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.03	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.05	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.06	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
		6.07	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Datasets AVG				0.30	0.21	0.26	0.28	0.33	0.23	0.26	0.26	0.23	0.27	0.30	0.33	0.23	0.23	0.33	0.30	0.24	0.39	0.27	0.26	0.24	0.28	0.27	0.23	0.27	
Charateristics		c.01	0.44	0	1	1	1	0	0	1	0	1	0	0	1	1	1	0	0	0	0	0	1	0	1	0	1	0	
		c.02	0.92	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	
		c.03	0.16	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	
		c.04	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

F Additional Tables and Figures

It is possible to observe the different distributions among repositories of the *Section Presence Averages* in Figure 2. It can be observed that in all repositories the *Motivation* section is the one with higher variance in presence, followed by *Collection processes* and *Data processing procedures*: the fact suggests that compensation effects between high values and low values should be taken into account when analysing the average values. At the opposite extreme, the *Maintenance* and *Composition* sections have less dispersion, indicating more homogeneous results of the sampled data within each repository. We also observe that *Composition* in Huggingface and *Maintenance* in UCI have most measurements close to the average value (respectively, 0.28 and 0.15).

Finally, the sections of the datasets were plotted (Figure 3) in a plane in which the x-axis represents the *Dataset Presence Average* and the y-axis the *Section Presence Average*. This type of representation is useful in order to understand how the distribution of values in each individual section varies in relation to the overall completeness of the dataset documentation. Looking at the data from this point of view, the focus on utilisation-related information once again emerges: Figure 3e shows the high quantity of individual datasets with a low value of *Dataset Presence Average* and a high value of *Uses Section Presence Average*.

In addition, frequency histograms of the *Dataset Presence Average* and of the *Section Presence Average* can be observed in the plot. These histograms, on one hand, show a Gaussian distribution for *Dataset Presence Average* and for the *Section Presence Average* of sections *Motivation* (fig. 3a), *Composition* (fig. 3b) and *Uses* (fig. 3e). On the other hand, sections *Data processing procedures* (fig. 3d) and *Maintenance* (fig. 3f) show a decreasing frequency as their *Presence Average* increases.

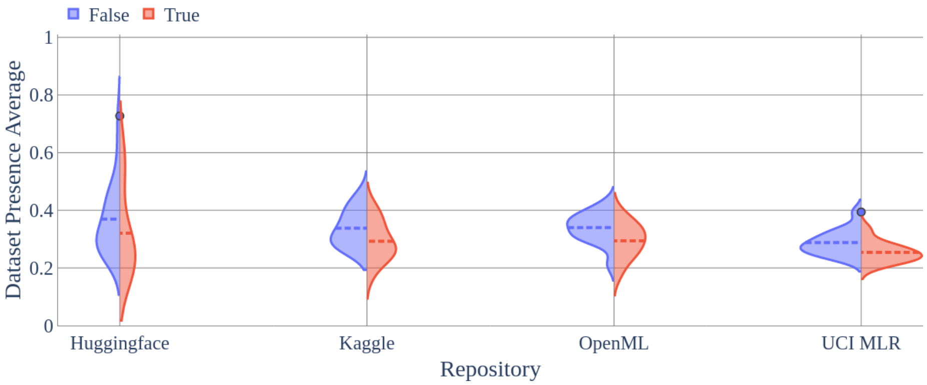
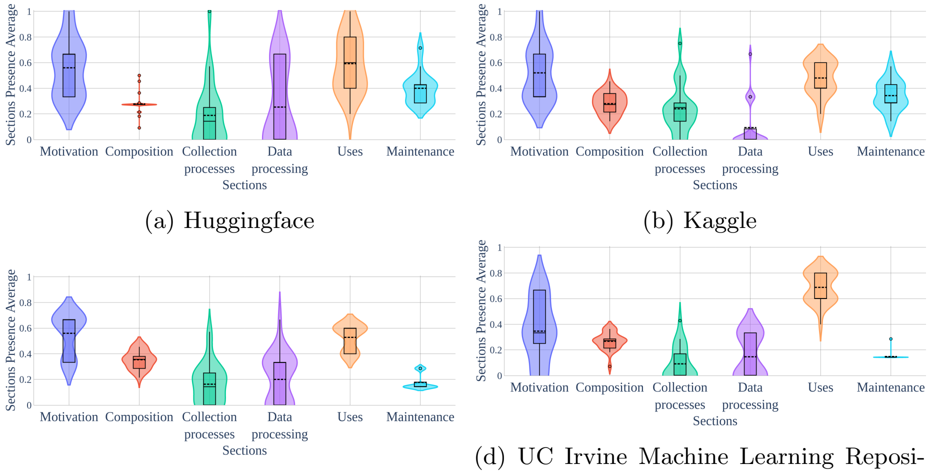


Fig. 1: Distribution of *Dataset Presence Averages* values, grouped by repository, according to the presence of people-related data. The dotted line indicates the mean value.



