R Source Code for assessment

B209510

Link to git hub repository

Loading NHSR Dataset

Loading required libraries and data

```
# Loading required libraries
library(tidyverse)
library(NHSRdatasets)
library(here)
library(knitr)
library(scales)
library(caret)
library(dataMeta)
```

The dataset selected is LOS_model from the NHSRdatasets package and it contains an artificially created patient dataset with age, length of hospital stay and death status for 300 patients across 10 hospitals.

```
#Load the LOS_model data.
data(LOS_model)

LOS_data <- LOS_model
```

Data inspection

An initial observation of the dataset will be carried out to evaluate its characteristic and to observe if there is any missing data.

```
# Initial exploration of dataset

class(LOS_data)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
LOS_data
```

```
## # A tibble: 300 x 5
##       ID Organisation   Age   LOS Death
##    <int> <ord>        <int> <int> <int>
##  1     1 Trust1          55     2     0
##  2     2 Trust2          27     1     0
##  3     3 Trust3          93    12     0
##  4     4 Trust4          45     3     1
##  5     5 Trust5          70    11     0
##  6     6 Trust6          60     7     0
##  7     7 Trust7          25     4     0
##  8     8 Trust8          48     4     0
##  9     9 Trust9          51     7     1
## 10    10 Trust10         81     1     0
## # ... with 290 more rows
```

```
glimpse(LOS_data)
```

```
## Rows: 300
## Columns: 5
## $ ID           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ Organisation <ord> Trust1, Trust2, Trust3, Trust4, Trust5, Trust6, Trust7, T~
```

```
## $ Age          <int> 55, 27, 93, 45, 70, 60, 25, 48, 51, 81, 58, 16, 21, 82, 1~
## $ LOS          <int> 2, 1, 12, 3, 11, 7, 4, 4, 7, 1, 4, 3, 1, 9, 12, 1, 4, 3, ~
## $ Death        <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, ~
```

```
head(LOS_data)
```

```
## # A tibble: 6 x 5
##      ID Organisation   Age   LOS Death
##   <int> <ord>        <int> <int> <int>
## 1     1 Trust1          55     2     0
## 2     2 Trust2          27     1     0
## 3     3 Trust3          93    12     0
## 4     4 Trust4          45     3     1
## 5     5 Trust5          70    11     0
## 6     6 Trust6          60     7     0
```

```
tail(LOS_data,10)
```

```
## # A tibble: 10 x 5
##       ID Organisation   Age   LOS Death
##    <int> <ord>        <int> <int> <int>
##  1   291 Trust1          17     3     0
##  2   292 Trust2          53     3     0
##  3   293 Trust3          81     8     1
##  4   294 Trust4          75    11     0
##  5   295 Trust5          32     4     0
##  6   296 Trust6          32     6     0
##  7   297 Trust7          55     6     1
##  8   298 Trust8          21     3     0
##  9   299 Trust9          54     1     0
## 10   300 Trust10         93    15     0
```

```
nrow(LOS_data)
```

```
## [1] 300
```

```
LOS_data %>%
  map(is.na) %>%
  map(sum)
```

```
## $ID
## [1] 0
##
## $Organisation
## [1] 0
##
## $Age
## [1] 0
##
## $LOS
## [1] 0
##
## $Death
## [1] 0
```

```
summary(LOS_data)
```

```
##        ID          Organisation      Age             LOS
```

```
##  Min.   :  1.00   Trust1 : 30   Min.   : 5.00   Min.   : 1.000
##  1st Qu.: 75.75   Trust2 : 30   1st Qu.:24.00   1st Qu.: 2.000
##  Median :150.50   Trust3 : 30   Median :54.00   Median : 4.000
##  Mean   :150.50   Trust4 : 30   Mean   :50.66   Mean   : 4.937
##  3rd Qu.:225.25   Trust5 : 30   3rd Qu.:75.25   3rd Qu.: 7.000
##  Max.   :300.00   Trust6 : 30   Max.   :95.00   Max.   :18.000
##                   (Other):120
##      Death
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1767
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

```
# vignette("LOS_model")
```

Data preparation

An index will be created to allow identification of the observations that will be used for analyse. A transformation to a factor of categorical data will take place and the dataset will be saved into raw data folder.

