Sl. No	Questions
1	Explain why unary ratings are significantly different from other types of ratings in the design of recommender systems.
2	Discuss cases in which content-based recommendations will not perform as well as ratings-based collaborative filtering.
3	Suppose you set up a system, where a guided visual interface is used in order to determine the product of interest to a customer. What category of recommender system does this case fall into?
4	Discuss a scenario in which location plays an important role in the recommendation process.
5	The chapter mentions the fact that collaborative filtering can be viewed as a generalization of the classification problem. Discuss a simple method to generalize classification algorithms to collaborative filtering. Explain why it is difficult to use such methods in the context of sparse ratings matrices.
6	Suppose that you had a recommender system that could predict raw ratings. How would you use it to design a top-*k* recommender system? Discuss the computational complexity of such a system in terms of the number of applications of the base prediction algorithm. Under what circumstances would such an approach become impractical?
7	Use CART trees shown in Figures .1 and 2. to answer Questions (a) and (b). Figure .1 is a tree that has been developed to predict the success of a movie (Y = 1) using the predictors budget and YouTube likes (YouTube-L). In Figure 2. the predators are actor category A (Actor_A) and Movie Content Masala.

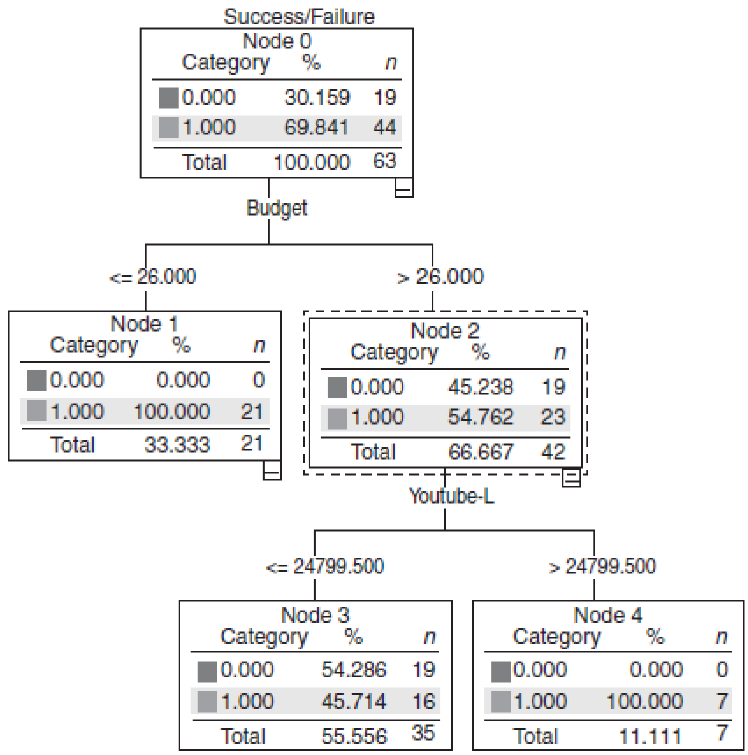

Figure 1: Classification and regression tree with budget and YouTube likes as predictors.

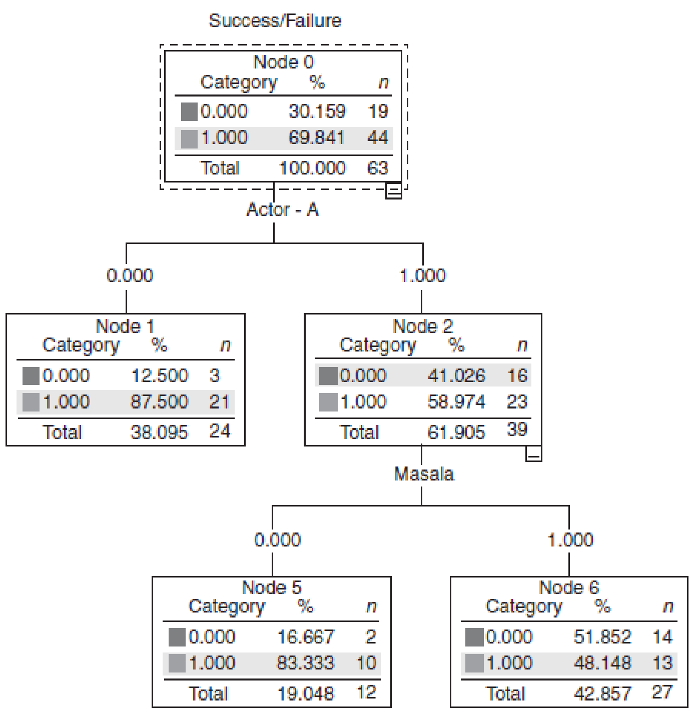

Figure 2: Classification and regression tree with actor and genre as predictors.

