Problems Inclass 8_2. You can comment in this document and submit a pdf of your work. Please mark clearly all your answers and answer problems in the order provided.

1. Think through and answer the following problems to the best of your abilities.

   a) Valentine Day is approaching. A restaurant is trying to decide if to organize a singles' night or if to offer a special romantic menu. The restaurant has an established base of customers and collects demographic, income, social media and behavioral information on its customers. They decide to use the help of a data scientist to make sense of their Valentine's day menu in order to maximize sales (Valentine's days tend to be cash cows for restaurants). What algorithm would you use?

      a. I would use a decision tree algorithm because it would make it easier to visualize the restaurants customer base and decide what algorithm to use to determine what kind of menu should be offered on Valentine's Day.

   b) Describe the type of information you would collect (what features) to decide if an email is spam or non-spam and what machine learning algorithm you would use.

      a. I would collect information such as the title of the email, as well as the content of that email. Spam email tends to contain key phrases in the title and body that indicate it is spam. I would use the Naïve Bayes algorithm because it is suitable for text classification, which is what this is.

   c) Describe the type of information you would collect (what features) and from what sources to decide if to buy or sell a stock (financial investment). What machine learning algorithm can you use?

      a. I would collect stock value and stock growth to determine if a stock should be bought or sold. I would use a classification machine learning algorithm.

   d) How would you use Facebook to recommend certain products to people and what machine learning algorithm would you use?

      a. I would use the pages that people "like" to determine what products they would be most likely to be interested in purchasing. I would use a clustering machine learning algorithm for this.

2. A classification algorithm classifies emails into spam and non-spams. The following confusion matrix was returned by using the classifier on the testing set:

264	14
22	158

Consider “non-spam” = “positive” class. The matrix has the organization described in class. Calculate and interpret the following:

1) Accuracy rate
    a. 422/458 = .92
2) Precision
    a. 264/(264+22) = .92
3) Recall
    a. 264/(278) = .95
4) F1
    a. 2*(.92)*(.95/.92+.95) = 2*.92*.51 = 93.5%
5) Sensitivity
    a. 264/(278) = .95
6) Specificity
    a. 158/180 = .88
7) In your opinion, is it more important to have good recall or precision?
    a. I think it is more important to have good recall because recall is the percentage of items correctly predicted by the classifier, which in my opinion is more important than the percentage the classifier assigns to the positive class.