Assignment – Classification and resampling methods

In this problem, you will develop machine learning models to predict car acceptability based on multiple features. The used dataset "DataAssign2.csv" is uploaded on Blackboard. The different features with their respective attribute values are described below:

- Column 1: the buying price of the car (v-high, high, med, low)
- Column 2: price of the maintenance (v-high, high, med, low)
- Column 3: number of doors (2, 3, 4, 5-more)
- Column 4: capacity in terms of persons to carry (2, 4, more)
- Column 5: the size of luggage boot (small, med, big)
- Column 6: estimated safety of the car (low, med, high)

- Last column: the **response variable**, the acceptability of the car (bad, good)

This is a **teamwork of 2-3 students max** and a **single R Notebook per team** (containing the scripts, the graphs, the analyses, the interpretation of your findings and finally a conclusion) must be uploaded on Blackboard. Your script must be very well commented and do not forget to mention your names at the beginning of the notebook.

You are highly encouraged to explore and use libraries that we didn't cover in class.

Deadline: November 16, 2023 @ 11:59 PM

a) Explore the data graphically in order to investigate the association between the response and the other features. Which of the features seem most likely to be useful in predicting the response. Describe your findings.

b) Split the data into a training set and a test set using the validation set approach (Do not forget to set a random seed before beginning your analysis).

c) Perform logistic regression on the training data in order to predict the response using the variables that seemed most associated with it. What is the test error of the model obtained? Do any of the predictors appear to be statistically significant?

   Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

d) Repeat the same for LDA and QDA. Compare the performance of logistic regression, LDA and QDA by drawing their ROC curves and computing their AUC values.

e) Perform KNN on the training data, with several values of K, in order to predict the response by using only the variables that seemed most associated with it. What test errors do you obtain? Which value of K seems to perform the best on this data set? You can present your results graphically.

Last, perform a 5-fold cross validation to estimate the test error and a bootstrap to compute the estimated standard errors for one of the methods above. Compare the obtained results of the 5-fold cross validation to the validation set approach.

Plagiarism and cheating policy

Cheating is a serious offense. If a student is found to have copied part or all of the assignment, he or she will receive a zero on the assignment. NO EXCUSES WILL BE ACCEPTED. The same applies to the person providing others with material.

Deadline policy

Deadlines for submitting assignments will be announced in class or on the assignments. They must be respected. Students failing to meet the deadline will get 20% of the grade deducted per day late.