```
# Adding index
LOS_data <- rowid_to_column(LOS_data,"index")

# Saving raw data
#write.csv(LOS_data,here("Rawdata","LOS_data.csv"))

# Data manipulation: Death variable will need to be converted to a factor

LOS_data$Death <- as.factor(LOS_data$Death)
class(LOS_data$Death)
```

```
## [1] "factor"
```

```
levels(LOS_data$Death)
```

```
## [1] "0" "1"
```

```
LOS_data <- LOS_data %>%
  mutate(Outcome=fct_collapse(Death,
                              "Discharge"="0",
                              "Death"="1"))

LOS_data
```

```
## # A tibble: 300 x 7
##    index    ID Organisation   Age   LOS Death Outcome
##    <int> <int> <ord>        <int> <int> <fct> <fct>
##  1     1     1 Trust1          55     2 0     Discharge
##  2     2     2 Trust2          27     1 0     Discharge
##  3     3     3 Trust3          93    12 0     Discharge
##  4     4     4 Trust4          45     3 1     Death
##  5     5     5 Trust5          70    11 0     Discharge
##  6     6     6 Trust6          60     7 0     Discharge
```

```
##  7     7     7 Trust7          25     4 0     Discharge
##  8     8     8 Trust8          48     4 0     Discharge
##  9     9     9 Trust9          51     7 1     Death
## 10    10    10 Trust10         81     1 0     Discharge
## # ... with 290 more rows
```

```
class(LOS_data$Outcome)
```

```
## [1] "factor"
```

```
levels(LOS_data$Outcome)
```

```
## [1] "Discharge" "Death"
```

Subsetting the data

The desired variables will be kept in the subseted dataset for posterior analysis. The subseted dataframe will be stored into raw data.

```
# Subsetting dataframe with the variables of interest

subset_LOS <- LOS_data %>%
  select(index,LOS,Age,Outcome)

# Summary statistics of subseted dataframe
summary(subset_LOS)
```

```
##      index             LOS              Age             Outcome
##  Min.   :  1.00   Min.   : 1.000   Min.   : 5.00   Discharge:247
##  1st Qu.: 75.75   1st Qu.: 2.000   1st Qu.:24.00   Death    : 53
##  Median :150.50   Median : 4.000   Median :54.00
##  Mean   :150.50   Mean   : 4.937   Mean   :50.66
##  3rd Qu.:225.25   3rd Qu.: 7.000   3rd Qu.:75.25
##  Max.   :300.00   Max.   :18.000   Max.   :95.00
```

```
# Saving sunbset raw data
#write_csv(subset_LOS, here("RawData", "subsetLOS.csv"))
```

Dividing the working dataset into training and testing sets

A train and test data partition will be performed to allow validation of models created from the input data. The data will be respectively saved into data folder as test and train data and also one observation is going to be saved for assessment marking purposes.

```
subset_LOS
```

```
## # A tibble: 300 x 4
##    index   LOS   Age Outcome
##    <int> <int> <int> <fct>
##  1     1     2    55 Discharge
##  2     2     1    27 Discharge
##  3     3    12    93 Discharge
##  4     4     3    45 Death
##  5     5    11    70 Discharge
##  6     6     7    60 Discharge
##  7     7     4    25 Discharge
##  8     8     4    48 Discharge
##  9     9     7    51 Death
```

```
## 10    10     1    81 Discharge
## # ... with 290 more rows
```

```
nrow(subset_LOS)
```

```
## [1] 300
```

```
prop<-(1-(15/nrow(subset_LOS)))

print(prop)
```

```
## [1] 0.95
```

```
set.seed(333)

trainIndex <- createDataPartition(subset_LOS$index, p = prop,
                                  list = FALSE,
                                  times = 1)
head(trainIndex)
```

```
##      Resample1
## [1,]         1
## [2,]         2
## [3,]         3
## [4,]         4
## [5,]         5
## [6,]         6
```

