Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Fernández Alvarez, Raul; Giner Miguelez, Joan, dir. Open Data for Machine Learning. 2023. (Enginyeria Informàtica)

This version is available at https://ddd.uab.cat/record/280712

under the terms of the  license

Open Data for Machine Learning

Raúl Fenández Álvarez

Resum— Aquest treball explora el potencial dels grans models de llenguatge (LLMs), concretament el GPT-3.5, en millorar la qualitat de les dades en portals de dades obertes. Estudis recents a la comunitat de Machine Learning (ML) així com Iniciatives Legislatives com l'European AI ACT, apunten a la necessitat de documentar els datasets usat per entrenar models de ML en un seguit de dimensions per garantir la seva equitat i seguretat. En aquestes iniciatives s'hi destaca la importa de documentar el context de creació de dades, així com els equips i infraestructura que han participat en la col·lecció i anotació de dades. En el cas dels Open Data portals, els estàndard de metadades com DCAT, no ofereixen suport per anotar d'aquesta informació i aquesta, en cas de ser-hi, només la podem trobar a la documentació adjunta dels dataset en format de text natural.

En aquest treball s'explora l'ús de LLM per extreure de forma estructurada d'aquesta informació de la documentació dels datasets. Amb aquest fi, s'ha identificat els tipus de documentació presents susceptibles de funcionar amb el mètode proposat i s'ha explorat diferents estratègies de prompting per optimitzar l'ús LLM. Els resultats d'aquest estudi mostren bon resultat en format de documentació estructurada de dades presents als Open Data portals, com el Data Mangement Plans (DMP), i obren possibilitat a desenvolupar eines i mètodes per millorar la qualitat de les dades en aquests portals.

Paraules clau— Grans Models de Llenguatge, Open Data portal, Metadades, Data Management Plan (DMP), User Guide, Prompting Strategies, Extracció de Dades.

Abstract— This work explores the potential of large language models (LLMs), specifically GPT-3.5, in improving the quality of data in Open Data portals. Recent studies in the machine learning (ML) community and legislative initiatives like the European AI ACT emphasize the need to document the datasets used to train ML models across various dimensions to ensure their fairness and safety. These initiatives highlight the importance of documenting the data creation context, as well as the teams and infrastructure involved in data collection and annotation. In the case of Open Data portals, metadata standards like DCAT do not provide support for annotating this information, and if present, it can only be found in the accompanying documentation of the dataset in natural language format.

This work explores the use of LLMs to extract this information from the documentation of datasets in a structured manner. To this end, the types of susceptible documentation present for the proposed method have been identified, and different prompting strategies have been explored to optimize the use of LLMs. The results of this study demonstrate good performance in generating structured documentation of data present in Open Data portals, such as Data Management Plans (DMPs), and open up possibilities for developing tools and methods to improve the quality of data in these portals.

Index Terms— Large Language Models, Open Data portals, Metadata, Data Management Plan (DMP), User Guide, Prompting Strategies, Data Extraction

1 INTRODUCTION

Society has evolved to be more data-driven, which means that data is produced and used in order to make most decisions in life. The advancements in the field of AI have led to these applications having an increasingly significant role in society and, precisely, these applications are heavy consumers of data. Recent studies have pointed out that data is not as accurate as it should be [1]. For example, facial analysis datasets with a low number of darker skin faces registered could reduce the accuracy of facial analysis models in that particular group, which represents social harm for them [2].

The Machine Learning community has a need to annotate datasets with the context of their creation. From this documentation, information such as the dataset creator, how it was created, whether any pre-processing was done

- *E-mail de contacte: raul.fernandeza@uab.cat*
- *Menció realitzada: Tecnologies de la Informació*
- *Treball tutoritzat per: Joan Giner-Miguelez*
- *Curs 2022/23*

before publishing, etc., can be extracted. This is valuable information for someone who wants to use the dataset to decide whether using that data or not is suitable.

Open Data portals use standards to describe the data, such as DCAT, which does not support this context of creation, something important for preparing data for Machine Learning. Despite this, scientific datasets upload documentation with information about the context, but it is in natural language text format and difficult to compute. Moreover, recent legislative initiatives like the European AI ACT ask for annotations of the data of those dimensions.

In this work, we explore the use of new large language models, such as GPT-3.5, for extracting the dimensions demanded by the community, through the documentation associated with datasets from Open Data portals.

