Project

The exercise is meant to assess your ability to apply data mining to predictive develop models to interpret data and enable decision-making.

Data is provided in a file named “xy1.xlsx” and your task is to develop predictive models that can be used for determining the response for future predictor data points.

The dataset has a column named “Response”, which is the dependent variable. The other 18 variables are the predictors (i.e., independent variables).

Split the data into two subsets: training set (70%) and test set (30%).

Perform the following tasks.

1. Perform a multicollinearity test to reduce the number of independent variables to no more than 5. You can use the R “vif” function from the library “car”.
2. Factorize the “Response” data column into six categories. Use the *ctree* function or otherwise to determine a classification model, first as a decision tree and then convert to *if-then* rules. The rules should assign future instances of predictors to a class (or response). Test the accuracy of the rules developed (using the training set) on the test set.
3. Determine the predictors that are significant in predicting the response. Develop a linear regression model between the response and the regressors (using the training set). Test the accuracy using the test set.

Provide your results in a single R Markdown file (and knit to HTML). Briefly document each step.