```
LOStrain <- subset_LOS[ trainIndex,]
nrow(LOStrain)
```

```
## [1] 288
```

```
#write_csv(LOStrain, here("Data", "Los_subset_train.csv"))

LOStest  <- subset_LOS[-trainIndex,]
nrow(LOStest)
```

```
## [1] 12
```

```
LostestMarker <- LOStest[1,]

LostestMarker
```

```
## # A tibble: 1 x 4
##   index   LOS   Age Outcome
##   <int> <int> <int> <fct>
## 1    43     6    24 Death
```

```
#write_csv(LostestMarker, here("Data", "subset_Los_test_marker.csv"))

LosTest  <- LOStest[2:nrow(LOStest),]

#write_csv(LosTest, here("Data", "LosTest.csv"))
```

Data Capture tool

Performed in Python programming language and avaliable in the provided git hub repository

Three variables were selected : Age and LOS are integer numerical variables and a widget intText was used for data capture and the variable Outcome is a character categorical variable and a radiobutton widget was used for data capture. A consent variable was added into the tool as a boolean variable with a checkbox widget and index is inputed manually as integer numerical variable as an user input query.

Data Dictionary

Importing and inspecting collected data

The collected data will be assigned to an object called collected_data and a brief inspection and visualization of this will take place to ensure it was no erroneous data.

```
# Importing collected data
collected_data <- read_csv(here("Rawdata","CollectedDataLOSFinal.csv"))
```

```
## Rows: 11 Columns: 5
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Outcome
## dbl (3): Index, Lenght of stay, Age
## lgl (1): Consent
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Inspecting collected data

head(collected_data)
```

```
## # A tibble: 6 x 5
##   Index `Lenght of stay`   Age Outcome   Consent
##   <dbl>            <dbl> <dbl> <chr>     <lgl>
## 1    56                1    11 Discharge TRUE
## 2    72                6    79 Discharge TRUE
## 3    76                7    66 Discharge TRUE
## 4    85               12    90 Discharge TRUE
## 5   147                3    38 Discharge TRUE
## 6   162                2    17 Discharge TRUE
```

```
glimpse(collected_data)
```

```
## Rows: 11
## Columns: 5
## $ Index            <dbl> 56, 72, 76, 85, 147, 162, 202, 204, 252, 275, 298
## $ `Lenght of stay` <dbl> 1, 6, 7, 12, 3, 2, 6, 2, 4, 6, 3
## $ Age              <dbl> 11, 79, 66, 90, 38, 17, 69, 27, 90, 44, 21
## $ Outcome          <chr> "Discharge", "Discharge", "Discharge", "Discharge", "~
## $ Consent          <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE,~
```

```
variable_description <- c("Index column allows identification of the observation regarding the original
                          "The number of days spend in Hospital", "Age of the individual in years", "Ou
```

```
                          "Consent from the end-user to process and share the data collected with the da
variable_type <- c(0,0,0,1,1)

linker <- build_linker(collected_data,variable_description,variable_type)
print(linker)
```

```
##         var_name
## 1          Index
## 2 Lenght of stay
## 3            Age
## 4        Outcome
## 5        Consent
##
## 1 Index column allows identification of the observation regarding the original LOS dataset in the ra
## 2                                                                    The number of days spend in
## 3                                                                          Age of the individual
## 4                                                                          Outcome of the hospit
## 5               Consent from the end-user to process and share the data collected with the data cap
##   var_type
## 1        0
## 2        0
## 3        0
## 4        1
## 5        1
```

```
data_dictionary <- build_dict(my.data = collected_data,linker = linker)
```

```
## Enter description for variable 'Age' and option '11 to 90':
## Enter description for variable 'Consent' and option 'TRUE':
## Enter description for variable 'Index' and option '56 to 298':
## Enter description for variable 'Lenght of stay' and option '1 to 12':
## Enter description for variable 'Outcome' and option 'Discharge':
## Enter description for variable 'Outcome' and option 'Death':
```

