Package ‘mlr3data’

August 3, 2020

Title Collection of Machine Learning Data Sets for 'mlr3'

Version 0.2.0

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. Currently contains the following data sets: (1) housing prices in Kings County, and (2) Titanic passenger survival data. 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 bibtex, covr, mlr3

Encoding UTF-8

LazyData true

NeedsCompilation no

RoxygenNote 7.1.1

Author Michel Lang [cre, aut] (<https://orcid.org/0000-0001-9754-0393>)

Maintainer Michel Lang <michellang@gmail.com>

Repository CRAN

Date/Publication 2020-08-03 16:30:02 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. Currently contains the following data sets: (1) housing prices in Kings County, and (2) Titanic passenger survival data. Data sets are automatically added to the dictionary of tasks if 'mlr3' is loaded.

Author(s)

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

See Also

Useful links:

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

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)
```

`penguins` *Palmer Penguins*

Description

Classification data to predict the species of penguins from the **palmerpenguins** package (CRAN version 0.1.0). Promoted as an alternative to the iris data set.

Pre-processing

- The unit have been removed from the column names. Lengths are given in millimeters (mm), weight in gram (g).

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", package = "mlr3data")
str(penguins)
```

`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