POLITECNICO DI TORINO

Master's Degree in
Computer Engineering - Data Analytics

Politecnico
di Torino

Master's Degree Thesis

Ethical Manufacturing of Datasets for Artificial Intelligence: an Empirical Investigation into the State of Documentation Practice

Supervisor

Prof. Antonio VETRÒ

Co-Supervisor

Prof. Juan Carlos DE MARTIN

Candidate

Marco RONDINA

Academic year 2021-2022

Abstract

Artificial Intelligence research and industrial developments have made great strides in recent years becoming increasingly pervasive within society, given the diffusion of AI applications in relevant fields (like medicine, banking, welfare, insurances, etc.) especially with the aim of automating processes and decisions.

One of the key elements of AI-based technologies is data, which play a central role in the quality of software outcomes [27]. It is therefore becoming increasingly important to ensure that AI practitioners are fully aware of the quality of datasets and of the process generating them [61], in such a way that all the –typically implicit– assumptions, ethical issues, modelling choices [11] clearly and transparently emerge, and their impact to downstream effects can be tracked, analysed, and possibly mitigated [67].

One of the tools that can be useful in this perspective is dataset documentation [12]. Dataset documentation represents the key form of communication between datasets builders and datasets users. Through this tool it is indeed possible to let the context in which data was produced and transformed emerge, bringing to the attention of all stakeholders relevant facts on data properties and data procedures. Therefore, dataset documentation can reveal potential data ethical issues, making it possible to manage them and reinforcing users' trust. In this perspective, an accurate dataset documentation can promote the achievement of transparency, accountability and auditability targets, foster reproducibility and avoid data cascade effects on the entire AI pipeline.

The first aim of this work of thesis was to draw up a scheme of the relevant information that should always be attached to a dataset, in order to ensure a proper choice of dataset and an informed use by professionals. To this aim, a set of essential information useful to build a good dataset documentation has been identified through a review of the available literature. The research started from published proposals for standardizing documentation. The structure of the scheme mainly follows the subdivision presented in *Datasheets for Datasets* [26], which contains a list of questions useful to drive the documentation writing by dataset builders. These questions were further integrated with additional contents proposed in *Data Statements for Natural Language Processing* [8] and *The Dataset Nutrition Label*

[33]. The resulting questions were summarized and collapsed in "fields", each representing an essential piece of information to be included in dataset documentation. Moreover, particular attention has been paid to the generalizability of the proposed scheme to any type of documentation, so that it can be used for datasets pertaining different areas of artificial intelligence. Fields were constructed in a short and easily readable formulation to make it easier the check of the presence (or not) of the relative information in the analysed documentation. The presence has been translated in a 1 value, while the absence in a 0 value. The scheme produced according to these principles has been named **Documentation Test Sheet**. This tool provides a measure of the completeness of the documentation based on the average value of the information present in the grouping under investigation (repository, dataset, section or field), thus providing a good and easy indicator to understand one aspect of dataset documentation quality.

Next step consisted in the application of the Documentation Test Sheet to the most popular datasets in the AI community. To this aim, four different repositories were selected (Huggingface, Kaggle, OpenML and UC Irvine Machine Learning Repository). Within each of them, the top 25 datasets were chosen, using dataset popularity as sorting criterion. This criterion was adapted to the specific metadata available from each repository. Duplicates were excluded, discarding datasets already present in other repositories. The aim was to assess how readily accessible this information was in the very same place where the data can be accessed. For this reason, the research was focused on the analysis of the dataset description pages in the hosting repositories. Automatic assessment was investigated, checking the match between repositories metadata fields and documentation test sheets fields. Since it led to inaccurate or incomplete results, it was integrated with manual checking.

The results were analysed with mixed methods (qualitative and quantitative) that allowed the identification of some correlations between the available documentation and dataset characteristics. First, the most documented section was that relative to the use of datasets, reinforcing the hypothesis that the main focus of AI practitioners is on models. Next sections of information, ordered by completeness, resulted dataset purposes, data characteristics and maintenance over time. On the contrary, sections about data collection procedures and data preprocessing procedures resulted very poorly documented.

Some of the best documented fields concerned features description, the number of the instances and description of tasks in which the dataset has already been used. On average, datasets containing people-related data showed equal or even less detailed documentation compared to other datasets. Nevertheless, it is possible to observe some positive correlations between the presence of people-related data and the presence of information on subjects that maintain the dataset over time.

Overall, a lack of relevant information was observed, highlighting a paucity

of transparency. This observation is even more significant considering that the analysis was restricted to some of the most popular and well-known datasets, although no correlation between the popularity ranking of the datasets and the completeness of the documentation were identified. Another evidence that emerged from the data concerned the potential of repositories to help dataset publishers to produce better documentation. Indeed, when a specific piece of information was present in approximately all the datasets of a repository, but it was hardly found in other repositories, the causes were often ascribable to the structure of the datasets' metadata scheme offered by the repository itself.

The scheme here proposed represents a useful tool to improve transparency and accountability. On the one hand it can be used by dataset hosts and dataset users to check the completeness of a documentation quickly and simply. On the other hand, it can serve as a guideline for dataset creators, helping them to improve their documentation so that dataset consumer can verify the underlying choices and assumptions, data procedures and, more generally, the context in which the dataset was produced. The recommended path should be supported by the investigation and experimentation of techniques to fully integrate documentation models and processes into the AI pipeline, in order to promote transparency, accountability, auditability and avoid data cascade effects on the entire AI pipeline.

Quantitative expansion of this research could be put in place expanding the preliminary work on the feasibility of an automatic system capable of controlling the information presence. From the qualitative point of view, it might be fascinating to expand the Documentation Test Sheet in order to include (measuring them) other aspects of documentation quality (e.g., sparsity).

Altogether, these results show that huge efforts of the AI community in devoting more attention to the dataset documentation process are urgent and necessary. There are no purely technical aspects, and every technical choice that led to the construction of a given model hides a set of ethical considerations, regardless of whether the context is considered or not.

Table of Contents

List of Tables

List of Figures

Acronyms

AI Artificial Intelligence. 3, 4, 11–16, 19–21, 50, 91

APIs Application Programming Interfaces. 31, 42, 44, 51, 52, 62

DTS Documentation Test Sheet. 24, 40, 47, 48, 58, 91–93

NLP Natural Language Processing. 14, 21, 32, 33

UCI MLR UC Irvine Machine Learning Repository. 7, 48, 49, 55, 60–62

Chapter 1

Introduction

Artificial Intelligence research and industrial developments have made great strides in recent years becoming increasingly pervasive within society, given the diffusion of AI applications in relevant fields, especially with the aim of automating processes and decisions. One of the key elements of AI-based technologies is data, which play a central role in the quality of software outcomes [27]. It is therefore becoming increasingly important to ensure that AI practitioners are fully aware of the quality of datasets and of the process generating them [61], in such a way that all the –typically implicit– assumptions, ethical issues, modeling choices [11] clearly and transparently emerge, and their impact to downstream effects can be tracked, analysed and possibly mitigated [67]. One of the tools that can be useful in this perspective is dataset documentation [12].

This thesis will investigate the relevant information that should always be attached to a dataset in order to ensure a correct choice of dataset and an informed use by practitioners, with the aim of conducting an empirical investigation on dataset documentation state of practice.

This chapter will introduce the research work providing the context in which it was constructed. Section 1.1 will provide a background overview of what is artificial intelligence. Then, will be presented an overview of AI industry from different points of view in section 1.2, followed by a focus on the data containers at the basis of the whole pipeline: datasets (section 1.3). Once clarified the background, research questions at the basis of the research design will be presented in section 1.4. Finally, section 1.5 will describe the thesis structure.

1.1 Background: What is Artificial Intelligence?

The digital revolution radically changed our society. Since time immemorial, mankind has been looking for ever more efficient ways to calculate. Over the centuries, the set of scientific discoveries and technological applications led to an increasing computing power, paving the way for the processing of a growing amount of information. This type of evolution has broadened the scope of computing machines, making it possible to imagine increasingly intelligent automation. It is possible to observe thoughts related to the general idea of a 'reasoning machine' many centuries ago, for instance in the works of Ramon Llull and Gottfried Willhelm von Leibniz [30]. Although this, in the mid-20th century Artificial Intelligence began to take shape as a field of academic research. Machine learning begin to flourish in the 1980s, while the new millennium paved the way to new techniques capable to automatically learn representations from data, enabling the deep learning revolution[1]. Over the decades, research and technological progress have enabled huge advances in this area, allowing computers to «changing the way they carry out tasks by learning from new data, without a human being needing to give instructions in the form of a program» [2].

The impact of artificial intelligence, however, goes far beyond academia. The economic world was not slow to exploit it in order to reduce costs and increase profits. An increasing number of commercial applications are nowadays based on AI systems, impacting on power relations within society.

To introduce a universal definition of Artificial Intelligence is not an immediate task. The European Commission, in the *Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, define an AI systems as «software that is developed with one or more of the techniques and approaches listed [...] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with» [64]. This formulation helps demystify the sci-fi conception of AI systems that can often be encountered in public media communication, policies and public narratives [56, 48, 59, 57]. Indeed, when we talk about artificial intelligence, we are therefore referring to software designed and developed by human software developers. As an artifact, characteristics of artificial intelligence are human crafted, with some uses enabled (intended or not), and others inhibited [24].

Without going into the details of machine learning techniques and theories,

[1] `https://developer.nvidia.com/deep-learning`

[2] `https://dictionary.cambridge.org/dictionary/english/machine-learning`

their operation is based on data pattern recognition. The way in which this data pattern work (in terms of theory, data structures, algorithm, and so on) constitutes the *model*. The internal representation of data pattern will then be useful for performing a given task on other data instances not already encountered. This basic fundamental principle at the basis of all AI software production introduce the basic raw materials needed to feed the algorithms: data.

1.2 AI industry

Before going into specifics regarding data and datasets, this section will present a general overview of the artificial intelligence industry. This has been done for several reasons: firstly, because it makes it easier to approach the topics addressed by this thesis, and secondly, because several fundamental aspects of the research are grounded in the very structure of this industry.

Artificial Intelligence industry is far from being abstract and immaterial. As mentioned in the previous section, the two main components can be individuated in data and model. But these are not enough. Data must be produced first, then collected and finally transformed in order to adapt them to the specific context. Model must be developed, trained and then prepared to be put into production.

The first characteristic of all these processes is that they are strongly based on human work. Casilli [13] talks about the different types of digital labour and the capitalist coordination behind them, oriented to the automation of production processes and to foster human actions able to produce value through data. Most models, for example, require labelled data. Micro-task crowd-sourced platform such as Amazon Mechanical Turk are increasingly becoming the answer, promoting the *taskification* of work: parcelling out, outsourcing, precarisation. This has certainly reduced costs (i.e. increased profits), at the expense of a new class of workers to whom intermediation has made it easier to reduce already acquired rights. The data quality itself is directly affected by compensation rates [47, 71]. From a geographic point of view, moreover, often this exploitation is mainly concentrated in poorest area of the planet, configuring a new form of Western imperialism. As can be guessed, the impacts go beyond the industry of producing artificial intelligence applications: these transformations of digital work are leading to major changes in society as a whole.

The second characteristics concerns the AI sustainability. As narrated by Crawford in *The Atlas of AI*, the "artificial intelligence" lexical expression may lead to imagine a set of predominantly intangible artefacts, such as algorithms, data, cloud architectures. Nevertheless, «none of that can function without the minerals and

resources that build computing's core components» [15]. In addition to capital and labour, Earth's resources are one fundamental element of the AI supply chain. Devices and data center able to provide the fundamental computational power exploit a great number of rare minerals which required billions of year to be formed. Furthermore, the energy required to train model is significant: Dhar estimated that the carbon footprint of training procedures related to a large NLP model can be approximated to 125 round-trip flights between New York and Beijing [18]. During the lifecycle of an AI system, the waste treatment final phase is not less problematic, both environmentally and geopolitically. This, considering the wide spread of e-waste dumping grounds in places like Ghana and Pakistan [15]. All these aspects play an even more important role when contextualised in the midst of the most serious climate crisis of our history [43, 39], making it necessary to adopt policies to reduce AI's negative impacts on the climate [1].

In addition to human labor and on planet resources, AI industry relies heavily on data [3].

One first aspect concerns the massive extraction of data. Morozov introduced the concept of the «lastest incarnation of capitalism» referring to the centralisation of power conducted by the greatest digital companies. Indeed, personal data are one of the biggest sources of profit in the digital era, with considerable consequences in terms of privacy and freedom. The power relationship reflects the fact that in the vast majority of cases, data sovereignty does not belong to those who produce it, but to those who exploit it. From this point of view interesting proposals in terms of digital public goods are taking shape in activist and academic communities around the world [19].

A further relevant aspect concerns all those who are affected, positively or negatively, by the decisions of an automated system managed by some model of artificial intelligence. Technicalities, and more in general every decision during the workflow, can be affected by subjectivity and can incorporate implicit values and beliefs. To make an example, in the context of a dataset with the purpose of train a facial recognition software, the way in which photos in a dataset are taken can have a significant impact on the results, as described by Scheuerman et al. in [68] about the work of Wu and Zhang [84]. The impact of technicalities does not concern only data, but also algorithms: for this reason, impact assessments of algorithms are considered increasingly necessary, especially when these algorithms are exploited by public institutions [37]. Here again, despite the supposed intangibility of this industry, the impacts on people can be very practical, painful and discriminatory [22]. «Algorithms are opinions embedded in code» as stated by O'Neil [58].

For this reason, different approaches can be pursued to identify and reduce risks, such as the data quality one [82, 81, 50]. It is necessary to adopt a power-aware perspective, able to put together data quality, data work and data documentation

[52]. In order to do so, interdisciplinary competencies in the fields of computer science and ethics are useful in order to obtain better ethical evaluations [29], using available analytical framework like the one presented by Hanley et al. [32]. During the developing of technologies with such high stakes on users, it is important to understand the concept of fairness in that specific context [10] and to take into account all the information needed in order to achieve fair results [70], from both technical and non-technical point of view. This, even if the outcome of these operations converges on the fact that the best solution to the problem is not technological [79]. It is necessary also to imagine the consequences of the so-called «Ripple effect», as introduced by Selbst et al. [70]: the insertion of technology into an existing social system changes the embedded behaviours and values of the pre-existing system.

All these aspects, the exploitation of data, earth resources and human labor, in addition to the risks and impacts for people, make it necessary regulation policies of all the AI ecosystem [16, 64, 38, 2, 78].

1.3 Raw materials: datasets

In the pipeline underlying the artificial intelligence industry, data is one of the most significant elements, since they allow models to learn patterns useful to refine prediction on not already known data. They can be considered the core element, since simple models with more data achieve better results rather than complex model with less data [31]. As already described in the previous sections, the amount of data needed is always bigger and this data hunger raises new issues. Datasets are not simply operational instruments of digital knowledge production and, for this reason, it is necessary to «bring people back in» [76].

In order to obtain the fair results described in the previous section, data should be [68]:

- Diverse and varied;
- Unbiased;
- High-quality;
- Realistic;
- Challenging;
- Comprehensive and large-scale;

In most cases, except unsupervised learning, data should be annotated in order to train model that want to perform a classification task. To annotate data means

classifying them, providing a *label* that the model can consider the *ground truth*. One desired property of labeling is objectivity: however, the subjectivity of human annotation and the power relationships related to the work context of annotators make this a very ambitious goal [83, 53]. Not only because the implicit personal bias can influence that work, but also because the very fact of considering a reading of reality as objective is often misleading [85]. The way the translation from high-level strategic objective to practical application is carried out can lead to different results, with the strong influence of the decision-maker [60].

Many issues related to fairness, transparency and accountability in AI systems are rooted in the data collection and annotation procedures, to the extent that proposals have emerged for professionals specifically dedicated to these delicate stages of development [40]. The information accompanying a dataset plays a very significant role in uncovering data issues [12], in fostering reproducibility and auditability [42], in ensuring accountability [36], users' trust [5] and in avoiding *data cascading* effects on the entire AI pipeline [67]. With documentation, it is possible to understand training data characteristics and use this information in order to at least partially mitigate attested and unknown risks [9]. Documentation production should be considered a significant part of datasets manufacturing, as a place to disclose basic decisions and facts in parallel with what is proposed to be documented with respect to models [54, 66] or rankings [86, 88]. While interesting proposal on post hoc documentation are catching on [21, 6], this procedure it's not always feasible, especially in case of very large datasets, and can lead to documentation debt situations [87].

For these reasons, this thesis work will focus on datasets' documentation state of practice. The aim is to understand which information should always be clear to datasets' stakeholders in order to mitigate risks and then measure how much this information is present in the documentation of the most popular (and influential [4]) datasets. Datasets cover a very important role in the AI ecosystem, since they are the first tangible result of the translation from the problem formulation to the practice application [7]. Often the focus on ethical manufacturing of datasets for artificial intelligence, require compromises to other desired values during development, as described by Scheuerman et al. in «Do Datasets Have Politics?» [68]. Indeed, documentation care, debiasing, quality control and labor rights for data labor require a *care* focus over *efficiency* in some cases. To make another example, to recognize that it is impossible to standardize a classification about the world [15] require that *contextuality* acquire more importance than *universality*. All these efforts are necessary in order to put people, rights and democracy before innovation, efficiency and technology [44]. The actual practice of industry practitioners is slightly different from the fair ML research literature [34], but in order to guarantee people safety probably «there is another way to make data sets better» [77].

1.4 Research questions

On the basis of the principles analysed in the previous sections, the research work started from the definition of the following research questions.

RQ1: What information should be transparent to dataset users?
To answer this research question, it is first necessary to analyse the relevance of the documentation accompanying a dataset. It is necessary to identify what information should be present in the documentation, in order to facilitate a conscious use of the dataset. This set of information must be structured in a way useful to simplify the information presence check within the documentation, typically by turning the description of the individual field of information into a question that can be answered with yes, if this information is present or not, if this information is not present in the analysed documentation.

RQ2: Which of the information that should be transparent to dataset users, is present in the most popular datasets?
To answer this research question, it is first necessary to define quantitatively the meaning of 'most popular'. Next, the test derived from RQ1 will be applied to the most popular datasets. The aim is to understand the extent to which the most popular datasets in the AI community are complete (or lacking) in terms of documentation.

1.5 Thesis structure

Following this introductory chapter, the thesis unfolds along four chapters.

Chapter 2 contains the analysis to answer RQ1. Here, the intrinsic value of the documentation that accompanies a dataset will first be studied, with an overview of the main standardisation proposals in the academic literature. Subsequently, a possible adaptation of the aforementioned standardisation proposals will be shown, in a scheme of information useful for composing a test of documentation completeness.

Chapter 3 contains the analysis related to RQ2. Within this chapter, there will be a general overview of the documentation analysis method, as well as details related to the selection of repositories from which to select datasets and the method of data collection. Reasoning related to the assumptions of automation of the documentation completeness test will be presented and the manual integration of the analysis itself will be discussed. Next, the criteria for selecting datasets will

be presented, detailing the sorting criteria and the management of duplicates, i.e. datasets present in more than one repository. Finally, the method of 'reading' the documentation and thus conducting the completeness test will be presented.

Chapter 4 contains the analysis of the results. First, the raw data collected will be presented. Next, we will go into detail using quantitative and qualitative methods, analysing the peculiarities of the datasets in their entirety, the analysed documentation sections and finally the individual fields of information. Lastly, will be presented the results, linked to a subset of fields of a single repository, of a comparison between manual checking (based only on the selected datasets) and automatic checking (based on the entire repository under investigation).

Finally, Chapter 5 will contain conclusions drawn from the entire thesis work and possibilities for future work to expand this work.

Chapter 2

Study of a recommended information scheme

This chapter presents an outline of recommended information that should be present in a dataset documentation, in order to ensure a proper choice of dataset and an informed use by professionals (developers, data scientists, etc). The aim is to recognize, with a literature study, which information is really important for building a dataset documentation in order to achieve transparency, accountability and to avoiding data cascades on AI pipeline [67]. Since the ultimate goal of this research is to understand the state of practice of dataset documentation, the information identified as significant will be outlined in a scheme. The objective of this scheme is to make it easy to measure the completeness of the documentation of a dataset. The interest in a measure of completeness stems from the fact that this property can be a good indicator to measure a qualitative aspect of dataset documentation.

The novelty of this work, indeed, concerns the fact that the individual fields representing specific information, have been transformed and summarised into a concept represented by few words to which it is easy to answer "yes" or "no", depending on the presence or absence of this information in the documentation under analysis. In addition, an attempt has been made to make this scheme as generalisable as possible to any type of documentation, so that it can be used for datasets pertaining different areas of artificial intelligence.

The construction of this list is strongly based on «Datasheets for Datasets»[26] by Gebru et al., «The Dataset Nutrition Label»[33] by Holland et al. and «Data Statements for Natural Language Processing»[8] by Holland et al. Information is organized following the categorisation presented by Gebru et al., i.e. information is grouped into 6 categories: *Motivation*, *Composition*, *Collection processes*,

Preprocess, cleaning, labelling procedures, *Uses*, *Maintenance*.

2.1 Discussion

Section **1** *Motivation*, is intended to make explicit the basic context around which the dataset was created and published. To do this, the presence of information concerning the purpose of the dataset is considered important (**1.01** *Purpose for the dataset creation*), as is information concerning creators (**1.02** *Dataset creators*) and funders (**1.02** *Dataset creators*). These type of information could be useful to create a basic context around the dataset, in a such a way that an interested users can immediately understand whether the dataset is compatible with its purpose and avoid potential harms from using an inappropriate one [72].

Section **2** *Composition* pursues the objective of showing how data are made, in order to bring out all statistically relevant aspects. Three different groups of information can be identified: the first part describe the characteristics of the data; the second part presents additional information in the case of people related data; the third part shows - with the graphic approach used in [33] - statistics of the data.

In order to continue the discussion, it is necessary to define what is meant by people-related data. Are people-related data all data produced by humans? Or are the ones which contains specific information about individuals? As can be guessed, there is no single convention and this question leaves room for interpretation. In order to promote greater caution by dataset producers, Gebru et al. recommend to take a broad interpration of whether a dataset relates to people. During this research work this principle was partially followed, albeit with certain limitations discussed in chapter 3.5. Regardless of the interpretation one chooses to maintain in order to determine whether a given dataset is considered people-related, it is considered that a distinction in this direction is useful and could help the community to pay more attention with more sensitive data.

Field **2.01** *What do the instances that comprise the dataset represent* aims to clarify the content of each record, through a feature description. Field **2.02** *Number of the instances* represent a simple quantity value of the number of instances contained within the dataset. The intent of fields **2.03** *Information about missing values* and **2.05** *Description of errors, noise or redundancies* is to describe the eventual presence of missing values or errors and how they are highlighted. The transparency of these aspects could avoid a lot of problems also in the subsequent steps of data transformation. Field **2.04** *Recommended data splits* refers to the possibility that dataset author already provides a data subdivision specifically adopted for training, validation and evaluation of AI models. **2.06** *Information*

about data confidentiality and **2.07** *Information about possible data dangerousness (offensive, insulting, threatening or cause anxiety) or biases* are intended to expose some hazard warning: the fact that some information contained within the dataset may be confidential and that it is possible that someone might feel threatened in any way by them, respectively. Given this definition of data dangerousness, it was natural to include the presence of any bias in the data among the information reported by this field.

Field **2.08** *Information about people involved in data production and their compensation (if people related)* is the first peoplerelated data specific field. It refers to information about people produced data or from whom was they extracted. This field was suggested by [8]: while they deisgned it in a NLP specific way, it was generalised in order to achieve the same targets. Thai aim is to ensure that datasets users are fully aware of data provenance, because it can highlight some type of harmful biases perhaps difficult to detect otherwise. This can be achieved by describing the persons producing data, either through a general description or by specifying some useful demographic information, e.g. job, dialect or ethnicity. Moreover, a specific reference about the compensation of the people who produced data was added. As widely argued by Casilli in [13], the production of data is nothing but a form of labour, nowadays almost never remunerated. That is why it would be important to start taking this aspect into account to avoid exploitation of free labour. The recommended information scheme proposed mentions the remuneration of all workers that contribute to the creation of the dataset (with due differences). Section **2** contains two more people related information: **2.09** *Description of identifiability for individuals (if people related)* and **2.10** *Description of data sensitivity (if people related).* The first, it aims to make explicit if it is possible to identify individuals and in which measure. The second, instead, reflects the needing of explicit if data can be considered sensitive in any way. This definition includes personal information such as ethnicity, political views, religious and philosophical beliefs, union affiliation, financial data, health or sexual life data, sexual orientation and genetic or biometric data [26, 65]. Since the misuse of such data can cause serious damage, it is important to include every useful information about them in dataset documentation.

