**Homework\_2: Modelling with the doubs data using the CARET package**

This assignment bases on the dataset used in Homework\_01. The goal is to learn how to training a model with the caret package. It consists of two major sections. The first one is to preProcess data. The second involves training a tree-based model to answer the question raised by the dataset.

**Submission**

1. Export your R script as a Homework\_02.R file

2. Submit your homework through your own github

3. Due Date: April 5, 2023

Section I: Read data and data pre-Process – preliminary. This section covers:

- reading env and fish data, and summarizing fish abundance data by sites, and combining env and total fish to a new data frame named as "env\_fish".

代码

- visualizing the features of the new env\_fish set using scatterplot(). Do you find any linear relationships between environmental variables and the total fish abundance at the sites?

代码

- having the sites with no fishes? If yes, deleting such sites. Having null values or outliers? If yes, removing all rows where any column contains an outlier.

代码

- identifying near zero-variance, outlies of the env variables. If yes, excluding them for analysis.

代码

- detecting the collinearity among env variables or removing highly correlated features (with an absolute correlation of 0.75 or higher)

代码

In short, the env variables should be scaled and centered before training a model.

Section II: Building a regression model. This section covers:

- splitting data into training and test sets, and visualizing the features and targets of the training set

代码

- Creating and evaluating a baseline model between the environmental variables and the total fish abundance with the tree-based algorithm

代码

Feature selection: depending on what algorithm learn between X and Y, i.e., the variable prove useful in a tree-based algorithm, but may be less helpful in a regression-based model.

In general, you are encouraged to experiment with any machine learning algorithms to try to make them as accurate as you can.