Question 1 (3 points):

Classifier	Briefly describe how a model is built (Enter “N/A” if the classifier does not build a model)	Briefly describe how the model is applied to a new data instance	“Ideal” Input Feature Type (discrete or continuous)
Naïve Bayes	Calculate prior probabilities and likelihood values.	Calculate posterior probabilities using prior probabilities and likelihood	Discrete
Support Vector Machine	Find hyperplane that maximizes the margin between classes	Determine where the new instance lies based on the hyperplane(s)	Continuous
Nearest Neighbor	N/A (Doesn’t really have a model, just a set of labeled data points)	Find k nearest neighbors and classify via class count	Continuous
Decision Trees	Find ‘purest’ set of nodes that represent splits in variables	Traverse the tree based on the new instance’s variables to reach a decision	<answer>

Question 2 (1 point): You have built a Naïve Bayes classifier model and it produces the following confusion matrix for a test set with 1000 data instances. Do you consider this model’s performance to be acceptable? Why or why not?

Actual class		**Predicted class**	
		Class 1	Class 2
	Class 1	850	0
	Class 2	150	0

I do not consider this model’s performance to be acceptable. Although this model produces a relatively acceptable precision and a perfect recall, the model seems to predict every single instance as Class 1. The acceptable precision is only a result of the dataset containing a high amount of Class 1 instances.

Question 3 (2 points): You have built a Decision Tree model with a max depth of 15. The following two confusion matrices have been generated using the model. The first confusion matrix denotes model performance using the training set, while the second confusion matrix is the performance using the test set. Why do you think the model has produced these results?

Actu		**Predicted class**		
		High	Medium	Low

al cla ss	High	750	5	10
	Medium	7	550	12
	Low	9	8	350

Ac tu al cla ss		**Predicted class**		
		High	Medium	Low
	High	140	150	175
	Medium	87	45	76
	Low	104	101	99

The model has produced these results because it has been overfit. It has high accuracy for classification within the training data, but cannot accurately predict new data (the testing set).

Question 4 (1 point): The age of patients in a medical data set range from 18 years old to 75 years old. There are 1000 patients in the data set. Describe how you would discretize the “Age” feature into 3 separate categories, such that there is a relatively even distribution of patients across the 3 categories.

I would sort the data by age, and then find the values of the 33rd and 66th percentiles. After that, I would find appropriate labels for those ages below the 33rd percentile, between the 33rd and 66th percentiles, and above the 66th percentile.