Fields from **2.11** to **2.14** are part of the diagnostic framework suggested by Holland et al. in order to provide a comprehensive overview of dataset ingredients before AI model development [33]. They consist in different modules: **2.11** *Statistics*, as the name suggests, refers to basic statistics such as row counts, unique entries, most and least frequent items and number of missing values; **2.12** *Pair plots* refers to comparisons between two variables values; **2.13** *Probabilistic model* refers to values distribution for a given variable and, finally, **2.14** *Ground truth correlations* refers to investigations on correlations between ground truth values and other features.

Section **3** *Collection processes* is designed to describe how (**3.01** *Description of instances acquisition and data collection processes*), when (**3.03** *Time frame of data collection*) and by whom (**3.02** *Information about people involved in the data collection process and their compensation*) the data collection work was carried out, with some additional information in case of people related data. The facts that people involved in data production (literally, who produce the data) know about the data collection (**3.05** *Information on individuals' knowledge of data collection (if people related)*) and give their consent to it (**3.06** *Information on individuals' consent for data collection (if people related)*) are relevant in order to promote responsible dataset production. Moreover, this information concerns the way in which data subjects know about the data collection and how they eventually gave their consent, broadening the context around data collection. In this way it is possible to make the origin of the data explicit, a source of further ethical questions [75]. These aspects are really relevant in terms of privacy and should be clear and understandable to every stakeholder, on the base of the same principles of «privacy nutrition label» by Kelley et al. [41].

Fields **3.04** *Information about ethical review processes* and **3.07** *Analysis of potential impact of dataset and its use on data subjects* refers to ethical reviews processes about data (how it has been conducted, which outcomes has been produced) and to analysis of how data could impact data subjects.

Section **4** *Preprocess, cleaning, labelling procedures* aims to showcase all data processing, sampling, cleaning and labeling procedures, whose effects may go beyond technical processing. The issue of the non-neutrality of technical choices is even more relevant when it comes to labelling, where the categorisation process is carried out on the basis of one's own perception (influenced by experience, subjectivity and the socio-economic context). This reasoning can be generalised to most preprocessing activities. For this reason, a descriptive reference to who performs these activities has also been added to this section: **4.02** *Information about people involved in the data sampling, preprocessing, cleaning procedures* is based on *annotator demographic* field present in [8].

Information represented by field **4.01** *Description of sampling, preprocessing, cleaning, labeling procedures* could be useful in order to address a technical transparency about these data processing procedures. In this section, it is possible to introduce a new aspect in addition to the description of the activities performed on the data: the suggestion of other data procedures that could be performed on them. Indeed, can be very useful for a dataset user to be able to access a description of other possible recommended transformations (field **4.03** *Description of other possible preprocessing, sampling, cleaning, labeling procedures*).

Section **5** *Uses* provides useful information about the uses that should and should not be made of the dataset, in order to avoid misuse that might produce undesirable or even dangerous results. Field **5.01** *Description of the tasks in which the dataset has already been used and their results*, in addition to include a description of the tasks for which the dataset has already been used (as present in «Datasheets for Datasets»), makes explicit request for information regarding any results obtained. This is similar to what is shown in the 'eval statistics' section of some dataset page in the *UCI Machine Learning Repository* repository. In that case, accuracy and precision values are shown. This integration was done because to know the performance of other models on same data could positively redirect subsequent developments. Similarly, the field **5.04** *Repository that links to papers or system that use the datasets* represents another way of making the uses of the dataset transparent. In this case, however, by referring to the fact that there is some way of accessing the scientific applications that have exploited the dataset.

Fields **5.02** *Description of recommended uses or tasks* and **5.03** *Description of not recommended uses* refers to description about how to use or not to use data and play a key role in reducing ethical risks. Such information, however, should not be limited to a description of the tasks for which they were designed (although this is an important part) but should also tell more generally whether such data would be suitable for certain types of riskier applications, such as people-related automatic decision-making systems.

Field **5.05** *Description of license and terms of use* in [26] has been inserted in the "distribution" section. Here, the "distribution" section has been excluded because the analysis is focused on generalist repositories of already shared datasets, so it is not very applicable. For this reason, field **5.05** has also been placed in the "Uses" section.

Finally, section **6** *Maintenance* provides information about the maintenance of the datasets. This section is significant to avoid misuses: the active stewardship (with appropriate updates of documentation and even availability) is an essential part of the production phase of a machine learning dataset [17], since ethics concerns can evolve over time [61]. Field **6.01** *Information about subject supporting, hosting, maintaining the dataset* represent information useful to understand who is in charge of the active stewardship of the dataset, while **6.02** *Contact of the owner* refers to presence of some specific way to contact the owner of the dataset. In order to achieve a correct stewardship, **6.03** *DOI* cover a very important role because it helps to uniquely identify the dataset.

Another relevant maintenance aspect concerns deprecated datasets: in order to avoid the perpetuation of the risks that possibly led to their withdrawal, it is decisive to manage with transparency and awareness "zombie datasets" in all the place they were publicized [14]. Fields **6.04** *Erratum*, **6.05** *Information about*

dataset updates and **6.06** *Information about management of older dataset versions* find their justification in avoiding the negative effects of zombie datasets or even zombie versions of datasets. Finally, field **6.07** *Information about mechanism to extend, augment, build on, contribute to the dataset* refers to information about how it is possible to contribute (in different ways) to the dataset. These instructions should not be underestimated because they help to improve the dataset on the base of the experience gained by using it: they can be considered as a way of communication between the users of the dataset user (and sometimes toward the builders of the dataset).

2.2 Documentation Test Sheet

All the information represented by the fields discussed in the previous section, has been arranged in a Documentation Test Sheet (DTS). This test aims to measure the dataset documentation completeness. The following list summarises the fields described: for each field there is a reference to the paper from which it was selected. Table 2.1 presents fields in a Sheet intended to make it easier to evaluate a dataset documentation.

1 Motivation

- 1.1 *Purpose for the dataset creation* [26]
- 1.2 *Dataset creators* [26]
- 1.3 *Dataset funders* [26]

2 Composition

- 2.1 *What do the instances that comprise the dataset represent* [26]
- 2.2 *Number of the instances* [26]
- 2.3 *Information about missing values* [26]
- 2.4 *Recommended data splits* [26]
- 2.5 *Description of errors, noise or redundancies* [26]
- 2.6 *Information about data confidentiality* [26]
- 2.7 *Information about possible data dangerousness (offensive, insulting, threatening or cause anxiety) or biases* [26]
- 2.8 *Information about people involved in data production and their compensation (if people related)* [8]
- 2.9 *Description of identifiability for individuals (if people related)* [26]

2.10 *Description of data sensitivity (if people related)* [26]

2.11 *Statistics* [33]

2.12 *Pair plots* [33]

2.13 *Probabilistic model* [33]

2.14 *Ground truth correlations* [33]

3 Collection process

3.1 *Description of instances acquisition and data collection processes* [26]

3.2 *Information about people involved in the data collection process and their compensation* [26]

3.3 *Time frame of data collection* [26]

3.4 *Information about ethical review processes* [26]

3.5 *Information on individuals' knowledge of data collection (if people related)* [26]

3.6 *Information on individuals' consent for data collection (if people related)* [26]

3.7 *Analysis of potential impact of dataset and its use on data subjects* [26]

4 Preprocess, sample, cleaning, labeling

4.1 *Description of sampling, preprocessing, cleaning, labeling procedures* [26]

4.2 *Information about people involved in the data sampling, preprocessing, cleaning procedures* [8]

4.3 *Description of other possible preprocessing, sampling, cleaning, labeling procedures* [26]

5 Uses

5.1 *Description of the tasks in which the dataset has already been used and their results* [26]

5.2 *Description of recommended uses or tasks* [26]

5.3 *Description of not recommended uses* [26]

5.4 *Repository that links to papers or system that use the datasets* [26]

5.5 *Description of license and terms of use* [26]

6 Maintenance

6.1 *Information about subject supporting, hosting, maintaining the dataset* [26]

6.2 *Contact of the owner* [26]

6.3 *DOI* [26]

6.4 *Erratum* [26]

6.5 *Information about dataset updates* [26]

6.6 *Information about management of older dataset versions* [26]

6.7 *Information about mechanism to extend, augment, build on, contribute to the dataset* [26]

Table 2.1: Documentation Test Sheet

Dataset:		
Field ID	**Field Name**	**Presence**
1.01	*Purpose for the dataset creation*	
1.02	*Dataset creators*	
1.03	*Dataset funders*	
	1 *Motivation* Presence Average	
2.01	*What do the instances that comprise the dataset represent*	
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 (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	
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 impact of dataset and its use on data subjects*	
	3 *Collection processes* Presence Average	
		Continue on next page

Table 2.1 – continued from previous page

Dataset:		
Field ID	**Field Name**	**Presence**
4.01	*Description of sampling, preprocessing, cleaning, labeling procedures*	
4.02	*Information about people involved in the data sampling, preprocessing, cleaning procedures*	
4.03	*Description of other possible preprocessing, sampling, cleaning, labeling procedures*	
	4 *Preprocess, cleaning, labelling 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	
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 mechanism to extend, augment, build on, contribute to the dataset*	
	6 *Maintenance* Presence Average	
	Dataset Presence Average	

Chapter 3

Documentation analysis of the most popular datasets

One of the main objectives of this thesis work concerns the documentation analysis of the most popular datasets in the artificial intelligence community. More specifically, the aim of this work is to understand whether ethically relevant information is present in the same place where the data is publicly accessible. For this reason, the pages of the repositories hosting the datasets were analysed. Thanks to this analysis work, it is possible to get an overview of the completeness of the documentation (one of relevant quality aspects) of the most used datasets, especially from the point of view of ethically relevant information.

This chapter will discuss the details of data collection for the research. First, a descriptive overview of this research phase will be presented, then the selection of repositories under investigation will be discussed, showing the metadata each repository offers. The section 3.3 will show the hypothesis of work automation, i.e. the automatic collection of data and how this information can be used for this research. However, due to accuracy problems (which will be detailed in section 3.3.2), human checking was assessed as essential. It was therefore necessary to select a subset of the datasets offered by the selected repositories: details, such as sorting criteria and duplicate management, are presented in Section 3.4. Finally, in section 3.5, the conduct of the information presence check is presented.

3.1 Description of the analysis

The analysis of dataset documentation is structured with the aim of understand how much the documentation under analysis are complete and which sections of information are most complete and most lacking.

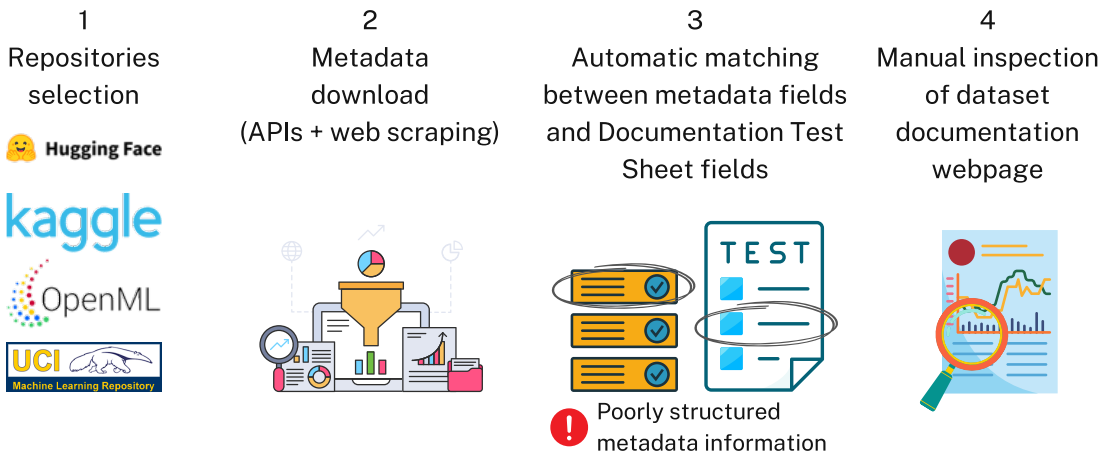
The starting point of this work is represented by the set of ethically relevant information shown and explained in the previous chapter. As mentioned above, the fields of that set have been arranged in a completeness test, i.e. in such a way that was easily and immediate to answer with a yes (True value: 1), if the information represented by the field is present, or no (False value: 0), if not present.

In addition to the fields presented in the chapter 2, in order to trace some peculiar characteristics of the individual dataset, the following values were recorded for better indexing of the results, depending on the type of the dataset:

- **c.01** *Data is people related* (True or False)
- **c.02** *Presence of label (target variable)* (True or False)
- **c.03** *Dataset is a sample(rows)/reduction(columns) of a larger set* (True or False)
- **c.04** *Recently updated* (True or False)

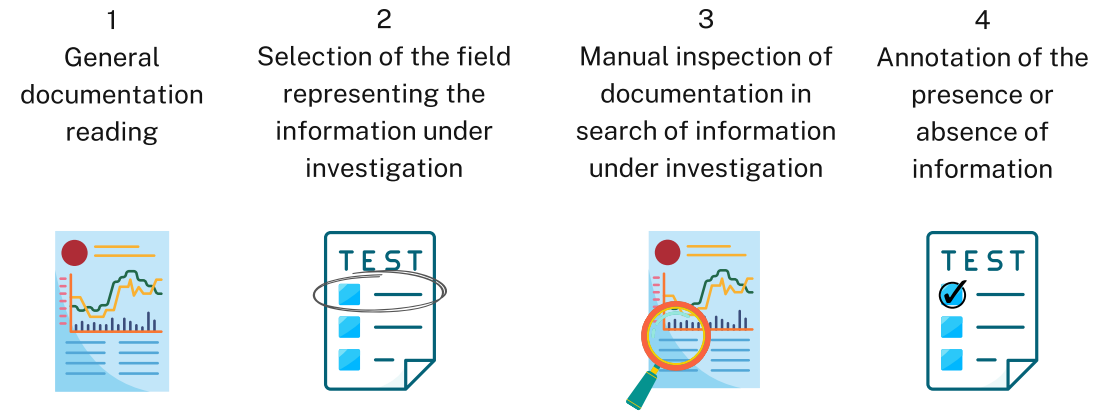
These fields has been added in a section named *Characteristics*.

Figure 3.1: Documentation analysis steps



The steps of the documentation analysis are described in figure 3.1. The data collection phase of this research is represented by the collection of all metadata presented on the dataset description pages in the hosting repositories. First of all, it is necessary to select the repositories under analysis. The second step concerns metadata download and access to datasets webpage. This step, detailed in section 3.3.1, is was partly carried out through web scraping algorithms and partly by querying the APIs offered by the platforms. Obviously, this procedure is strictly dependent on repositories selection, both in terms of the organisation of information and of the data access. Once the metadata have been obtained, the work can focus on matching the metadata fields with the test fields discussed in the previous chapter. Due to the non-overlap between the metadata offered by the repositories and the completeness test presented, a manual analysis of the web pages of the datasets was carried out, following the procedure described in figure 3.2. This approach to work makes research less scalable with the available resources. During the analysis of a specific dataset, for each test field, the presence (1) or absence (0) of the information represented (in the repository hosting page) has been recorded.

Figure 3.2: Documentation reading procedure



The desired end result is therefore a table capable of showing, for each dataset, the presence or absence of each field presented in the completeness test. These raw results, and all subsequent analyses, are presented in Chapter 4.

The values representing the presence (or absence) of the information represented by the single field, in a single dataset under analysis, are then combined in a new introduced measure: the **information presence measure**. It is a completeness measure: it represent the average obtained grouping by field, section, dataset or repository (as well as the various combinations). To give an example, for a single

field it is obtained by dividing the number of datasets containing the information represented by the field, by the total number of analysed datasets. This measure allows us to evaluate, in a quantitative way, the average presence of the information represented by a single field, the average completeness of a documentation section, the average completeness of a dataset documentation or the average completeness or a whole repository.

3.2 Repositories under analysis

In order to study and measure the completeness of the most popular datasets in AI community, once the set of information to be checked has been determined, it is necessary to choose which datasets (and relative documentation) to analyse.

One of the possible ways is to select datasets from a corpus of academic papers, as seen in works like «Do Datasets Have Politics?»[68] by Scheuerman et al., «Mitigating Dataset Harms Requires Stewardship»[61] by Peng et al. or «Garbage In, Garbage Out?»[28] by Geiger et al. One of the new elements represented by this work, however, concerns the analysis of the documentation and information in the very same place where the data can be accessed. For this reason, the focus has been shifted to generalist repositories hosting datasets for artificial intelligence. Following this decision, the documentation being analysed is not the scientific article linked to the dataset (if any), but rather the webpage where it can be downloaded or the metadata downloadable attached to data. In this way, it is possible to obtain a partial hint of *opacity*, i.e. the documentation uncompleteness of a single dataset, or *sparsity*, i.e. the fact the ethically relevant information exists but that does not reach all stakeholders, following the tassonomy presented by Fabris et al. [21].

This shift brought out the additional difficulty represented by the fact that each repository had different set of metadata attached to datasets. Different repositories means different websites and different metadata.

The choice of repositories was determined by the need to want to analyse repositories that are fairly well known and commonly used by ai practitioners. By scientific choice, due to the considerable scientific production in the field, it was decided to exclude repositories specifically related to the field of computer vision. A further element taken into account for the choice was the free access to the datasets. Furthermore, for reasons of time required for analysis, it was decided to limit the number of repositories to 4.

On the basis of these elements, three generalist repositories[1] and one repository

[1]The term 'generalist repository' refers to repositories not focused on a single sector of artificial intelligence applications (e.g. computer vision)

related to the world of Natural Language Processing (NLP) were selected. The decision to include a repository related to the NLP world was dictated by the fact that this type of sector is booming and applications from this sector require special attention to ethical issues [8].

The selected repositories are:

- Huggingface[2], NLP-specific
- Kaggle[3], generalist
- OpenML[4], generalist
- UCI Machine Learning Repository[5], generalist

In the remaining space of this section, the details of each platform are discussed.

HuggingFace

Huggingface is a platform that let users to create, discover and collaborate on Machine Learning models and datasets[6] [46].

In the HuggingFace repository[35], each dataset is associated with the following information:

- Dataset structure
  - Data instances
  - Data fields
  - Data splits
- Dataset creation
  - Curation rationale
  - Source data
  - Annotations
  - Personal and sensitive information

[2] `https://huggingface.co/datasets`

[3] `https://www.kaggle.com/datasets`

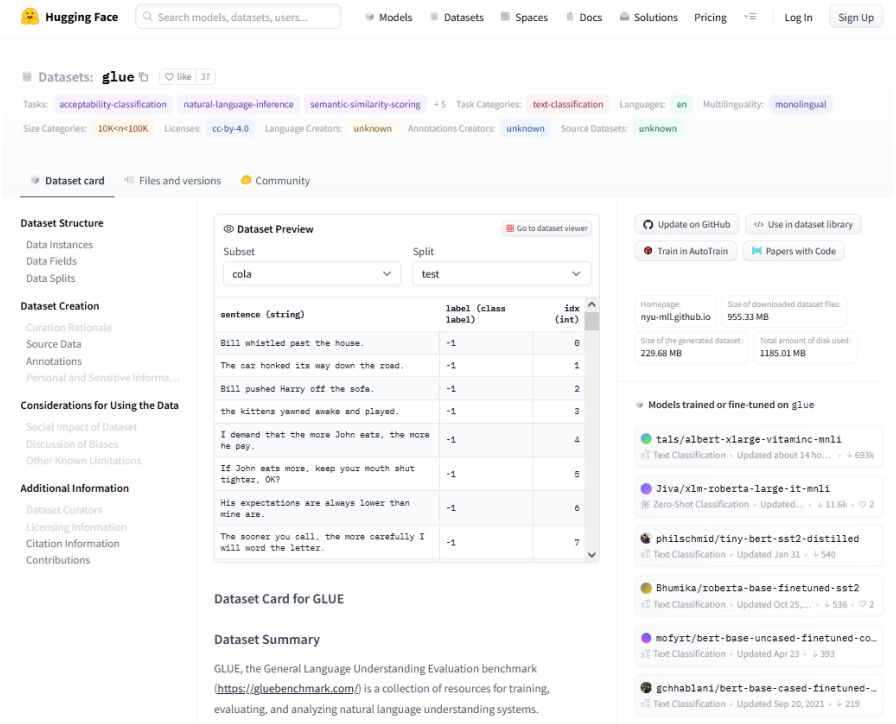
[4] `https://www.openml.org/search?type=data`

[5] `https://archive-beta.ics.uci.edu/ml/datasets`

[6] `https://huggingface.co/`

- Considerations for using the data
  - Social impact of datasets
  - Discussion of biases
  - Other known limitations
- Additional information
  - Dataset curators
  - Licensing information
  - Citation information
  - Contributions

Figure 3.3: Huggingface dataset page example

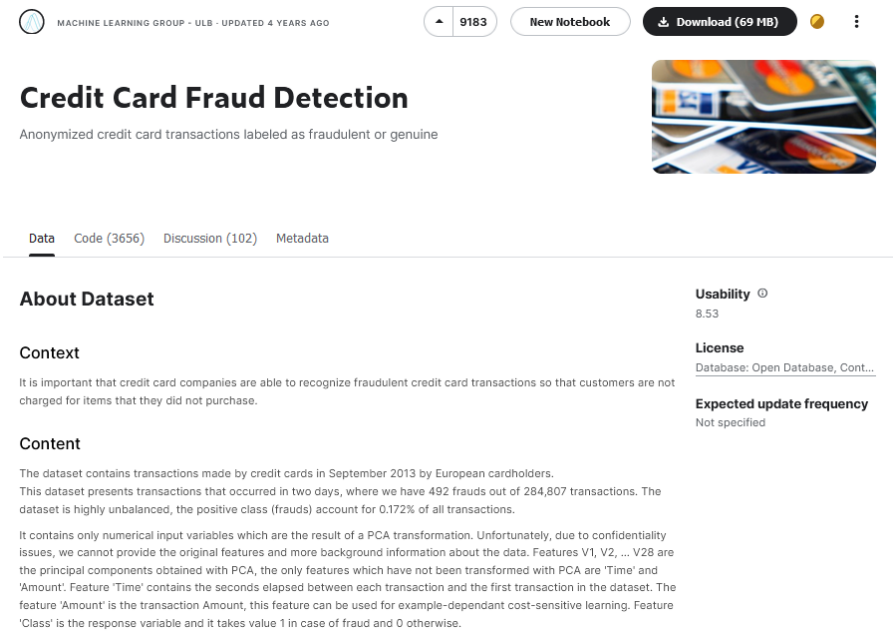


Huggingface declared to own 1.920 datasets (as at 18/11/2021)

Kaggle

Kaggle, owned by Google[7], is a platform which aim to simplify the work of data scientists, allowing to find datasets, build models and interact with other users in "the world largest data science and machine learning community"[8].

Figure 3.4: Kaggle dataset page example



In the Kaggle repository, each dataset is associated with the following information:

- Title
- Subtitle
- Data
  - Usability score
  - Tags
  - Overview description
- Related tasks

[7] https://en.wikipedia.org/wiki/Kaggle

[8] https://www.youtube.com/watch?v=TNzDMOg_zsw

- Related code
- Discussion
- Activity
  - Views
  - Downloads
  - Unique contributors
  - Top techniques
  - Top complementary datasets
  - Discussion stats
- Metadata
  - Usage information
    - License
    - Visibility
  - Provenance
    - Sources
    - Collection methodology
  - Maintainers
    - Dataset owner
    - Collaborators
  - Updates
    - Expected update frequency
    - Last updated
    - Date created
    - Current version
- File description(s)
- Column description(s)

The "Overview description" field is represented by a single free text field, but users are invited to specify several pieces of information, including the context from which the dataset originates the content present and which questions it should be able to answer.

