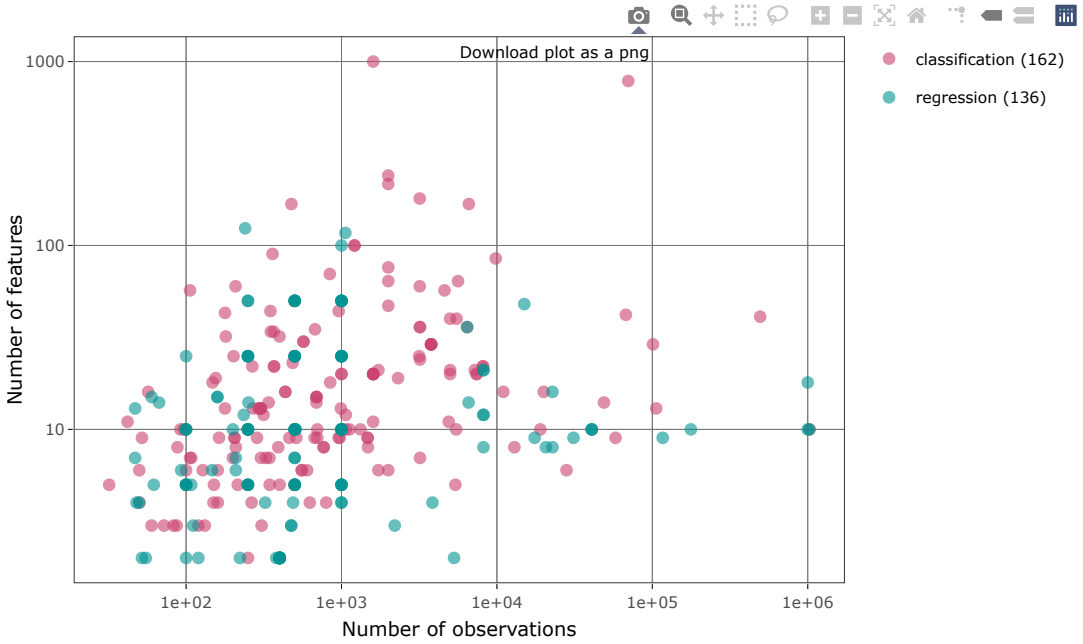
# Penn Machine Learning Benchmarks

Penn Machine Learning Benchmarks (PMLB) is a large collection of curated benchmark datasets for evaluating and comparing supervised machine learning algorithms. These datasets cover a broad range of applications including binary/multi-class classification and regression problems as well as combinations of categorical, ordinal, and continuous features.

## Summary statistics

In the interactive plotly (https://plotly.com/) chart below, each dot represents a dataset colored based on its associated task (classification vs. regression). In log scale, the *x* and *y* axis shows the number of observations and features respectively. Please click on the legend to hide/show the groups of datasets. Click on each dot to access the dataset's pandas-profiling (https://pandas-profiling.github.io/pandas-profiling/docs/master/rtd/) report.

*Note*: If a dataset has more than 20 features, we randomly chose 20 to be displayed in its profiling report. Therefore, please disregard the `Number of variables` in the corresponding report and, instead, use the correct `n_features` in the chart and table below.




Browse, sort, filter and search the complete table of summary statistics below.

- Click on the dataset's name to access its pandas-profiling (https://pandas-profiling.github.io/pandas-profiling/docs/master/rtd/) report.
- Click on the GitHub Octocat to access its metadata.
- To filter, please type in the box at the bottom of each numeric column in the format `low ... high` . For example, if you want to see all *classification datasets with 80 to 100 observations*, select `classification` at the bottom of `Task` and type `80 ... 100` at the bottom of the `n_observations` column.

