R Package name : SENDQSAR

Please create a copy able, nice , excellent, well constructed, and well explained .rmd document file/code for this function

This package is designed to fetch data from SEND nonclinical data base and perform several task. This packge has several function where are as follows :

The must input is the path of the database---------------------Database contains nonclinical study results for each STUDYID.

1. get\_compile\_data  
    : It fetch the database a create a final data frame
2. get\_bw\_score   
   :it calculate for each animal’s bw zscore
3. get\_livertobw\_zscore  
   : it calculate for each animal’s livertobw zscore
4. get\_lb\_score  
   : it calculate for each animal’s laboratory test (lb) zscore
5. get\_mi-score  
   : it calculate for each animal’s microscopic findings (mi) zscore
6. get\_liver\_om\_lb\_mi\_tox\_score\_list  
   : it combine the score of lb, mi, livertobw for each animal’s laboratory test (lb) zscore and create a combined dataframe where each row represents a STUDYID and columns the calculated zscore.
7. get\_col\_harmonized\_scores\_df  
   : each STUDYID can have different number of mi zscore columns based on the nonclinical findings and also for the same finding there may be different naming. This function do some work harmonize the columns name across the STUDYIDs.
8. get\_ml\_data\_and\_tuned\_hyperparameters  
   : This function transform the column harmonized data frame to format suitable for classification modeling and also do hyperparameter tuning to get optimized parameters for subsequent machine learning operations.
9. get\_rf\_model\_with\_cv  
   : This function performs random classification on the data frame and do cross validation using defined number of repetitions and give a confusion matrix of the results containing sensitivity, specificity, ppv etc.
10. get\_zone\_exclusioned\_rf\_model\_with\_cv  
    : This function is also do the similar work like get\_rf\_model\_with\_cv , except that, in this function, based on the lower and upper threshold used for the predicted value to created a undeterminate zone to get finer results.
11. get\_imp\_features\_from\_rf\_model\_with\_cv  
    : The function calculate the importance of the features using the data and randomForest::importance function.
12. get\_auc\_curve\_with\_rf\_model  
    : Using the built random forest model and the prediction result on the test set, auc curve can be calculated using this function employing ROCR package.
13. get\_histogram\_barplot  
    : using the ggplot, colum harmonized data, a barplot can be generated for the two class of target variables.
14. get\_reprtree\_from\_rf\_model  
    : This function create a random forest tree using the rf model and the reprtree packages for a particular tree.
15. get\_prediction\_plot  
    : This function create histogram plot for the prediction value which is the probability values predicted by the model.
16. get\_Data\_formatted\_for\_ml\_and\_best.m  
    : This is helper function can be used by the other plot. The function generated a data frame from the database in a format suitable for the subsequent machine learning pipeline.
17. get\_rf\_input\_param\_list\_output\_cv\_imp  
    : The function will give the same output as get\_rf\_model\_with\_cv function but key difference is that, get\_rf\_model\_with\_cv is simple function where all the argument variable has to calculate before passing to the get\_rf\_model\_with\_cv function. Basically, function 1 -8 is required to call the get\_rf\_model\_with\_cv function.   
    But , get\_rf\_input\_param\_list\_output\_cv\_imp designed in a way that, if just path of database is provided it will perform all the necessary function call to prepare the data suitable for classification machine learning and built the a random forest model which was random forest and do prediction on the test set and provide a confusion matrix.
18. get\_zone\_exclusioned\_rf\_model\_cv\_imp  
    : This function is pretty similar with the get\_zone\_exclusioned\_rf\_model\_cv\_imp function except that, this function remove some the prdicton based on the defined threshold values essentially creating an undeteminate zone and gives better result.

# SENDQSAR R Package Documentation

# SENDQSAR: QSAR Modeling with SEND Database

## About

This package facilitates developing Quantitative Structure-Activity Relationship (QSAR) models using the SEND database. It streamlines data acquisition, preprocessing, descriptor calculation, and model evaluation, enabling researchers to efficiently explore molecular descriptors and create robust predictive models.

---

## Features

- \*\*Automated Data Processing\*\*: Simplifies data acquisition and preprocessing steps.

- \*\*Comprehensive Analysis\*\*: Provides z-score calculations for various parameters such as body weight, liver-to-body weight ratio, and laboratory tests.