The experimentation consists in identifying the different formats of documentation present in Open Data portals, where two relevant ones have been found: Data Management Plans and User Guides [3]. After this, a sample of datasets containing both types of documents were extracted. Prompting strategies were applied to extract the information required by the community. This has allowed for the extraction of different data where the potential of LLMs in this task has been verified.

LLM have great potential, but a structured format is needed, such as DMP. In a less structured format, such as User Guide, LLM are less effective. As we have seen, this is a start to developing new LLM tools that improve the metadata of the Open Data portals.

2 OBJECTIVES

The objective of this work is to explore the use of large language models (LLM) to improve the metadata of Open Data portals. The aim is to investigate how LLM can be leveraged to enhance the quality and accuracy of metadata associated with datasets available in these portals.

In order to achieve the main objective, we need to identify the different formats of documentation that are present in Open Data portals and are suitable for this experimentation. This involves understanding the types of documentation, such as Data Management Plans (DMP) and User Guides, that exist within the portals and can be utilized to improve metadata.

We also need to identify the most effective prompting strategies for extracting relevant information from the available documentation. The goal is to determine the best approaches and techniques for interacting with LLM in order to prompt them to generate structured and accurate metadata from the identified documentation formats.

The final step we need to achieve the main objective is evaluating the performance of LLM using sample datasets to draw conclusions about their viability for enhancing metadata in Open Data portals. In order to get to that point, we should apply LLM-based methods to datasets and analyze the outcomes to assess the potential benefits and limitations of using LLM in this context.

By pursuing these objectives, this study aims to contribute to the advancement of metadata quality and documentation practices in Open Data portals by taking advantage of the capabilities of LLM.

3 STATE OF THE ART

Currently, to know where the data in a dataset comes from, one must search for its technical documentation, which sometimes is very dense and unreadable, and often impossible to find on the same Open Data portal website.

This issue has attracted interest within the community in search of data standardization or the need for documentation. Recently, thanks to this interest, works such as "Datasheets for Datasets"[4] and "Data Nutrition Labels"[5] have been published, aiming to create guidelines for standardized documentation in datasets.

The study gets the idea of datasheets for documentation purposes. Throughout each phase of the dataset description process, the authors identify data aspects that could impact how the dataset could be used. Moreover, they ask for a discussion about the potential harms and bias in the data as part of their description.

3.1 Datasets documentation practices for ML

The dimensions that are the most relevant for this work are [16]:

- **Description** Purposes of the dataset, and their recommended applications.
- **Distribution** Link of the repository and third parties in charge of the licenses.
- **Provenance** Aspects of the gathering and the labeling process can be expressed. Also, the requirements of the processes or the information of who labeled the data.
- **Social Concerns** Aspects about social issues of the data and relate them to provenance aspects.

3.2 Documentation in Open Data Portals

European Union's Open Data portal [6] and the United States Open Data portal [7] were analyzed. They are very similar with both recollecting datasets from all over the

territory with the goal of promoting access to data. However, European one does not have access to the technical documentation making it difficult finding information. On the other side, the American one displays all the information need from every dataset making it easy to be found.

The European Union's Open Data portal [6] collects datasets from the following sources:

- European Union institutions
- European agencies
- International organizations
- National governments
- EU open data

This Open Data portal gathers information from sources belonging to the European Union and offers easy, fast, and transparent access to the data.

However, it does not provide the technical documentation of the datasets, only displaying the data and metadata, usually in CSV and JSON formats, respectively. To access the technical documentation of the datasets, you need to visit the corresponding institution and search for what you are looking for.

On the other hand, the United States Open Data portal [7] collects information from:

- Federal agencies
- State governments
- Local governments
- Non-profit organizations
- Private companies

This Open Data portal gathers information from US sources. Its main objectives are to promote transparency and accessibility to data in order to encourage the development of applications that benefit society.

Unlike the European Open Data portal, this one offers a wider variety of documents beyond data and metadata. You can find technical documentation in PDF format, such as User Guides and Data Management Plans.

LLM are used for information extraction. The newer versions are capable of extracting information in scientific fields. Also, they show good capabilities while extracting information for generating technical documents.

Prompting strategies are also a way to use LLM. These strategies refer to the way of communicating with LLM to have a desired outcome. The methods used can vary a lot, so they require heavy experimentation [8].

