# **I. Data and Metadata Profile**

# **Introduction**

Exploring data sets is a way to view statistics and trends that can be useful in creating data visualizations. There is so much data on the internet, that you can use the data that is available to use instead of reinventing “the wheel” by collecting your own data. There are many repositories now on the internet that hold data that you collect or that someone else collects. By accessing the data around you can gain an important perspective on what data you can find online to support your data analysis. For this data curation project, I will focus on one data set from a repository. After reviewing several repositories such as Figshare, Zendodo, Kaggle, and Data.gov, I found an interesting data set about mushrooms from Kaggle entitled “Mushroom Classification Safe to Eat or Deadly Poison. Being new to curating data sets, I found this data set informative, clear, concise, user friendly, which made it easy for a beginner to read, comprehend, and locate the information that I needed. This data set not only benefits the scientific community, but it is a useful data set in distinguishing between edible and poisonous mushrooms for the average “shrooming” individual.

# **Data Information**

The Mushroom Classification data set contains hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family from The Audubon Society Field Guide to North American Mushrooms (1981). The mushrooms were further broken down into two categories: edible and non-edible (definitely poisonous) including those of unknown edibility and therefore, not recommended. The data set, Mushroom Classification Safe to Eat or Deadly Poison, is owned and maintained by UCI (University of California -Irving) Center of Machine Learning. The most current version of this data set was created December 1, 2016. It was originally donated to UCI Machine Learning by Jef Schlimmer on April 27, 1987. Although this data set was created over thirty years ago, it is still used often today. On Kaggle, it has received over 900 kernels, which shows how much interest there is about this data set. The Mushroom Classification data set is in public domain (CCO), which means there is no copyright. Anyone can copy, modify, distribute, and use the information even for commercial purposes without asking permission for use. The data set file is a zip file containing one file, which can be opened by using the notepad app or the Excel program on the computer. There are 8,124 rows describing the 22 different attributes of mushrooms. Each attribute is further divided into various descriptive words. For example, the attribute describing the cap-shape is further broken down into bell=b, conical=c, convex=x, flat=f, knobbed-k, and sunken=s.

# **Metadata Information**

Datasets that use a common metadata standard helps to develop interoperability. This also makes discovery of existing data easier. There are several reasons why metadata standards should be used in datasets. Datasets that use metadata standards can:

* build aggregators to run the same search over
* put the data in context so it easier to read
* help see relationships in the data
* save the next person time and effort without rebuilding the datasets
* allow others to build on and integrate the data