(a) Calculate the total reduction in Gini index in the model (that is, between root node and all the leaf nodes).
(b) Calculate the value of entropy at node 0.
(c) Calculate the sensitivity and specificity of the CART model shown in Figure.2
(d) Write all the business rules that can be used for predicting earnings manipulator using the CART tree in Figure .2.

8 The CHAID tree for the prediction of earnings manipulator (Y = 1) is shown in Figure .3. Calculate the value of chi-square statistic and the corresponding p-value for the split used in the tree.

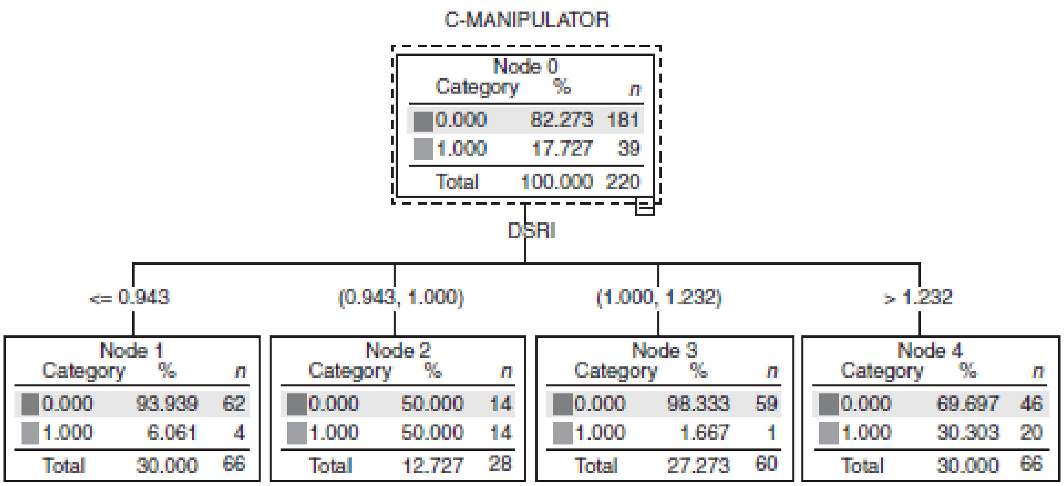

FIGURE .3 CHAID tree for earnings manipulations.

9 CHAID tree is developed to classify a tumour as benign (Y = 0) or malignant (Y = 1).3 CHAID tree in Figure.4 is developed using the variable, Perimeter-W (perimeter of the tumour).
(a) Calculate the p-value of chi-square test of independence used to create the internal nodes.
(b) What can you infer from the CHAID tree in Figure .4?