4 METHODOLOGY

In order to find out the type of documentation needed to make the experiment we carried out a field study.

After finding out the documentation needed, a sample selection was made by mixing DMP and User Guide.

While carrying out the experiment, an exploratory study was done, where a mixed methodology of quantitative and qualitative research methods was used.

The results of the experiments helped draw conclusions about the quality of the information, taking into account the number of fields that could be recovered and also the quality of the responses generated based on the context.

The hypothesis of the experiment is:

- LLM can be useful for improving metadata in Open Data portals.

Also, the sub-hypothesis of this experiment is:

- Using LLM and prompting strategies, answers will be of quality in a large number of the dimensions asked.

In order to measure the quality of the answers, a selected sample of different datasets was selected, where different types of files were chosen. The prompts were directly asked to OpenAI [9] from an API call in a Jupyter Notebook in Python.

The answers were manually reviewed and annotated in three different categories: good, bad, and hallucinate.

5 DEVELOPMENT

Firstly, we searched the various Open Data portals for datasets that could potentially contain what we were looking for. The main objective was to find annotated datasets. This way, the human factor becomes an important component.

Next, the steps followed during the development of the research and experimentation are detailed.

5.1 Identification of documentation types

Various files were found, of different types and for different purposes. It should be noted that it was difficult to find technical documentation during the searches, which had to be done manually since the various portals do not provide any assistance in searching for documentation.

Three types of documents were found, divided into 8 different datasets:

- Data Management Plan
- User Guide
- Log File

5.1.1 Data Management Plan

A Data Management Plan is a document that describes how data will be collected, how it will be stored or processed. Additionally, it also describes who will have access to the data and how it will be shared with other researchers or the public. Finally, it addresses topics such as data quality, security, or data preservation.

5.1.2 User Guide

A User Guide is a document that provides users with detailed information in order to understand the structure of a dataset and its contents. It also includes information about data collection, processing, and storage.

5.1.3 Log File

A log file is a file that records all the necessary information to understand a dataset. It includes all the activities or events performed to extract the data.

5.2 Development of the extraction pipeline

After the search, the experimentation with the documents had to begin. All the documents were in PDF format, so they were converted to TXT format using an online tool. Next, a cleaning process was performed to remove those text fragments that were not relevant to the experimentation.

Once the cleaning process was completed, and thanks to the code from the SOM Research group [10], the initial steps of the experimentation were carried out. The experimentation involves drawing conclusions about the answers to various questions of interest, seeking both quality and quantity in the results. The code, written in Python format and executed in a notebook, makes an API call to OpenAI [9] and enables working in a similar manner to its playground. Furthermore, the code includes prompts that will be used to draw conclusions about the texts. We can determine that the process was as seen in Figure 1.

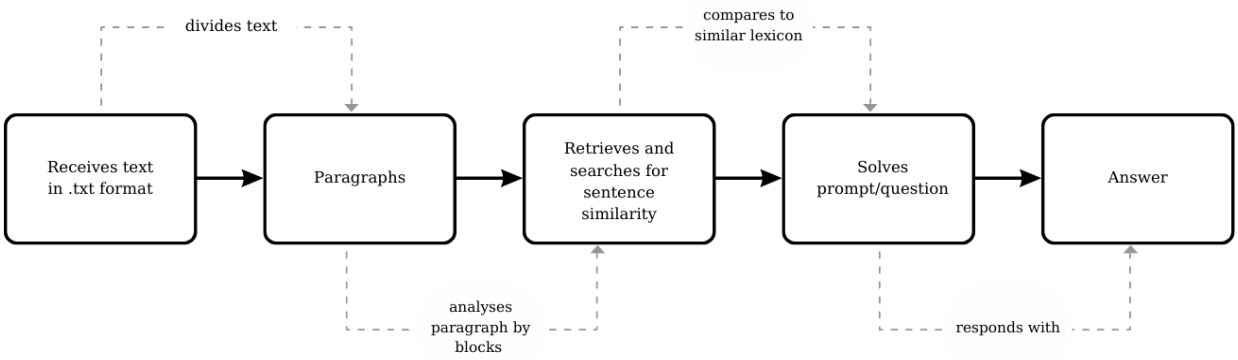

Figure 1: Code workflow

For DMP files, being a more structured file type, no modifications were made to the already provided TXT files. To obtain better results, the code was inspected to identify its weaknesses in reading the file. This required understanding how the code worked and how it processed the provided texts.