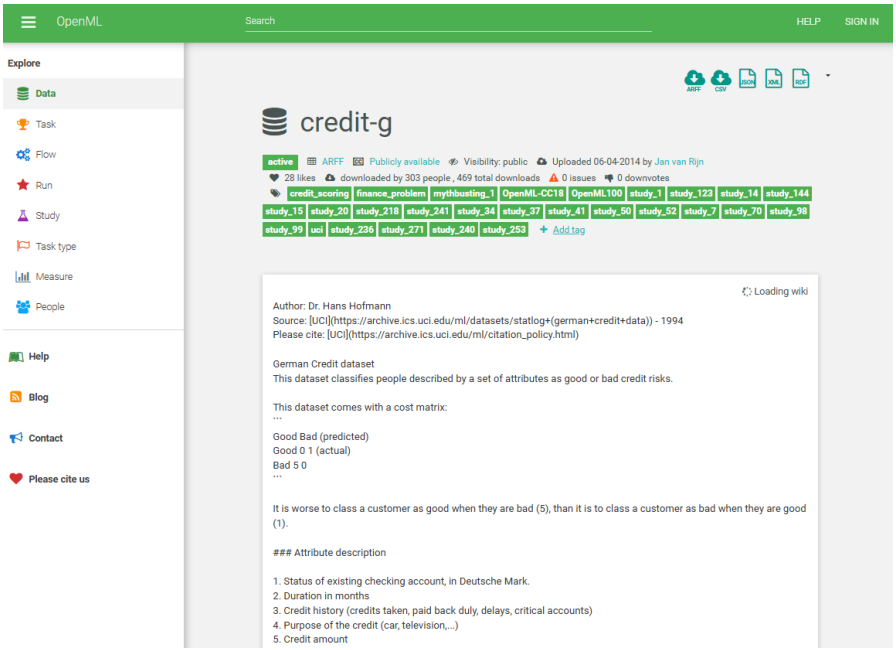
The site offers a usability score that grows, up to a maximum of 10.00, in proportion to some parameters like level of documentation, availability of related public content as references, file types and metadata coverage[9].

Kaggle declared to own 117.788 datasets (as at 18/11/2021).

OpenML

OpenML[80] is a platform useful to "share datasets, algorithms and experiments results"[10] in the field of Machine Learning. The platform is a project supported by the non-for-profit Open Machine Learning Foundation organization[11].

Figure 3.5: OpenML dataset page example



The repository is organized in:

- datasets;
- tasks, i.e. a specific problem that can be solved with the attached dataset[12];

[9] https://www.kaggle.com/product-feedback/93922

[10] https://www.openml.org/about

[11] https://docs.openml.org/Governance/

[12] https://www.openml.org/search?type=task

- flows, i.e. all the information to build a model[13];
- runs, i.e. flows applied to a given task, with all the hyperparameters used[14];

In the OpenML repository, each dataset is associated with the following information:

- Status
- Format
- License
- Visibility
- Upload date
- User
- Number of likes
- Number of download
- Wiki
- Features
- Related tasks

The wiki field is a free text field.
The platform declared to own 3.452 datasets (22/11/2021).

UC Irvine Machine Learning Repository

"The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms"[15].

In the UC Irvine Machine Learning Repository, each dataset is associated with the following information:

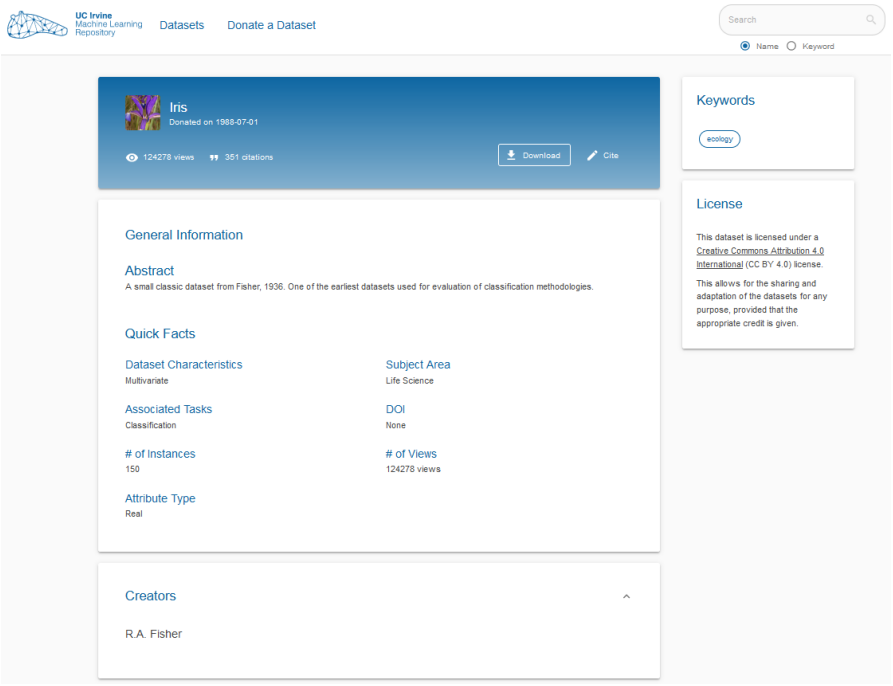
- General information

[13] `https://www.openml.org/search?type=flow`

[14] `https://www.openml.org/search?type=run`

[15] `https://archive.ics.uci.edu/ml/about.html`

Figure 3.6: UCI ML Repository dataset page example



  - Abstract
  - Quick facts
    - Dataset characteristics
    - Subject area
    - Associated tasks
    - DOI
    - Number of instances
    - Number of views
- Creators
- Descriptive questions
  - For what purpose was the dataset created?
  - Who funded the creation of the dataset?
  - What do the instances that comprise the dataset represent?
  - Are there recommended data splits?
  - Does the dataset contain data that might be considered sensitive in any way?

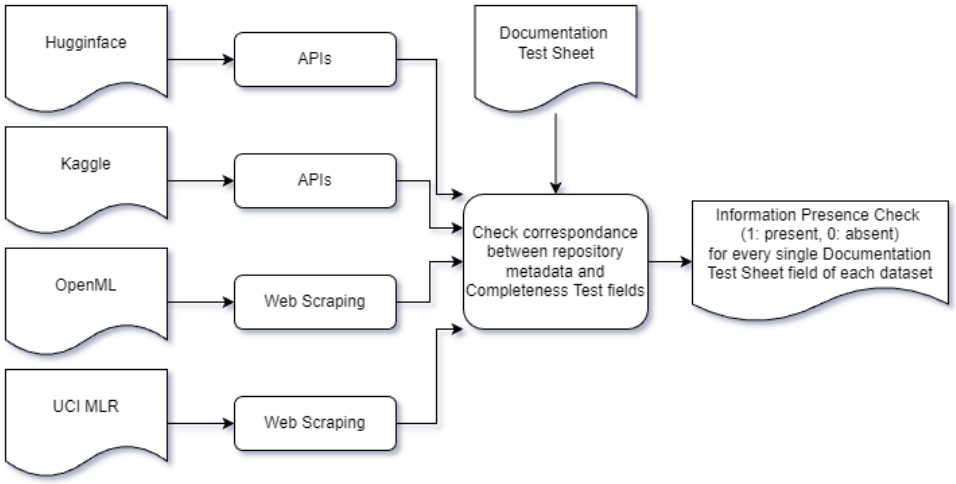
  - Was there any data preprocessing performed?
  - Has the dataset been used for any tasks already?
  - Additional information
  - Citation, Requests/Acknowledgments
- Tabular data properties
  - Does this dataset contain missing values?
  - What symbol is used to indicated missing data?
  - Number of Attributes
- Features
  - Attribute name
  - Role
  - Type
  - Description
  - Units
  - Missing Values
- Papers citing this datasets
- Keywords
- License

Within this repository, there was 593 datasets (18/11/2021).

3.3 Exploratory study for the automation of data extraction

The goal of this research phase consists of collect dataset metadata and matching them with the set of ethically relevant information presented in the previous chapter. With the aim of develop a scalable system able to measure the documentation completeness of a great number of dataset in a short period of time, the structure of a software capable of performing this task automatically was explored, as described in figure 3.7. The desired output of such a system is a vector of 1/0 representing the information presence/absence of every documentation completeness test fields for each dataset.

Figure 3.7: Automation hypothesis flow chart



Section 3.3.1, will discuss the details of data collection, showing the code that made it possible to collect data.

Section 3.3.2 will discuss the correspondence between repository metadata and DTS fields.

Finally, Section 3.3.3 will show results of the information presence check on selected metadata from a selected repository.

This phase of the research work did not provide the desired results. It turned out that it is very difficult to obtain a satisfactory match between metadata fields offered by repositories and the completeness test fields, such as described in table 3.1.

This is partly attributable to the opacity of the documentation, but mainly to the fact the most information is localised in generic textual fields. Indeed, in cases where the repository offers fields that perfectly match those proposed in the documentation completeness test, the absence of values in those fields does not imply the absence of the specific information elsewhere in the documentation. This is why, at the end, it was decided to opt for a manual verification of the presence of the information.

3.3.1 Data collection

In order to collect data needed for this part of the research work, a set of script a set of script was developed. It is possible to find the detailed listings useful to collect data from each repository in appendix A.

Each repository has it own peculiarities. From a technical point of view, two

approaches were used: on one hand, in order to download data from Huggingface and Kaggle platforms, their public APIs was used; on the other, web scraping tool were used to collect data from UCI Machine Learning Repository and OpenML. The aim of this operation was both to check the feasibility of an automation system, at least for certain information fields, and to collect parameters useful to rank datasets. The latter have been collected to facilitate a datasets subset selection, in case a manual check for the presence of information is necessary.

Huggingface

Huggingface exposes public APIs through which it is possible to browse the list of dataset published on the plaform[16].

After installed the required package, the following commands let to browse platform datasets:

Listing 3.1: Commands to retrieve the Huggingface datasets list with metadata

```
1 from datasets import list_datasets
2 dataset_with_details = list_datasets(with_details=True)[id]
```

It is possible to extrapolate the following information, for each dataset:

1 id: dataset identification;

2 *key*;

3 *lastModified*;

4 link;

5 description: dataset general description;

6 citation: bibtex for cite dataset;

7 *size*;

8 *etag*;

9 *siblings*;

10 author;

11 private: boolean value representing the public availability of the dataset;

12 paperswithcode_id: dataset identification name on paperswithcode website;

[16] `https://huggingface.co/docs/datasets/index`

13 downloads: number of downloads;

14 annotations_creators: tag describing who created annotations;

15 language_creators: tag describing who created language expressions;

16 languages;

17 licenses;

18 multilinguality: tag describing the quantity and type of languages supported by the dataset;

19 size_categories: categorical representation of the dataset number of instances;

20 source_datasets: tag describing the dataset origin;

21 task_categories: tag describing the category of tasks for which the dataset was designed;

22 task_ids: tag describing the specific task for which the dataset was designed;

23 pretty_name: fancy representation of the dataset name;

24 *extended*

25 *type*

26 *language*

27 *license*

28 *thumbnail*

29 *benchmark*

30 *task*

31 *submission_name*

32 *multilingualism*

33 *metrics*

34 *datasets*

Fields highlighted in italics represent fields that have values in none or very few datasets.

It is possible to implement an algorithm able to exploit the command showed in listing 3.1, in order to download the metadata of every dataset present in the repository. A possible implementation is accessible in listing A.1.

Kaggle

Kaggle exposes public APIs through which it is possible to browse the list of dataset published on the platform[17]. In order to pursue the purpose of the research, it was decided to exploit these public APIs rather then develop a web page scraping tool. This is also due to the specific structure of the dataset web pages, consisting of several 'tabs' (Data, Code, Discussion, Metadata).

After installed the required packages, with the command

Listing 3.2: Command to retrieve the Kaggle datasets list

```
1 kaggle datasets list --sort-by votes -p 1 --min-size 1
```

it is possible to display the list of datasets present on the platform, with a few
.

In addition, with the command

Listing 3.3: Command to retrieve the metadata of a Kaggle dataset

```
1 kaggle datasets metadata -p path_to_folder dataset_name
```

it is possible to download a json summarizing the metadata of the dataset called *dataset_name*, obtaining some additional information.

the set of information that can be obtained is summarised in the following list:

1 kaggle datasets list (listing 3.2)

- 1.1 ref: dataset name (part of the url);
- 1.2 title: name of the dataset;
- 1.3 size: dataset dimension;
- 1.4 lastUpdated: last update date;
- 1.5 downloadCount: number of dataset download;
- 1.6 voteCount: number of dataset upvotes from users;
- 1.7 usabilityRating: usability score

2 kaggle datasets metadata (listing 3.3)

- 2.1 id: equivalent to *ref* field from listing 3.2;
- 2.2 id_no: dataset identification number;
- 2.3 datasetId: equivalent to id_no;
- 2.4 datasetSlug: dataset title part of the id;

[17] `https://www.kaggle.com/docs/api`

2.5 ownerUser: username of the owner, first part of the id;

2.6 usabilityRating: equivalent to *usabilityRating* from listing 3.2;

2.7 totalViews: number of views;

2.8 totalVotes: equivalent to *voteCount* from listing 3.2;

2.9 totalDownloads: equivalent to *downloadCount* from listing 3.2

2.10 title: equivalent to *title* from listing 3.2;

2.11 subtitle;

2.12 description;

2.13 isPrivate: boolean value representing the public availability of the dataset;

2.14 keywords: array of keywords;

2.15 licenses: array of licenses;

2.16 collaborators: array of users who collaborated on dataset;

2.17 data;

Using an appropriate script it is possible to browse the datasets grouped in pages of approx 20 datasets (some pages contain fewer results). Without search words it is possible to retrieve only a subset of all datasets: this happens because pages beyond the 500th are empty.

The first step concerns the retrieving of the complete list of datasets, exploiting the command showed in listing 3.2. Once obtained the list of datasets, with their *ref*, it is possible to download the metadata json file for each of them exploiting the command showed in listing 3.3. A possible implementation of this algorithm is available in listings A.2. An example of a metadata json file is represented in listing A.3.

OpenML

OpenML exposes public APIs accessible through an ad-hoc python library[23], but there is no access to metadata useful for possible datasets sorting. The website, instead, allows access to useful values in this respect. For this reason, then, data are scraped from website, collecting the following fields:

1 name: dataset name;

2 link;

3 abstract: brief generic text description of the dataset;

4 runs: number of runs (see section 3.2 for further details);

5 likes: number of user likes;

6 downloads: number of downloads;

7 reach;

8 impact;

9 instances: number of the instances in the dataset;

10 features: number of data features;

11 classes: number of classes of the target variable;

12 missing_values: quantity of missing values

A possible implementation of a script collecting the above-mentioned data is shown in listing A.4.

UCI Machine Learning Repository

The UC Irvine Repository does not expose public APIs. For this reason, data are scraped from the web site. In order to achieve the goal of this section, it has been used the new beta version of the website[18].

Using a script like the one shown in listing A.5, the following information was collected:

1 name: dataset name;

2 dataCharacteristics: data type category (e.g. sequential, time series, etc.);

3 subjectArea: dataset science area (e.g. social, life, financial, etc.);

4 task: tag describing the category of tasks for which the dataset was designed;

5 donated: date of donation, i.e. the publication date of the dataset on the site;

6 instances_list: number of data instances;

7 attributes: number of data features;

8 views: number of dataset views;

9 abstract: generic textual field with a brief dataset description;

[18] `https://archive-beta.ics.uci.edu/ml/datasets`

10 link;

11 doi;

12 citations: number of dataset citations;

13 creators_str: dataset creators names;

14 *purpose*: textual field describing the purpose of the dataset;

15 *funding*: textual field describing who funding the dataset;

16 instances_meta: description of the instances;

17 *data_splits*: suggested data split;

18 *contain_sensitive*: description of data sensitivity;

19 *preprocess*: description of preprocessing steps performed on the data;

20 *already_used_task*: description of task that already used the dataset;

21 additional_info: generic textual field with additional information;

22 *acknowledgements*: acknowledgements and citation info;

23 missing_values: boolean value describing if dataset contain missing values;

24 symbol_missing_values: symbol representing any missing values;

25 license;

Fields highlighted in italics represent fields that have values in none or very few datasets.

3.3.2 Matching between collected data fields and Documentation Test Sheet fields

The data collected as described in the previous section, are then analyzed in order to understand if an automatic information presence checker is feasible (with an adequate accuracy) or not. The most relevant fact emerging from this step, is that in general a lot of information is distributed in general text fields. This fact, makes it very complex to implement an automatic scalable system able to check the presence or absence of the selected information.

One example able to justify the reasoning just expressed is represented by the UCI Machine Learning Repository. Although some fields of the repository metadata structure match perfectly with some fields of the DTS, that fields results

almost always empty. Specifically, UCI MLR fields *purpose* (14), *preprocess* (19) and *already_used_task* (20), respectively overlapping with DTS fields **1.01** *Purpose for the dataset creation*, **4.01** *Description of sampling, preprocessing, cleaning, labeling procedures* and **5.01** *Description of the tasks in which the dataset has already been used and their results*, turn out to be almost always empty. A manual analysis on some selected datasets (which will be described in section 3.5), however, showed that it is possible to find some reference to these information elsewhere on the webpage. The manual analysis performed on the top 25 datasets of each repository, the results of which will be shown in chapter 4, they show that the three aforementioned informations (DTS fields no. **1.1**, **4.1** and **5.1**) have an information presence measure equal to 0.44, 0.44 and 0.88 respectively.

Based on this evidence, the choice was made to automate the valorisation of only those fields that could be used automatically beyond any reasonable doubt of correctness, as described in table 3.1. The numbers written in the repository columns refer to the data fields collected from the repositories, as described in section 3.3.1.

Table 3.1: Matching between documentation test fields and collected data fields

Test Field ID	Test Field Name	Hugging face	Kaggle	Open ML	UCI MLR
4.02	*Information about people involved in the data sampling, preprocessing, cleaning procedures*	14			
5.02	*Description of recommended uses or tasks*	21			
5.04	*Repository that links to papers or system that use the datasets*	12			
5.05	*Description of license and terms of use*	17			
c.04	*Recently updated*		1.4		5

3.3.3 Results of the automatic check for Documentation Test Sheet fields

A further test was conducted with a view to automation: an extended selection of data collected fields was used to calculate the information presence measure in all datasets of the UCI MLR repository. This test can be also useful looking forward to possible extensions of this research work, making it possible to calculate the information presence measure of a larger set of fields from a greater number of datasets.

Within the UCI MLR have been taken into consideration all the fields overlapping with a field in DTS. The results presented in table 3.2 is computed on all

the 598 datasets of the repository. The presence average is computed by dividing the number of dataset with the field marked as 'present' with the total number of dataset.

Table 3.2: Matching between documentation test fields and UCI MLR collected data fields

Test Field ID	**Test Field Name**	**UCI MLR Field**	**Presence average**	**Presence**
1.01	*Purpose for the dataset creation*	14	0,01	3
1.02	*Dataset creators*	13	0,55	329
1.03	*Dataset funders*	15	0,00	2
2.01	*What do the instances that comprise the dataset represent*	16	0,99	593
2.02	*Number of the instances*	6	1,00	598
2.03	*Information about missing values*	23	0,57	338
2.04	*Recommended data splits*	17	0,00	2
2.10	*Description of data sensitivity (if people related)*	18	0,00	0
4.01	*Description of sampling, preprocessing, cleaning, labeling procedures*	19	0,00	1
5.01	*Description of the tasks in which the dataset has already been used and their results*	20	0,00	2
5.05	*Description of license and terms of use*	25	0,99	593
6.03	*DOI*	11	0,00	1

3.4 Dataset selection

On the basis of the outcome presented in section 3.3.2, it was decided to perform a manual check of information presence for all the fields excluded from table 3.1. The downside of this decision was the scalability of the research process. This decision, indeed, made the analysis of all datasets in the various repositories unfeasible. In addition, made it compulsory to focus the analysis on a subset of datasets.

In order to do so, the first step concerned the dataset selection criteria. As mentioned in the previous sections, it was decided to select the most popular datasets of each repository. With the aim to perform that selection, it was needed to individuate the proxy variables able to represent the dataset popularity. The reasoning about these choices are discussed in section 3.4.1. Based on the available time, in order to obtain a sufficiently significant sample, it was decided to select the top 25 datasets for each repository.

Once performed that selection, it emerged clearly that some datasets are popular in more then one repository. This fact on the one hand confirmed the validity of the popularity proxy variable, because it is credible that the most popular datasets are present in multiple platforms, on the other made it necessary to define a protocol useful to manage duplicates, as described in section 3.4.2.

3.4.1 Sorting criteria

The research goal, at the basis of this thesis, is to understand the state-of-art of documentation in the very same place which data are available to AI practitioners. Based on the issues raised in the previous sections, it was necessary to make a selection of the datasets offered by the various repositories under analysis. Faced with this need, it was decided to orient the work toward to the concept of "popularity", in terms of most used datasets. The aim is to focus the analysis on datasets that have a detectable impact on the AI community.

From a practical point of view, a value is required to proxy the concept of popularity. That parameter, should have the following properties:

- transparency: i.e. the clarity of how it is calculated;
- comparability: i.e. the value should be applicable to different repositories;
- orderability: i.e. it is possible to sort datasets on the basis of that value;

When available, it was identified that the number of downloads as the best proxy variable. Naturally, each repository has different parameters and it is organized in a different way, so it was needed to take specific decisions for every single repository. In the following sections, the choices taken for each repository will be discussed.

The selected datasets are listed in the appendix B.

HuggingFace

The only one parameter that can be used for popularity ranking, within data extracted by APIs, is the number of downloads. The web version also offers a sorting by likes, but the maximum value shows that the usage of like button is not enough popular among users and is therefore unable to adequately represent a ranking among almost 2000 elements.

Table 3.3: Huggingface most popular datasets

id	downloads
glue	719 706
super_glue	490 789
anli	171 100
wikitext	114 761
wino_bias	102 485
squad	98 446
imdb	93 646
trec	71 906
adversarial_qa	70 084
race	68 456
duorc	67 179
squad_v2	66 750
winogrande	58 213
hellaswag	54 372
common_voice	54 179
cnn_dailymail	53 207
piqa	53 155
xsum	50 393
cosmos_qa	50 151
mlqa	49 740
quail	49 413
paws	48 998
wmt16	48 694
ai2_arc	47 424
rotten_tomatoes	46 131
ropes	45 861
ag_news	45 359
amazon_polarity	44 621
cos_e	43 434

Continue on next page

Table 3.3 – continued from previous page

name	downloads
wiki_qa	42 721

Kaggle

Kaggle APIs directly offers the following sorting methods[19], accessible with code described in listing 3.2:

- **hottest**, trending datasets (18/11/2021-9969 records): this is the default sorting, that take into account either recently released Dataset and historical popular datasets;
- **votes**, sorting by number of upvotes expressed by users (18/11/2021-9986 records): this option "surfaces the most popular Datasets of all time";
- **updated**, sort by date of last update (18/11/2021-9981 records)
- **active**, recently active datasets (18/11/2021-9982 records)
- **published**, sort by date of publication (from most recent to oldest) (18/11/2021-9982 records)

Hotness has been documented as being a very good parameter for determining popularity[20], with even the most recent datasets achieving high engagement being taken into account. The problem lies in the algorithm opacity: a possible modification (undetectable) would make the results not reproducible. Moreover, it is a platform-specific parameter, without the comparability property. Since there is no possibility (through APIs) to extract directly by number of downloads, in order to have a deterministic and transparent parameter for the selection, it was decided to select by "votes" and then sort by number of downloads ("downloadCount" field).

The limitation of this approach is represented by the possibility that we lost some dataset with low votes and high download count. On the other hand, since we can retrieve about 10k datasets by number of votes, that risk can be considered negligible.

[19] `https://github.com/Kaggle/kaggle-api`

[20] `https://www.kaggle.com/docs/datasets#datasets-listing`

Table 3.4: Kaggle most popular datasets

title	downloadCount	voteCount
Credit Card Fraud Detection	360 828	8503
Novel Corona Virus 2019 Dataset	347 779	5627
Video Game Sales	264 773	4073
Heart Disease UCI	256 102	5379
Pima Indians Diabetes Database	235 317	2734
Iris Species	228 045	2688
World Happiness Report	202 882	3316
Netflix Movies and TV Shows	183 020	5804
The Movies Dataset	178 101	2655
Breast Cancer Wisconsin (Diagnostic) Data Set	177 162	2445
TMDB 5000 Movie Dataset	174 636	2855
COVID-19 Dataset	168 132	1449
Google Play Store Apps	166 169	3803
Trending YouTube Video Statistics	158 082	4275
Wine Reviews	148 561	3202
Chest X-Ray Images (Pneumonia)	143 227	4477
European Soccer Database	140 647	3580
COVID-19 in India	137 213	1679
COVID-19 Open Research Dataset Challenge (CORD-19)	134 256	9634
Students Performance in Exams	134 210	2883
FIFA 19 complete player dataset	130 521	3660
Avocado Prices	126 093	2587
House Sales in King County, USA	111 243	1626
Suicide Rates Overview 1985 to 2016	111 006	2723
New York City Airbnb Open Data	110 090	2453
Red Wine Quality	108 089	1817
Amazon Fine Food Reviews	108 069	1750
Fashion MNIST	102 001	1962
Telco Customer Churn	101 166	1812
Bitcoin Historical Data	98 823	2800

OpenML

OpenML offers APIs accessible through the python library but does not expose parameters useful for a popularity ranking. The web version, on the other hand,

offers the following values useful to a dataset ranking:

- runs
- likes
- downloads
- reach
- impact

The reach and impact values are not sufficiently documented and so they have not the transparency property. The Likes field has a very low maximum value (a symptom of a low use of this tool by users). The runs parameter takes into account the executions within the platform (see the OpenML structure peculiarities in section 3.2 for further details). Despite the fact that the number of downloads has a low maximum value (only four dataset exceed 100 downloads), it has been considered the best parameters able to represent the dataset usage level.

