Package 'EpidigiR'

November 5, 2025

Type Package

Title Digital Epidemiological Analysis and Visualization Tools

Version 0.1.2

Description Integrates methods for epidemiological analysis, modeling, and visualization, including functions for summary statistics, SIR (Susceptible-Infectious-Recovered) modeling, DALY (Disability-Adjusted Life Years) estimation, age standardization, diagnostic test evaluation, NLP (Natural Language Processing) keyword extraction, clinical trial power analysis, survival analysis, SNP (Single Nucleotide Polymorphism) association, and machine learning methods such as logistic regression, k-means clustering, Random Forest, and Support Vector Machine (SVM). Includes datasets for prevalence estimation, SIR modeling, genomic analysis, clinical trials, DALY, diagnostic tests, and survival analysis. Methods are based on Gelman et al. (2013) <doi:10.1201/b16018> and Wickham et al. (2019, ISBN:9781492052040>.

License MIT + file LICENSE

Encoding UTF-8

Language en-US

LazyData true

RoxygenNote 7.3.2

Depends R (>= 4.0.0)

Imports deSolve, sp, tm, glmnet, caret, survival

Suggests kernlab, randomForest, stats, knitr, rmarkdown, quarto, usethis, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

NeedsCompilation no

Author Esther Atsabina Wanjala [aut, cre]

Maintainer Esther Atsabina Wanjala <digitalepidemiologist23@gmail.com>

Repository CRAN

Date/Publication 2025-11-05 20:30:13 UTC

Contents

`clinical_data` *Clinical Trials Data for Epidemiological Analysis*

Description

A dataset containing simulated clinical trial data for analyzing treatment outcomes, suitable for power calculations, logistic regression, Random Forest, and SVM.

Usage

```
clinical_data
```

Format

A data frame with 200 rows and 6 columns:

trial_id Character, unique identifier for each trial participant.

arm Character, treatment arm (e.g., Treatment, Control).

outcome Numeric, binary outcome (0 = no response, 1 = response).

age Numeric, patient age (years).

health_score Numeric, baseline health score (0 to 100).

dose Numeric, treatment dose level (e.g., 0 for control, 1 for low dose, 2 for high dose).

Source

Simulated data for demonstration purposes.

Examples

```
data("clinical_data")
clinical_data$outcome <- as.factor(clinical_data$outcome)
epi_model(clinical_data, formula = outcome ~ age + health_score + dose, type = "logistic")
epi_model(clinical_data, formula = outcome ~ age + health_score + dose, type = "rf")
epi_visualize(clinical_data, x = "arm", y = "outcome", type = "boxplot")
```

---

daly_data *DALY Data for Global Health Burden*

Description

A dataset containing simulated data for calculating Disability-Adjusted Life Years (DALY) in epidemiological studies.

Usage

```
daly_data
```

Format

A data frame with 20 rows and 3 columns:

group Character, population group (e.g., region or age group).

yll Numeric, years of life lost due to premature mortality.

yld Numeric, years lived with disability.

Source

Simulated data for demonstration purposes.

Examples

```
data("daly_data")
epi_analyze(daly_data, outcome = NULL, type = "daly")
```

diagnostic_data *Diagnostic Test Data for Evaluation*

Description

A dataset containing simulated data for evaluating diagnostic tests in epidemiological studies.

Usage

```
diagnostic_data
```

Format

A data frame with 10 rows and 5 columns:

test_id Character, unique identifier for each test.

true_positives Numeric, number of true positive results.

false_positives Numeric, number of false positive results.

true_negatives Numeric, number of true negative results.

false_negatives Numeric, number of false negative results.

Source

Simulated data for demonstration purposes.

Examples

```
data("diagnostic_data")
epi_analyze(diagnostic_data, outcome = NULL, type = "diagnostic")
```

epi_analyze *Performs summary statistics, SIR modeling, DALY calculation, age standardization, diagnostic test evaluation, or NLP keyword extraction.*

Description

Performs summary statistics, SIR modeling, DALY calculation, age standardization, diagnostic test evaluation, or NLP keyword extraction.

Usage

```
epi_analyze(
  data,
  outcome,
  population,
  group = NULL,
  type = c("summary", "sir", "daly", "age_standardize", "diagnostic", "nlp"),
  ...
)
```

Arguments

data	Input data frame with relevant columns (e.g., cases, population, yll, yld, text).
outcome	Outcome column name (character, e.g., "cases").
population	Population column name (character, e.g., "population", required for summary).
group	Grouping column name (character, e.g., "region", optional).
type	Analysis type: "summary", "sir", "daly", "age_standardize", "diagnostic", "nlp".
...	Additional parameters (e.g., N, beta, gamma for SIR).

Value

A data frame with analysis results.

---

epi_model *Unified Epidemiological Modeling*

Description

Performs clinical trial power calculation, survival analysis, SNP association, logistic regression, k-means clustering, Random Forest, or SVM.

Usage

```
epi_model(
  data,
  formula = NULL,
  type = c("power", "survival", "snp", "logistic", "kmeans", "rf", "svmRadial"),
  ...
)
```

Arguments

data	Input data frame with relevant columns (e.g., outcome, genotypes).
formula	Model formula (optional, for survival/logistic/rf/svmRadial, e.g., "outcome ~ x").
type	Model type: "power", "survival", "snp", "logistic", "kmeans", "rf", "svmRadial".
...	Additional parameters (e.g., n, effect_size for power; k for kmeans).

Value

A data frame or list with model results.

`epi_prevalence` *Disease Prevalence Data by Region and Age Group*

Description

A dataset containing disease prevalence data across different regions and age groups, including spatial coordinates.