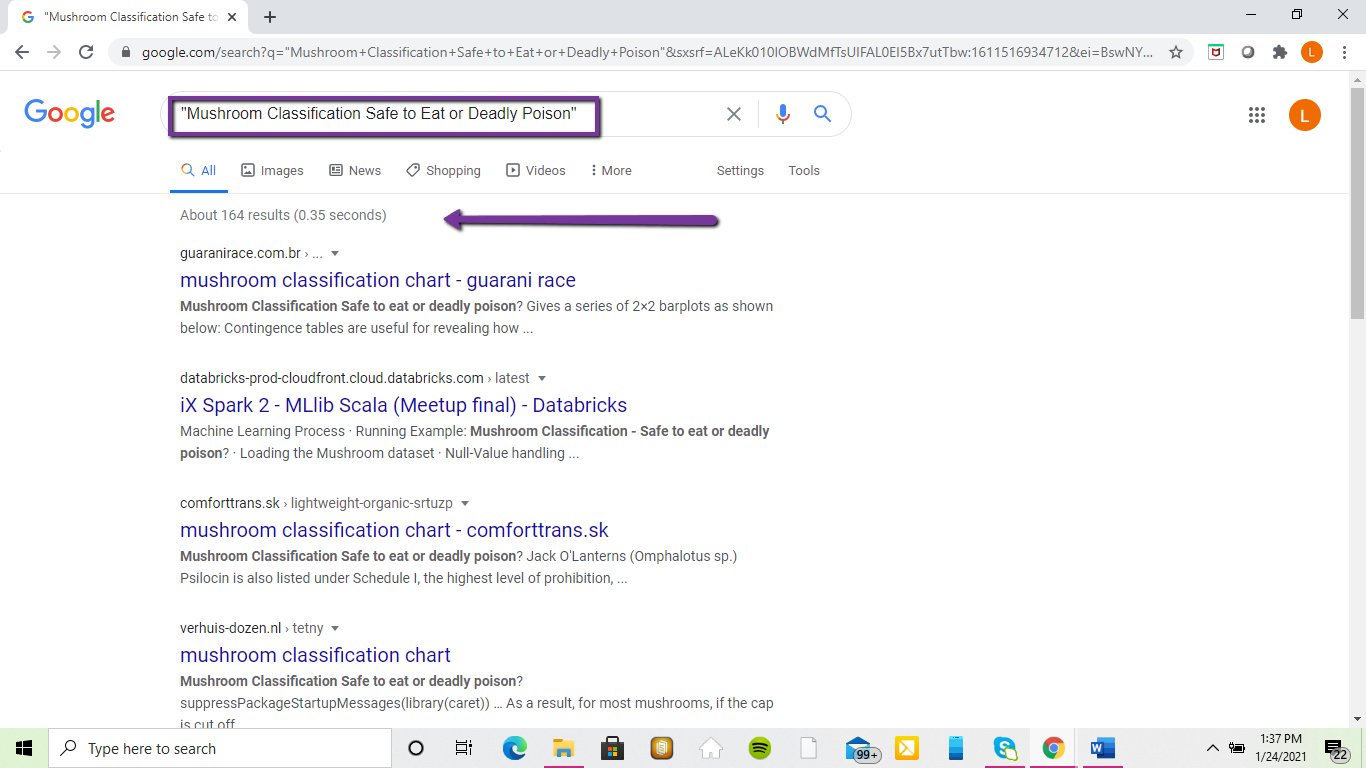
The Mushroom Classification data set is a machine learning data set which uses a structured metadata standard called a tabular data package. It uses the extension csv (comma-separated values), for example in the data set it is labeled as mushrooms.csv. This tabular format uses rows and columns, which corresponds to the record and the features. It also uses the kernel method, which means in machine learning that it uses algorithms for pattern analysis. This is used to study and find the relationships in correlation, classification, and ranking in the data set. This means that the metadata is included in the data files.

**Improving the Current Data and Metadata**

The data set is easy to understand and offers essential metadata. It contains file descriptions, tags, subtitles, and has a public kernel which is fairly active. The notebook feature that allows people to share their work with others. It also offers a discussion thread that allows people to ask questions and make comments or suggestions. Some suggestions for further improvement are to enhance the visualization and readability. This data set would benefit with better data visualizations. The presentation of the data could also be presented in the form of bar graphs, pie graphs, and scatter graphs, just to name a few data visualization enhancements. Showing the pictures of the mushrooms themselves will aid in visualizing the different species of mushrooms and more easily understand the data better, for example, pictures of the various types of mushroom caps. Visualization of the poisonous and non-poisonous mushrooms would help increase recognition of the mushrooms if the data is used for the “shrooming” individual.

**Publications that used the Data Set**

Kaggle does not show or list any publications that have cited this data set. The owners (University of California, Irving Machine Learning Repository) of the data set, Mushroom Classification Safe to Eat or Deadly Poison, used rexa.info to show how many and what publications cited this data set in their research. Forty-seven papers have cited this data set in their publications. In doing a google scholar advance search, I typed in the exact phrase box the name of the data set. The search resulted in one source (Sabrina Ingrid Davita) used the data set at rbubs.com on December 1, 2019. However, when I did a google search I typed “Mushroom Classification Safe to Eat or Deadly Poison” it yielded 164 results. See the screen shot below.