Table 3.5: OpenML most popular datasets

name	**downloads**
credit-g (1)	289
SpeedDating (1)	168
iris (1)	155
diabetes (1)	101
blood-transfusion-service-center (1)	99
eeg-eye-state (1)	94
tic-tac-toe (1)	93
spambase (1)	93
mnist_784 (1)	79
letter (1)	73
isolet (1)	71
Satellite (1)	70
one-hundred-plants-texture (1)	66
creditcard (1)	58
soybean (1)	56
waveform-5000 (1)	54
gisette (2)	53
glass (1)	52
arrhythmia (1)	50
Continue on next page	

Table 3.5 – continued from previous page

name	downloads
steel-plates-fault (1)	50
mammography (1)	49
amazon-commerce-reviews (1)	48
electricity (1)	44
spectrometer (1)	44
kr-vs-kp (1)	43
mushroom (1)	42
covertype (3)	40
Titanic (1)	40
bank-marketing (1)	40
one-hundred-plants-shape (1)	40

UC Irvine Machine Learning Repository

The UCI Machine Learning Repository site offers the following sorting methods (ascending and descending):

- Name
- Donation date
- Number of instances
- Popularity
- Number of attributes

The popularity sorting method, in his two version "Most Popular" or "Least Popular", is based on the number of views. The number of views is a transparent parameter, which is able to reflect the popularity of the datasets and it is reasonable to think that it is a good proxy for number of downloads.

Table 3.6: UCI MLR most popular datasets

name	views
Iris	117 923
Diabetes	82 902
Adult	78 136
Heart Disease	73 679
Wine	60 390

Continue on next page

Table 3.6 – continued from previous page

name	**views**
Car Evaluation	57 614
Breast Cancer Wisconsin (Diagnostic)	53 226
Abalone	44 346
Breast Cancer	43 632
Mushroom	43 005
Glass Identification	39 370
Census Income	33 757
Breast Cancer Wisconsin (Original)	33 457
Statlog (German Credit Data)	33 068
Thyroid Disease	27 802
Liver Disorders	27 738
Optical Recognition of Handwritten Digits	27 025
Ionosphere	26 431
Auto MPG	25 825
Pen-Based Recognition of Handwritten Digits	25 782
Image Segmentation	24 724
Congressional Voting Records	24 373
Zoo	23 785
Letter Recognition	22 791
Lung Cancer	22 469
Yeast	21 464
Spambase	21 314
Hepatitis	21 005
Internet Advertisements	20 165
Statlog Project	19 837

3.4.2 Duplicates

Once the datasets had been identified from the repositories, according to the criteria analysed in the previous section, a search was made for any duplicates. According to the purposes of the analysis, it was decided to eliminate any duplicates (within the same repository, or within different repositories) to avoid the duplication of information related to the same dataset. Moreover, to compare the information accompanying the same dataset in different repositories, was not a central objective of this research. This one, however, could be an interesting starting point for future analysis.

As a selection criterion between duplicates, it was decided to pick up the highest one based on the sorting criteria previously described. In the case of two datasets at the same ranking position, it was observed whether one of them was the primary source of the other.

In tables B.1, B.2, B.4 and B.3 in appendix B it is possible to observe the removed dataset with a reference to the dataset which led to its exclusion.

3.5 Dataset documentation reading principles

The web page of the selected datasets, containing dataset metadata, were analyzed through manual inspection. This type of analysis became necessary as a result of the issues outlined in section 3.3.2. The fact that much of the information resides in general text fields, indeed, leads to possible ambiguities. For this reason, an interpretation of the documentation is necessary and the outcomes 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 DTS field (detailed in chapter 2), in order to facilitate the reproducibility of this research work.

Motivation

The motivation section aims to provide general background information about who created the datasets, who founded it and for which purposes.

The field **1.01** *Purpose for the dataset creation* was checked as present when it emerged a clear goal for the dataset (e.g. benchmarking) or it was possible to derive the main task for which the dataset had been designed (e.g. credit scoring).

The **1.02** *Dataset creators* field represent the check for the presence of information about people or institution that create the dataset. On this aspect, some ambiguities arises in platform that let users upload datasets in a unmediated manner, such as Kaggle. In platforms such as this, it is not rare to encounter datasets that represent an easier way of accessing data created elsewhere (e.g. datasets with information on youtube videos) or even other datasets. In such situations, it is difficult to distinguish in a clear way the creator and the subject supporting/hosting/maintaining the dataset (discussed in field **6.01**). During the analysis an attempt was made to interpret ambiguous situations, trying to identify the subjects or institutions that create the dataset. This, defining the creation of the dataset as a process requiring some kind of collection and/or transformations in order to be used in some way.

1.03 *Dataset funders* field suffer the same issues illustrated above. This field, indeed, required an interpretation of institutions and companies that funded in some way the creation of the dataset. One of the elements used to determine the value field is related to companies/institutions acknowledgements.

Composition

The composition section aims to provide information about data characteristics.

In order to set a value for the field **2.01** *What do the instances that comprise the dataset represent*, information about the value comprises in a single instances was sought.

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

2.03 *Information about missing values* represent the presence of some explicit referrals to missing value absence/presence and eventually details about symbols representing these missing values. An element that should be taken into account is the fact that, if the repository does not expose an explicit field that require this info, authors may tend to discuss missing values only if submitted datasets contain them. That warning should be considered for all the following fields representing data issues as well, like **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)* and **2.09** *Description of identifiability for individuals (if people related).*

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

Fields **2.8**, **2.9**, **2.10** are only applicable to those datasets that contain people-related data. **2.07** *Information about possible data dangerousness (offensive, insulting, threatening or cause anxiety) or biases* refers to information on the subjects from which the data were produced, like demographics and whether they were remunerated in any way.

In order to check **2.10** *Description of data sensitivity (if people related)*, **2.11** *Statistics*, **2.12** *Pair plots* and **2.13** *Probabilistic model*, graphical elements useful to clearify data statistical properties have been searched for.

Collection Processes

Fields related to collection processes section, aims to check the presence of information concerning how data has been collected.

In order to check field **3.01** *Description of instances acquisition and data collection processes*, description of how the single data instances had been collected, e.g. questionnaiers, automatic scraping tool, etc were sought.

Field **3.02** *Information about people involved in the data collection process and their compensation* is similar to field **2.7** except the fact that in this case the focus was not on subjects who produced data, but rather on subjects working on data collection.

3.03 *Time frame of data collection* was checked as present if emerges clearly the time period in which the data collection was carried out.

3.04 *Information about ethical review processes* searched for details about any presence of discussion and results of some ethical review related to the dataset.

Fields **3.5**, **3.6** are only applicable to those datasets that contain people-related data. Field **3.05** *Information on individuals' knowledge of data collection (if people*

related) was checked as presence when documentation discuss in some way the fact that data producers knew about the data collection, while field **3.06** *Information on individuals' consent for data collection (if people related)* was checked when it was clear that the subjects 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 that they gave 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 impact of dataset and its use on data subjects* reflects presence of discussions about how the datasets can impact on those subjects who produced the data.

Preprocess / Cleaning / Labelling

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

Field **4.01** *Description of sampling, preprocessing, cleaning, labeling procedures* was checked as present when in the documentation materials emerged some hints about how data had been sampled, preprocessed, cleaned or labeled.

Field **4.02** *Information about people involved in the data sampling, preprocessing, cleaning procedures*, instead, concerns information about people who perform the procedures described in field **4.1**, with the same principles adopted for fields **2.7** e **3.2**.

The field **4.03** *Description of other possible preprocessing, sampling, cleaning, labeling procedures* represent the presence of details about how data can be further sampled, preprocessed, cleaned or labeled.

Uses

Fields related to Uses section, aims to check the presence of information concerning 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 within the dataset, and possibly to the results obtained. In the case of platform that let users to share their model like Kaggle, OpenML and Huggingface, users uploaded model related to the datasets were considered as an example of tasks that used the dataset. In the case of UCI MLR repositories, the sometimes present Evals section showing accuracy and precision of some classification algorithm (such as Support Vector Classification, Random Forest Classification, Logistic Regression, etc.) were 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 presence of some hint about the suggested tasks to use the dataset or to not use it, respectively.

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 used 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 UCI MLR 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 checked as present when at least the name of the licence was clearly displayed within the documentation.

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 dataset creators, rather then about the subject hosting and maintaining it. Within Kaggle and OpenML repositories, the subject had been uploaded the dataset was considered as the subject that is maintaining it. Within UCI MLR, since there is a mechanism for donating datasets and the repository management is similar to a library, those responsible for the site itself were considered to be the subjects maintaining the datasets.

In order to check the presence a **6.02** *Contact of the owner*, contacts of any kind to the creators of the dataset (UCI MLR) or to the party that made it public (Kaggle, OpenML, Huggingface) were sought. Any means of sending a message to such persons, such as email, in-platform messages, was considered as contact.

Field **6.03** *DOI* was checked present if a DOI number was provided.

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

One relevant aspect of **6.05** *Information about dataset updates* field check, concerns the fact that for Huggingface repository has been considered the commit information present in the "Files and versions" tab.

Field **6.06** *Information about management of older dataset versions* refers to the presence of details about older version of the dataset still available or retired.

Field **6.07** *Information about mechanism to extend, augment, build on, contribute to the dataset* was checked present if it is noticeable the presence of some mechanism for these purposes, like the GitHub link present in datasets within Huggingface repository.

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. [26], to take a broad interpretation of whether a dataset relates to people. The authors, to make an example, suggest that any dataset containing text that was written by people relates to people.

During the documentation analysis, that recommendation was not taken literally, especially for NLP specific dataset (typical Huggingface). In order to clarify the interpretation method some examples will follow. Datasets containing any kind of users text production such as reviews, search engines queries, and so on, were considered as people related data. Datasets containing wikipedia texts, newspapers articles, book corpus, school exams texts or ad-hoc text generation, instead, were not. The provenance is not always clear, but an attempt has been made to deduce the specificity of the data from the context. Image recognition of handwritten letter or numbers were not considered as people related data. Datasets containing patients medical data were considered as people related data.

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

Field **c.03** *Dataset is a sample(rows)/reduction(columns) 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 indicated in the *Files and versions* tab of the dataset web page. The Kaggle update date is the one obtained by the *lastUpdated* field obtained through APIs. The OpenML update date refers to the publishing date indicated on the website. The UCI MLR update date refers to the *Donated on* date indicated on the website.

Chapter 4

Results analysis

The research work design presented in chapter 3 represented the preliminary work to answer RQ2 discussed in section 1.4. This chapter will discuss the result obtained by the documentation analysis, with the aim of understand the extent to which the most popular datasets in the AI community are complete (or lacking) in terms of documentation.

First of all, raw results of each dataset field, for each dataset, will be presented. Subsequently, the level of detail under investigation will be progressively deepened. In fact, it was decided to approach the analysis of the results in such a way that as much information as possible could be obtained on the repositories as a whole, on the datadasets under investigation, on the information sections and finally on the individual fields.

In order to make this publication as accessible and inclusive as possible, for the illustrations contained in this chapter an attempt was made to choose a colour palette that would ensure, where possible, proper usability for colour-blind people, while trying to preserve digital and print graphic qualities[1].

[1]The colours palette was selected with the help of `https://colorbrewer2.org/`, by selecting the 4-class Paired and the 6-class Dark2 qualitative palettes depending on the number of classes represented

4.1 Raw results

This Section will present, for each repository, the result obtained from the documentation analysis detailed and discussed in section 3. For each repository, a figure of a summary table will be presented. The core of this table is represented by datasets columns and field rows. The cell referring to dataset x and field y may contains one of the three possible values:

- 1: whether it is possible to retrieve the information represented by the y-field within the documentation of the dataset x;
- 0: whether it is not possible to retrieve the information represented by the y-field within the documentation of the dataset x;
- NA: if the information represented by field y is not applicable to dataset x

The table also contains other useful information, such as the section to which the field belongs (**Section**) and whether the value of that field was obtained with the help of automatic systems (**Field automation**). The values of the latter one follow the Boolean convention already presented elsewhere: 1 equals 'yes', 0 equals 'no'. There are also summary rows and columns representing the average values of a particular data slicing, such as the section completeness percentage (**Sec %**), the field completeness percentage (**Field %**), the dataset completeness average (**Dataset AVG** and the global repository completeness average (**Repo AVG**)).

The field name has been omitted for reasons of space and readability: it is possible to retrieve the description for each **Field ID** in section 2. The same concept applies to the dataset name: it is possible to retrieve it in the appendix B.

Quantitative values are represented by different color shades, in order to facilitate a more immediate graphical reading of the data.

The data shown here and the code that enabled their analysis are available on GitHub[2].

The Huggingface repository, represented in figure 4.1, is the repository with the higher presence of information. Indeed, is characterised by the presence of the three most complete dataset documentation found throughout the analysis, i.e. **hug16**, **hug09** and **hug15** respectively the *cnn_dailymail* (an english news articles dataset), the *common_voice* (the crowdsourced speech recognition oriented datasets from Mozilla Foundation[3]) and the *adversial_qa* datasets. These three

[2] `https://github.com/RondinaMR/datasets-documentation-practice-mtcode`

[3] More information can be found at `https://commonvoice.mozilla.org/en/datasets`

datasets can be considered as a positive example of documentation. The dataset with the minimum average is **hug02**, the pretty famous *super_glue* dataset: this empirical result may be a first hint at the fact that popularity does not imply quality of documentation.

Within this repository four different field values obtained in an automatic way: **4.02** *Information about people involved in the data sampling, preprocessing, cleaning procedures*, **5.02** *Description of recommended uses or tasks*, **5.04** *Repository that links to papers or system that use the datasets* and **5.05** *Description of license and terms of use*.

The most present information, i.e. that information represented by fields present in the qualified majority (greater than two thirds) of the repository datasets, are **1.2**, **2.1**, **2.2**, **2.4**, **2.8**, **2.9**, **2.10**, **3.5**, **3.6**, **5.1**, **5.4**, **6.5**. Half of these fields belong to the *Composition* section, but the most complete section is the *Uses* one, while the *Collection processes* section is the least complete. All datasets in this repository were updated after 01/01/2021.

Within the Kaggle repository, represented in figure 4.2, the most complete dataset documentation is the one attached to **kag17** dataset, the *European Soccer Database*. The dataset with the minimum average is **kag27**, the *Amazon Fine Food Reviews* dataset.

This repository contains four fields calculated in an automatic way: **5.05** *Description of license and terms of use* and the **c.04** *Recently updated* characteristics.

The most present information, i.e. that information represented by fields present in the qualified majority (greater than two thirds) of the repository datasets, are **1.2**, **2.1**, **2.2**, **2.11**, **5.1**, **5.2**, **5.5**, **6.1**, **6.2**. One third of these fields belongs to the *Composition* section and another one third belongs to *Uses* section. The latter results to be the most complete section, on an equal footing with the *Motivation* one, while the *Preprocess, cleaning, labelling procedures* section is the least complete.

Within the OpenML repository, represented in figure 4.3, the most complete dataset documentation is the one attached to **oml08** dataset, i.e. *mnist_784*, an handwritten digits database. The dataset with the minimum average is **oml19**, the *mammography* dataset, the least documented dataset of all the datasets analysed in the survey.

This repository doesn't contain fields calculated in an automatic way.

The most present information, i.e. that information represented by fields present in the qualified majority (greater than two thirds) of the repository datasets, are **1.1**,**1.2**, **2.1**, **2.2**, **2.3**, **2.8**, **2.11**, **5.1**, **5.5**, **6.1**. Half of these fields belongs to the *Composition* section. Nevertheless, the *Uses* section is the most complete, while the *Maintenance* section is the least complete.

The UCI Machine Learning Repository is represented in figure 4.4. The most complete dataset documentation is the one attached to **uci20** dataset, the *Pen-based recognition of handwritten digits*. The dataset with the minimum average is **uci02**, the *Diabetes* dataset.

This repository contains one field calculated in an automatic way: the **c.04** *Recently updated* characteristics.

The most present information, i.e. that information represented by fields present in the qualified majority (greater than two thirds) of the repository datasets, are **2.1**, **2.2**, **2.3**, **5.1**, **5.4**, **5.5**, **6.1**. Three of these fields belongs to the *Composition* section and other three belongs to *Uses* section. The latter results to be the most complete section, while the *Collection processes* section is the least complete.

Figure 4.1: Huggingface raw data

Recommended information					Datasets																										
Section	**Sec %**	**Field ID**	**Field automation**	**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	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		
		1.02	0	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	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	0	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	1	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	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	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	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	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	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	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	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	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	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	1	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	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	1	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	1	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	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	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	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	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	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	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	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	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	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	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	0	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		

Figure 4.2: Kaggle raw data

Recommended information					Datasets																										
Section	**Sec %**	**Field ID**	**Field automation**	**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	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		
		1.02	0	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	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	0	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	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	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	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	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	0,08	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0		
		2.07	0	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	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	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	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	0	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	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	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	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	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	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	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		0,31
		3.04	0	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	0,00	0	0	NA	0	NA	0	NA	NA	NA	0	0	NA	0	0	0	NA	NA	NA	0	NA	NA	FALSO	NA	0	NA		
		3.06	0	0,00	0	0	NA	0	NA	0	NA	NA	NA	0	0	NA	0	0	0	NA	NA	NA	0	NA	NA	FALSO	NA	0	NA		
		3.07	0	0,00	0	0	NA	0	NA	0	NA	NA	NA	0	0	NA	0	0	0	NA	NA	NA	0	NA	NA	FALSO	NA	0	NA		
Preprocess / cleaning / labelling	0,09	4.01	0	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	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	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	0	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	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	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	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	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	0	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	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	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	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	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	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	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	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	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	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	1	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		

Figure 4.3: OpenML raw data

Recommended information					Datasets																										
Section	**Sec %**	**Field ID**	**Field automation**	**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	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	0	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	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	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	0	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	0	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	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	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	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	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	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	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	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	0	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	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	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	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	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	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	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	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	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	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	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	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	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	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	0	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	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	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	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	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	0	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	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	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	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	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	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	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	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	0	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	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	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	

Figure 4.4: UCI Machine Learning Repository raw data

Recommended information					Datasets																									
Section	**Sec %**	**Field ID**	**Field automation**	**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	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	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	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	0	
Composition	0,26	2.01	0	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	0	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	0	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	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	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	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	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	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	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	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	0,88	1	0	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	
		c.03	0	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	1	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	

4.2 Datasets level

On the basis of the raw data shown in the previous section, it is possible to begin an analysis aimed at attempting to explicate some data properties obtained, in order to draw useful conclusions for the research work. This section will analyse the data from the perspective of datasets, first from a general point of view, then on the basis of datasets characteristics. The average measure is calculated by summing the values (0 or 1) of all fields within the same grouping under investigation (repository, dataset, section, field).

The dataset with the most comprehensive documentation results **hug16**, the *cnn_dailymail*[4] one. It contains over 300k unique news articles written by journalists at CNN and Daily Mail. His *Dataset Card* results very comprehensive in all the different sections and it can be considered a positive reference point from the point of view of documentation practice.

4.2.1 Repositories

The repository with the higher average is Huggingface, while the one with the lower average is UCI Machine Learning Repository. Figure 4.5 shows that Kaggle and OpenML achieve a very similar result. The fact that the three most complete datasets (from a documentation point of view) are datasets from the Huggingface repository certainly contributes to the Huggingface best average result. Despite the ranking observations, data shows small variations between repositories.

In addition to the repository average, it is possible to analyse the distribution of the dataset mean. Figure 4.6 shows that Huggingface mean is certainly influenced by the top one dataset outlier. OpenML, on the other hand, shows that his average value is influenced by the worst one dataset outlier.

Moreover, an attempt was made to understand whether there is any correlation between popularity and completeness of documentation. However, the data, shown in figure 4.7, disprove this hypothesis. One justification could be found in the fact that this analysis was conducted on the 25 most popular datasets of each repository: future work could deepen this by analysing more datasets.

[4] `https://huggingface.co/datasets/cnn_dailymail`

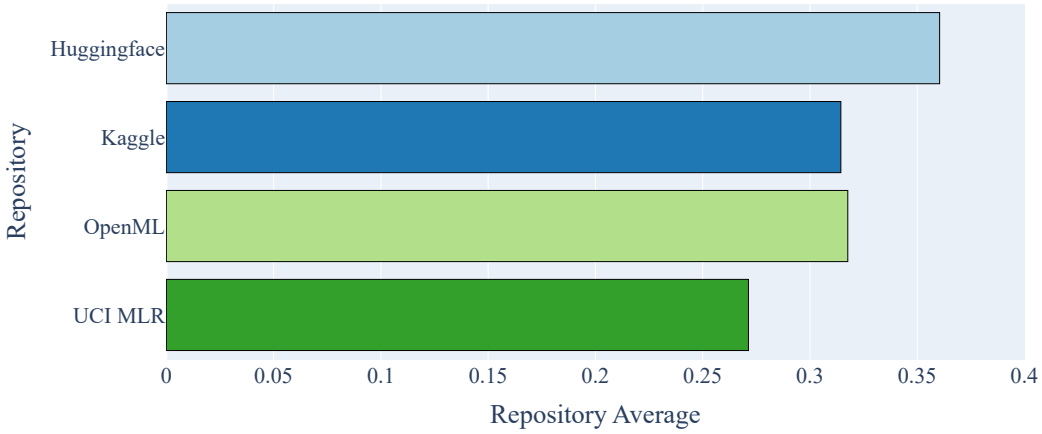
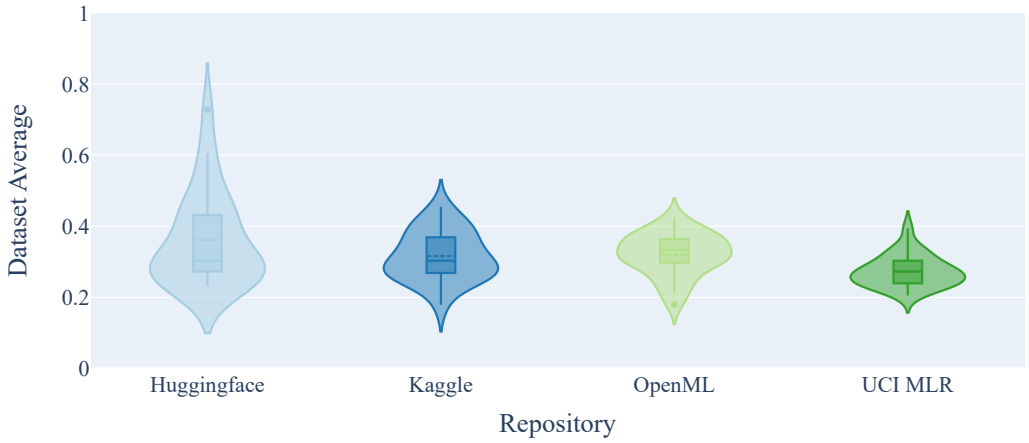


Figure 4.5: Datasets mean of available information per repository

Figure 4.6: Distribution of mean available information per dataset grouped by repository



4.2.2 Completeness according to dataset characteristics

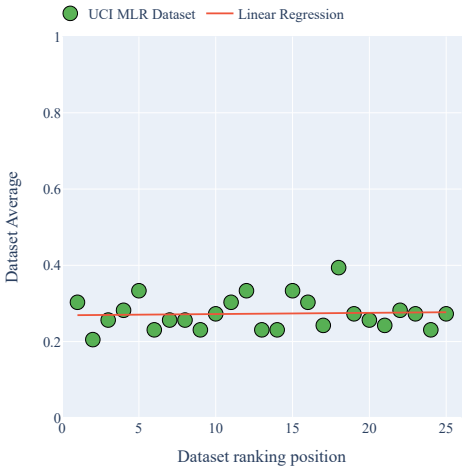
An interesting result emerge from the analysis of the mean amount of information accompanying the dataset according to different characteristics of the dataset itself

Figure 4.7: Correlation between the dataset average and the ranking of the dataset within its repository

(a) Huggingface

(b) Kaggle

(c) OpenML



(d) UCI Machine Learning Repository

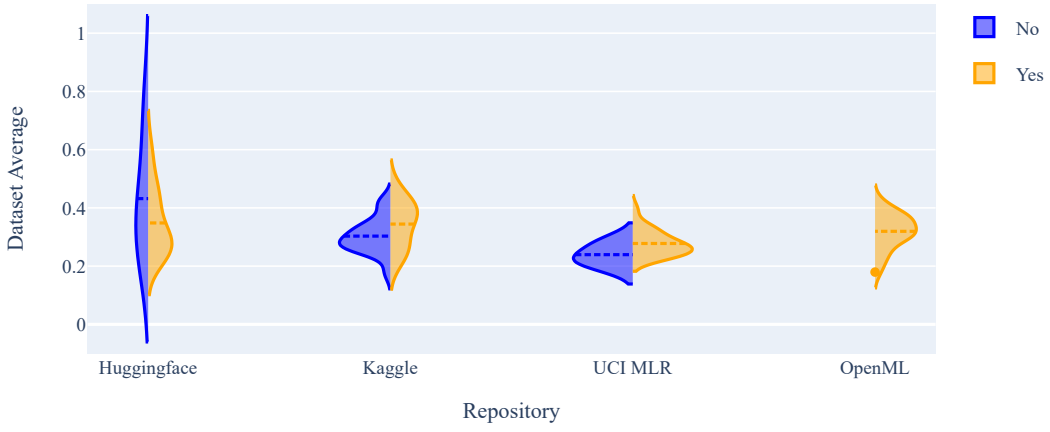
(figure 4.8). Figure 4.8a shows that datasets containing people-related data present on average a lower amount of documentation information.