Usage

```
epi_prevalence
```

Format

A data frame with 12 rows and 7 columns:

region Character, region name (e.g., North, South, East, West).

age_group Character, age group (e.g., 0-19, 20-59, 60+).

cases Numeric, number of disease cases.

population Numeric, population size in the region and age group.

prevalence Numeric, prevalence percentage (cases / population * 100).

lat Numeric, latitude for spatial mapping.

lon Numeric, longitude for spatial mapping.

Source

Simulated data for demonstration purposes.

Examples

```
data("epi_prevalence")
library(sp)
coordinates(epi_prevalence) <- ~lon+lat
epi_visualize(epi_prevalence, x = "prevalence", type = "map")
epi_analyze(epi_prevalence,outcome   = "cases",population = "population",type     = "summary")
if (interactive()) {
epi_prevalence$region_id <- as.numeric(factor(epi_prevalence$region))
epi_visualize(epi_prevalence, x = "region_id", y = "prevalence", type = "scatter")
with(epi_prevalence, axis(1, at = unique(region_id), labels = levels(factor(region))))
}
```

epi_visualize *Flexible Epidemiological Visualization*

Description

Creates visualizations for prevalence mapping, epidemic curves, or general plots (scatter, boxplot).

Usage

```
epi_visualize(
  data,
  x,
  y = NULL,
  type = c("map", "curve", "scatter", "boxplot"),
  ...
)
```

Arguments

data	Input data frame or SpatialPolygonsDataFrame with relevant columns.
x	X-axis column name (character, e.g., "region").
y	Y-axis column name (character, e.g., "prevalence", optional).
type	Plot type: "map", "curve", "scatter", "boxplot".
...	Additional plotting parameters (e.g., main, xlab).

Value

A plot (spplot for maps, base R for others).

geno_data *Genomic SNP-Case Data*

Description

A dataset containing simulated genotypes and case-control status for SNP association analysis.

Usage

```
geno_data
```

Format

A data frame with 100 rows and 2 columns:

genotypes Numeric, genotype (0 = AA, 1 = Aa, 2 = aa).

cases Numeric, case (1) or control (0) status.

Source

Simulated data for demonstration purposes.

Examples

```
data("geno_data")
epi_model(geno_data, type = "snp")
```

`ml_data` *Machine Learning Data for Disease Risk Prediction*

Description

A dataset containing simulated patient data for predicting disease risk, suitable for logistic regression, clustering, Random Forest, and SVM.

Usage

```
ml_data
```

Format

A data frame with 100 rows and 5 columns:

outcome Numeric, binary disease status (0 = healthy, 1 = diseased).

age Numeric, patient age (years).

exposure Numeric, exposure level (0 to 1, e.g., environmental risk).

genetic_risk Numeric, genetic risk score (0 to 1).

region Character, region name (e.g., North, South, East, West).

Source

Simulated data for demonstration purposes.

Examples

```
data("ml_data")
ml_data$outcome <- as.factor(ml_data$outcome)
epi_model(ml_data, formula = outcome ~ age + exposure + genetic_risk, type = "logistic")
epi_model(ml_data, formula = outcome ~ age + exposure + genetic_risk, type = "rf")
epi_visualize(ml_data, x = "age", y = "outcome", type = "scatter")
```

nlp_data *NLP Data for Epidemiological Text Analysis*

Description

A dataset containing simulated epidemiological text data, such as outbreak reports or health alerts, for NLP analysis.

Usage

```
nlp_data
```

Format

A data frame with 100 rows and 2 columns:

id Character, unique identifier for each text entry.

text Character, text content (e.g., outbreak descriptions, health reports).

Source

Simulated data for demonstration purposes.

Examples

```
data("nlp_data")
epi_analyze(nlp_data, outcome = NULL, type = "nlp", n = 5)
```

sir_data *SIR Model Simulation Data*

Description

A dataset containing simulated SIR model outputs for a population of 1000.

Usage

```
sir_data
```

Format

A data frame with 50 rows and 4 columns:

time Numeric, time point (1 to 50 days).

Susceptible Numeric, number of susceptible individuals.

Infected Numeric, number of infected individuals.

Recovered Numeric, number of recovered individuals.

Source

Generated using `epi_analyze(type = "sir", N = 1000, beta = 0.3, gamma = 0.1, days = 50)`.

Examples

```
data("sir_data")
epi_visualize(sir_data, x = "time", y = "Infected", type = "curve")
```

`survey_data` *Survey Data for Age Standardization*

Description

A dataset containing simulated survey data for age standardization in epidemiological studies.

Usage

```
survey_data
```

Format

A data frame with 20 rows and 3 columns:

age_group Character, age group (e.g., 0-19, 20-39, 40-59, 60+).

rates Numeric, disease rates (e.g., cases per 1000).

pop_weights Numeric, population weights for standardization.

Source

Simulated data for demonstration purposes.

Examples

```
data("survey_data")
epi_analyze(survey_data, outcome = NULL, type = "age_standardize")
```

`survival_data` *Survival Analysis Data*

Description

A dataset containing simulated data for survival analysis in epidemiological studies.

Usage

```
survival_data
```

Format

A data frame with 100 rows and 3 columns:

id Character, unique identifier for each individual.

time Numeric, time to event (e.g., years until death or censoring).

status Numeric, event status (0 = censored, 1 = event occurred).

Source

Simulated data for demonstration purposes.

Examples

```
data("survival_data")
epi_model(survival_data, type = "survival")
epi_visualize(survival_data, x = "time", y = "status", type = "scatter")
```

Index