The code divides the text into chunks, which are fragments of the text passed in TXT format but divided for better reading and interpretation by the language model. After investigation, it was determined that the code was dividing the text into very small chunks. Therefore, the chunk_size field was modified from 200 to 500. The chunk_size field determines the length and extent of a chunk. By doing this, we have fewer chunks but of larger size, which means the text is divided into smaller texts. This allows chunks without clear context to be combined with others to provide a larger context for answering questions.

For User Guides, as they contain free-form text, it was decided to restructure the TXT file to remove unnecessary paragraphs, such as section titles. This is because, during the text separation process, they were treated as separate paragraphs, which made it impossible for the language model to understand the text. Additionally, the code was modified multiple times to find an optimal solution that improved the responses obtained after experimenting with default values. To achieve this improvement, the chunk_size field was modified. However, it was modified several times because an optimal value that consistently produced accurate responses was not found. Eventually, the chunk_size value was set to 500. Lastly, in order to continue experimenting, the chunk_overlap value was changed from 10 to 20. This was done to provide better context for the responses, considering previous answers and maintaining a connection between chunks.

5.3 Review

After analyzing the datasets of each document type, it was determined that the process would not work well with the log file due to the file typology, which is not optimal for experimentation.

However, good conclusions were obtained with the DMP. Since it is a structured text with questions, the answers tend to be more complete obtaining better results.

The User Guide, on the other hand, proved to be the document type that worked worse. Since it contains natural text, a disfavorable context can be generated for the language mode. This provides errors and generates confusion while answering the prompts.

So, DMP provided high-quality answers, and the conclusions for User Guide drew poorer answers to the questions, concluding in less-quality answers.

6 RESULTS

In this section, the results of the experiments will be shown. With these results, a conclusion will be made for later interpretation and explanation. The answers to the prompts that will be shown are in Attachment A1.

6.1 Used datasets

In order to perform the experiments, a data source is needed first. Below are detailed five datasets that will be used in the experimentation. These datasets have been downloaded through the data.gov website [7]:

- **Biologically Important Areas for Cetaceans within U.S. Waters** DMP file about areas where cetaceans have their biologically important areas in the U.S. Waters.[11]
- **Biscayne Bay Dolphin Photo ID System** DMP file about the study of the population stock structure of bottlenose dolphins.[12]
- **Habitat Mapping Camera (HABCAM)** DMP file about the imagery collected by HabCam underwater vehicle.[13]
- **Glacier Photograph Collection** User guide file about the digitalization of the Glacier Photograph Collection Index project.[14]
- **School Neighborhood Poverty Estimates** User guide file about the economic conditions of neighborhoods where schools are located.[15]

6.2 Description Analysis

For every dataset, as mentioned in the experiments section, some tests were made in order to achieve the best results. For comparing the results, the prompts that were asked to LLM were also answered by a human. In Table 1, we can see some of the results done with the prompt 'Which are the purposes of the dataset?', which is a dimension of the description part:

DATASET	DOCUMENT TYPE	EVALUATION
Biologically Important Areas for Cetaceans within U.S. Waters	DMP	Good
Biscayne Bay Dolphin Photo ID System	DMP	Good
Habitat Mapping Camera (HABCAM)	DMP	Good
Glacier Photograph Collection	User Guide	Good
School Neighborhood Poverty Estimates	User Guide	Good

Table 1: Evaluation of prompt in description

As we can see, for the selected prompt, the evaluation was good, which means that in every dataset, the answer given by the LLM was as good as the one given by humans.

6.3 Distribution Analysis

In Table 2, we can see some of the results done with the prompt 'Which are the rights of the stand-alone dataset?', which is a dimension of the distribution part:

DATASET	DOCUMENT TYPE	EVALUATION
Biologically Important Areas for Cetaceans within U.S. Waters	DMP	Good
Biscayne Bay Dolphin Photo ID System	DMP	Good
Habitat Mapping Camera (HABCAM)	DMP	Good
Glacier Photograph Collection	User Guide	Good
School Neighborhood Poverty Estimates	User Guide	Bad

Table 2: Evaluation of prompt in distribution

With this prompt, we can see that every dataset is working properly except for the User Guide for the School

Neighborhood Poverty Estimates. The answer is wrong because the LLM mixes some terms, which makes the answer not consistent.