No particular conclusions can be drawn from the results on the basis of the other characteristics. On the one hand, because in some cases they show divergent results depending on the repository (figure 4.8b and figure 4.8c), highlighting how

these characteristic is not always so relevant in terms of the completeness of the documentation. On the other hand, because in some cases, such as the fact that the dataset was recently updated, the number of datasets belonging to the two identified classes is not always adequate to draw conclusions.

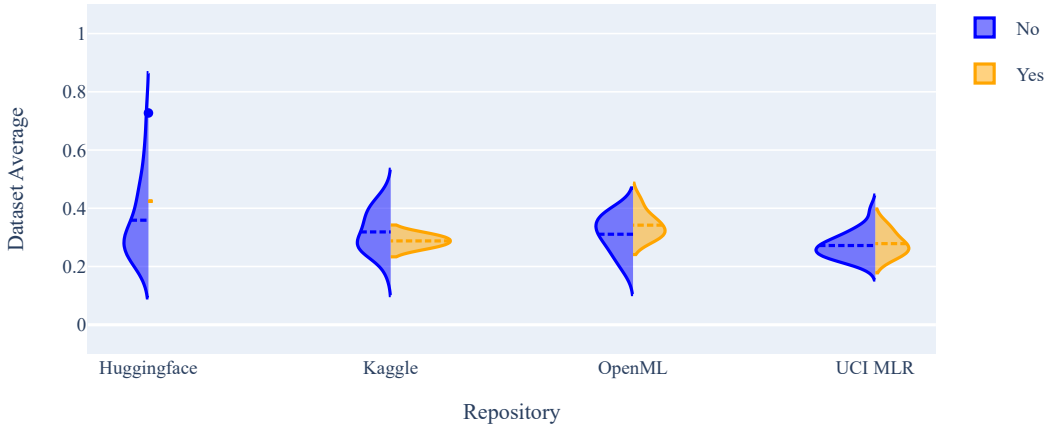
Figure 4.8: Mean amount of information accompanying the dataset according to different characteristics of the dataset itself

(a) Data is people related

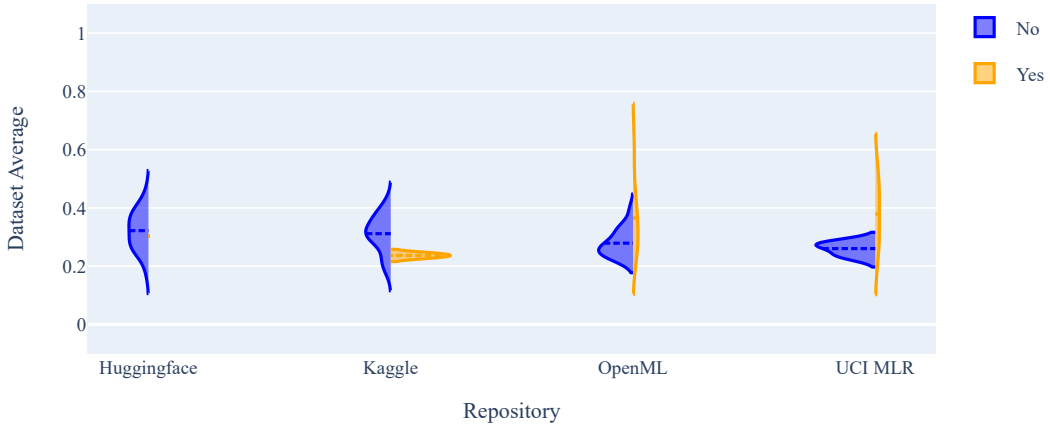


(b) Presence of explicit label

Figure 4.8: Mean amount of information accompanying the dataset according to different characteristics of the dataset itself



(c) Dataset is a sample/reduction of a larger set



(d) Recently updated

4.3 Sections level

A further level of analysis is the analysis by category of information, i.e. sections. This type of analysis is useful to dig deeply in the documentation information presence. While the datasets averages is useful to extract peculiarities of a documentation as a whole, with sections it is possible to understand which aspects of documentation are better elaborated and which not, allowing us to focus on the parts most affected by opacity (or sparsity).

Figure 4.9 (data are presented in table 4.1) make it possible to understand at a glance which parts receive the most attention in the analysed repositories. Data shows that dataset publishers pay the majority of their focus on information related to how use data contained within them. This result testifies to the extent to which the documentation production of public datasets is currently purely utilisation-oriented, without paying too much attention to various ethically relevant aspects.

It is indeed very difficult to find information concerning all the various data collection processes and data transformation steps: all extremely delicate phases in which the failure to take certain contextual aspects into account can lead to various problems in the models trained on these data, as described by Sambasivan et al. [67]. The high result of the *Motivation* section can also be partially justified by a greater focus on purely 'technical' aspects: the purpose of the dataset often encapsulates the meaning of why the data within it should be used. The result of the *Composition* section is often influenced by the fact that repositories implement special sections on their website that display certain data characteristics (more or less) automatically for the dataset publisher: this underlines the importance of the role played by dataset hosts, who have the tools and possibilities to trigger virtuous mechanisms from the perspective of documentation quality. The Maintenance section, finally, shows the lack of attention paid to what happens after the dataset is published, justifying the attention shown by works such as that of Corry et al. [14].

Table 4.1: Mean of available information per section and repository

Section	**Sec ID**	**Sec AVG**	**Hugging face**	**Kaggle**	**Open ML**	**UCI MLR**
Motivation	1	0,50	0,56	0,52	0,56	0,35
Composition	2	0,29	0,28	0,28	0,35	0,26
Collection processes	3	0,16	0,19	0,22	0,16	0,09
Preprocess / cleaning / labelling	4	0,17	0,25	0,09	0,20	0,15
Uses	5	0,57	0,59	0,48	0,53	0,69
Maintenance	6	0,27	0,40	0,34	0,18	0,15

Figure 4.9: Mean of available information per section

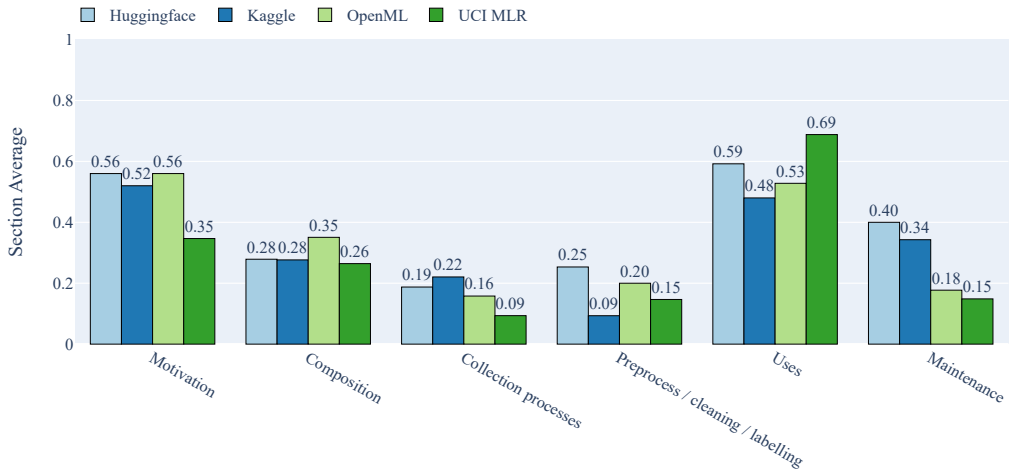
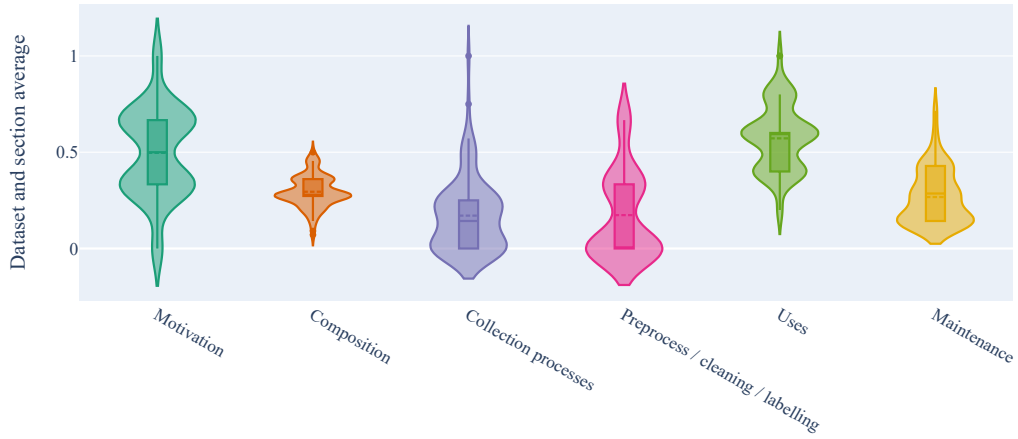


Figure 4.10 allows us to observe the distribution of averages obtained by grouping the fields of a given dataset in a given section (without subdividing by repository). The characteristic hourglass shape is due to the fact that the number of average values obtainable for each section of each dataset is equal to the number of fields plus one. For example, since the *Motivation* section consists of three fields, each dataset may have a mean value of this section of 0, 0.33, 0.67 or 1.

Further research was carried out into the distribution of dataset section averages, this time keeping the different repositories divided. Figure 4.11 show the single dataset section values, paired with the relative box with the quartile division, the mean value and the standard deviation (dotted line forming the shape of a triangle). This graph allows us to observe that in roughly all sections there are several positive outliers, representing fairly complete portions of documentation. This is not the case in the *Composition* section, where no dataset exceeds 50% of the searched information.

Finally, the sections of the datasets were graphed (figure 4.12) in a plane in which the x-axis represents the global average of the dataset and the y-axis the average of the individual section of the dataset. This type of representation is useful for understanding 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

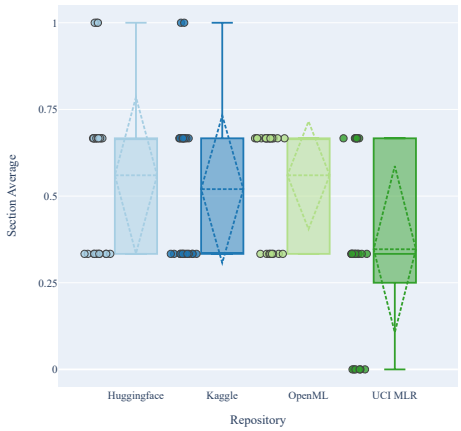
Figure 4.10: Distribution of mean available information grouped by section and dataset



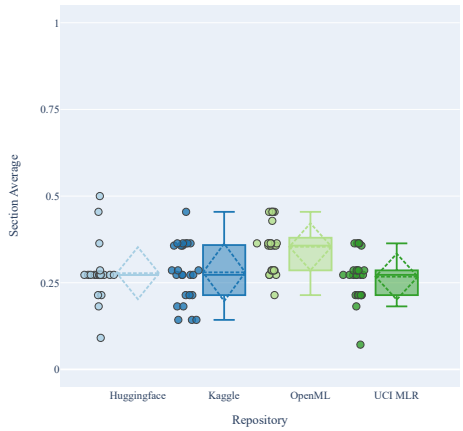
once again emerges: figure 4.12e shows the high quantity of individual datasets with a low global average and a high average of the *Uses* section.

In addition, frequency histograms of the dataset mean and of the dataset section mean can be observed in the graph. These histograms show a Gaussian distribution for the global average of the datasets and for the sections *Motivation*, *Composition* and *Uses*. *Preprocess, cleaning, labelling procedures* and *Maintenance*, on the other hand, show a decreasing frequency as the section average increases.

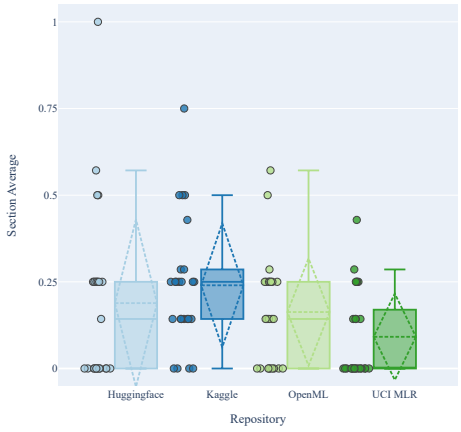
Figure 4.11: Distribution of mean available information grouped by section, repository and dataset



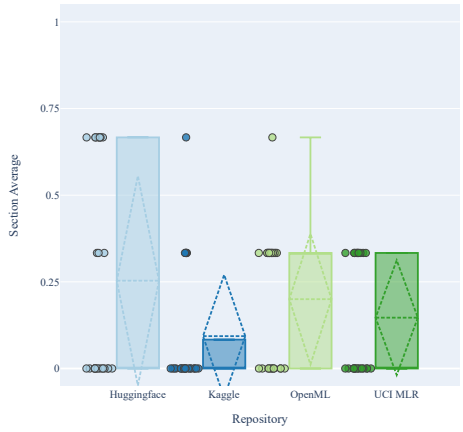
(a) Motivation



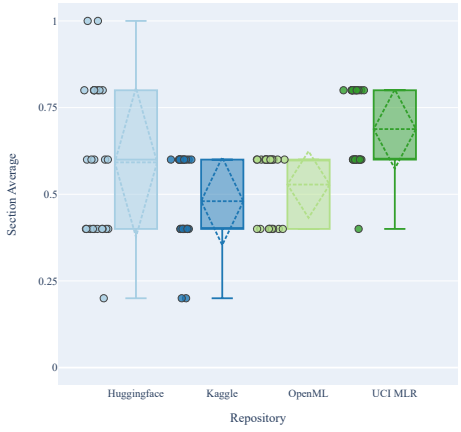
(b) Composition



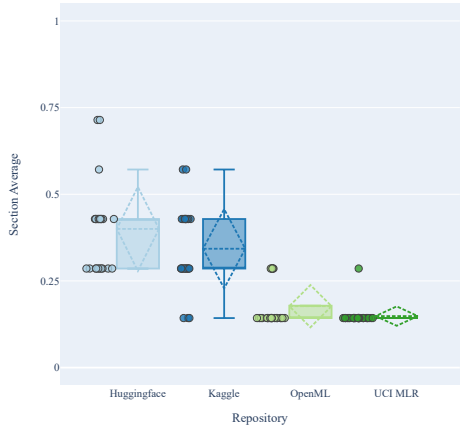
(c) Collection process



(d) Preprocess, sample, cleaning, labeling

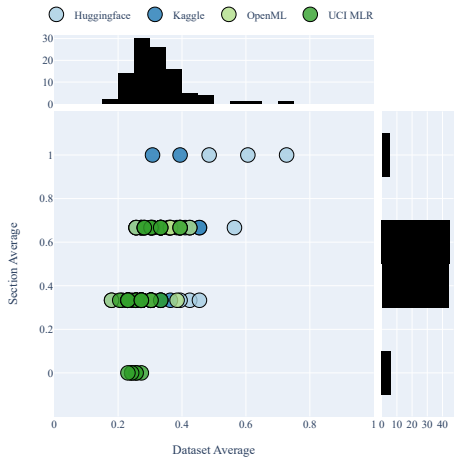


(e) Collection process

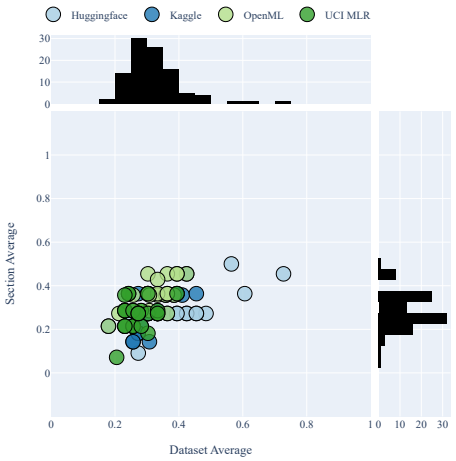


(f) Preprocess, sample, cleaning, labeling

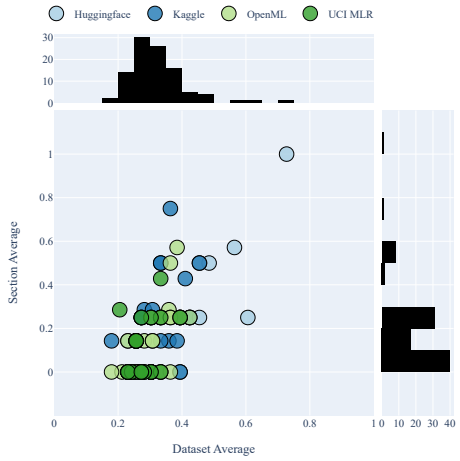
Figure 4.12: Dataset and section mean of available information



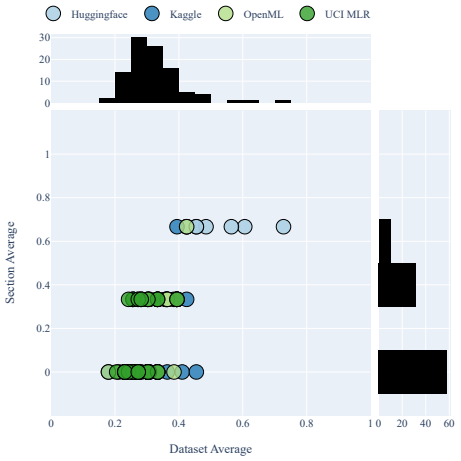
(a) Motivation



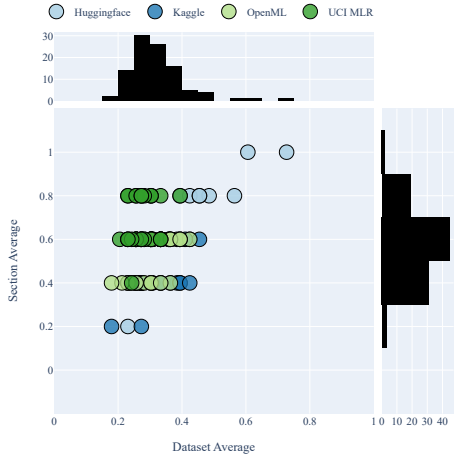
(b) Composition



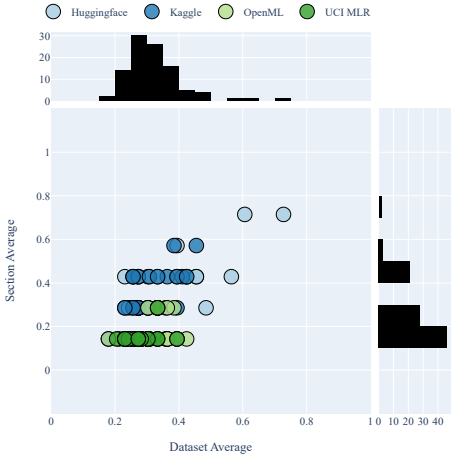
(c) Collection process



(d) Preprocess, sample, cleaning, labeling



(e) Uses



(f) Maintenance

4.4 Fields level

The last level of investigation is the individual fields of information. This part of the discussion of the results is divided into two different sections: a first part focused on the analysis of the averages of information presence of individual fields, and a second part on the search for possible correlations between two distinct fields.

4.4.1 Field completeness

Table 4.2 highlight the average value of each information field, globally and by repository. One fact that emerges from the graphical observation of the table is that there are certain fields whose frequent presence (or vice versa, whose frequent absence) is common to all repositories. This evidence can be explained by the fact that some information are really common, such as the **2.01** *What do the instances that comprise the dataset represent*, **2.02** *Number of the instances* and **5.01** *Description of the tasks in which the dataset has already been used and their results*.

Other fields, instead, reveal the ability of a repository to foster the presence of a specific piece of information, as described in chapter 4.3. Some examples to illustrate this concept are: **2.04** *Recommended data splits*, **2.11** *Statistics*, **5.04** *Repository that links to papers or system that use the datasets*, **5.05** *Description of license and terms of use*, **6.05** *Information about dataset updates* and **6.07** *Information about mechanism to extend, augment, build on, contribute to the dataset*. These fields are very much present in repositories that structurally exposes the information their represent in the metadata schema of the repository. Conversely, they are almost totally absent in repositories that do not include such information in their metadata schema.

Table 4.2: Fields averages for each repository

Field ID	Field AVG	Huggingface	Kaggle	OpenML	UCI MLR
1.01	0,57	0,64	0,52	0,68	0,44
1.02	0,86	0,88	0,96	1,00	0,60
1.03	0,06	0,16	0,08	0,00	0,00
2.01	0,92	1,00	1,00	0,80	0,88
2.02	0,90	0,92	0,72	1,00	0,96
2.03	0,50	0,00	0,12	1,00	0,88
2.04	0,31	0,92	0,08	0,12	0,12
2.05	0,13	0,00	0,16	0,08	0,28
2.06	0,04	0,08	0,08	0,00	0,00

Continue on next page

Table 4.2 – continued from previous page

Field ID	Field AVG	Huggingface	Kaggle	OpenML	UCI MLR
2.07	0,03	0,12	0,00	0,00	0,00
2.08	0,43	0,25	0,42	0,64	0,31
2.09	0,15	0,50	0,17	0,09	0,08
2.10	0,03	0,25	0,00	0,00	0,00
2.11	0,50	0,00	1,00	1,00	0,00
2.12	0,00	0,00	0,00	0,00	0,00
2.13	0,00	0,00	0,00	0,00	0,00
2.14	0,00	0,00	0,00	0,00	0,00
3.01	0,53	0,52	0,60	0,64	0,36
3.02	0,08	0,16	0,12	0,00	0,04
3.03	0,19	0,04	0,48	0,12	0,12
3.04	0,01	0,04	0,00	0,00	0,00
3.05	0,05	0,25	0,00	0,09	0,00
3.06	0,05	0,25	0,00	0,09	0,00
3.07	0,00	0,00	0,00	0,00	0,00
4.01	0,39	0,32	0,24	0,56	0,44
4.02	0,11	0,44	0,00	0,00	0,00
4.03	0,02	0,00	0,04	0,04	0,00
5.01	0,95	0,92	1,00	1,00	0,88
5.02	0,62	0,56	0,72	0,64	0,56
5.03	0,02	0,08	0,00	0,00	0,00
5.04	0,48	0,92	0,00	0,00	1,00
5.05	0,79	0,48	0,68	1,00	1,00
6.01	0,84	0,36	1,00	1,00	1,00
6.02	0,30	0,20	0,80	0,16	0,04
6.03	0,09	0,24	0,04	0,08	0,00
6.04	0,00	0,00	0,00	0,00	0,00
6.05	0,38	1,00	0,52	0,00	0,00
6.06	0,00	0,00	0,00	0,00	0,00
6.07	0,26	1,00	0,04	0,00	0,00
c.01	0,38	0,16	0,48	0,44	0,44
c.02	0,76	0,84	0,32	1,00	0,88
c.03	0,14	0,04	0,08	0,28	0,16
c.04	0,30	1,00	0,20	0,00	0,00

4.4.2 Correlations

The data presented and discussed in the previous chapters may raise the following question: *are there correlations between individual fields?* This turns out to be an interesting question, and although this question may require further specific investigation, we went in search of associative rules between fields. Given the complexity due to the number of hypotheses to be examined, in order to limit the search space, only association rules between two individual fields were considered. In other words, correlations were sought between a field A and a field B other than A.

