The City of Somerville sends out a happiness survey to a random sample of Somerville residents asking them to rate their personal happiness and their satisfaction with City services every second year since 2011. Every year, the survey is refined. We will use data of year 2015. The dataset includes six features (X1 – X6), each of which take value 1 to 5, and one decision attribute (D) with values 0 (unhappy) and 1 (happy).

Variable Information:
D = decision attribute (D) with values 0 (unhappy) and 1 (happy)

Attributes X1 to X6 have values 1 to 5.
X1 = the availability of information about the city services
X2 = the cost of housing
X3 = the overall quality of public schools
X4 = your trust in the local police
X5 = the maintenance of streets and sidewalks
X6 = the availability of social community events

Randomly select 100 observations for training and the remaining observation will serve as test data. Conduct the following classification analysis and then identify the best classifier based on area under the curve (AUC)

a. Logistic regression (12 points)
b. Classification Tree (12 points)
c. Linear Discriminant Analysis (10 points)
d. Quadratic Discriminant Analysis (10 points)

Compile a table to compare the AUC for these approaches and write a report which method you would recommend based on your analysis.

Classification Method	Area Under Curve
Logistic Regression	0.6645022
Classification Tree	0.6277056
Linear Discriminant Analysis	0.6590909
Quadratic Discriminant Analysis	0.5508658

Out of the four classification methods, the classification rule created by logistic regression gave the largest area under the ROC curve, 0.6645, so would have the least misclassification errors.

Perform the classification analysis using the whole dataset (i.e. without splitting as training and test) by implementing the approach you recommended in Question 2 with default threshold or majority vote (i.e. probability of 0.5 as threshold). Then

a. Construct a confusion matrix. A confusion matrix is a table describing the performance of a classification model (or "classifier"), i.e. a 2x2 Table in this example. (8 points)
b. What's misclassification rate and accuracy? (8 points)
c. What's your sensitivity and specificity? (8 points)
   What's your positive predicted value and negative predicted value?

Part A

Misclassification Error Table		Predicted Class		Total True/False
		0	**1**	
True Class	**0**	44 true negative	22 false positive	66
	1	27 false negative	50 true positive	77
Predicted True/False		71	72	143

Part B

accuracy $= \frac{44+50}{143} = 0.6573$

misclassification rate $= \frac{22+27}{143} = 0.3427$

Part C

false positive $= \frac{22}{66} = 0.3333$

false negative $= \frac{27}{77} = 0.3506$

sensitivity $= \frac{50}{77} = 0.6494$

specificity $= \frac{44}{66} = 0.6667$

Part D

positive predictive value (PPV) $= \frac{50}{72} = 0.6944$

negative predictive value (NPV) $= \frac{44}{71} = 0.6197$