CSV Show 10 entries Search:

| Dataset | n_observations | n_features | n_classes | Endpoint | Imbalance |
|---|---|---|---|---|---|
| adult (https://epistasislab.github.io/pmlb/profile/adult.html) | 48842 | 14 | 2 | categorical | 0.27 |
| agaricus_lepiota (https://epistasislab.github.io/pmlb/profile/agaricus_lepiota.html) | 8145 | 22 | 2 | categorical | 0 |
| allbp (https://epistasislab.github.io/pmlb/profile/allbp.html) | 3772 | 29 | 3 | categorical | 0.88 |
| allhyper (https://epistasislab.github.io/pmlb/profile/allhyper.html) | 3771 | 29 | 4 | categorical | 0.93 |
| allhypo (https://epistasislab.github.io/pmlb/profile/allhypo.html) | 3770 | 29 | 3 | categorical | 0.78 |
| All | All | All | All | All | All |

| Dataset | n_observations | n_features | n_classes | Endpoint | Imbalance |
|---|---|---|---|---|---|
| allrep (https://epistasislab.github.io/pmlb/profile/allrep.html) | 3772 | 29 | 4 | categorical | 0.91 |
| analcatdata_aids (https://epistasislab.github.io/pmlb/profile/analcatdata_aids.html) | 50 | 4 | 2 | categorical | 0 |
| analcatdata_asbestos (https://epistasislab.github.io/pmlb/profile/analcatdata_asbestos.html) | 83 | 3 | 2 | categorical | 0.01 |
| analcatdata_authorship (https://epistasislab.github.io/pmlb/profile/analcatdata_authorship.html) | 841 | 70 | 4 | categorical | 0.08 |
| analcatdata_bankruptcy (https://epistasislab.github.io/pmlb/profile/analcatdata_bankruptcy.html) | 50 | 6 | 2 | categorical | 0 |
| All | All | All | All | All | All |

Showing 1 to 10 of 298 entries

Previous 1 2 3 4 5 … 30 Next

The complete table (https://github.com/EpistasisLab/pmlb/blob/master/pmlb/all_summary_stats.tsv) of dataset characteristics is also available for download. Please note, in our documentation, a feature is considered:

- "binary" if it is of type integer and has 2 unique values (equivalent to pandas profiling's "boolean")
- "categorical" if it is of type integer and has *more than* 2 unique values (equivalent to pandas profiling's "categorical")
- "continuous" if it is of type float (equivalent to pandas profiling's "numeric").

# Dataset format

All datasets are stored in a common format:

- First row is the column names
- Each following row corresponds to one observation of the data
- The dependent variable/endpoint/outcome column is named `target`
- All columns are tab ( `\t` ) separated
- All files are compressed with `gzip` to conserve space

# Citing PMLB

If you use PMLB in a scientific publication, please consider citing one of the following papers:

Le, Trang T., William La Cava, Joseph D. Romano, John T. Gregg, Daniel J. Goldberg, Praneel Chakraborty, Natasha L. Ray, Daniel Himmelstein, Weixuan Fu, and Jason H. Moore. PMLB v1. 0: an open source dataset collection for benchmarking machine learning methods (https://arxiv.org/abs/2012.00058). *arXiv preprint arXiv:2012.00058* (2020).

```
@article{romano2021pmlb,
  title={PMLB v1.0: an open source dataset collection for benchmarking machine learning methods},
  author={Romano, Joseph D and Le, Trang T and La Cava, William and Gregg, John T and Goldberg, Daniel J and Chakraborty, Prane
        el and Ray, Natasha L and Himmelstein, Daniel and Fu, Weixuan and Moore, Jason H},
  journal={arXiv preprint arXiv:2012.00058v2},
  year={2021}
}
```

Olson, Randal S., William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz, and Jason H. Moore. PMLB: a large benchmark suite for machine learning evaluation and comparison (https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0154-4). BioData mining 10, no. 1 (2017): 1-13. *BioData Mining* **10**, page 36.

BibTeX entry:

```
@article{Olson2017PMLB,
    author="Olson, Randal S. and La Cava, William and Orzechowski, Patryk and Urbanowicz, Ryan J. and Moore, Jason H.",
    title="PMLB: a large benchmark suite for machine learning evaluation and comparison",
    journal="BioData Mining",
    year="2017",
    month="Dec",
    day="11",
    volume="10",
    number="36",
    pages="1--13",
    issn="1756-0381",
    doi="10.1186/s13040-017-0154-4",
    url="https://doi.org/10.1186/s13040-017-0154-4"
}
```

# Support for PMLB

PMLB was developed in the Computational Genetics Lab (http://epistasis.org/) at the University of Pennsylvania (https://www.upenn.edu/) with funding from the NIH (http://www.nih.gov/) under grant AI117694, LM010098 and LM012601. We are grateful for the support of the NIH and the University of Pennsylvania during the development of this project.