To do this, a 2-way contingency table for variables A and B was first constructed, with the structure shown in table 4.3. In the context of this research work, f_{11} represent the number of times where information represented by field A and information represented by field B are both present in a dataset documentation. For the purposes of this exposition, each dataset represent a *transaction*.

Table 4.3: Contingency table for A and B

		B	
		0	1
A	0	f_{00}	f_{01}
	1	f_{10}	f_{11}

Due to the binary representation characteristic of the data subject of this research, we opted to measure the strength of the association rules in terms of support (definition 4.4.1) and confidence (definition 4.4.2), as presented by Larose and Larose [45]. Supports provides a measure of how much two fields A and B are present in the same dataset documentation. Confidence measures the measures how likely it is that field B is present in a dataset documentation that contains field A.

Definition 4.4.1 (Support) *The support of A is the fraction of transaction containing A. Given A:*

$$s(A) = P(A) = \frac{\textit{Number of transaction containing } A}{\textit{Number of transactions}}$$

The support of A and B is the fraction of transaction containing both A and B. Given A,B:

$$s(A \cap B) = P(A \cap B) = \frac{\textit{Number of transaction containing } A \textit{ and } B}{\textit{Number of transactions}}$$

Definition 4.4.2 (Confidence) *The confidence is the frequency of B in transactions containing A. Given A,B:*

$$c(A \to B) = \mathbb{P}(B|A) = \frac{s(A \cap B)}{s(A)} = \frac{\textit{Number of transactions containing A and B}}{\textit{Number of transactions containing A}}$$

Based on these two definitions, support and confidence measures was calculated for all pairs of fields, as described in equations 4.1 and 4.2.

$$s(A \cap B) = \frac{f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \tag{4.1}$$

$$c(A \to B) = \frac{f_{11}}{f_{10} + f_{11}} \tag{4.2}$$

Results are shown graphically in figure 4.13. The support matrix is symmetrical by definition, since the number of transactions containing A and B is equal to the number of transactions containing B and A. Figure 4.13a display the support of each pair of fields. Figure 4.13b display instead the confidence value of the pairs of fields above the support and confidence thresholds. In order to limit the search space, were set a minimum support threshold equals to 0.35 and a minimum confidence threshold equals to 0.9.

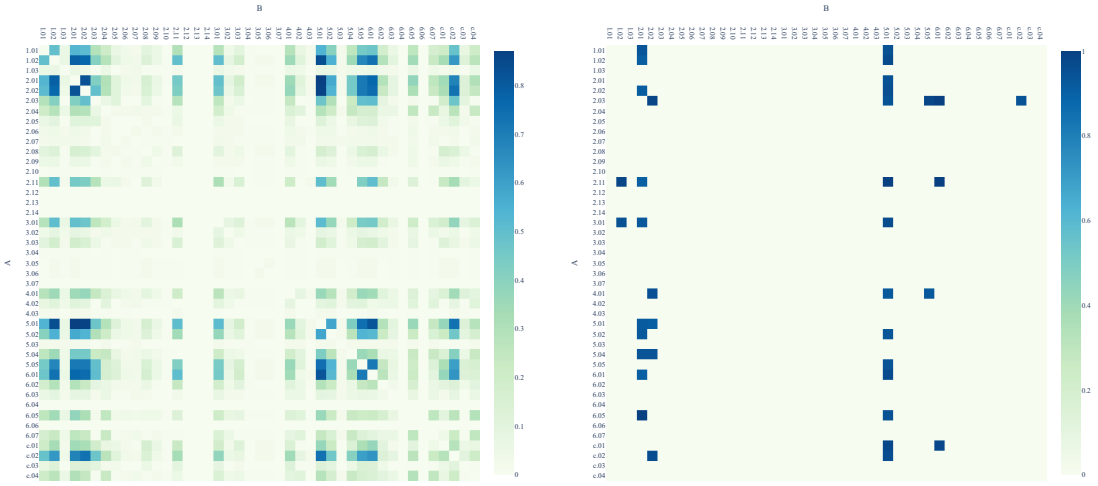
Table 4.4 highlights the residual field pairs after pruning as just described. Figure 4.14 offer a graph representation based on association rules: the arrows dimension is determined by the confidence measure between the two fields.

An important premise about the results obtained in the search for associative rules is that the inference from a field A to another field B does not necessarily imply a causal relationship between the two fields and may simply explicate a strong co-occurence relationship [73]. The most relevant relationships results to be the ones oriented toward the field **5.01** *Description of the tasks in which the dataset has already been used and their results.* As stated above, it sometimes difficult to see a causal relationship in some pairs.

On the basis of the results obtained, the following rules were deemed interesting.

c.01 → 5.01 (support=0.37, confidence=0.97) and *c*.01 → 6.01 (support=0.37, confidence=0.97). These relationship from **c.01** *Data is people related* can be a symptom of a correlation between the fact that dataset contains people related-data and the presence of some particular type of information. Taking a closer look at the data it is possible to realise that fields **5.01** *Description of the tasks in which the dataset has already been used and their results* and **6.01** *Information about subject supporting, hosting, maintaining the dataset* are present in most datasets, in 95% and 84% respectively. Thus, it could be assumed that the high confidence

Figure 4.13: Support and Confidence of contemporary information presence between two fields



(a) Support **(b)** Confidence

measure is simply due to the large presence of the latter two fields. On the other hand, there was no guarantee that in the case of data on persons, these fields were almost always present. What can be revealed from these results is that, although on average the datasets consisting of people-related data do not possess a greater completeness of documentation, it is possible to observe a relevant focus on the description of the tasks that have already used the dataset and on the subjects involved in supporting, hosting and maintaining the dataset.

1.02 → 5.01 (support=0.83, confidence=0.97). The relationship between **1.02** *Dataset creators* and **5.01** *Description of the tasks in which the dataset has already been used and their results* reveal that when it was written in a clear way the dataset creators, it si very likely that will be present in the same documentation some referrals to tasks that have already used the datasets. As mentioned above, field **5.01** is the field with the higher presence average, therefore, it is needed to take these conclusions with caution and keep in mind that this correlation may be due to the simple statistical distribution of the data.

3.01 → 1.02 (support=0.50, confidence=0.94). The relationship between **3.01** *Description of instances acquisition and data collection processes* and **1.02** *Dataset creators* exposes the fact that when, within the documentation, it is described in sufficient detail how the data were collected, it is very likely that attention will also be paid to exposing some kind of reference to the creators of the dataset.

Table 4.4: Fields association rules

Field A	Field B	Support	Confidence
2.03	6.01	0,50	1,00
2.11	5.01	0,50	1,00
2.11	6.01	0,50	1,00
6.05	2.01	0,38	1,00
2.03	2.02	0,49	0,98
2.03	5.05	0,49	0,98
2.11	1.02	0,49	0,98
c.01	5.01	0,37	0,97
c.01	6.01	0,37	0,97
1.02	5.01	0,83	0,97
6.01	5.01	0,81	0,96
3.01	5.01	0,51	0,96
c.02	2.02	0,73	0,96
c.02	5.01	0,73	0,96
2.02	5.01	0,86	0,96
4.01	2.02	0,37	0,95
1.01	5.01	0,54	0,95
6.05	5.01	0,36	0,95
2.01	5.01	0,87	0,95
3.01	1.02	0,50	0,94
2.03	5.01	0,47	0,94
2.03	c.02	0,47	0,94
5.04	2.01	0,45	0,94
5.04	2.02	0,45	0,94
5.05	5.01	0,74	0,94
5.02	5.01	0,58	0,94
1.01	2.01	0,53	0,93
3.01	2.01	0,49	0,92
4.01	5.01	0,36	0,92
4.01	5.05	0,36	0,92
5.02	2.01	0,57	0,92
5.01	2.01	0,87	0,92
2.02	2.01	0,82	0,91
1.02	2.01	0,78	0,91
5.01	2.02	0,86	0,91

5.02 → 5.01 (support=0.58, confidence=0.94). The relationship between **5.02** *Description of recommended uses or tasks* and **5.01** *Description of the tasks in which the dataset has already been used and their results* makes explicit the fact that when it is present some specific hint about recommended uses or tasks, it is pretty common to find some referrals to tasks that have already used the dataset.

To sum up, the search for correlations between fields has provided hints about possible implications between one type of information and another, but these are often due to the statistical properties of the data. It might be interesting to explore this further by implementing special associative rule generation algorithms, such as the Apriori algorithm and the FP-growth algorithm.

4.5 Comparison between manual and automatic check

One of the main aspects taken into account during research design concerns the feasibility of an automated systems able to check information presence in dataset metadata. As described in chapter 3.3.3 a test was conducted with a view to automation.

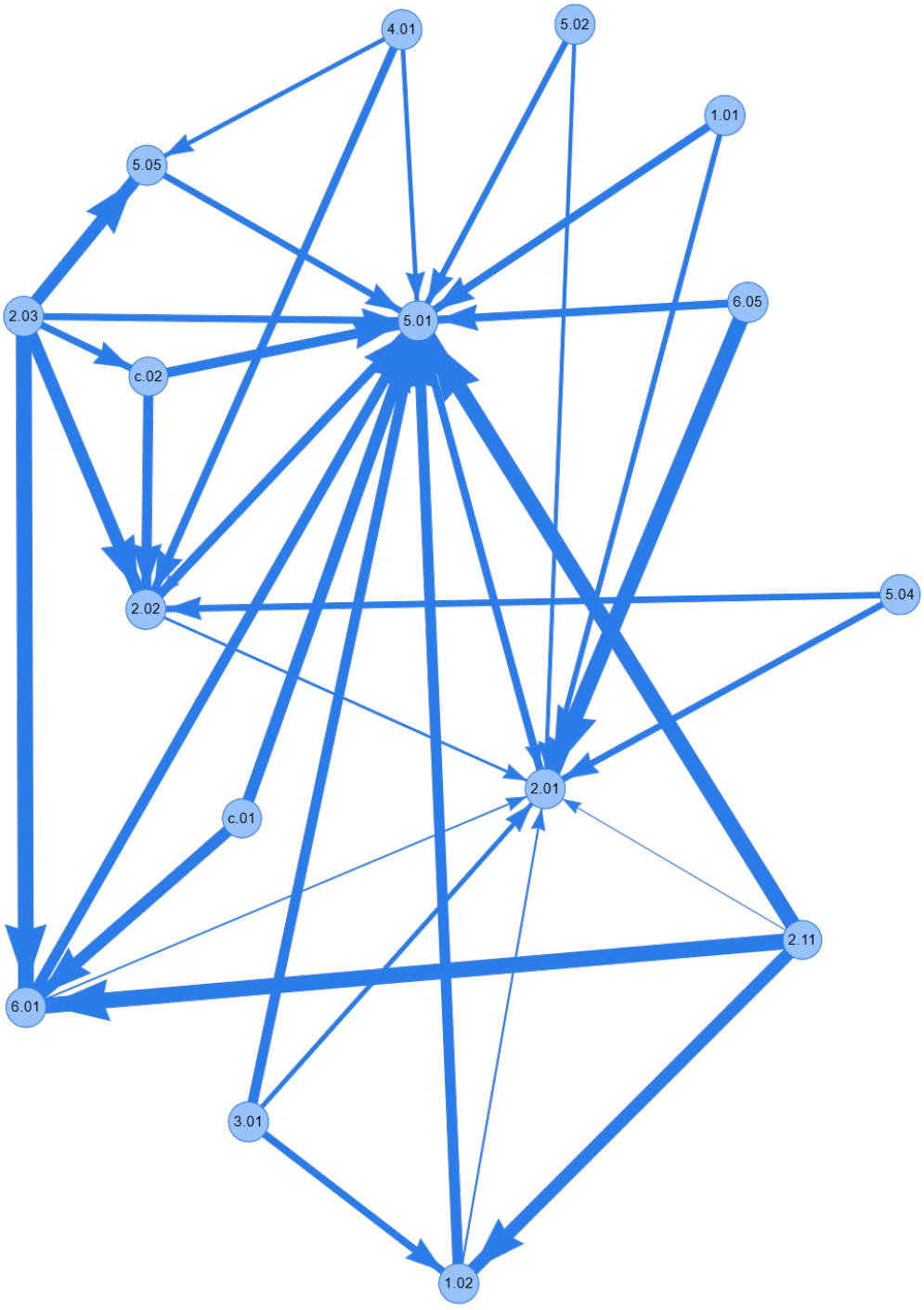
Table 4.5 expands data presented in table 3.2 comparing the results of the implemented automated system on dataset metadata with results obtained from the manual inspection of dataset documentation. Automated check results is obtained from the metadata analysis of the whole repository, while the manual inspection focused on the top 25 datasets.

On one hand, it is possible to notice a very similar results on fields **1.02** *Dataset creators*, **2.01** *What do the instances that comprise the dataset represent*, **2.02** *Number of the instances* and **5.05** *Description of license and terms of use*. On the other hand, results from **1.01** *Purpose for the dataset creation*, **2.03** *Information about missing values*, **4.01** *Description of sampling, preprocessing, cleaning, labeling procedures* and **5.01** *Description of the tasks in which the dataset has already been used and their results* highlight very different results. Moreover, fields **1.03** *Dataset funders*, **2.10** *Description of data sensitivity (if people related)* and **6.03** *DOI* results empty in both the inspection types. By restricting the list of fields to the set in this table, results of acceptable quality can be revealed. The problem lies in the fact that the fields shown here represent only the 28% of the number of information fields taken into account for the UCI repository. These results give us a further proof of the only limited applicability of these metadata fields to create an automatic system capable of verifying the presence of information.

Table 4.5: Comparison between automated system and manual inspection

Test Field ID	UCI MLR Field	Automated Inspection	Manual Inspection
1.01	14	0,01	0,44
1.02	13	0,55	0,60
1.03	15	0,00	0,00
2.01	16	0,99	0,88
2.02	6	1,00	0,96
2.03	23	0,57	0,88
2.04	17	0,00	0,12
2.10	18	0,00	0,00
4.01	19	0,00	0,44
5.01	20	0,00	0,88
5.05	25	0,99	1,00
6.03	11	0,00	0,00

Figure 4.14: Oriented graph of the fields association rules



Chapter 5

Conclusions and future work

This thesis work on ethical manufacturing of datasets for artificial intelligence focused on an empirical investigation on the state of documentation practice. The aim was to understand which information should always be clear to datasets' stakeholders in order to achieve transparency and accountability, and measure how much this information is present in the documentation of the most popular datasets in AI community. In this research work, a set of relevant information that should be transparent to users of datasets was presented and adapted into a Documentation Test Sheet (DTS) capable of measuring the completeness of documentation. The results of applying this test to the most popular datasets within the Huggingface, Kaggle, OpenML and UCI ML repositories were presented.

In order to address **RQ1: What information should be transparent to dataset users?**, documentation standardization proposals from Gebru et al., Holland et al. and Bender and Friedman has been used as starting point in order to develop the DTS. It covers different category of information: *Motivation*, *Composition*, *Collection processes*, *Preprocess, cleaning, labelling procedures*, *Uses*, and *Maintenance*. The aim is to ensure that AI practitioners are fully aware of the characteristics of the data, the processes by which it was produced and transformed, how to use it and the potential risks associated with it. Put another way, it is crucial that the context that led to the final form of the dataset as an artefact is transparent. For the purposes of this work, it was decided to take this context into account as broadly as possible, for instance by including explicit references to the remuneration of the persons producing or processing the data.

DTS can be useful to measure one specific aspects of dataset documentation quality: completeness. On one hand it can be used by dataset hosts and dataset consumers to quickly and simply check the completeness of a documentation. On

the other it can serve as a guideline for dataset creators, helping them to improve their documentation so that dataset consumers can verify the underlying choices and assumptions. Moreover, completeness reinforces accountability, making the DTS useful also from this point of view.

In order to address **RQ2: Which of the information that should be transparent to dataset users, is present in the most popular datasets?**, automatic and manual inspections has been performed on datasets metadata and documentation, focusing on the information available in the very same place where data can be accessed. This investigation brought out some relevant results about the state of practice of documentation of datasets manufacturing. First it emerges that dataset publishers pay the majority of their focus on information related to how use data contained within them. On the contrary, maintenance over time or processes behind the data generation result very poorly documented. In general, a lack of relevant information was observed, highlighting a paucity of transparency. All these observations are even more relevant when considering that the analysis was restricted to some of the most popular and well-known datasets.

Moreover, datasets containing people related data present on average a lower amount of documentation information. On the other hand, however, some association rules was observed from the presence of people related data. Support and confidence metrics reveal an implication from that to the presence of descriptive information on tasks that have already used the dataset. Another one association rule was founded from the presence of people related data to the presence of information on subjects that maintain the dataset over time.

Another interesting fact that can be detected from the data obtained, concerns the repositories potential in order to help dataset producer to produce better documentation. Indeed, when a specific piece of information is present in approximately all the datasets of a repository, but it is very difficult to find the same information in other repositories, the causes are often related to the structure of the datasets' metadata scheme offered or not by the repository.

One of the main limitations of this research work can be found in the partial non-scalability of the proposed procedure, which was primarily based on manual inspection of dataset documentation. Moreover, the number and the types of the selected repositories was focused on some artificial intelligence sectors, neglecting other relevant ones, such as computer vision. A further limiting aspect is represented by the fact that the procedures exposed in this thesis tried to evaluate only a single quality aspect of dataset documentation - completeness - leaving out other quality aspects such as correctness, relevance, etc.

The obtained results and the limitations highlighted provide insights and suggestions on possible expansions of this thesis. A first hypothesis of future work is

related to increasing the number of datasets under investigation. This would provide a more complete view of the state of the art and could allow us to investigate whether there is any form of correlation between the popularity of a dataset and the completeness of the documentation related to it (correlation not revealed by the data obtained through the datasets selected for this work).

Quantitative expansion of the research could be put in place starting with investigations into the feasibility of an automatic system capable of controlling the presence of information. A first possibility is an expansion, in quantitative and qualitative terms, of the metadata useful to confirm or not the presence of a given piece of information. This can be done expanding the preliminary work discussed in section 3.3.3 and in section 4.5. A further perspective would be to study the feasibility of a natural language processing model specifically trained to check the presence of information represented by the DTS fields, in a given text.

Remaining within the confines of this work, section 4.4.2 concerning the search for possible correlations between individual fields of information could be further expanded. In the process of generating associative rules, it might be interesting to also search for correlations between more than two fields. For this purpose, algorithms specifically designed to make this search feasible, such as the Apriori algorithm and the FP-growth algorithm, could be considered.

From the qualitative point of view, it might be fascinating to expand the DTS in order to also take into account (measuring them) other aspects of documentation quality, as mentioned above. Based on the work and taxonomy proposed by Fabris et al. could be interesting to introduce a *sparsity* measure, e.g. by trying to compare the information found in the repositories and the information that can instead be retrieved by analysing the papers with which the datasets have been published, if any. Moreover, this research could be useful in identifying the most popular and at the same time most problematic datasets from a documentation point of view. Retrospective documentation work could be carried out on those datasets in order to reduce documentation debt, as shown in [21, 6, 25].

Altogether these results show that huge efforts of the AI community in devoting more attention to the dataset documentation process are urgent and necessary. There are no purely technical aspects, and every 'technical' choice that led to the construction of a given model hides behind it a set of ethical considerations, regardless of whether the context is to be taken into account or not. The recommended path should be supported by the investigation and experimentation of techniques to fully integrate documentation models and processes into the AI pipeline, in order to reduce discriminations and to facilitate human-respectful AI innovations.

Appendix A

Metadata download scripts

Listing A.1: Hugginface metadata download script

```
from datasets import list_datasets
import pandas as pd
from datetime import datetime
datasets_list = list_datasets()
print(f"Currently {len(datasets_list)} datasets are available on the hub:")
datasets = []
numrecord = 0
for i in range(0,len(datasets_list)+1):
    # You can access various attributes of the datasets before downloading them
    dataset_with_details = list_datasets(with_details=True)[i]
    datasets.append(dataset_with_details.__dict__)
    numrecord+=1
    if numrecord%50==0:
        print(f'[{datetime.now().strftime("%H:%M:%S")}] records @ {numrecord}')

print(f'[{datetime.now().strftime("%H:%M:%S")}] done #{numrecord} records')

djn = pd.json_normalize(datasets)
datasets_df = pd.DataFrame.from_dict(djn, orient='columns')
result = datasets_df.to_json(orient="columns")
file=open('huggingface.json',"w",encoding='utf-8')
file.write(result)
file.close()
```

Listing A.2: Kaggle metadata download script

```
from datetime import datetime
#sortby Valid options are 'hottest', 'votes', 'updated', and 'active'
def kaggle_datasets_list(sortby):
    file=open(f'kaggle-{sortby}.csv',"w",encoding='utf-8')
    file.write("ref,title,size,lastUpdated,downloadCount,voteCount,
        usabilityRating\n")
    file.close()
    file=open(f'kaggle-{sortby}.csv',"a",encoding='utf-8',errors='
        backslashreplace')
    datasets_raw_csv = "ref,title,size,lastUpdated,downloadCount,
        voteCount,usabilityRating\n"
    numrecord = 0
    page = 0
    for page in range(1,501):
        command = f' datasets list --sort-by {sortby} -p {page} --csv --
            min-size 1'
        datasets_page = !kaggle{command}
        datasets_page.remove('ref,title,size,lastUpdated,downloadCount,
            voteCount,usabilityRating')
        if page%25==0:
            print(f'[{sortby}-{datetime.now().strftime("%H:%M:%S")}]
                page@{page}')
        numrecordpage = 0
        for record in datasets_page:
            if record:
                file.write("{}\n".format(record))
                numrecord += 1
                numrecordpage +=1
        if numrecordpage!=20:
            print(f'[{sortby}-{datetime.now().strftime("%H:%M:%S")}] !:
                page {page} has {numrecordpage} records')
    file.close()
    print(f'[{sortby}-{datetime.now().strftime("%H:%M:%S")}] done #{page}
         pages (#{numrecord} records)')

def download_metadata(dataset,index,base_folder_path):
    folder_path = fr'{base_folder_path}{index}_{dataset.replace("/", "_")
        }'
    command = f' datasets metadata -p \"{folder_path}\" {dataset}'
    datasets_page = !kaggle{command}
    file_path = fr'{folder_path}\dataset-metadata.json'
    print(f'{index} - {dataset}')
    return
```

```
35
36 def load_metadata(dataset,index):
37     file_path = fr'{base_folder_path}{index}_{dataset.replace("/", "_")}\
           dataset-metadata.json'
38     # Opening JSON file
39     try:
40         f = open(file_path)
41         # returns JSON object as a dictionary
42         kaggle_metadata.append(json.load(f))
43         # print(f'{index} - {dataset} - found')
44     except:
45         print(f'{index} - {dataset} - not found')
46     return
47
48 kaggle_datasets_list('votes')
49 kaggle_votes_df = pd.read_csv('kaggle-votes.csv')
50 kaggle_df = kaggle_votes_df.sort_values(by=['downloadCount'],ascending=
       False)
51
52 base_folder_path = 'C:\kaggle-datasets\'
53 i=0
54 for dataset in kaggle_df['ref']:
55     download_metadata(dataset,i,base_folder_path)
56     load_metadata(dataset,i,base_folder_path)
57     i+=1
58     if(i>1000): break
59
60 kaggle_metadata = list()
61 # Creates DataFrame
62 kaggle_metadata_df = pd.DataFrame(kaggle_metadata)
63 kaggle_join_df = kaggle_df.join(kaggle_metadata_df.set_index('id'), on='
       ref', lsuffix='_list', rsuffix='_meta')
64 kaggle_join_df.head(1000).to_csv('kaggle_withmeta.csv')
```

