Tae Coding
Introduction to Data Science: CS61
Summer 2018
Homework#8

Date Given: July 5, 2018 Due Date:

===

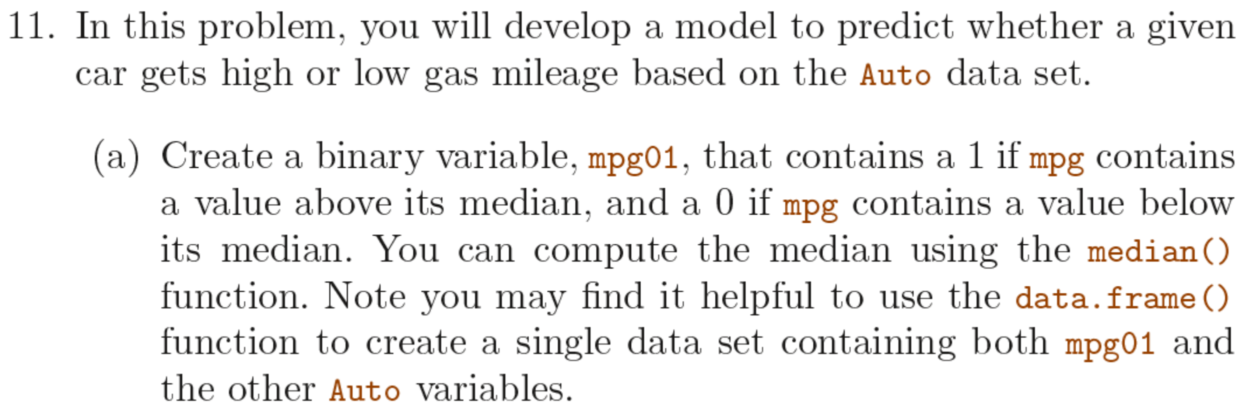
Please use Python's Scikit-Learn package to solve this problem.

Text Book: "An Introduction to Statistical Learning" (ISLR).
By James, Witten, Hastie, Tibshirani
Chapter 4: Classification: Page 171/172, Problem#11.

There is no need to buy this text book. I have copied the problems from the PDF version of this book.

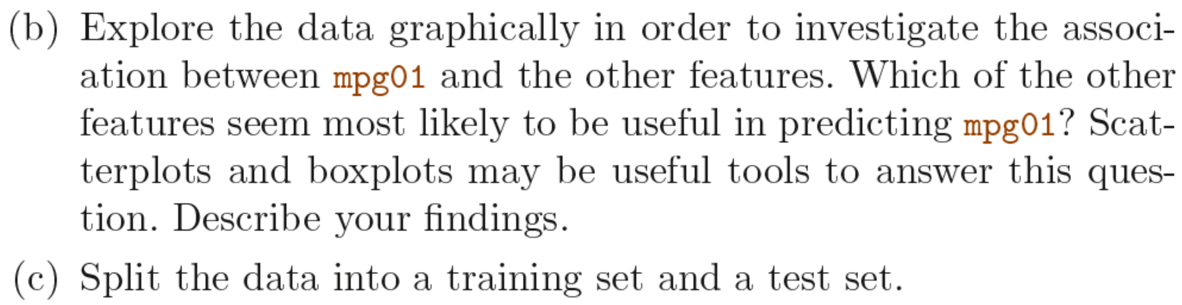
==

Problem#1



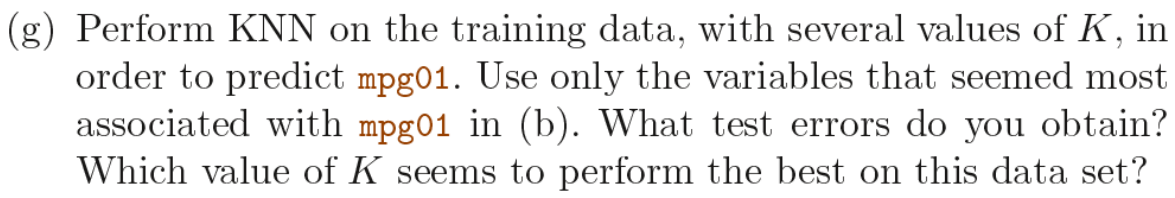
11. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

    (a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.



(b) Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

(c) Split the data into a training set and a test set.



(g) Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

Compute the Confusion Matrix and accuracy for both training and testing dataset.
Plot the ROC Curves.