heart-disease-dataset

October 2, 2025

Title heart-disease-dataset

Author Barbara Wieckowska

Version 0.1.0

Description A processed clinical dataset with synthetically generated variables for evaluating binary classification methods.

License MIT

`heart_disease`	*Complete Heart Disease Dataset with Generated Variables for Method Validation*

Description

A processed clinical dataset (from UCI Machine Learning Repository) with synthetically generated variables for evaluating binary classification methods. Combines real cardiac data with controlled random variables to test model robustness. Contains 661 complete cases (347 healthy, 314 with coronary artery disease).

Usage

```
heart_disease
```

Format

A data frame with 661 rows, 56 columns, and the following variables:

disease Coronary artery disease status (factor: 0 = <50% stenosis, 1 = >50% stenosis)

location Data source (factor: 'cl' (Cleveland), 'hu' (Hungarian), 'sw' (Switzerland), 'va' (VA))

age Age in years (numeric)

sex Sex (0 = female, 1 = male)

cp Chest pain type (factor: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)

bp Resting systolic blood pressure (mmHg)

chol Serum cholesterol (mg/dl)

glu Fasting blood sugar >120 mg/dl (1 = yes, 0 = no)

ecg Resting ECG results (factor: 0 = normal, 1 = ST-T abnormality, 2 = LV hypertrophy)

hr Maximum heart rate achieved (bpm)

exang Exercise-induced angina (1 = yes, 0 = no)

stde Exercise-induced ST depression (mm)

rnd_normal Non-stratified random variable: N(0,1)

rnd_uniform Non-stratified random variable: U(0,10)

rnd_exp Non-stratified random variable: Exp(1)

rnd_bernoulli Non-stratified random variable: Bernoulli(0.8)

rnd_binomial Non-stratified random variable: Binomial(6,0.8)

rnd_poisson Non-stratified random variable: Poisson(1)

strat_rnd_normal Stratified random variable: N(10,2) for controls | N(12,2) for cases

strat_rnd_uniform Stratified random variable: U(0,6) | U(2,8)

strat_rnd_exp Stratified random variable: Exp(0.5) | Exp(1)

strat_rnd_bernoulli Stratified random variable: Bern(0.5) | Bern(0.2)

strat_rnd_binomial Stratified random variable: Binom(7,0.6) | Binom(7,0.5)

strat_rnd_poisson Stratified random variable: Pois(1) | Pois(1.6)

hlt_slight_asym Asymmetric stratified variable: N(1,2) for controls | N(0,1) for cases

ill_slight_asym Asymmetric stratified variable: N(0,1) for controls | N(1,2) for cases

hlt_high_asym Asymmetric stratified variable: N(1,4) for controls | N(0,1) for cases

ill_high_asym Asymmetric stratified variable: N(0,1) for controls | N(1,4) for cases

rnd_normal0_1 Age-correlated variable: N(0,1) with r=0.1 to age

rnd_normal0_2 Age-correlated variable: N(0,1) with r=0.2 to age

rnd_normal0_3 Age-correlated variable: N(0,1) with r=0.3 to age

rnd_normal0_4 Age-correlated variable: N(0,1) with r=0.4 to age

rnd_normal0_5 Age-correlated variable: N(0,1) with r=0.5 to age

rnd_normal0_6 Age-correlated variable: N(0,1) with r=0.6 to age

rnd_normal0_7 Age-correlated variable: N(0,1) with r=0.7 to age

rnd_normal0_8 Age-correlated variable: N(0,1) with r=0.8 to age

rnd_normal0_9 Age-correlated variable: N(0,1) with r=0.9 to age

strat_rnd_normal0_1 Stratified age-correlated variable: N(10,2.5)|N(11,2.5) with r=0.1 to age

strat_rnd_normal0_2 Stratified age-correlated variable: N(10,2.5)|N(11,2.5) with r=0.2 to age

strat_rnd_normal0_3 Stratified age-correlated variable: N(10,2.5)|N(11,2.5) with r=0.3 to age

strat_rnd_normal0_4 Stratified age-correlated variable: N(10,2.5)|N(11,2.5) with r=0.4 to age

strat_rnd_normal0_5 Stratified age-correlated variable: N(10,2.5)|N(11,2.5) with r=0.5 to age

strat_rnd_normal0_6 Stratified age-correlated variable: N(10,2.5)|N(11,2.5) with r=0.6 to age

strat_rnd_normal0_7 Stratified age-correlated variable: N(10,2.5)|N(11,2.5) with r=0.7 to age

strat_rnd_normal0_8 Stratified age-correlated variable: N(10,2.5)|N(11,2.5) with r=0.8 to age

strat_rnd_normal0_9 Stratified age-correlated variable: N(10,2.5)|N(11,2.5) with r=0.9 to age

Data Processing

- Combined datasets from 4 sources (Cleveland, Hungarian, Swiss, VA)
- Removed cases with missing values or biologically implausible measurements (BP/chol = 0)
- Generated variables using `faux::rnorm_pre()` for correlated variables
- Categorical variables (`disease`, `cp`, `ecg`) converted to factors with reference levels set.

