Package 'mlr3data'

October 13, 2022

Title Collection of Machine Learning Data Sets for 'mlr3'

Version 0.6.1

Description A small collection of interesting and educational machine learning data sets which are used as examples in the 'mlr3' book (<https://mlr3book.mlr-org.com>), the use case gallery (<https://mlr3gallery.mlr-org.com>), or in other examples. All data sets are properly preprocessed and ready to be analyzed by most machine learning algorithms. Data sets are automatically added to the dictionary of tasks if 'mlr3' is loaded.

License LGPL-3

URL https://github.com/mlr-org/mlr3data

BugReports https://github.com/mlr-org/mlr3data/issues

Depends R (>= 3.1.0)

Suggests mlr3 (>= 0.13.3)

Encoding UTF-8

LazyData true

NeedsCompilation no

RoxygenNote 7.2.1

Author Michel Lang [cre, aut] (<https://orcid.org/0000-0001-9754-0393>), Marc Becker [ctb] (<https://orcid.org/0000-0002-8115-0400>)

Maintainer Michel Lang <michellang@gmail.com>

Repository CRAN

Date/Publication 2022-08-15 07:30:05 UTC

R topics documented:

`mlr3data-package` *mlr3data: Collection of Machine Learning Data Sets for 'mlr3'*

Description

A small collection of interesting and educational machine learning data sets which are used as examples in the 'mlr3' book (https://mlr3book.mlr-org.com), the use case gallery (https://mlr3gallery.mlr-org.com), or in other examples. All data sets are properly preprocessed and ready to be analyzed by most machine learning algorithms. Data sets are automatically added to the dictionary of tasks if 'mlr3' is loaded.

Author(s)

Maintainer: Michel Lang <michellang@gmail.com> (ORCID)

Other contributors:

- Marc Becker <marcbecker@posteo.de> (ORCID) [contributor]

See Also

Useful links:

- https://github.com/mlr-org/mlr3data
- Report bugs at https://github.com/mlr-org/mlr3data/issues

`bike_sharing` *Bike Sharing Demand*

Description

Regression data to predict the total count of bikes rented. Contains 13 features and 17379 observations. Target column is `"count"`.

Pre-processing

- All columns have been renamed.
- `instant`, `"registered"` and `"casual"` column have been removed.
- `"season"` and `"weather"` have been converted to `factor()`.
- `"holiday"` and `"working_day"` have been converted to `logical()`.

Source

https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

Examples

```
data("bike_sharing", package = "mlr3data")
str(bike_sharing)
```

---

ilpd *Indian Liver Patient Dataset*

Description

Classification data to predict whether or not a person is a liver patient. Obtained using the **mlr3oml** package. Contains 538 observations and 10 features. Target column is `"diseased"`.

Pre-processing

- All variables have been renamed.
- The target variable has been re-encoded to `"yes"` and `"no"`.

Source

https://www.openml.org/d/1480

Examples

```
data("ilpd", package = "mlr3data")
str(ilpd)
```

---

kc_housing *House Sales in King County*

Description

Regression task to predict house sale prices for King County, including Seattle, between May 2014 and May 2015.

Contains 19 features and 21613 observations. Target column is `"price"`.

Pre-processing

- Id column has been removed.
- Dates in column `"date"` have been converted from strings to POSIXct.
- Values `0` in feature `"yr_renovated"` have been replaced with `NA`.
- Values `0` in feature `"sqft_basement"` have been replaced with `NA`.
- Feature `"waterfront"` has been converted to logical.

Source

https://www.kaggle.com/harlfoxem/housesalesprediction

Examples

```
data("kc_housing", package = "mlr3data")
str(kc_housing)
```

`moneyball` *Major League Baseball Statistics 1962-2012*

Description

Regression data to predict the number of runs scored. Obtained using the **mlr3oml** package.

Contains 14 features and 1232 observations. Target column is `"rs"`.

Pre-processing

- All variable names have been converted from upper case to lower case.
- The variables `"year"`, `"rs"`, `"ra"`, `"w"` have been coerced to integers.

Source

https://www.openml.org/d/41021

Examples

```
data("moneyball", package = "mlr3data")
str(moneyball)
```

`optdigits` *Optical Recognition of Handwritten Digits*

Description

Classification data to predict handwritten digits. Obtained using the **mlr3oml** package.

Binarized version of the original data set. The multi-class target column has been converted to a two-class nominal target column by re-labeling the majority class as positive (`"P"`) and all others as negative (`"N"`). Originally converted by Quan Sun.

Contains 64 features and 5620 observations. Target column is `"binaryclass"`.

Pre-processing

- All feature variables `"input1"`, ..., `"input64"` (number of on pixels in each block) have been coerced to integers.
- The target variable has been renamed from `"binaryClass"` to `"binaryclass"`.

Source

https://www.openml.org/d/980

Examples

```
data("optdigits", package = "mlr3data")
str(optdigits)
```

`penguins_simple` *Simplified Palmer Penguins Data Set*

Description

Classification data to predict the species of penguins from the **palmerpenguins** package. A better alternative to the iris data set.

Pre-processing

- The unit of measurement have been removed from the column names. Lengths are given in millimeters (mm), weight in gram (g).
- Observations with missing values have been removed.
- Factor variables are one-hot encoded.

Source

palmerpenguins

References

Gorman KB, Williams TD, Fraser WR (2014). "Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis)." *PLoS ONE*, **9**(3), e90081. doi:10.1371/journal.pone.0090081.

https://github.com/allisonhorst/palmerpenguins

Examples

```
data("penguins_simple", package = "mlr3data")
str(penguins_simple)
```

`titanic` *Titanic*

Description

Classification data to predict the fate of passengers on the ocean liner "Titanic". Contains 10 features and 1309 observations. Target column is `"Survived"`.

Pre-processing

- All column names have been changed to `snake_case`.
- training and test set have been joined. Observations of the test set have a missing value in the target column `"survived"`.
- Column '"survived"' has been re-encoded to a factor with levels '"yes"' and '"no"'.
- Id column has been removed.
- Passenger class `"pclass"` has been converted to an ordered factor.
- Features `"sex"` and `"embarked"` have been converted to factors.
- Empty strings in `"cabin"` and `"embarked"` have been encoded as missing values.

Source

titanic and https://www.kaggle.com/c/titanic/data

Examples

```
data("titanic", package = "mlr3data")
str(titanic)
```

Index