Listing A.3: Kaggle metadata json example

```
{
  "id": "gregorut/videogamesales",
  "id_no": 284,
  "datasetId": 284,
  "datasetSlug": "videogamesales",
  "ownerUser": "gregorut",
  "usabilityRating": 0.5882352941176471,
  "totalViews": 1121341,
  "totalVotes": 4282,
  "totalDownloads": 284182,
  "title": "Video Game Sales",
  "subtitle": "Analyze sales data from more than 16,500 games.",
  "description": "This dataset contains a list of video games with sales
      greater than 100,000 copies. It was generated by a scrape of [
     vgchartz.com][1].\n\nFields include\n\n* Rank - Ranking of overall
     sales\n\n* Name - The games name\n\n* Platform - Platform of the
     games release (i.e. PC,PS4, etc.)\n\n* Year - Year of the game's
     release\n\n* Genre - Genre of the game\n\n* Publisher - Publisher
     of the game\n\n* NA_Sales - Sales in North America (in millions)\n\
     n* EU_Sales - Sales in Europe (in millions)\n\n* JP_Sales - Sales
     in Japan (in millions)\n\n* Other_Sales - Sales in the rest of the
     world (in millions)\n\n* Global_Sales - Total worldwide sales.\n\
     nThe script to scrape the data is available at https://github.com/
     GregorUT/vgchartzScrape.\nIt is based on BeautifulSoup using Python
     .\nThere are 16,598 records. 2 records were dropped due to
     incomplete information.\n\n\n [1]: http://www.vgchartz.com/",
  "isPrivate": false,
  "keywords": [
    "games",
    "video games"
  ],
  "licenses": [
    {
      "name": "unknown"
    }
  ],
  "collaborators": [],
  "data": []
}
```

Listing A.4: OpenML metadata download script

```
from datetime import datetime
def openml_download_list(n):
    driver = webdriver.Chrome()
    driver.get(f'https://www.openml.org/search?type=data&size={n}')
    driver.maximize_window()
    time.sleep(60)

    date = datetime.now().strftime("%Y%m%d")
    filename = f'openml-list-{date}.csv'
    file=open(filename,"w",encoding='utf-8')
    file.write("name,link,abstract,runs,likes,downloads,reach,impact,
        instances,features,classes,missing_values\n")
    file.close()
    file=open(filename,"a",encoding='utf-8',errors='backslashreplace')
    numrecord = 0
    for i in range(1,n+1):
        name = try_find_element(driver,f'#itempage > div:nth-child({i}) >
             div.itemhead > a')
        link = try_find_elementlink(driver,f'#itempage > div:nth-child({i
            }) > div.itemhead > a')
        info = driver.find_element_by_css_selector(f'#itempage > div:nth-
            child({i}) > div.runStats.statLine').text
        abstract = try_find_element(driver,f'#itempage > div:nth-child({i
            }) > div.teaser')

        if len(info.splitlines())==2:
            [line1,line2] = info.splitlines()
            line2 = line2.split(' ')
            instances = line2[0]
            features = line2[3]
            classes = line2[6]
            missing_values = line2[9]
        else:
            print(f'[{datetime.now().strftime("%H:%M:%S")}] !record@{
                numrecord}')
            [line1] = info.splitlines()
            instances = 0
            features = 0
            classes = 0
            missing_values = 0
        line1 = line1.split(' ')
        runs = line1[0]
        likes = int(line1[1][4:])
```

```
38          downloads = int(line1[2][5:])
39          reach = int(line1[3][9:])
40          impact = int(line1[4][5:])
41
42          file.write(f'\"{name}\",\"{link}\",\"{abstract}\",{runs},{likes
                },{downloads},{reach},{impact},{instances},{features},{classes
                },{missing_values}\n')
43          numrecord+=1
44          if numrecord%100==0:
45              print(f'[{datetime.now().strftime("%H:%M:%S")}] record@{
                    numrecord}')
46
47      file.close()
48      print(f'[{datetime.now().strftime("%H:%M:%S")}] done #{numrecord}
            records')
49      return filename
50
51
52  openml_filename_list = openml_download_list(3458)
```

Listing A.5: UCIMLR metadata download script

```
import pandas as pd
from datetime import datetime

def sanitize(string):
    string = string.replace("\"","\'").replace(";",",")
    return string

def css_click(driver,selector):
    driver.find_element_by_css_selector(selector).click()
    return

def try_find_element(driver,selector):
    try:
        value = sanitize(driver.find_element_by_css_selector(selector).
           text)
    except NoSuchElementException:
        value = ''
    return value

def ucimlr_datasets_list_download(n):
    driver = webdriver.Chrome()
    driver.implicitly_wait(10)
    driver.get('https://archive-beta.ics.uci.edu/ml/datasets?&p%5Boffset
       %5D=0&p%5Blimit%5D=596&p%5BorderBy%5D=NumHits&p%5Border%5D=desc')
    time.sleep(10)
    date = datetime.now().strftime("%Y%m%d")
    filename = f'ucimlr-list-{date}.csv'
    file=open(filename,"w",encoding='utf-8')
    file.write("name,dataCharacteristics,subjectArea,task,donated,
       instances,attributes,views,abstract,link\n")
    file.close()
    file=open(filename,"a",encoding='utf-8',errors='backslashreplace')
    numrecord = 0
    css_click(driver,f'#root > div:nth-child(2) > div > div > div > div >
        div.MuiGrid-root.MuiGrid-container.MuiGrid-direction-xs-column >
        div.MuiGrid-root.MuiGrid-container.MuiGrid-spacing-xs-2.MuiGrid-
       align-items-xs-center > div:nth-child(2) > button')
    time.sleep(3)
    for i in range(1,n):
```

```
34          # css_click(driver,f'#root > div:nth-child(2) > div > div > div > div.MuiTableContainer-root > table > tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > span > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-4.MuiGrid-grid-sm-2 > button')
35          name = try_find_element(driver,f'#root > div:nth-child(2) > div > div > div > div.MuiTableContainer-root > table > tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > span > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-8.MuiGrid-grid-sm-10 > p > a')
36          link = try_find_elementlink(driver,f'#root > div:nth-child(2) > div > div > div > div > div.MuiTableContainer-root > table > tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > span > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-8.MuiGrid-grid-sm-10 > p > a')
37          data_characteristics = try_find_element(driver,f'#root > div:nth-child(2) > div > div > div > div > div.MuiTableContainer-root > table > tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > p > div > table > tbody > tr > td:nth-child(1) > p')
38          subject_area = try_find_element(driver,f'#root > div:nth-child(2) > div > div > div > div > div.MuiTableContainer-root > table > tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > p > div > table > tbody > tr > td:nth-child(2) > p')
39          task = try_find_element(driver,f'#root > div:nth-child(2) > div > div > div > div > div.MuiTableContainer-root > table > tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > p > div > table > tbody > tr > td:nth-child(3) > p')
40          # task = task_mapping(task)
41          donated = try_find_element(driver,f'#root > div:nth-child(2) > div > div > div > div > div.MuiTableContainer-root > table > tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > p > div > table > tbody > tr > td:nth-child(4) > p')
42          instances = try_find_element(driver,f'#root > div:nth-child(2) > div > div > div > div > div.MuiTableContainer-root > table > tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > p > div > table > tbody > tr > td:nth-child(5) > p')
```

```
43          attributes = try_find_element(driver,f'#root > div:nth-child(2) >
                 div > div > div > div > div.MuiTableContainer-root > table >
                tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > p >
                 div > table > tbody > tr > td:nth-child(6) > p')
44          views = try_find_element(driver,f'#root > div:nth-child(2) > div
                > div > div > div > div.MuiTableContainer-root > table > tbody
                 > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > p >
                 div > table > tbody > tr > td:nth-child(7) > p')
45          abstract = try_find_element(driver,f'#root > div:nth-child(2) >
                div > div > div > div > div.MuiTableContainer-root > table >
                tbody > tr:nth-child({i}) > div > li > div > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-11 > div > p >
                 div > div > p')[10:]
46          file.write(f'\"{name}\",\"{data_characteristics}\",\"{
                subject_area}\",\"{task}\",{donated},{instances},{attributes
                },{views},\"{abstract}\",\"{link}\"\n')
47          numrecord+=1
48          if numrecord%25==0:
49              print(f'list-record@{numrecord}')
50      print(f'list-done #{numrecord} records')
51      file.close()
52      return filename
53
54  def ucimlr_datasets_metadata_download(filename):
55      list_df = pd.read_csv(filename)
56      driver = webdriver.Chrome()
57      # driver.implicitly_wait(10)
58      date = datetime.now().strftime("%Y%m%d")
59      filename = f'ucimlr-meta-{date}.csv'
60      file=open(filename,"w",encoding='utf-8')
61      file.write(f'name,doi,citations,creators_str,purpose,funding,
            instances,data_splits,contain_sensitive,preprocess,
            already_used_task,additional_info,acknowledgements,missing_values
            ,symbol_missing_values,num_attributes,features_description,
            license\n')
62      file.close()
63      file=open(filename,"a",encoding='utf-8',errors='backslashreplace')
64      numrecord = 0
65      for dataset_link in list_df['link']:
66          # print(dataset_link)
67          driver.get(dataset_link)
68          time.sleep(3)
69          # Quick facts
```

```
70          doi = try_find_element(driver,f'#root > div:nth-child(2) > div >
                div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-xs-flex
                -start.MuiGrid-justify-xs-center > div.MuiGrid-root.MuiGrid-
                item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div.MuiPaper-root.
                MuiCard-root.jss12.MuiPaper-elevation3.MuiPaper-rounded > div.
                MuiCardContent-root.MuiGrid-root.MuiGrid-container.MuiGrid-
                justify-xs-space-between > div.MuiGrid-root.MuiGrid-container.
                MuiGrid-spacing-xs-3 > div:nth-child(5) > p')
71          citations = try_find_element(driver,f'#root > div:nth-child(2) >
                div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-
                xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div.
                MuiPaper-root.MuiCard-root.jss16.MuiPaper-elevation3.MuiPaper-
                rounded > div.MuiCardContent-root.MuiGrid-root.MuiGrid-
                container.MuiGrid-align-items-xs-center.MuiGrid-justify-xs-
                space-between > div.MuiGrid-root.MuiGrid-container.MuiGrid-
                spacing-xs-2.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-sm-6
                 > div:nth-child(2) > p')
72          cit_num = citations.split(" ")[0]
73          # Creators
74          name = try_find_element(driver,f'#root > div:nth-child(2) > div >
                 div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-xs-
                flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div.
                MuiPaper-root.MuiCard-root.jss16.MuiPaper-elevation3.MuiPaper-
                rounded > div.MuiCardHeader-root > div.MuiCardHeader-content >
                 h5')
75          creators = try_find_list(driver,f'#root > div:nth-child(2) > div
                > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-xs-
                flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:nth-
                child(3) > div.MuiCollapse-container.MuiCollapse-entered > div
                 > div > div > div > ul')
76          separator = ';'
77          creators_str = separator.join(creators)
78          # Descriptive questions
79          purpose = try_find_element(driver,f'#root > div:nth-child(2) >
                div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-
                xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:nth-
                child(4) > div.MuiCollapse-container.MuiCollapse-entered > div
                 > div > div > div > table > tbody > tr:nth-child(1) > td:nth-
                child(2) > p')
```

```
80          funding = try_find_element(driver,f'#root > div:nth-child(2) >
                div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-
                xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:nth-
                child(4) > div.MuiCollapse-container.MuiCollapse-entered > div
                 > div > div > div > table > tbody > tr:nth-child(2) > td:nth-
                child(2) > p')
81          instances = try_find_element(driver,f'#root > div:nth-child(2) >
                div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-
                xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:nth-
                child(4) > div.MuiCollapse-container.MuiCollapse-entered > div
                 > div > div > div > table > tbody > tr:nth-child(3) > td:nth-
                child(2) > p')
82          data_splits = try_find_element(driver,f'#root > div:nth-child(2)
                > div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items
                -xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:nth-
                child(4) > div.MuiCollapse-container.MuiCollapse-entered > div
                 > div > div > div > table > tbody > tr:nth-child(4) > td:nth-
                child(2) > p')
83          contain_sensitive = try_find_element(driver,f'#root > div:nth-
                child(2) > div > div.MuiGrid-root.MuiGrid-container.MuiGrid-
                align-items-xs-flex-start.MuiGrid-justify-xs-center > div.
                MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9
                 > div:nth-child(4) > div.MuiCollapse-container.MuiCollapse-
                entered > div > div > div > div > table > tbody > tr:nth-child
                (5) > td:nth-child(2) > p')
84          preprocess = try_find_element(driver,f'#root > div:nth-child(2) >
                 div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-
                xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:nth-
                child(4) > div.MuiCollapse-container.MuiCollapse-entered > div
                 > div > div > div > table > tbody > tr:nth-child(6) > td:nth-
                child(2) > p')
85          already_used_task = try_find_element(driver,f'#root > div:nth-
                child(2) > div > div.MuiGrid-root.MuiGrid-container.MuiGrid-
                align-items-xs-flex-start.MuiGrid-justify-xs-center > div.
                MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9
                 > div:nth-child(4) > div.MuiCollapse-container.MuiCollapse-
                entered > div > div > div > div > table > tbody > tr:nth-child
                (7) > td:nth-child(2) > p')
```

```
86          additional_info = try_find_element(driver,f'#root > div:nth-child
                (2) > div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-
                items-xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-
                root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:
                nth-child(4) > div.MuiCollapse-container.MuiCollapse-entered >
                 div > div > div > div > table > tbody > tr:nth-child(8) > td:
                nth-child(2) > p')
87          acknowledgements = try_find_element(driver,f'#root > div:nth-
                child(2) > div > div.MuiGrid-root.MuiGrid-container.MuiGrid-
                align-items-xs-flex-start.MuiGrid-justify-xs-center > div.
                MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9
                 > div:nth-child(4) > div.MuiCollapse-container.MuiCollapse-
                entered > div > div > div > div > table > tbody > tr:nth-child
                (9) > td:nth-child(2) > p')
88          # Tabular data properties
89          missing_values = try_find_element(driver,f'#root > div:nth-child
                (2) > div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-
                items-xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-
                root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:
                nth-child(5) > div.MuiCollapse-container.MuiCollapse-entered >
                 div > div > div > div > table > tbody > tr:nth-child(1) > td:
                nth-child(2) > p')
90          symbol_missing_values = try_find_element(driver,f'#root > div:nth
                -child(2) > div > div.MuiGrid-root.MuiGrid-container.MuiGrid-
                align-items-xs-flex-start.MuiGrid-justify-xs-center > div.
                MuiGrid-root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9
                 > div:nth-child(5) > div.MuiCollapse-container.MuiCollapse-
                entered > div > div > div > div > table > tbody > tr:nth-child
                (2) > td:nth-child(2) > p')
91          num_attributes = try_find_element(driver,f'#root > div:nth-child
                (2) > div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-
                items-xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-
                root.MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:
                nth-child(5) > div.MuiCollapse-container.MuiCollapse-entered >
                 div > div > div > div > table > tbody > tr:nth-child(3) > td:
                nth-child(2) > p')
92          # Features
93          feat_str = try_find_element(driver,f'#root > div:nth-child(2) >
                div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-
                xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-9 > div:nth-
                child(6) > div.MuiButtonBase-root.MuiAccordionSummary-root.
                jss24.jss13.Mui-expanded.jss26 > div.MuiAccordionSummary-
                content.jss25.Mui-expanded.jss26 > h5')
94          if feat_str== 'Features':
```

```
95              features_description = 1
96          else:
97              features_description = 0
98          # License
99          license = try_find_element(driver,f'#root > div:nth-child(2) >
                div > div.MuiGrid-root.MuiGrid-container.MuiGrid-align-items-
                xs-flex-start.MuiGrid-justify-xs-center > div.MuiGrid-root.
                MuiGrid-item.MuiGrid-grid-xs-12.MuiGrid-grid-md-3 > div:nth-
                child(2) > div.MuiCardContent-root > p:nth-child(1)')
100         file.write(f'\"{name}\",\"{doi}\",{cit_num},\"{creators_str
                }\",\"{purpose}\",\"{funding}\",\"{instances}\",\"{data_splits
                }\",\"{contain_sensitive}\",\"{preprocess}\",\"{
                already_used_task}\",\"{additional_info}\",\"{acknowledgements
                }\",\"{missing_values}\",\"{symbol_missing_values}\",{
                num_attributes},{features_description},\"{license}\"\n')
101         numrecord+=1
102         if numrecord%25==0:
103             print(f'meta-record@{numrecord}')
104     print(f'meta-done #{numrecord} records')
105     file.close()
106     return filename
107
108 filename_list_df = ucimlr_datasets_list_download(597)
109 filename_meta_df = ucimlr_datasets_metadata_download(filename_list_df)
110 filename_join_df = f'ucimlr-join-{filename_list_df.split("-")[2].split
        (".")[0]}.csv'
111 ucimlr_list_df = pd.read_csv(filename_list_df)
112 ucimlr_meta_df = pd.read_csv(filename_meta_df)
113 ucimlr_join_df = ucimlr_list_df.join(ucimlr_meta_df.set_index('name'),
        on='name', lsuffix='_list', rsuffix='_meta')
114 ucimlr_join_df.to_csv(filename_join_df)
```

Appendix B

Selected datasets

Table B.1: Huggingface selected datasets (03/02/2022)

ID	Name	Download	Duplicate of	URL
hug01	glue	7 197 060		https://huggingface.co/datasets/glue
hug02	super_glue	4 907 890		https://huggingface.co/datasets/super_glue
hug03	anli	1 711 000		https://huggingface.co/datasets/anli
hug04	wikitext	1 147 610		https://huggingface.co/datasets/wikitext
hug05	wino_bias	1 024 850		https://huggingface.co/datasets/wino_bias

Continue on next page

Table B.1 – continued from previous page

ID	Name	Download	Duplicate of	URL
hug06	squad	984 460		https://huggingface.co/datasets/squad
hug07	imdb	936 460		https://huggingface.co/datasets/imdb
hug08	trec	719 060		https://huggingface.co/datasets/trec
hug09	adversarial_qa	700 840		https://huggingface.co/datasets/adversarial_qa
hug10	race	684 560		https://huggingface.co/datasets/race
hug11	duorc	671 790		https://huggingface.co/datasets/duorc
hug12	squad_v2	667 500		https://huggingface.co/datasets/squad_v2
hug13	winogrande	582 130		https://huggingface.co/datasets/winogrande
hug14	hellaswag	543 720		https://huggingface.co/datasets/hellaswag
hug15	common_voice	541 790		https://huggingface.co/datasets/common_voice
hug16	cnn_dailymail	532 070		https://huggingface.co/datasets/cnn_dailymail
hug17	piqa	531 550		https://huggingface.co/datasets/piqa
hug18	xsum	503 930		https://huggingface.co/datasets/xsum

Continue on next page

Table B.1 – continued from previous page

ID	Name	Download	Duplicate of	URL
hug19	cosmos_qa	501 510		https://huggingface.co/datasets/cosmos_qa
hug20	mlqa	497 400		https://huggingface.co/datasets/mlqa
hug21	quail	494 130		https://huggingface.co/datasets/quail
hug22	paws	489 980		https://huggingface.co/datasets/paws
hug23	wmt16	486 940		https://huggingface.co/datasets/wmt16
hug24	ai2_arc	474 240		https://huggingface.co/datasets/ai2_arc
hug25	rotten_tomatoes	461 310		https://huggingface.co/datasets/rotten_tomatoes

Table B.2: Kaggle selected datasets (26/01/2022)

ID	Name	Download	Duplicate of	URL
kag01	Credit Card Fraud Detection	360 828		https://www.kaggle.com/mlg-ulb/creditcardfraud
kag02	Novel Corona Virus 2019 Dataset	347 779		https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
kag03	Video Game Sales	264 773		https://www.kaggle.com/gregorut/videogamesales
kag04	Heart Disease UCI	256 102	uci04	https://www.kaggle.com/ronitf/heart-disease-uci
kag05	Pima Indians Diabetes Database	235 317	oml04	https://www.kaggle.com/uciml/pima-indians-diabetes-database
kag06	Iris Species	228 045	uci01	https://www.kaggle.com/uciml/iris
kag07	World Happiness Report	202 882		https://www.kaggle.com/unsdsn/world-happiness
kag08	Netflix Movies and TV Shows	183 020		https://www.kaggle.com/shivamb/netflix-shows
kag09	The Movies Dataset	178 101		https://www.kaggle.com/rounakbanik/the-movies-dataset
kag10	Breast Cancer Wisconsin (Diagnostic) Data Set	177 162	uci07	https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
kag11	TMDB 5000 Movie Dataset	174 636		https://www.kaggle.com/tmdb/tmdb-movie-metadata
kag12	COVID-19 Dataset	168 132	kag02	https://www.kaggle.com/imdevskp/corona-virus-report

Continue on next page

Table B.2 – continued from previous page

ID	Name	Download	Duplicate of	URL
kag13	Google Play Store Apps	166 169		https://www.kaggle.com/lava18/google-play-store-apps
kag14	Trending YouTube Video Statistics	158 082		https://www.kaggle.com/datasnaek/youtube-new
kag15	Wine Reviews	148 561		https://www.kaggle.com/zynicide/wine-reviews
kag16	Chest X-Ray Images (Pneumonia)	143 227		https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
kag17	European Soccer Database	140 647		https://www.kaggle.com/hugomathien/soccer
kag18	COVID-19 in India	137 213		https://www.kaggle.com/sudalairajkumar/covid19-in-india
kag19	COVID-19 Open Research Dataset Challenge (CORD-19)	134 256		https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
kag20	Students Performance in Exams	134 210		https://www.kaggle.com/spscientist/students-performance-in-exams
kag21	FIFA 19 complete player dataset	130 521		https://www.kaggle.com/karangadiya/fifa19
kag22	Avocado Prices	126 093		https://www.kaggle.com/neuromusic/avocado-prices
kag23	House Sales in King County, USA	111 243		https://www.kaggle.com/harlfoxem/housesalesprediction

Continue on next page

Table B.2 – continued from previous page

ID	Name	Download	Duplicate of	URL
kag24	Suicide Rates Overview 1985 to 2016	111 006		https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
kag25	New York City Airbnb Open Data	110 090		https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data
kag26	Red Wine Quality	108 089		https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
kag27	Amazon Fine Food Reviews	108 069		https://www.kaggle.com/snap/amazon-fine-food-reviews
kag28	Fashion MNIST	102 001		https://www.kaggle.com/zalando-research/fashionmnist
kag29	Telco Customer Churn	101 166		https://www.kaggle.com/blastchar/telco-customer-churn
kag30	Bitcoin Historical Data	98 823		https://www.kaggle.com/mczielinski/bitcoin-historical-data

Table B.3: UC Irvine Machine Learning Repository selected datasets (27/01/2022)

ID	Name	Download	Duplicate of	URL
uci01	Iris	122 721		https://archive-beta.ics.uci.edu/ml/datasets/iris
uci02	Diabetes	85 583		https://archive-beta.ics.uci.edu/ml/datasets/diabetes
uci03	Adult	81 206		https://archive-beta.ics.uci.edu/ml/datasets/adult
uci04	Heart Disease	77 318		https://archive-beta.ics.uci.edu/ml/datasets/heart+disease
uci05	Wine	63 145		https://archive-beta.ics.uci.edu/ml/datasets/wine
uci06	Car Evaluation	60 395		https://archive-beta.ics.uci.edu/ml/datasets/car+evaluation
uci07	Breast Cancer Wisconsin (Diagnostic)	54 593		https://archive-beta.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+diagnostic
uci08	Abalone	45 356		https://archive-beta.ics.uci.edu/ml/datasets/abalone
uci09	Breast Cancer	44 779		https://archive-beta.ics.uci.edu/ml/datasets/breast+cancer
uci10	Mushroom	44 738		https://archive-beta.ics.uci.edu/ml/datasets/mushroom
uci11	Glass Identification	40 148		https://archive-beta.ics.uci.edu/ml/datasets/glass+identification

Continue on next page

Table B.3 – continued from previous page

ID	Name	Download	Duplicate of	URL
uci12	Census Income	34 569	uci03	https://archive-beta.ics.uci.edu/ml/datasets/census+income
uci13	Breast Cancer Wisconsin (Original)	33 993		https://archive-beta.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+original
uci14	Statlog (German Credit Data)	33 688	oml01	https://archive-beta.ics.uci.edu/ml/datasets/statlog+german+credit+data
uci15	Thyroid Disease	28 521		https://archive-beta.ics.uci.edu/ml/datasets/thyroid+disease
uci16	Liver Disorders	28 141		https://archive-beta.ics.uci.edu/ml/datasets/liver+disorders
uci17	Optical Recognition of Handwritten Digits	27 391		https://archive-beta.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits
uci18	Ionosphere	26 767		https://archive-beta.ics.uci.edu/ml/datasets/ionosphere
uci19	Auto MPG	26 543		https://archive-beta.ics.uci.edu/ml/datasets/auto+mpg
uci20	Pen-Based Recognition of Handwritten Digits	26 233		https://archive-beta.ics.uci.edu/ml/datasets/pen+based+recognition+of+handwritten+digits

Continue on next page

Table B.3 – continued from previous page

ID	Name	Download	Duplicate of	URL
uci21	Image Segmentation	24 985		https://archive-beta.ics.uci.edu/ml/datasets/image+segmentation
uci22	Congressional Voting Records	24 684		https://archive-beta.ics.uci.edu/ml/datasets/congressional+voting+records
uci23	Zoo	24 438		https://archive-beta.ics.uci.edu/ml/datasets/zoo
uci24	Letter Recognition	23 194	oml10	https://archive-beta.ics.uci.edu/ml/datasets/letter+recognition
uci25	Lung Cancer	23 075		https://archive-beta.ics.uci.edu/ml/datasets/lung+cancer
uci26	Spambase	21 980	oml08	https://archive-beta.ics.uci.edu/ml/datasets/spambase
uci27	Yeast	21 867		https://archive-beta.ics.uci.edu/ml/datasets/yeast
uci28	Hepatitis	21 415		https://archive-beta.ics.uci.edu/ml/datasets/hepatitis
uci29	Internet Advertisements	20 802		https://archive-beta.ics.uci.edu/ml/datasets/internet+advertisements

Table B.4: 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

Continue on next page

Table B.4 – continued from previous page

ID	Name	Download	Duplicate of	URL
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

Bibliography

[1] *A Responsible AI Strategy for Climate Action*. URL: https://oecd.ai/en/wonk/a-responsible-ai-strategy-for-climate-action (cit. on p. 14).

