Package ‘CGMissingDataR’

January 20, 2026

Title Missingness Benchmark (MICE Imputation, Random Forest, kNN)

Version 0.0.0.9000

Description Evaluates predictive performance under feature-level missingness in repeated-measures continuous glucose monitoring-like data. The benchmark injects missing values at user-specified rates, imputes incomplete feature matrices using an iterative chained-equations approach inspired by multivariate imputation by chained equations (MICE) (Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis and Philip J. Leaf (2011) <doi:10.1002/mpr.329>), fits Random Forest regression models (Leo Breiman (2001) <doi:10.1023/A:1010933404324>) and k-nearest-neighbor regression models (Zhongheng Zhang (2016) <doi:10.21037/atm.2016.03.37>), and reports mean absolute percentage error (MAPE) and R-squared (R2) across missingness rates.

License GPL (>= 2)

Encoding UTF-8

Roxygen list(markdown = TRUE)

Depends R (>= 4.3)

RoxygenNote 7.3.3

Imports reticulate

Suggests testthat (>= 3.0.0),
spelling,
knitr,
rmarkdown

Config/testthat/edition 3

NeedsCompilation no

Language en-US

URL `https://github.com/saraswatsh/CGMissingDataR`

BugReports `https://github.com/saraswatsh/CGMissingDataR/issues`

LazyData true

Contents

`CGMExampleData` *Example dataset for CGMissingData*

Description

A small synthetic dataset intended for examples and tests of `run_missingness_benchmark()`.

Usage

```
CGMExampleData
```

Format

A data frame with 250 rows and 6 variables:

LBORRES Laboratory Observed Result for Glucose (numeric).

TimeSeries Numeric feature representing time series data.

TimeDifferenceMinutes Time difference in minutes between measurements (numeric).

USUBJID Numeric subject identifier.

SiteID Site identifier (character).

Visit Visit label (character).

Examples

```
data("CGMExampleData")
```

`run_missingness_benchmark` *Run missingness benchmark*

Description

Loads a CSV, splits train/validation, masks feature values at various rates, imputes via an Iterative Imputer (MICE-style), trains Random Forest and kNN regressors, and returns MAPE and R2 per model and mask rate.

This function is a thin R wrapper over the Python implementation shipped in `inst/python/CGMissingData`.

Usage

```
run_missingness_benchmark(
  data_path,
  target_col = "LBORRES",
  feature_cols = c("TimeSeries", "TimeDifferenceMinutes", "USUBJID"),
  mask_rates = c(0.05, 0.1, 0.2, 0.3, 0.4),
  test_size = 0.2,
  random_state = 42,
  imputer_random_state = 42,
  rf_n_estimators = 200,
  knn_k = 5
)
```

Arguments

`data_path` Path to a CSV file.

`target_col` Name of the target column.

`feature_cols` Character vector of feature column names.

`mask_rates` Numeric vector of missingness rates (0-1).

`test_size` Validation split fraction.

`random_state` Random seed for train/val splitting and model seeding.

`imputer_random_state` Random seed for the iterative imputer.

`rf_n_estimators` Number of trees for the random forest.

`knn_k` Number of neighbors for kNN.

Value

A data.frame with columns MaskRate, Model, MAPE, R2.

Author(s)

Shubh Saraswat, Hasin Shahed Shad, and Xiaohua Douglas Zhang

Examples

```
data("CGMExampleData")
tmp <- tempfile(fileext = ".csv")
write.csv(CGMExampleData, tmp, row.names = FALSE)
results <- run_missingness_benchmark(tmp, mask_rates = c(0.05, 0.10))
head(results)
```

Index