CS795: Topics in Data Mining and Security
Sample Exam I

1. **[Points 35]** Using data in **Table 1**, answer the following.
   a. Using each attribute (1R), form rudimentary rules to predict the decision.
   b. Construct a decision tree (with all relevant attributes) to model the decision process. Due to short time, do the work only for the 1st level of the tree. Clearly show work to justify your choice for the root attribute.
2. **[Points 30]**
   (a) Given the data in **Table 2**, calculate: (i) % of True positives (ii) % of True Negatives (iii) % of False positives (iv) % of False negatives (v) True positive rate (vi) False positive rate. Each percentage (i-iv) is computed with respect to the total observations.
   (b) Suppose we have used XYZ.com's search engine and PQR.com's search engine to search for the key word "data mining." XYZ.com produced 10,000 entries out of which 8,000 were found to be relevant. PQR.com produced 15,000 entries out of which 10,000 were found to be relevant. Compute the recall, precision, and F-measure for both the search engines for this search.
   (c) A marketing firm wants to do a door-to-door survey. It costs the company $10 for each person contacted. They have information on 10 customers (**Table 3**). Create a lift chart (e.g., Fig. 5.1). Describe the chart as a table with %Sample size as x-value and the number of sales as y-value.

3. **[Points 35]**
   (a) The decision tree in **Figure 1** currently has a depth of 3 with the maximum path having three internal nodes. Prune the tree (using subtree replacement) so the decision point is at the root (attribute A1). Draw the confusion matrix prior to pruning and after pruning. Show your work.

   (b) Using the data in Table 1, derive the rules with exceptions. Draw graph similar to Figure 6.7 (in the textbook).

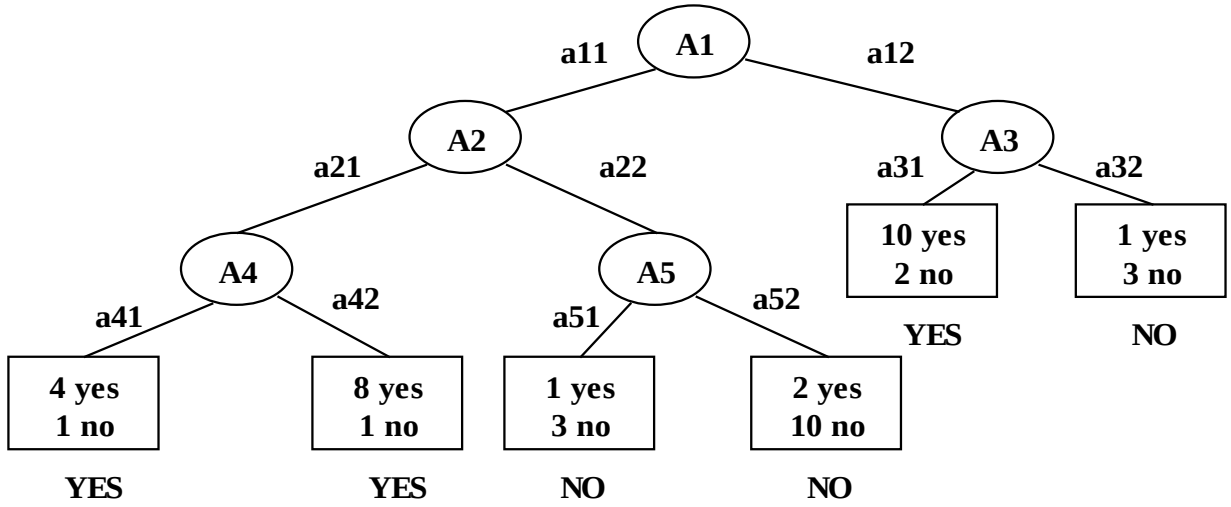

Figure 1. Decision tree to be pruned (A1-A5 are the attributes; a11-a52 are values for the respective attributes; yes/no in the leaf nodes represent the number of training instances that reach that leaf node; YES/NO represent the final decision for test instances that reach that node.)

Instance#	GPA	GRE	Experience (Years)	Decision
1	3.8	1450	0	PhD Admit
2	3.5	1300	0	MS with TA
3	2.9	1250	1	MS Admit
4	2.8	1300	2	MS Admit
5	2.95	1200	0	Reject
6	2.99	1240	1	Reject
7	3.0	1200	0	MS Admit
8	3.4	1300	1	MS Admit
9	3.2	1400	0	MS Admit
10	3.5	1200	1	MS Admit

Table 1. Past data of admission decisions (PhD Admit, MS with TA, MA Admit, Reject are possible decisions)

		Predicted class	
		Yes	**No**
Actual Class	**Yes**	50	10
	No	10	30

Table 2. Data for the marketing firm

Customer Name	Height	Age	Prob of Yes response	Actual Response
Alan	70	39	61	N
Bob	72	21	79	Y
Jessica	65	25	75	Y
Elizabeth	62	30	70	N
Hilary	67	19	81	N
Fred	69	48	52	Y
Alex	65	12	88	Y
Margot	63	51	49	N
Sean	71	65	35	Y
Chris	73	42	58	N

Table 3. Data for the marketing firm