Advanced techniques with R

Statement

Datasets

The compressed file contains several simulated datasets. These files contain input variables or predictors named X1, X2, etc., and an output variable named Y.

Practice:

Load the different datasets into R. For each dataset:

- Identify if the dataset corresponds to a regression or classification problem. In the latter, identify if it is a binary or multiclass problem. Convert the output variable to factors if needed.
- Use ggplot2 for plotting a 2D scatterplot of the data. If Y is categorical, make sure to change the color or the shape of the data depending on the values of Y.
- Can you find the solution of the problem? Can you express the solution mathematically?
- Does the data need preprocessing? Check the following options:
    - Outliers.
    - Missing values. If any, how many are there?

Practice with ggplot2 and the dataset SimData7.csv

- Plot the histograms of the input and output variables.
- Divide the histogram plot into subplots based on the values of the output variable.
- Apply ggpairs plot dividing the results based on the values of the output variable.