- \*\*Machine Learning Integration\*\*: Supports classification modeling, hyperparameter tuning, and performance evaluation.

- \*\*Visualization Tools\*\*: Includes histograms, bar plots, and AUC curves for better data interpretation.

- \*\*Flexible Customization\*\*: Harmonizes column names across studies and adapts to variable database structures.

---

## Functions Overview

### Data Acquisition and Processing

1. \*\*get\_compile\_data\*\* - Fetches data from the database and creates a final data frame.

2. \*\*get\_bw\_score\*\* - Calculates body weight (BW) z-scores for each animal.

3. \*\*get\_livertobw\_zscore\*\* - Computes liver-to-body weight z-scores.

4. \*\*get\_lb\_score\*\* - Calculates z-scores for laboratory test (LB) results.

5. \*\*get\_mi\_score\*\* - Computes z-scores for microscopic findings (MI).

6. \*\*get\_liver\_om\_lb\_mi\_tox\_score\_list\*\* - Combines z-scores of LB, MI, and liver-to-BW into a single data frame.

7. \*\*get\_col\_harmonized\_scores\_df\*\* - Harmonizes column names across studies.

### Machine Learning Preparation and Modeling

8. \*\*get\_ml\_data\_and\_tuned\_hyperparameters\*\* - Prepares data and tunes hyperparameters for machine learning.

9. \*\*get\_rf\_model\_with\_cv\*\* - Builds a random forest model with cross-validation and outputs performance metrics.

10. \*\*get\_zone\_exclusioned\_rf\_model\_with\_cv\*\* - Introduces an indeterminate zone for improved classification accuracy.

11. \*\*get\_imp\_features\_from\_rf\_model\_with\_cv\*\* - Computes feature importance for model interpretation.

12. \*\*get\_auc\_curve\_with\_rf\_model\*\* - Generates AUC curves to evaluate model performance.

### Visualization and Reporting

13. \*\*get\_histogram\_barplot\*\* - Creates bar plots for target variable classes.

14. \*\*get\_reprtree\_from\_rf\_model\*\* - Builds representative decision trees for interpretability.

15. \*\*get\_prediction\_plot\*\* - Visualizes prediction probabilities with histograms.

### Automated Pipelines

16. \*\*get\_Data\_formatted\_for\_ml\_and\_best.m\*\* - Formats data for machine learning pipelines.

17. \*\*get\_rf\_input\_param\_list\_output\_cv\_imp\*\* - Automates preprocessing, modeling, and evaluation in one step.

18. \*\*get\_zone\_exclusioned\_rf\_model\_cv\_imp\*\* - Similar to the above function, but excludes uncertain predictions based on thresholds.

---

## Workflow

1. \*\*Input Database Path\*\*: Provide the database path containing nonclinical study results for each `STUDYID`.

2. \*\*Preprocessing\*\*: Use functions 1-8 to clean, harmonize, and prepare data.

3. \*\*Model Building\*\*: Employ machine learning functions (9-18) for training, validation, and evaluation.

4. \*\*Visualization\*\*: Generate plots and performance metrics for better interpretation.

---

## Dependencies

- `randomForest`

- `ROCR`

- `ggplot2`

- `reprtree`

---

## Installation

```R

# Install from CRAN (if available)

install.packages("SENDQSAR")

# Install from GitHub

devtools::install\_github("<github-username>/SENDQSAR")

```

---

## Examples

### Example 1: Basic Data Compilation

```R

library(SENDQSAR)

data <- get\_compile\_data("/path/to/database")

```

### Example 2: Z-Score Calculation

```R

bw\_scores <- get\_bw\_score(data)

liver\_scores <- get\_livertobw\_zscore(data)

```

### Example 3: Machine Learning Model

```R

model <- get\_rf\_model\_with\_cv(data, n\_repeats=10)

print(model$confusion\_matrix)

```

### Example 4: Visualization

```R

get\_histogram\_barplot(data, target\_col="target\_variable")

```

---

## Contribution

Contributions are welcome! Feel free to submit issues or pull requests via GitHub.

---

## License

This project is licensed under the MIT License - see the LICENSE file for details.

---

## Contact

For more information, visit the project [GitHub Page](https://<github-username>.github.io/SENDQSAR/) or contact [email@example.com](mailto:email@example.com).