[2] «AI in the UK: Ready, Willing and Able». In: (), p. 183 (cit. on p. 15).

[3] *Anatomy of an AI System*. Anatomy of an AI System. URL: http://www.anatomyof.ai (cit. on p. 14).

[4] Martin Anderson. *A Cartel of Influential Datasets Is Dominating Machine Learning Research, New Study Suggests*. Unite.AI. Dec. 6, 2021. URL: https://www.unite.ai/a-cartel-of-influential-datasets-are-dominating-machine-learning-research-new-study-suggests/ (cit. on p. 16).

[5] Matthew Arnold et al. «FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity». Feb. 7, 2019. arXiv: 1808.07261 [cs]. URL: http://arxiv.org/abs/1808.07261 (cit. on p. 16).

[6] Jack Bandy and Nicholas Vincent. «Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus». May 11, 2021. DOI: 10.48550/arXiv.2105.05241. arXiv: 2105.05241 [cs]. URL: http://arxiv.org/abs/2105.05241 (cit. on pp. 16, 93).

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL: https://fairmlbook.org/datasets.html (cit. on p. 16).

[8] Emily M. Bender and Batya Friedman. «Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science». In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 587–604. DOI: 10.1162/tacl_a_00041. URL: https://aclanthology.org/Q18-1041 (cit. on pp. 3, 19, 21, 22, 24, 25, 33, 91).

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?» In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association

for Computing Machinery, Mar. 3, 2021, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922 (cit. on p. 16).

[10] Elena Beretta, Antonio Santangelo, Bruno Lepri, Antonio Vetrò, and Juan Carlos De Martin. *The Invisible Power of Fairness. How Machine Learning Shapes Democracy*. Mar. 22, 2019. DOI: 10.48550/arXiv.1903.09493. arXiv: 1903.09493 [cs, stat]. URL: http://arxiv.org/abs/1903.09493 (cit. on p. 15).

[11] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. «The Values Encoded in Machine Learning Research». June 29, 2021. arXiv: 2106.15590 [cs]. URL: http://arxiv.org/abs/2106.15590 (cit. on pp. 3, 11).

[12] Karen L. Boyd. «Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data». In: *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2 Oct. 18, 2021), 438:1–438:27. DOI: 10.1145/3479582. URL: https://doi.org/10.1145/3479582 (cit. on pp. 3, 11, 16).

[13] Antonio A. Casilli. *Schiavi Del Clic*. Feltrinelli Editore (cit. on pp. 13, 21).

[14] Frances Corry, Hamsini Sridharan, Alexandra Sasha Luccioni, Mike Ananny, Jason Schultz, and Kate Crawford. «The Problem of Zombie Datasets:A Framework For Deprecating Datasets». Oct. 18, 2021. arXiv: 2111.04424 [cs]. URL: http://arxiv.org/abs/2111.04424 (cit. on pp. 23, 77).

[15] Kate Crawford. *The Atlas of AI*. Yale University Press (cit. on pp. 13, 14, 16).

[16] Kate Crawford. «Time to Regulate AI That Interprets Human Emotions». In: *Nature* 592.7853 (7853 Apr. 6, 2021), pp. 167–167. DOI: 10.1038/d41586-021-00868-5. URL: https://www.nature.com/articles/d41586-021-00868-5 (cit. on p. 15).

[17] Maria J. Cruz, Shalini Kurapati, and Yasemin Turkyilmaz-van der Velden. «The Role of Data Stewardship in Software Sustainability and Reproducibility». In: *2018 IEEE 14th International Conference on E-Science (e-Science)*. 2018 IEEE 14th International Conference on E-Science (e-Science). Oct. 2018, pp. 1–8. DOI: 10.1109/eScience.2018.00009 (cit. on p. 23).

[18] Payal Dhar. «The Carbon Impact of Artificial Intelligence». In: *Nature Machine Intelligence* 2.8 (8 Aug. 1, 2020), pp. 423–425. ISSN: 2522-5839. DOI: 10.1038/s42256-020-0219-9. URL: https://www.nature.com/articles/s42256-020-0219-9 (cit. on p. 14).

[19] *Digital Public Goods*. URL: https://ethics.harvard.edu/digital-public-goods (cit. on p. 14).

[20] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml (cit. on p. 23).

[21] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. *Algorithmic Fairness Datasets: The Story so Far*. May 6, 2022. arXiv: 2202.01711 [cs]. URL: http://arxiv.org/abs/2202.01711 (cit. on pp. 16, 32, 93).

[22] Alessandro Fabris et al. *Algorithmic Audit of Italian Car Insurance: Evidence of Unfairness in Access and Pricing*. May 21, 2021. arXiv: 2105.10174 [cs]. URL: http://arxiv.org/abs/2105.10174 (cit. on p. 14).

[23] Matthias Feurer et al. «OpenML-Python: An Extensible Python API for OpenML». June 23, 2021. DOI: 10.48550/arXiv.1911.02490. arXiv: 1911.02490 [cs, stat]. URL: http://arxiv.org/abs/1911.02490 (cit. on p. 45).

[24] Luciano Floridi and Federico Cabitza. *Intelligenza Artificiale*. Bompiani. 192 pp. ISBN: 978-88-301-0938-4 (cit. on p. 12).

[25] Christian Garbin, Pranav Rajpurkar, Jeremy Irvin, Matthew P. Lungren, and Oge Marques. «Structured Dataset Documentation: A Datasheet for CheXpert». May 6, 2021. DOI: 10.48550/arXiv.2105.03020. arXiv: 2105.03020 [cs, eess]. URL: http://arxiv.org/abs/2105.03020 (cit. on p. 93).

[26] Timnit Gebru et al. «Datasheets for Datasets». Dec. 1, 2021. arXiv: 1803.09010 [cs]. URL: http://arxiv.org/abs/1803.09010 (cit. on pp. 3, 19–21, 23–26, 62, 91).

[27] R. Stuart Geiger et al. «"Garbage In, Garbage Out" Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data?» In: *Quantitative Science Studies* 2.3 (Nov. 5, 2021), pp. 795–827. ISSN: 2641-3337. DOI: 10.1162/qss_a_00144. arXiv: 2107.02278. URL: http://arxiv.org/abs/2107.02278 (cit. on pp. 3, 11).

[28] R. Stuart Geiger et al. «Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?» In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Jan. 27, 2020, pp. 325–336. DOI: 10.1145/3351095.3372862. arXiv: 1912.08320 [cs]. URL: http://arxiv.org/abs/1912.08320 (cit. on p. 32).

[29] Anne Gerdes. «A Participatory Data-Centric Approach to AI Ethics by Design». In: *Applied Artificial Intelligence* 0.0 (Dec. 8, 2021), pp. 1–19. ISSN: 0883-9514. DOI: 10.1080/08839514.2021.2009222. URL: https://doi.org/10.1080/08839514.2021.2009222 (cit. on p. 15).

[30] Jonathan Gray. «"Let Us Calculate!": Leibniz, Llull, and the Computational Imagination». In: *Public Domain Review* (Nov. 10, 2016). URL: http://publicdomainreview.org/2016/11/10/let-us-calculate-leibniz-llull-and-computational-imagination/ (cit. on p. 12).

[31] Alon Halevy, Peter Norvig, and Fernando Pereira. «The Unreasonable Effectiveness of Data». In: *IEEE Intelligent Systems* 24.2 (Mar. 2009), pp. 8–12. ISSN: 1941-1294. DOI: 10.1109/MIS.2009.36 (cit. on p. 15).

[32] Margot Hanley, Apoorv Khandelwal, Hadar Averbuch-Elor, Noah Snavely, and Helen Nissenbaum. «An Ethical Highlighter for People-Centric Dataset Creation». Nov. 27, 2020. arXiv: 2011.13583 [cs]. URL: http://arxiv.org/abs/2011.13583 (cit. on p. 15).

[33] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. «The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards». May 9, 2018. arXiv: 1805.03677 [cs]. URL: http://arxiv.org/abs/1805.03677 (cit. on pp. 4, 19–21, 25, 91).

[34] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. «Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?» In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, May 2, 2019, pp. 1–16. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300830. URL: https://doi.org/10.1145/3290605.3300830 (cit. on p. 16).

[35] *Hugging Face – The AI Community Building the Future.* URL: https://huggingface.co/datasets (cit. on p. 33).

[36] Ben Hutchinson et al. «Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure». In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 3, 2021, pp. 560–575. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445918. URL: https://doi.org/10.1145/3442188.3445918 (cit. on p. 16).

[37] AI Now Institute. *Algorithmic Impact Assessments: Toward Accountable Automation in Public Agencies*. Medium. Apr. 9, 2018. URL: https://medium.com/@AINowInstitute/algorithmic-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde (cit. on p. 14).

[38] *Intelligenza Artificiale: l'Italia lancia la strategia nazionale*. Ministro per l'innovazione tecnologica e la transizione digitale. URL: https://innovazione.gov.it/notizie/articoli/intelligenza-artificiale-l-italia-lancia-la-strategia-nazionale/ (cit. on p. 15).

[39] Fieke Jansen. *The EU Is Ignoring AI's Effect on the Climate Crisis*. www.euractiv.com. Feb. 17, 2020. URL: https://www.euractiv.com/section/digital/opinion/the-eu-is-ignoring-ais-effect-on-the-climate-crisis/ (cit. on p. 14).

[40] Eun Seo Jo and Timnit Gebru. «Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning». In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Jan. 27, 2020), pp. 306–316. DOI: 10.1145/3351095.3372829. arXiv: 1912.10389. URL: http://arxiv.org/abs/1912.10389 (cit. on p. 16).

[41] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. «A "Nutrition Label" for Privacy». In: *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*. The 5th Symposium. Mountain View, California: ACM Press, 2009, p. 1. ISBN: 978-1-60558-736-3. DOI: 10.1145/1572532.1572538. URL: http://portal.acm.org/citation.cfm?doid=1572532.1572538 (cit. on p. 22).

[42] Florian Königstorfer and Stefan Thalmann. «Software Documentation Is Not Enough! Requirements for the Documentation of AI». In: *Digital Policy, Regulation and Governance* 23.5 (Jan. 1, 2021), pp. 475–488. ISSN: 2398-5038. DOI: 10.1108/DPRG-03-2021-0047. URL: https://doi.org/10.1108/DPRG-03-2021-0047 (cit. on p. 16).

[43] *La Rete è di tutti / Ma l'intelligenza artificiale fa bene al pianeta?* Valigia Blu. URL: https://www.valigiablu.it/intelligenza-artificiale-emergenza-climatica/ (cit. on p. 14).

[44] *La Rete è di tutti. Prima dell'innovazione vengono le persone. Prima dell'efficienza, i diritti. Prima della tecnica, la democrazia*. Valigia Blu. URL: https://www.valigiablu.it/rete-futuro-tecnologie/ (cit. on p. 16).

[45] Daniel T. Larose and Chantal D. Larose. «Association Rules». In: *Discovering Knowledge in Data*. John Wiley & Sons, Ltd, 2014, pp. 247–265. ISBN: 978-1-118-87405-9. DOI: 10.1002/9781118874059.ch12. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118874059.ch12 (cit. on p. 84).

[46] Quentin Lhoest et al. «Datasets: A Community Library for Natural Language Processing». Sept. 6, 2021. DOI: 10.48550/arXiv.2109.02846. arXiv: 2109.02846 [cs]. URL: http://arxiv.org/abs/2109.02846 (cit. on p. 33).

[47] Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig. «The Relationship between Motivation, Monetary Compensation, and Data Quality among US- and India-based Workers on Mechanical Turk». In: *Behavior Research Methods* 47.2 (June 1, 2015), pp. 519–528. ISSN: 1554-3528. DOI: 10.3758/s13428-014-0483-x. URL: https://doi.org/10.3758/s13428-014-0483-x (cit. on p. 13).

[48] Molly McCue and Kat Holmes. «Myth and the Making of AI». In: *Journal of Design and Science* (July 16, 2018). DOI: 10.21428/d3a0f14d. URL: https://jods.mitpress.mit.edu/pub/holmes-mccue/release/2 (cit. on p. 12).

[49] Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. «Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards». In: *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. ACL-GEM-IJCNLP 2021. Online: Association for Computational Linguistics, Aug. 2021, pp. 121–135. DOI: 10.18653/v1/2021.gem-1.11. URL: https://aclanthology.org/2021.gem-1.11.

[50] Mariachiara Mecati, Antonio Vetrò, and Marco Torchiano. «Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data». In: *2021 IEEE International Conference on Big Data (Big Data)*. 2021 IEEE International Conference on Big Data (Big Data). Dec. 2021, pp. 4287–4296. DOI: 10.1109/BigData52589.2021.9671443 (cit. on p. 14).

[51] Michelle N. Meyer. «Practical Tips for Ethical Data Sharing». In: *Advances in Methods and Practices in Psychological Science* 1.1 (Mar. 1, 2018), pp. 131–144. ISSN: 2515-2459. DOI: 10.1177/2515245917747656. URL: https://doi.org/10.1177/2515245917747656.

[52] Milagros Miceli, Julian Posada, and Tianling Yang. «Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?» Sept. 16, 2021. arXiv: 2109.08131 [cs]. URL: http://arxiv.org/abs/2109.08131 (cit. on p. 15).

[53] Milagros Miceli, Martin Schuessler, and Tianling Yang. «Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision». In: *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW2 Oct. 14, 2020), 115:1–115:25. DOI: 10.1145/3415186. URL: https://doi.org/10.1145/3415186 (cit. on p. 16).

[54] Margaret Mitchell et al. «Model Cards for Model Reporting». In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Jan. 29, 2019), pp. 220–229. DOI: 10.1145/3287560.3287596. arXiv: 1810.03993. URL: http://arxiv.org/abs/1810.03993 (cit. on p. 16).

[55] Evgeny Morozov. *Silicon Valley: i Signori Del Silicio*. Codice Edizioni. ISBN: 978-88-7578-688-5 (cit. on p. 14).

[56] Roberto Musa Giuliano. «Echoes of Myth and Magic in the Language of Artificial Intelligence». In: *AI & SOCIETY* 35.4 (Dec. 1, 2020), pp. 1009–1024. ISSN: 1435-5655. DOI: 10.1007/s00146-020-00966-4. URL: https://doi.org/10.1007/s00146-020-00966-4 (cit. on p. 12).

[57] Simone Natale and Andrea Ballatore. «Imagining the Thinking Machine: Technological Myths and the Rise of Artificial Intelligence». In: *Convergence* 26.1 (Feb. 1, 2020), pp. 3–18. ISSN: 1354-8565. DOI: 10.1177/1354856517715164. URL: https://doi.org/10.1177/1354856517715164 (cit. on p. 12).

[58] Cathy O'Neil, director. *The Era of Blind Faith in Big Data Must End*. 2017. URL: https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end/transcript (cit. on p. 14).

[59] Marinus Ossewaarde and Erdener Gulenc. «National Varieties of Artificial Intelligence Discourses: Myth, Utopianism, and Solutionism in West European Policy Expectations». In: *Computer* 53.11 (Nov. 2020), pp. 53–61. ISSN: 1558-0814. DOI: 10.1109/MC.2020.2992290 (cit. on p. 12).

[60] Samir Passi and Solon Barocas. «Problem Formulation and Fairness». In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Jan. 29, 2019), pp. 39–48. DOI: 10.1145/3287560.3287567. arXiv: 1901.02547. URL: http://arxiv.org/abs/1901.02547 (cit. on p. 16).

[61] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. «Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers». Nov. 21, 2021. arXiv: 2108.02922 [cs]. URL: http://arxiv.org/abs/2108.02922 (cit. on pp. 3, 11, 23, 32).

[62] Dana Pessach and Erez Shmueli. «Algorithmic Fairness». Jan. 21, 2020. arXiv: 2001.09784 [cs, stat]. URL: http://arxiv.org/abs/2001.09784.

[63] Vinay Uday Prabhu and Abeba Birhane. «Large Image Datasets: A Pyrrhic Win for Computer Vision?» July 23, 2020. arXiv: 2006.16923 [cs, stat]. URL: http://arxiv.org/abs/2006.16923.

[64] *Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. 2021. URL: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:52021PC0206 (cit. on pp. 12, 15).

[65] *Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati) (Testo rilevante ai fini del SEE)*. Apr. 27, 2016. URL: http://data.europa.eu/eli/reg/2016/679/oj/ita (cit. on p. 21).

[66] John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. «A Methodology for Creating AI FactSheets». June 27, 2020. arXiv: 2006.13796 [cs]. URL: http://arxiv.org/abs/2006.13796 (cit. on p. 16).

[67] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. «"Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI». In: 2021 (cit. on pp. 3, 11, 16, 19, 77).

[68] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. «Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development». In: *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2 Oct. 13, 2021), pp. 1–37. ISSN: 2573-0142. DOI: 10.1145/3476058. arXiv: 2108.04308. URL: http://arxiv.org/abs/2108.04308 (cit. on pp. 14–16, 32).

[69] Terrence J. Sejnowski. *The Deep Learning Revolution*. Cambridge, MA, USA: MIT Press, Oct. 23, 2018. 352 pp. ISBN: 978-0-262-03803-4.

[70] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. «Fairness and Abstraction in Sociotechnical Systems». In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. New York, NY, USA: Association for Computing Machinery, Jan. 29, 2019, pp. 59–68. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287598. URL: https://doi.org/10.1145/3287560.3287598 (cit. on p. 15).

[71] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. «Responsible Research with Crowds: Pay Crowdworkers at Least Minimum Wage». In: *Communications of the ACM* 61.3 (Feb. 21, 2018), pp. 39–41. ISSN: 0001-0782. DOI: 10.1145/3180492. URL: https://doi.org/10.1145/3180492 (cit. on p. 13).

[72] Chenkai Sun, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. «MithraLabel: Flexible Dataset Nutritional Labels for Responsible Data Science». In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. New York, NY, USA: Association for Computing Machinery, Nov. 3, 2019, pp. 2893–2896. ISBN: 978-1-4503-6976-3. DOI: 10.1145/3357384.3357853. URL: https://doi.org/10.1145/3357384.3357853 (cit. on p. 20).

[73] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Nueva Delhi (India: Pearson, 2016. ISBN: 978-93-325-7140-2 (cit. on p. 85).

[74] *The Nooscope Manifested: AI as Instrument of Knowledge Extractivism*. The Nooscope Manifested: AI as Instrument of Knowledge Extractivism. URL: http://nooscope.ai/.

[75] Daniel R. Thomas, Sergio Pastrana, Alice Hutchings, Richard Clayton, and Alastair R. Beresford. «Ethical Issues in Research Using Datasets of Illicit Origin». In: *Proceedings of the 2017 Internet Measurement Conference*. IMC '17. New York, NY, USA: Association for Computing Machinery, Nov. 1, 2017, pp. 445–462. ISBN: 978-1-4503-5118-8. DOI: 10.1145/3131365.3131389. URL: https://doi.org/10.1145/3131365.3131389 (cit. on p. 22).

[76] Nanna Bonde Thylstrup. «The Ethics and Politics of Data Sets in the Age of Machine Learning: Deleting Traces and Encountering Remains». In: *Media, Culture & Society* (Apr. 28, 2022), p. 01634437211060226. ISSN: 0163-4437. DOI: 10.1177/01634437211060226. URL: https://doi.org/10.1177/01634437211060226 (cit. on p. 15).

[77] *Timnit Gebru Is Building a Slow AI Movement*. IEEE Spectrum. Mar. 31, 2022. URL: https://spectrum.ieee.org/timnit-gebru-dair-ai-ethics (cit. on p. 16).

[78] *UK National AI Strategy*. GOV.UK. URL: https://www.gov.uk/government/publications/national-ai-strategy (cit. on p. 15).

[79] Shannon Vallor, Dianne Mckenna Professor, and Brian Green. «Ethics in Technology Practice: An Overwiew». In: (), p. 8. URL: https://www.scu.edu/ethics-in-technology-practice/overview-of-ethics-in-tech-practice/ (cit. on p. 15).

[80] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. «OpenML: Networked Science in Machine Learning». In: *ACM SIGKDD Explorations Newsletter* 15.2 (June 16, 2014), pp. 49–60. ISSN: 1931-0145. DOI: 10.1145/2641190.2641198. URL: https://doi.org/10.1145/2641190.2641198 (cit. on p. 37).

[81] Antonio Vetrò. «Imbalanced Data as Risk Factor of Discriminating Automated Decisions: A Measurement-Based Approach». In: *JIPITEC* 12.4 (Dec. 12, 2021). ISSN: 2190-3387. URL: http://www.jipitec.eu/issues/jipitec-12-4-2021/5452 (cit. on p. 14).

[82] Antonio Vetrò, Marco Torchiano, and Mariachiara Mecati. «A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision Making Systems». In: *Government Information Quarterly* 38.4 (Oct. 1, 2021), p. 101619. ISSN: 0740-624X. DOI: 10.1016/j.giq.2021.101619. URL: https://www.sciencedirect.com/science/article/pii/S0740624X21000551 (cit. on p. 14).

[83] Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. *Disembodied Machine Learning: On the Illusion of Objectivity in NLP*. Jan. 28, 2021. DOI: 10.48550/arXiv.2101.11974. arXiv: 2101.11974 [cs]. URL: http://arxiv.org/abs/2101.11974 (cit. on p. 16).

[84] Xiaolin Wu and Xi Zhang. «Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of arXiv:1611.04135)». May 26, 2017. arXiv: 1611.04135 [cs]. URL: http://arxiv.org/abs/1611.04135 (cit. on p. 14).

[85] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. «Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy». In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Jan. 27, 2020), pp. 547–558. DOI: 10.1145/3351095.3375709. arXiv: 1912.07726. URL: http://arxiv.org/abs/1912.07726 (cit. on p. 16).

[86] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. «A Nutritional Label for Rankings». In: *Proceedings of the 2018 International Conference on Management of Data* (May 27, 2018), pp. 1773–1776. DOI: 10.1145/3183713.3193568. arXiv: 1804.07890. URL: http://arxiv.org/abs/1804.07890 (cit. on p. 16).

[87] Nico Zazworka, Rodrigo O. Spínola, Antonio Vetro', Forrest Shull, and Carolyn Seaman. «A Case Study on Effectively Identifying Technical Debt». In: *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering*. EASE '13. New York, NY, USA: Association for Computing Machinery, Apr. 14, 2013, pp. 42–47. ISBN: 978-1-4503-1848-8. DOI: 10.1145/2460999.2461005. URL: https://doi.org/10.1145/2460999.2461005 (cit. on p. 16).

[88] Meike Zehlike, Ke Yang, and Julia Stoyanovich. «Fairness in Ranking: A Survey». May 12, 2021. DOI: 10.48550/arXiv.2103.14000. arXiv: 2103.14000 [cs]. URL: http://arxiv.org/abs/2103.14000 (cit. on p. 16).

Acknowledgements

I would like to use this space to thank those who contributed to the realisation of this thesis.

A special thanks to my supervisor Prof. Antonio Vetrò who followed me, with his infinite availability, in every step of the realisation of this thesis.

Thanks also to my co-supervisor Prof. Juan Carlos De Martin for his valuable advices, right from the choice of the topic.

Over the years, their teachings have passed on a passion for this subject to me.

I am infinitely grateful to all my family for always believing in me and allowing me to embark on this path.

My heartfelt thanks go to Irene who has always been a solid point of reference and a fundamental support, always pushing me to do better.

Finally, I would like to thank all friends near and far, course colleagues and fellow Alter.POLIS comrades: they have been an important part of my life over the years.