CIS8695

**Homework II**

**What to Turn In:**

Please write your answers and paste the required results/reports into **an MS Word file**. **Note: Please do NOT submit any other file format because it will cause grading inconvenience. Failing to submit a correct file format will cause the loss of homework grades! You can use screenshots if it is convenient for you.**

# Problem 1. Naïve Bayes Classifier

Look at Problem 2 in Chapter 8 of the Textbook. (A picture of the problem is attached below.) Complete sections (a) and (c)

A document with text and images

Description automatically generated with medium confidence

# Problem 2: PCA

Use the dataset **spambase.csv** (in iCollege). Use PCA to reduce the number of predictors.

1. Using PCA, how many PCs do you think are appropriate to be used for prediction? What is the total proportion of variance they capture?

2. Using the PCs you generate and the outcome variable “spam,” follow the four below steps. Report the model outputs (confusion metrics) and compare the overall accuracy of all the individual and ensemble models.

1. Create a data frame with the PCs you generate and the “spam” column. Partition the data with 60% training and 40% validation.
2. Fit models to the data for (1) logistic regression, (2) Naïve Bayes, and (3) random forest. Use “spam” as the outcome variable. Report the validation confusion matrix for each of the three models.
3. Compare the accuracy of the three individual models and conclude which model is the best.

# Problem 3. Clustering

Look at Problem 4 in Chapter 15 of the Textbook. (A picture of the problem is attached below.) Complete sections (a), (b), (e), and (f)

