

# Package ‘pmlbr’

February 20, 2025

**Title** Interface to the Penn Machine Learning Benchmarks Data Repository

**Description** Check available classification and regression data sets from the PMLB repository and download them.  
The PMLB repository (<<https://github.com/EpistasisLab/pmlbr>>) contains a curated collection of data sets for evaluating and comparing machine learning algorithms. These data sets cover a range of applications, and include binary/multi-class classification problems and regression problems, as well as combinations of categorical, ordinal, and continuous features. There are currently over 150 datasets included in the PMLB repository.

**Version** 0.2.3

**BugReports** <https://github.com/EpistasisLab/pmlbr/issues>

**Depends** R (>= 3.2.0)

**Imports** utils, FNN, stats

**License** GPL-2 | file LICENSE

**URL** <https://github.com/EpistasisLab/pmlbr>

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.3.2

**NeedsCompilation** no

**Author** Trang Le [aut, cre] (<https://trang.page/>),  
makeyourownmaker [aut] (<https://github.com/makeyourownmaker>),  
Jason Moore [aut] (<http://www.epistasisblog.org/>),  
University of Pennsylvania [cph]

**Maintainer** Trang Le <[grixor@gmail.com](mailto:grixor@gmail.com)>

**Repository** CRAN

**Date/Publication** 2025-02-19 23:40:05 UTC

## Contents

classification_dataset_names . . . . .	2
compute_imbalance . . . . .	2
dataset_names . . . . .	3
fetch_data . . . . .	3
get_type . . . . .	4
graceful_download . . . . .	5
nearest_datasets . . . . .	5
pmlb . . . . .	7
regression_dataset_names . . . . .	8
summary_stats . . . . .	8

<b>Index</b>	<b>10</b>
--------------	-----------

---

classification\_dataset\_names  
*Names of available classification datasets*

---

### Description

A list of the names of available classification datasets

### Usage

```
classification_dataset_names
```

### Format

An object of class character of length 162.

### Source

<https://github.com/EpistasisLab/pmlb>

---

compute\_imbalance      *Computes imbalance value for a given dataset.*

---

### Description

Computes imbalance value for a given dataset.

### Usage

```
compute_imbalance(target_col)
```

**Arguments**

target\_col      Factor or character vector of target column.

**Value**

A value of imbalance metric, where zero means that the dataset is perfectly balanced and the higher the value, the more imbalanced the dataset.

---

dataset_names	<i>Names of all available datasets</i>
---------------	--

---

**Description**

A list of the names of available datasets

**Usage**

```
dataset_names
```

**Format**

An object of class character of length 284.

**Source**

<https://github.com/EpistasisLab/pmlb>

---

fetch_data	<i>fetch_data function</i>
------------	----------------------------

---

**Description**

Download a data set from the PMLB repository, (optionally) store it locally, and return the data set. You must be connected to the internet if you are fetching a data set that is not cached locally.

**Usage**

```
fetch_data(  
  dataset_name,  
  return_X_y = FALSE,  
  local_cache_dir = NA,  
  dropna = TRUE  
)
```

**Arguments**

dataset_name	The name of the data set to load from PMLB
return_X_y	Boolean. Whether to return the data with the features and labels stored in separate data structures or a single structure (can be TRUE or FALSE, defaults to FALSE)
local_cache_dir	The directory on your local machine to store the data files in (defaults to NA, indicating cache will not be used)
dropna	Boolean. Whether rows with NAs should be automatically dropped. Default to TRUE.

**See Also**

[summary\\_stats](#).

**Examples**

```
# Features and labels in single data frame
if (interactive()){
  penguins <- fetch_data("penguins")
  head(penguins)

  # Features and labels stored in separate data structures
  penguins <- fetch_data("penguins", return_X_y = TRUE)
  penguins$x # data frame
  penguins$y # vector
}
```

---

get\_type

*Get type/class of given vector.*

---

**Description**

Get type/class of given vector.

**Usage**

```
get_type(x, include_binary = FALSE)
```

**Arguments**

x	Input vector.
include_binary	Boolean. Whether binary should be counted separately from categorical.

**Value**

Type/class of 'x'.

---

graceful_download	<i>Download a File Gracefully with Retry Mechanism</i>
-------------------	--

---

### Description

Attempts to download a file from a specified URL, retrying a set number of times if the download fails. This function meets CRAN's requirement for gracefully handling the use of internet resources by catching errors and returning a warning message if the download ultimately fails.

### Usage

```
graceful_download(url, destfile, retries = 3)
```

### Arguments

url	Character. The URL of the file to download.
destfile	Character. The path to the destination file where the downloaded content will be saved.
retries	Integer. The maximum number of download attempts (default is 3).

### Value

Logical. Returns 'TRUE' if the download succeeds, 'FALSE' otherwise.