## 

Repositories and data sets are good resources for data collection. By using data sets that are already available saves time and allows for people to build on other concepts and information. Data sets can foster creativity and data sharing using the internet through online repositories that hold the data sets. In the past, without the internet this sharing was not possible. Data sharing allows people to create useful data visualizations that can benefit many different fields such as medicine, education, agriculture, statistics, and more.

**II. Repository Information**

**Introduction**

The Mushroom Observer is a repository that was started by Nathan Wilson in 2006. The purpose of this repository is to record observations about mushrooms, to help people identify mushrooms, and to provide scientific exploration of mushrooms. This repository is used worldwide and is available in many languages such as English, French, German, Greek, and more. The Mushroom Observer helps to bridge the gap between what is known and unknown about mushrooms. This repository provides a space for the amateur shroomer to the professional mycologist to learn, share, and record observations of mushrooms, fungi, and other similar plants. The Mushroom Observer’s owner states that he likes to think of this repository as a living field guide for mushrooms or a collaborative mushroom field journal.

**Why I Chose This Repository**

Mushroom Observer was chosen, because it corresponds with my data set from Kaggle entitled “Mushroom Classification Safe to Eat or Deadly Poison.” This data set not only benefits the scientific community, but it is a useful data set in distinguishing between edible and poisonous mushrooms for the average “shrooming” individual. The data set that I have would be beneficial because Mushroom Observer’s main purpose is to identify mushrooms and to record observations of mushrooms to a wide variety of individuals from the amateur to the professional.

**The Repository and Making Data Submissions**

The Mushroom Observer is an open repository that only requires a simple registration of username, full name, and email address. Once you are a registered member, you become part of the community. The next step is to create one observation. To create an observation the following information is needed about the mushroom you observed: when (date), where (location – city, state for United States). There is a locate button to bring the location up on the map. Now add the marker and drag it to specific latitude and longitude. There is also an interactive map with an optional GPS location. Next step is to identify the mushroom (common name and scientific name) if you do not know the name of the mushroom, you can leave it blank. Then you check a box to say how you recognized the mushroom – by sight, used references, microscopic features, chemical features. In addition you can add any notes such as distinctive texture, scent, bruising. Notes may also be formatted using the Textile Markup System. Ex. \_\_Amanita ocreata\_,\_A.ocreata\_-->***Amanita Ocreata.*** Last step is to upload the image. When uploading an image (jpeg), you are giving permission for anyone to use your pictures, it is considered public domain. If you did not take the image, but you have the explicit permission of the person to upload the image, then you must update the ‘Copyright Holder’ field to reflect the actual owner. This page also allows you to select the License you want to release the image under.

**Getting Help with an Unknown**

To get help with an unknown mushroom, the community will assist by proposing one or more names for it within 24 hours. Members may also post feedback via Comments, any comments you receive will be sent as an email to you. There is also an “Edit Observation” button located at the top of your observation page. Links may be added within a comment using standard HTML formatting.

**Special Features**

**Working on Projects**

There is a project page which allows a group of users to work together keeping their project out of the public eye until they are ready to publish it. This is open to any members. Often members use this function to work on creating new descriptions for a species. When a description is published it will become official only if there are no other descriptions.

**Creating Species Lists**

After completing registration and one observation, members can create species lists. To create a species list, fill in the various fields. This will allow members to see if there is a match in the database. If there is no matching name, then members can add unknown names to the database. These new names will be used to create new observations. Members can add anyone’s observations to their species list. Species list can be saved for later use and added to other species list.

**Indexes**

Under the Index heading, there are six types of indexes that can aid the member.

* Fungaria – list of Institutional Funaria with links to fungarium records
* Locations – maps with the observation locations
* Names – Scientific names with observations
* News – articles, ideas in development
* Projects – project titles with links
* Glossary – glossary of mycology terms can be created by anyone in the community. Some pictures have been added in collaboration with the Rhode Island School of Design. (Some future plans consist of adding links to terms, translations of terms and definitions, multiple images.)