6.4 Provenance Analysis

In Table 3, we can see some of the results done with the prompt 'The data was collected by an internal team, an external team, or a crowdsourcing team?', which is a dimension of the provenance part:

DATASET	DOCUMENT TYPE	EVALUATION
Biologically Important Areas for Cetaceans within U.S. Waters	DMP	Good
Biscayne Bay Dolphin Photo ID System	DMP	Good
Habitat Mapping Camera (HABCAM)	DMP	Good
Glacier Photograph Collection	User Guide	Hallucinate
School Neighborhood Poverty Estimates	User Guide	Bad

Table 3: Evaluation of prompt in provenance

We can see that with a structured file, this question is correctly answered. On the other hand, with a less structured file, LLM tend to answer incorrectly.

6.5 Social harm Analysis

In Table 4, we can see some of the results done with the prompt 'Is there any potential bias in the data?', which is a dimension of the social harm part:

DATASET	DOCUMENT TYPE	EVALUATION
Biologically Important Areas for Cetaceans within U.S. Waters	DMP	Good
Biscayne Bay Dolphin Photo ID System	DMP	Good
Habitat Mapping Camera (HABCAM)	DMP	Good
Glacier Photograph Collection	User Gudie	Good
School Neighborhood Poverty Estimates	User Guide	Good

Table 4: Evaluation of prompt in social harm

In this prompt, LLM has good answers in every dataset, making a reliable response to the question.

6.6 Dimensions Analysis

The dimensions mentioned in the State of the Art section helped us in seeing which type of file is best for those dimensions. While with the DMP file, LLM tend to make good answers to the questions, User Guide made both good and bad guesses. With these results, we see that DMP is a way better file for what we are looking for.

6.7 Results Analysis

Dataset	Good	Bad	Hallucinate	Total
Biologically Important Areas for Cetaceans within U.S. Waters	43	14	0	57
Biscayne Bay Dolphin Photo ID System	54	3	0	57
Habitat Mapping Camera (HABCAM)	53	8	0	61
Total	150	25	0	175

Table 5: Results of the quality of each dataset with DMP.

As we can see in Table 5, the results are pretty good. Approximately 85% of the prompts that were passed to the LLM were answered correctly. However, LLM fails in some of the prompts in each of the datasets. As seen in Table 4, when asked about a location, LLM tends to answer about something physical not about a geographical point.

Also, we can see that Biologically Important Areas for Cetaceans within the U.S. Waters dataset has nearly double of bad answers than the next one which is Habitat Mapping Camera (HABCAM). This happens because of the

completeness of the points of the file. As the first dataset has fewer points resolved, LLM fails to extract good answers to some questions.

Dataset	Good	Bad	Hallucinate	Total
Glacier Photograph Collection	51	6	2	59
School Neighborhood Poverty Estimates	53	8	0	61
Total	104	14	2	120

Table 6: Results of the quality of each dataset with User Guide.

The results of the analysis for the User Guide are bad. Even though nearly 87% of the answers were cataloged as good, 2 out of 120 of the answers were cataloged as hallucination. This is a bad result for LLM because each of the files analyzed made a bad guess. However, if prompts were clearer and more specific for each dataset, the number of hallucinations could be less.

7 CONCLUSIONS

During the course of this work, all the proposed objectives have been successfully achieved within the established timeframe outlined in the project plan. Therefore, the project has been developed correctly.

We can conclude that LLMs have proven to be highly effective tools for extracting valuable information from documentation. In the context of this work, they have demonstrated their potential in extracting relevant data from various types of documentation.

Also, among the documentation formats, DMP have shown to be particularly suitable for extracting the required information accurately. Moreover, DMP has the information demanded by the ML community. On the other hand, User Guide tends to generate incorrect answers, resulting in a higher likelihood of hallucination.

Scientifically documented data stands out as the category with better-documented information. These datasets often provide more comprehensive and reliable documentation, enhancing the quality of metadata extraction.

Open Data portals can be challenging to navigate, and locating associated documentation can be a complex task. However, the American data portal has been identified as comparatively better than others in terms of documentation availability and accessibility. This work provides new ways of study for developing tools and methods to improve them.