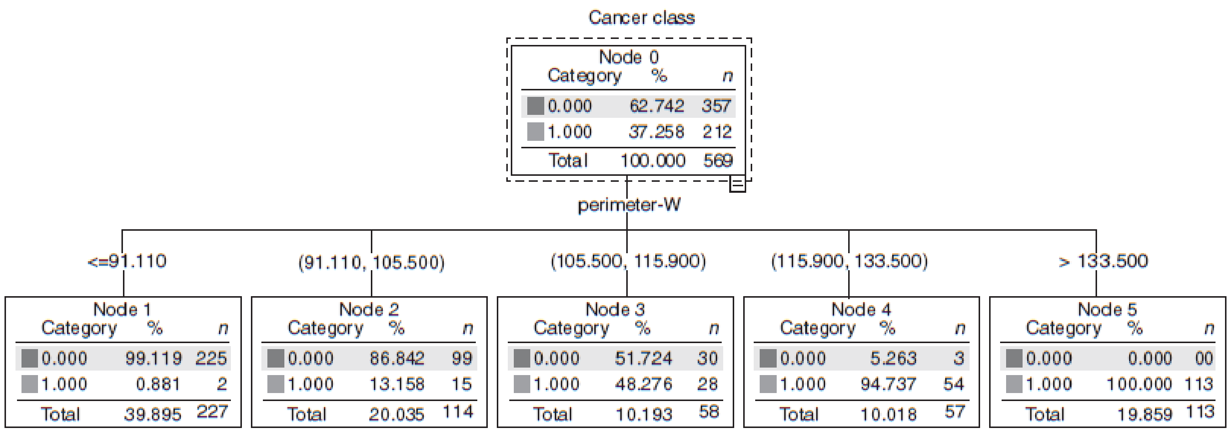

FIGURE .4 CHAID tree for cancer prediction

10 A CART tree for the prediction of cancer is developed using variables:
(a) Radius-W,
(b) Concave Points-W,
(c) Texture,
(d) Texture-W, and
(e) Concave points and is shown in Figure.5. Calculate the sensitivity and specificity of the CART Model
Compare and Comment on nodes 6 and 10.