**Downloadable Features**

There are several downloadable features that The Mushroom Observer offers for its members. In addition to creating and saving a Species List Report, a member may download their list. Another downloadable feature is that the member can print and download labels for observations. For example:

**MO #:** 417590  
**When:** 2020-07-15  
**Who:** I. G. Safonov  
**Where:** Washington Crossing State Park, Mercer Co., New Jersey, USA (40.3071°N 74.8627°W)  
**What:** ***Amanita*** sect. ***Amidella*** (E.-J. Gilbert) Konrad & Maubl.  
**Notes:** A single fruitbody growing on sloped land in deciduous woods (oaks and other hardwoods around, but no beech nearby). Found 15-20 yards away from sect. Amidella of obs 415288, but that one grew surrounded only by beech trees. Direct comparison between the two collection is impractical due to disparity in age and condition of basidiomata, but sequencing should provide an answer.  
Overall height = 9 cm  
Cap footprint = 4 cm; flesh above stipe = 0.5 cm thick  
Stipe = 8.5 cm long, 8 mm wide in apex, narrowing to 5 mm 2 cm below apex, gradually widening to 9 mm inside the volval sack  
Volva = 2 cm tall, thin, and soft  
Odor = none detected

Members may save a checklist of names as plain text, rich text, or spreadsheet. Members may also download their observations in different formats as CSV spreadsheet, The Adolf Special™

Darwin Core, Symbiota, and FunDiS. In addition to different formats members can also choose character encoding ASCII (no accents), WINDOWS-1252 (best for Excel), UTF-8 (most universal), and UTF-16 (best for Windows 7 and up).

In conclusion, the Mushroom Observer is an excellent repository that offers information for varying levels of mycology from the amateur to the professional. The Mushroom Observer continues to listen to the needs of their community members by allowing them to have a voice in discussions and suggestions to make the repository better, all the while fulfilling its mission and purpose to record observations about mushrooms, to help people identify mushrooms, and to provide scientific exploration of mushrooms.

**III Additional Information**

**Recommended Data Citation**

UCI Machine Learning. (2017). Mushroom Classification Safe to eat or deadly poison? 1987-2017 (Version1) [Data set]. Kaggle. <https://www.kaggle.com/uciml/mushroom-classification>

**Long-Term Preservation Concerns**

The Mushroom Classification data set has been around for thirty-four years in a csv format. This format can easily be opened in notepad and Excel. It is highly unlikely that there would be any long-term preservation concerns, since notepad and Excel are used in many programs and databases. The data set has been used by many researchers, students, and others in various ways and formats throughout the years.

**Copyright Statement**

The original data set was donated by Jeffrey Schlimmer to UCI Machine Learning on April 27, 1987, which means that it is in public domain. The copyright for this repository is Creative Commons Zero v1.0 Universal (public domain) the same as the original data set. It may be copied, modified, distributed, and used for all purposes without asking for permission to use.

**Human Subject Considerations**

Due to the nature of this dataset, no human subjects participated in the data set. Hence, no human subject consideration are needed for the data set.

**References**

Creative Commons Licenses. <http://creativecommons.org/licenses/>

DataOne Best Practices Primer (review section 5.4- Describe: Data Documentation); Available <http://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf>

Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Frye, Curt. (2019) Learning Public Data Sets <https://www.lynda.com/Excel-tutorials/Learning-Public-Data-Sets/5034173-2.html>

Introduction to Machine Learning Tutorials <https://www.educba.com/machine-learning-datasets/>

Mushroom Observer. <https://mushroomobserver.org/>

Research Data Alliance (RDA) Metadata Directory, <http://rd-alliance.github.io/metadata-directory/>

UCI Machine Learning. (2017). *Mushroom Classification Safe to eat or deadly poison?* 1987-2017 (Version1) [Data set]. Kaggle. <https://archive.ics.uci.edu/ml/datasets/mushroom> also located on <https://www.kaggle.com/uciml/mushroom-classification>