Homework 3

Due on 04/01/2025

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the dataset “auto.csv”. The dataset contains 392 observations. The response variable is “mpg_cat”, which indicates whether the miles per gallon of a car is high or low. The predictors include both continuous and categorical variables:

- cylinders: Number of cylinders between 4 and 8
- displacement: Engine displacement (cu. inches)
- horsepower: Engine horsepower
- weight: Vehicle weight (lbs.)
- acceleration: Time to accelerate from 0 to 60 mph (sec.)
- year: Model year (modulo 100)
- origin: Origin of car (1. American, 2. European, 3. Japanese)

Split the dataset into two parts: training data (70%) and test data (30%).

(a) Perform logistic regression analysis. Are there redundant predictors in your model? If so, identify them. If there are none, please provide an explanation.

(b) Train a multivariate adaptive regression spline (MARS) model. Does the MARS model improve prediction performance compared to logistic regression?

(c) Perform linear discriminant analysis using the training data. Plot the linear discriminant(s).

(d) Which model will you choose to predict the response variable? Plot its ROC curve and report the AUC. Next, select a probability threshold to classify observations and compute the confusion matrix. Briefly interpret what the confusion matrix indicates about your model's performance.