```
glimpse(data_dictionary)
```

```
## Rows: 6
## Columns: 4
## $ `variable name`        <chr> "Age", "Consent", "Index", "Lenght of stay", "O~
## $ `variable description` <chr> "Age of the individual in years", "Consent from~
## $ `variable options`     <chr> "11 to 90", "TRUE", "56 to 298", "1 to 12", "Di~
## $ notes                  <chr> "", "", "", "", "", ""
```

```
write_csv(data_dictionary,here("Rawdata","Collected_dataLOS_datadictionary.csv"))


# Appending data dictionary to collected data

main_string <- "This data is artificially created and fabricates a patient dataset with age, lenght of s

main_string
```

```
## [1] "This data is artificially created and fabricates a patient dataset with age, lenght of stay and
```

```
complete_collectedLOSdata <- incorporate_attr(my.data=collected_data,data.dictionary = data_dictionary,
```

```
attributes(complete_collectedLOSdata)$author[1] <- "First Last Name"

complete_collectedLOSdata
```

```
## # A tibble: 11 x 5
##    Index `Lenght of stay`   Age Outcome   Consent
##  * <dbl>            <dbl> <dbl> <chr>     <lgl>
##  1    56                1    11 Discharge TRUE
##  2    72                6    79 Discharge TRUE
##  3    76                7    66 Discharge TRUE
##  4    85               12    90 Discharge TRUE
##  5   147                3    38 Discharge TRUE
##  6   162                2    17 Discharge TRUE
##  7   202                6    69 Discharge TRUE
##  8   204                2    27 Discharge TRUE
##  9   252                4    90 Death     TRUE
## 10   275                6    44 Discharge TRUE
## 11   298                3    21 Discharge TRUE
```

```
attributes(complete_collectedLOSdata)
```

```
## $row.names
##  [1]  1  2  3  4  5  6  7  8  9 10 11
##
## $names
## [1] "Index"          "Lenght of stay" "Age"            "Outcome"
## [5] "Consent"
##
## $spec
## cols(
##   Index = col_double(),
##   `Lenght of stay` = col_double(),
##   Age = col_double(),
##   Outcome = col_character(),
##   Consent = col_logical()
## )
##
## $problems
## <pointer: 0x55dc8fd35cf0>
##
## $class
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
##
## $main
## [1] "This data is artificially created and fabricates a patient dataset with age, lenght of stay and
##
## $dictionary
##    variable name
## 1            Age
## 2        Consent
## 3          Index
## 4 Lenght of stay
## 5        Outcome
## 6
```

```
##                                                                                 variable des
## 1                                                                    Age of the individual
## 2                Consent from the end-user to process and share the data collected with the data capt
## 3 Index column allows identification of the observation regarding the original LOS dataset in the rau
## 4                                                    The number of days spend in
## 5                                                          Outcome of the hospita
## 6
##   variable options notes
## 1         11 to 90
## 2             TRUE
## 3        56 to 298
## 4          1 to 12
## 5        Discharge
## 6            Death
##
## $last_edit_date
## [1] "2022-06-18 20:42:05 UTC"
##
## $author
## [1] "First Last Name"
```

```
save_it(complete_collectedLOSdata,here("Rawdata","CollectedDataLOSFinal"))
```

Data description and linker

A variable description list with the detailed description of the selected variables will be allocated to an object that will be feed into the build_linker function to create a path to add this information to the dataset using the build_dict function. The data dictionary will be saved into raw data

```
variable_description <- c("Index column allows identification of the observation regarding the original
                          "The number of days spend in Hospital", "Age of the individual in years", "Ou
                          "Consent from the end-user to process and share the data collected with the da
variable_type <- c(0,0,0,1,1)

linker <- build_linker(collected_data,variable_description,variable_type)
print(linker)
```

```
##         var_name
## 1          Index
## 2 Lenght of stay
## 3            Age
## 4        Outcome
## 5        Consent
##
## 1 Index column allows identification of the observation regarding the original LOS dataset in the rau
## 2                                                    The number of days spend in
## 3                                                          Age of the individual
## 4                                                          Outcome of the hospita
## 5                Consent from the end-user to process and share the data collected with the data capt
##   var_type
## 1        0
## 2        0
## 3        0
## 4        1
## 5        1
```