Source

Clinical data from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

Examples

```
data(heart_disease)
# Check the structure of the dataset
str(heart_disease)
# Compare distributions of a stratified variable
boxplot(strat_rnd_normal ~ disease, data = heart_disease)
```

`heart_disease_test` *Balanced Test Set for the Heart Disease Dataset*

Description

A balanced stratified subset of the complete `heart_disease` dataset, intended for model testing and validation. Contains 330 complete cases with approximately equal distribution of disease cases. This set is complementary to `heart_disease_train` and together they form the complete dataset. All variables, both real and generated, are identical to those described in `heart_disease`.

Usage

```
heart_disease_test
```

Format

A data frame with 330 rows and 56 columns. The format and variables are identical to `heart_disease`.

See Also

heart_disease for the full dataset and detailed variable descriptions.

heart_disease_train for the complementary balanced training set.

Examples

```
data(heart_disease_test)
data(heart_disease_train)
# Train a model on the training set and predict on the test set
model <- glm(disease ~ age + chol, data = heart_disease_train, family = "binomial")
predictions <- predict(model, newdata = heart_disease_test, type = "response")
```

---

heart_disease_test_imbalanced_10

Complementary Test Set for 10% Imbalanced Training

Description

Test set containing the remaining cases after creating the 10% imbalanced training set. Reflects the natural distribution of the original dataset. Intended for validation of models trained on severely imbalanced data.

Usage

```
heart_disease_test_imbalanced_10
```

Format

A data frame with 330 rows and 56 columns. The format and variables are identical to heart_disease.

See Also

heart_disease_train_imbalanced_10 for the corresponding training set.

---

heart_disease_test_imbalanced_30

Complementary Test Set for 30% Imbalanced Training

Description

Test set containing the remaining cases after creating the 30% imbalanced training set. Reflects the natural distribution of the original dataset. Intended for validation of models trained on imbalanced data.

Usage

```
heart_disease_test_imbalanced_30
```

heart_disease_test_reduced_10

Format

A data frame with 330 rows and 56 columns. The format and variables are identical to heart_disease.

See Also

heart_disease_train_imbalanced_30 for the corresponding training set.

---

heart_disease_test_reduced_10

Reduced Test Set with 10% Disease Prevalence

Description

A modified test set with controlled class distribution (10% disease cases). Created by subsampling from the original test set to maintain specific prevalence. Useful for evaluating model performance under low prevalence scenarios.

Usage

```
heart_disease_test_reduced_10
```

Format

A data frame with variable rows (depending on available cases) and 56 columns. The format and variables are identical to heart_disease.

See Also

heart_disease_test_reduced_30 for 30% prevalence version.

---

heart_disease_test_reduced_30

Reduced Test Set with 30% Disease Prevalence

Description

A modified test set with controlled class distribution (30% disease cases). Created by subsampling from the original test set to maintain specific prevalence. Useful for evaluating model performance under specific clinical prevalence scenarios.

Usage

```
heart_disease_test_reduced_30
```

Format

A data frame with variable rows (depending on available cases) and 56 columns. The format and variables are identical to heart_disease.

See Also

heart_disease_test_reduced_10 for 10% prevalence version.

heart_disease_train *Balanced Training Set for the Heart Disease Dataset*

Description

A balanced stratified subset of the complete heart_disease dataset, intended for model training. Contains 331 complete cases with approximately equal distribution of disease cases. All variables, both real and generated, are identical to those described in heart_disease.

Usage

```
heart_disease_train
```

Format

A data frame with 331 rows and 56 columns. The format and variables are identical to heart_disease.

See Also

heart_disease for the full dataset and detailed variable descriptions.

heart_disease_test for the complementary balanced test set.

Examples

```
data(heart_disease_train)
# Train a model on the training set
model <- glm(disease ~ age + chol, data = heart_disease_train, family = "binomial")
summary(model)
```

heart_disease_train_imbalanced_10 *Imbalanced Training Set (10% Disease Cases)*

Description

A training set with severe class imbalance (10% disease cases, 90% healthy controls). Intended for testing classification methods under challenging imbalanced conditions. Contains 331 complete cases (approximately 33 disease, 298 healthy).

Usage

```
heart_disease_train_imbalanced_10
```

Format

A data frame with 331 rows and 56 columns. The format and variables are identical to heart_disease.

See Also

heart_disease_train_imbalanced_30 for moderate imbalance.

heart_disease_test_imbalanced_10 for the corresponding test set.

Examples

```
data(heart_disease_train_imbala
nced_10) # Check severe class
imbalance
prop.table(table(heart_disease_train_imbalanced_10$disease))
```

`heart_disease_train_imbalanced_30`
Imbalanced Training Set (30% Disease Cases)

Description

A training set with artificially induced class imbalance (30% disease cases, 70% healthy controls). Intended for testing classification methods under realistic imbalanced conditions. Contains 331 complete cases (approximately 99 disease, 232 healthy).

Usage

```
heart_disease_train_imbalanced_30
```

Format

A data frame with 331 rows and 56 columns. The format and variables are identical to `heart_disease`.

See Also

`heart_disease_train_imbalanced_10` for more extreme imbalance.

`heart_disease_test_imbalanced_30` for the corresponding test set.

Examples

```
data(heart_disease_train_imbala
nced_30) # Check class
distribution
table(heart_disease_train_imbalanced_30$disease)
```