About sampling - sample\_stratified: splits train/test and folds preserving the target variable proportion (stratification) per category.

Environment setup.

# Stratified sampling dataset  
  
# installation   
#install.packages("daltoolbox")  
  
# loading DAL  
library(daltoolbox)

Load dataset and view class distribution.

iris <- datasets::iris  
head(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

table(iris$Species)

##   
## setosa versicolor virginica   
## 50 50 50

Stratified train/test split preserving Species proportions.

# Split into train and test  
  
# using stratified sampling  
tt <- train\_test(sample\_stratified("Species"), iris)  
  
# training distribution  
print(table(tt$train$Species))

##   
## setosa versicolor virginica   
## 40 40 40

# test distribution  
print(table(tt$test$Species))

##   
## setosa versicolor virginica   
## 10 10 10

Create stratified folds and check distribution.

# Split the dataset into folds  
  
# using stratified sampling  
# preparing the dataset into four folds  
sample <- sample\_stratified("Species")  
folds <- k\_fold(sample, iris, 4)  
  
# folds distribution  
tbl <- NULL  
for (f in folds) {  
tbl <- rbind(tbl, table(f$Species))  
}  
print(tbl)

## setosa versicolor virginica  
## [1,] 13 13 13  
## [2,] 13 13 13  
## [3,] 12 12 12  
## [4,] 12 12 12

References - Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI. (advocates stratified sampling)