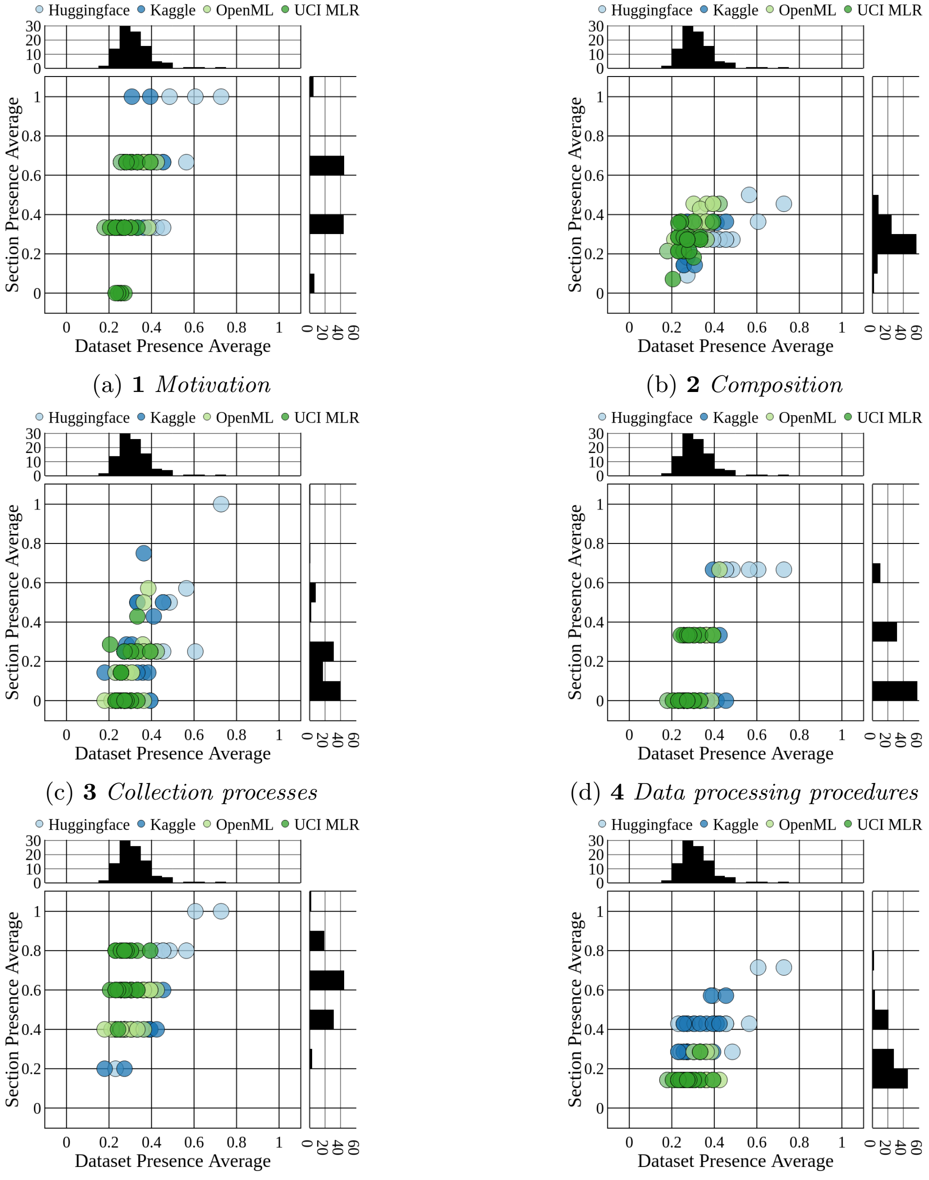
(a) Huggingface

(b) Kaggle

(c) OpenML

(d) UC Irvine Machine Learning Repository

Fig. 2: *Sections Presence Average* distributions. For each repository and section, the violin plot shows the frequency of each *Section Presence Average* value. The underside of the box plot indicate the q1 value, while the upper one represent the q3 value (linear interpolation method). The dotted line indicates the mean, while the solid line represents the median. The whiskers reach the minimum/maximum value below/above the q1/q3 value for a maximum distance of 1.5 of the interquartile range. Data points outside the latter ranges, are considered outliers.



(a) **1** *Motivation*

(b) **2** *Composition*

(c) **3** *Collection processes*

(d) **4** *Data processing procedures*

(e) **5** *Uses*

(f) **6** *Maintenance*

Fig. 3: *Dataset Presence Averages* and *Section Presence averages*. Each dot represents a section of a dataset, plotted with an x-value correspondent to the *Dataset Presence Average* (frequency histogram on the top of the main plot) and a y-value correspondent to the *Section Presence Average* (frequency histogram on the right of the main plot). Each plot contains exactly 100 dots (one for each dataset under analysis).

References

1. Bender, E.M., Friedman, B.: Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics **6**, 587–604 (2018). https://doi.org/10.1162/tacl_a_00041
2. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., III, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (Nov 2021). https://doi.org/10.1145/3458723
3. Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K.: The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677 [cs] (May 2018). https://doi.org/10.48550/arXiv.1805.03677