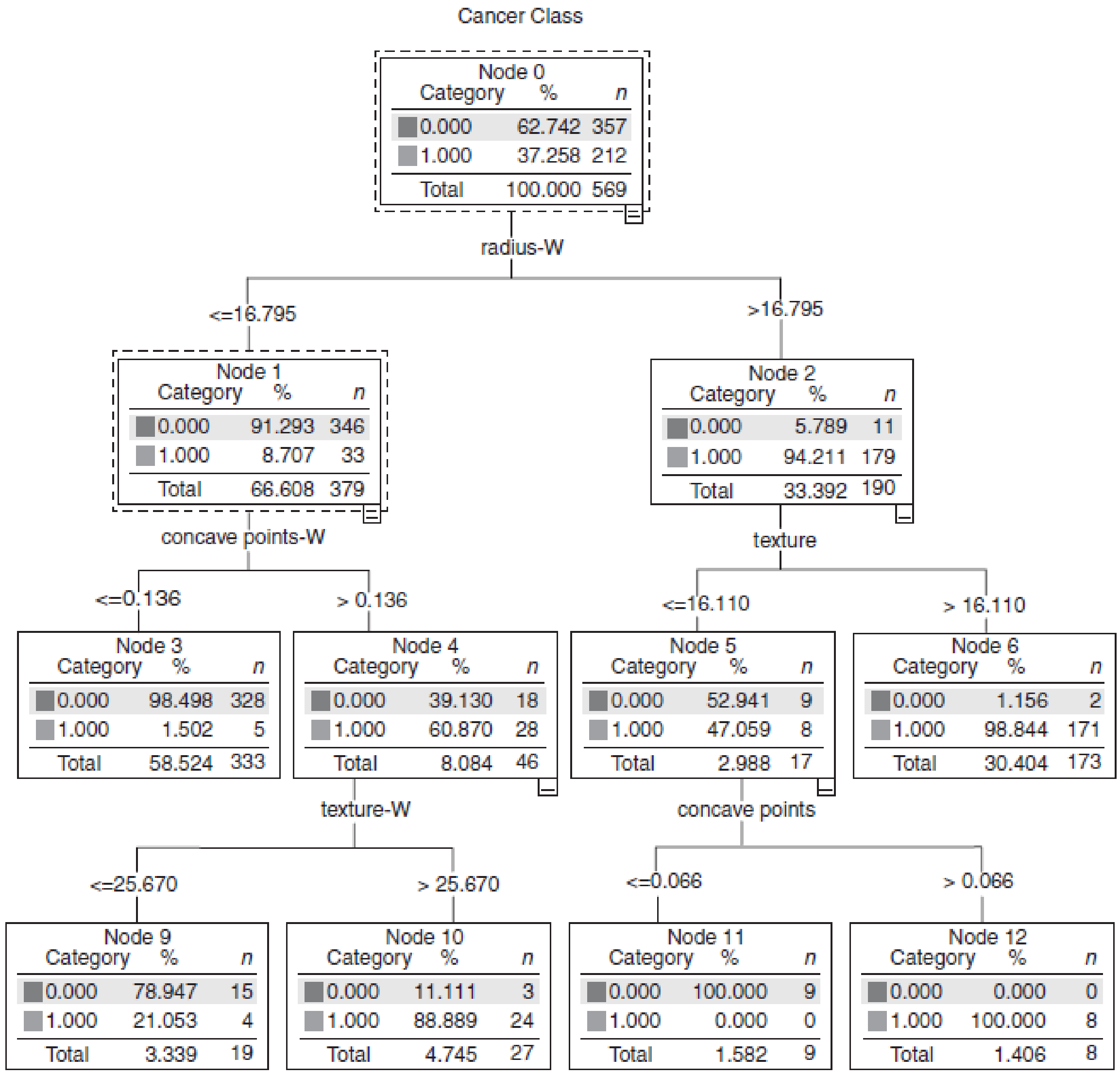

FIGURE .5 CART tree for cancer prediction.

11	The movie ratings given by 4 customers (C1, C2, C3, and C4) on five movies (A, B, C, D and E) are given in Table .1.
12	Use cosine similarity to find among customers C1, C2, and C3, who is the closest to customer C4. 2. An online store sells products under 8 categories labelled: A, B, …, H. The past purchase details of 7 customers are given in Table 14.18.
13	Suppose you have the set C of all frequent closed itemsets on a data set *D*, as well as the support count for each frequent closed itemset. Describe an algorithm to determine whether a given itemset *X* is frequent or not, and the support of *X* if it

	is frequent.
14	An itemset *X* is called a *generator* on a data set *D* if there does not exist a proper sub-itemset *Y* _ *X* such that *support*.*X*/ D *support*.*Y*/. A generator *X* is a *frequent generator* if *support*.*X*/ passes the minimum support threshold. Let G be the set of all frequent generators on a data set *D* (a) Can you determine whether an itemset *A* is frequent and the support of *A*, if it is frequent, using only G and the support counts of all frequent generators? If yes, present your algorithm. Otherwise, what other information is needed? Can you give an algorithm assuming the information needed is available? (b) What is the relationship between closed itemsets and generators?
15	The Apriori algorithm makes use of *prior knowledge* of subset support properties. (a) Prove that all nonempty subsets of a frequent itemset must also be frequent. (b) Prove that the support of any nonempty subset *s*0 of itemset *s* must be at least as great as the support of *s*. (c) Given frequent itemset *l* and subset *s* of *l*, prove that the confidence of the rule “*s*0).*l* □*s*0/” cannot be more than the confidence of “*s*).*l* □*s*/,” where *s*0 is a subset of *s*. (d) A *partitioning* variation of Apriori subdivides the transactions of a database *D* into *n* nonoverlapping partitions. Prove that any itemset that is frequent in *D* must be frequent in at least one partition of *D*.
16	Let *c* be a candidate itemset in *Ck* generated by the Apriori algorithm. How many length-.*k* -1/ subsets do we need to check in the prune step? Per your previous answer, can you give an improved version of procedure has infrequent subset in Figure 6.4?
17	Propose a more efficient method for *generating association rules* from frequent itemsets. Explain why it is more efficient than the one proposed there.
18	A database has five transactions. Let min sup D 60% and min conf D 80%. (a) Find all frequent itemsets using Apriori. Compare the efficiency of the two mining processes. (b) List all the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and itemi denotes variables representing items (e.g., “A,” “B,”): $\forall x \in transaction,\ buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)$

<table>
<tr><td></td><td>
<table>
<tr><th>TID</th><th>items_bought</th></tr>
<tr><td>T100</td><td>{M, O, N, K, E, Y}</td></tr>
<tr><td>T200</td><td>{D, O, N, K, E, Y }</td></tr>
<tr><td>T300</td><td>{M, A, K, E}</td></tr>
<tr><td>T400</td><td>{M, U, C, K, Y}</td></tr>
<tr><td>T500</td><td>{C, O, O, K, I, E}</td></tr>
</table>
</td></tr>
<tr><td>19</td><td>Discuss cases in which content-based recommendations will not perform as well as ratings-based collaborative filtering.</td></tr>
<tr><td>20</td><td>Propose an algorithm that uses random walks on a user-user graph to perform neighborhood-based collaborative filtering. [This question requires a background in
ranking methods.]</td></tr>
<tr><td>21</td><td>Discuss various ways in which graph clustering algorithms can be used to perform neighborhood-based collaborative filtering.</td></tr>
<tr><td>22</td><td>Implement the user-based and item-based collaborative filtering algorithms.</td></tr>
</table>