The experimentation conducted in this study has yielded positive results, particularly in the case of DMPs and specific dimensions that were queried to the LLM. These findings support the potential and efficacy of using LLMs for improving the metadata quality of Open Data portals.

As we have seen in the conclusions, the hypothesis that was made at the beginning of this work has been completely answered. LLM are very useful for improving metadata in Open Data portals.

Moreover, the sub-hypothesis has been as well answered. Using LLM and prompting strategies has been great and the answers provided are of quality in a large number of the dimensions that have been asked. However, we should also emphasize the quality of some of the answers that LLMs are doing. Doing a better-prompting strategy or a better paragraph split would help LLM understand better what the answer should be.

Finally, it would be very interesting if this work could be developed further in order to improve and be useful in the future. Some of the possible continuations of the projects are:

- LLM from an open source should be tested in order to avoid vendor-locking[1].
- Tools must be developed in order to automate the process and improve data in every Open Data portal.
- Better prompts that help LLM understand better what is being asked.

ACKNOWLEDGMENTS

I would like to thank all my family members that have always been by my side during these years. I would also like to extend my gratitude to my girlfriend for supporting me every day with this project and with every other thing in life. Finally, I would like to thank Sindic for the hours spent together and the suport they have given me. Finally I would like to thank my tutor for letting me in the world of investigation and experimentation.

1 Vendor-locking refers to a situation where the cost of switching to another vendor is so high that the customer is practically compelled to continue with the original vendor.

REFERENCES

[1] N. Nahar, S. Zhou, G. Lewis, C. Kästner, Collaboration challenges in building ML-enabled systems: Communication, documentation, engineering, and process, in: 44th International Conference on Software Engineering (ICSE '22), Vol. 1, 2022, p. 3.

[2] A. Khalil, S.G. Ahmed, A.M. Khattak, N. Al-Qirim, Investigating bias in facial analysis systems: A systematic review, IEEE Access 8 (2020) 130751–130761.

[3] Cynthia Hudson-Vitale and Heather Moulaison-Sandy. 2019. Data Management Plans: A Review. DESIDOC Journal of Library & Information Technology 39, 6(2019), 322–328.

[4] T. Gebru, J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H.D. Iii, K. Crawford, Datasheets for datasets, Commun. ACM 64 (12) (2021) 86–92.

[5] S. Holland, A. Hosny, S. Newman, J. Joseph, K. Chmielinski, The dataset nutrition label, in: Data Protection and Privacy, Volume 12: Data Protection and Democracy, 12, Bloomsbury Publishing, 2020, p. 1.

[6] European Open Data portal - Home, 2023, https://data.europa.eu, last accessed June 2023.

[7] Government's Open Data - Home, 2023, https://data.gov, last accessed June 2023.

[8] Weng, Lilian. (Mar 2023). Prompt Engineering. Lil'Log. [Online] Available: https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/.

[9] OpenAi - Home, 2023, https://openai.com, last accessed June 2023.

[10] Som Research - Home, 2023, https://som-research.uoc.edu/, last accessed June 2023.

[11] NOAA. Biologically Important Areas for Cetaceans within U.S. Waters. [Online]. Available: https://catalog.data.gov/dataset/biologically-important-areas-for-cetaceans-within-u-s-waters

[12] NOAA. Biscayne Bay Dolphin Photo ID System. [Online]. Available: https://catalog.data.gov/dataset/biscayne-bay-dolphin-photo-id-system

[13] NOAA. Habitat Mapping Camera (HABCAM). [Online]. Available: https://catalog.data.gov/dataset/habitat-mapping-camera-habcam

[14] NSIDC. Glacier Photograph Collection. [Online]. Available: https://nsidc.org/data/g00472/versions/1#qt-data_set_tabs

[15] Department of Education. School Neighborhood Poverty Estimates. [Online]. Available: https://catalog.data.gov/dataset/school-neighborhood-poverty-estimates-2019-20

[16] Giner-Miguelez, Joan & Gómez, Abel & Cabot, Jordi. (2022). DescribeML: A Tool for Describing Machine Learning Datasets. 10.1145/3550356.3559087.

ATTACHEMENTS

A1. Excel File

The excel file is attached in the dossier. It contains all the answers from table 1 to table 4.