```
data_dictionary <- build_dict(my.data = collected_data,linker = linker)
```

```
## Enter description for variable 'Age' and option '11 to 90':
## Enter description for variable 'Consent' and option 'TRUE':
## Enter description for variable 'Index' and option '56 to 298':
## Enter description for variable 'Lenght of stay' and option '1 to 12':
## Enter description for variable 'Outcome' and option 'Discharge':
## Enter description for variable 'Outcome' and option 'Death':
```

```
glimpse(data_dictionary)
```

```
## Rows: 6
## Columns: 4
## $ `variable name`        <chr> "Age", "Consent", "Index", "Lenght of stay", "O~
## $ `variable description` <chr> "Age of the individual in years", "Consent from~
## $ `variable options`     <chr> "11 to 90", "TRUE", "56 to 298", "1 to 12", "Di~
## $ notes                  <chr> "", "", "", "", "", ""
```

```
write_csv(data_dictionary,here("Rawdata","Collected_dataLOS_datadictionary.csv"))
```

Appending data dictionary to collected data

The data dictionary created will be added to the working dataset using the incorporate_attr function and an author name will be incorporated in the attributes of the dataset. The data will be saved into Rawdata folder.

```
# Appending data dictionary to collected data

main_string <- "This data is artificially created and fabricates a patient dataset with age, lenght of s

main_string
```

```
## [1] "This data is artificially created and fabricates a patient dataset with age, lenght of stay and
```

```
complete_collectedLOSdata <- incorporate_attr(my.data=collected_data,data.dictionary = data_dictionary,m

attributes(complete_collectedLOSdata)$author[1] <- "First Last Name"

complete_collectedLOSdata
```

```
## # A tibble: 11 x 5
##    Index `Lenght of stay`   Age Outcome   Consent
##  * <dbl>            <dbl> <dbl> <chr>     <lgl>
##  1    56                1    11 Discharge TRUE
##  2    72                6    79 Discharge TRUE
##  3    76                7    66 Discharge TRUE
##  4    85               12    90 Discharge TRUE
##  5   147                3    38 Discharge TRUE
##  6   162                2    17 Discharge TRUE
##  7   202                6    69 Discharge TRUE
##  8   204                2    27 Discharge TRUE
##  9   252                4    90 Death     TRUE
## 10   275                6    44 Discharge TRUE
## 11   298                3    21 Discharge TRUE
```

```
attributes(complete_collectedLOSdata)
```

```
## $row.names
##  [1]  1  2  3  4  5  6  7  8  9 10 11
##
## $names
## [1] "Index"          "Lenght of stay" "Age"            "Outcome"
## [5] "Consent"
##
## $spec
## cols(
##   Index = col_double(),
##   `Lenght of stay` = col_double(),
##   Age = col_double(),
##   Outcome = col_character(),
##   Consent = col_logical()
## )
##
## $problems
## <pointer: 0x55dc8fd35cf0>
##
## $class
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
##
## $main
## [1] "This data is artificially created and fabricates a patient dataset with age, lenght of stay and
##
## $dictionary
##    variable name
## 1            Age
## 2        Consent
## 3          Index
## 4 Lenght of stay
## 5        Outcome
## 6
##                                                                                                variable des
## 1                                                                                      Age of the individual
## 2                  Consent from the end-user to process and share the data collected with the data capt
## 3 Index column allows identification of the observation regarding the original LOS dataset in the rau
## 4                                                                                The number of days spend in
## 5                                                                                     Outcome of the hospita
## 6
##   variable options notes
## 1         11 to 90
## 2             TRUE
## 3        56 to 298
## 4          1 to 12
## 5        Discharge
## 6            Death
##
## $last_edit_date
## [1] "2022-06-18 20:42:05 UTC"
##
## $author
## [1] "First Last Name"
```

```
save_it(complete_collectedLOSdata,here("Rawdata","CollectedDataLOSFinal"))
```