### Examples

```
## Not run:
dataset_url <- "https://example.com/dataset.csv"
tmp <- tempfile(fileext = ".csv")
success <- download_file_gracefully(dataset_url, tmp)
if (!success) {
  message("Continuing gracefully without the dataset.")
}

## End(Not run)
```

---

nearest_datasets	<i>Select nearest datasets given input 'x'.</i>
------------------	---

---

### Description

If 'x' is a data.frame object, computes dataset characteristics. If 'x' is a character object specifying dataset name from PMLB, use the already computed dataset statistics/characteristics in 'summary\_stats'.

**Usage**

```
nearest_datasets(x, ...)

## Default S3 method:
nearest_datasets(x, ...)

## S3 method for class 'character'
nearest_datasets(
  x,
  n_neighbors = 5,
  dimensions = c("n_instances", "n_features"),
  target_name = "target",
  ...
)

## S3 method for class 'data.frame'
nearest_datasets(
  x,
  y = NULL,
  n_neighbors = 5,
  dimensions = c("n_instances", "n_features"),
  task = c("classification", "regression"),
  target_name = "target",
  ...
)
```

**Arguments**

<code>x</code>	Character string of dataset name from PMLB, or data.frame of <code>n_samples</code> x <code>n_features</code> (or <code>n_features+1</code> with a target column)
<code>...</code>	Further arguments passed to each method.
<code>n_neighbors</code>	Integer. The number of dataset names to return as neighbors.
<code>dimensions</code>	Character vector specifying dataset characteristics to include in similarity calculation. Dimensions must correspond to numeric columns of [all_summary_stats.tsv](https://github.com/EJCarroll/all_summary_stats.tsv). If 'all' (default), uses all numeric columns.
<code>target_name</code>	Character string specifying column of target/dependent variable.
<code>y</code>	Vector of target column. Required when 'x' does not contain the target column.
<code>task</code>	Character string specifying classification or regression for summary stat generation.

**Value**

Character string of names of most similar datasets to `df`, most similar dataset first.

**Examples**

```
if (interactive()){
```

```
nearest_datasets('penguins')
nearest_datasets(fetch_data('penguins'))
}
```

---

pmlb

*pmlb: R interface to the Penn Machine Learning Benchmarks data repository*

---

## Description

The **PMLB** repository contains a curated collection of data sets for evaluating and comparing machine learning algorithms. These data sets cover a range of applications, and include binary/multi-class classification problems and regression problems, as well as combinations of categorical, ordinal, and continuous features. There are approximately 290 data sets included in the PMLB repository and there are no missing values in these data sets.

## Details

This R library includes summaries of the classification and regression data sets but does NOT include any of the PMLB data sets. The data sets can be downloaded using the `fetch_data` function which is similar to the corresponding PMLB python function.

See `fetch_data`, `summary_stats` for usage examples and further information.

If you use PMLB in a scientific publication, please consider citing the following paper:

Randal S. Olson, William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz, and Jason H. Moore (2017).

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, page 36.

## Author(s)

**Maintainer:** Trang Le <[grixor@gmail.com](mailto:grixor@gmail.com)> (<https://trang.page/>)

Authors:

- [makeyourownmaker <makeyourownmaker@gmx.com>](mailto:makeyourownmaker@gmx.com) (<https://github.com/makeyourownmaker>)
- [Jason Moore <jhmoore@upenn.edu>](mailto:jhmoore@upenn.edu) (<http://www.epistasisblog.org/>)

Other contributors:

- University of Pennsylvania [copyright holder]

## See Also

Useful links:

- <https://github.com/EpistasisLab/pmlbr>
- Report bugs at <https://github.com/EpistasisLab/pmlbr/issues>

regression\_dataset\_names

*Names of available regression datasets*

---

### Description

A list of the names of available regression datasets

### Usage

```
regression_dataset_names
```

### Format

An object of class character of length 122.

### Source

<https://github.com/EpistasisLab/pmlb>

---

summary\_stats

*Summary statistics for the all datasets*

---

### Description

Summary statistics for the all datasets

### Usage

```
summary_stats
```

### Format

A data frame with 10 variables:

**dataset:** Dataset name

**n\_instances:** Number of data observations (equal to number of rows)

**n\_features:** Total number of features (number of columns - 1)

**n\_binary\_features:** Number of binary features

**n\_categorical\_features:** Number of categorical features

**n\_continuous\_features:** Number of continuous features

**n\_classes:** Number of classes in target variable

**endpoint\_type:** Value type of endpoint/target (can be binary, categorical or continuous)

**imbalance:** Imbalance metric, where zero means that the dataset is perfectly balanced and the higher the value, the more imbalanced the dataset

**task:** Type of problem/task. Can be classification or regression.

**Source**

<https://github.com/EpistasisLab/pmlb>

# Index

## \* datasets

- classification\_dataset\_names, 2
- dataset\_names, 3
- regression\_dataset\_names, 8
- summary\_stats, 8

- classification\_dataset\_names, 2
- compute\_imbalance, 2

- dataset\_names, 3

- fetch\_data, 3, 7

- get\_type, 4
- graceful\_download, 5

- nearest\_datasets, 5

- pm1b, 7
- pm1br (pm1b), 7

- regression\_dataset\_names, 8

- summary\_stats, 4, 7, 8