Guided Activity on Applying Random Forest in R

Start Assignment

- Due Dec 10 by 11:59pm
- Points 100
- Submitting a file upload
- File Types pdf
- Available Nov 17 at 12am - Dec 20 at 11:59pm

Objective: Build, evaluate, and optimize a Random Forest model to classify income levels using the ***Adult* dataset from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/2/adult)** .

Important: For guidance on how to approach the data analysis, refer to Chapter 8 of our textbook, with particular attention to Section 8.3.

Data Preparation

1. Download the *Adult* dataset from the UCI repository.
2. Import the dataset into R.
3. Check for missing values and impute them using median/mode imputation.
4. Clean and preprocess the dataset (e.g., remove leading/trailing spaces, convert categorical variables to factors, etc.).

Model Building

1. Split the dataset into training and test sets (e.g., 70%-30% split).
2. Train a Random Forest model using packages such as `randomForest` or `MLR`. Start by using default parameters to understand how the functions work.
3. Evaluate the model on the test set and compute metrics such as accuracy, precision, and recall.

Optimization

1. Experiment with different values for function arguments such as `mtry`, `ntree`, and `nodesize`.
2. Use cross-validation to find the optimal hyperparameters.
3. Re-train the model using the optimal hyperparameters and compare the performance to the initial model.

Analysis

1. Generate a variable importance plot and interpret the top predictors.
2. Compare the results of the Random Forest model to those of a simple decision tree.

Report

1. Summarize your findings: Which parameters had the greatest impact on the model's performance?
2. Discuss the advantages and limitations of using the Random Forest model for this dataset.