HW 2

Question 1:

Draw the full decision tree for the parity function of four Boolean attributes, A, B, C, and D. Is it possible to simplify the tree?

Question 2:

Consider the training examples show in the below table for a binary classification problem

(a). Compute the Gini index for the overall collection of training examples.

(b). Compute the Gini index for the Customer ID attribute.

(c). Compute the Gini index for the Gender attribute.

Question 3: Consider the following dataset for a binary class problem.

(a). Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

A picture containing text, device

Description automatically generated

Question 4:

C4.5 rules is an implementation of an indirect method for generating rules from a decision tree. RIPPER is an implementation of a direct method for generating rules directly from data.

(a). Discuss the strengths and weaknesses of both methods.

(b). Consider a dataset that has a large difference in the class size (i.e., some classes are much bigger than others). Which method (between C4.5 rules and RIPPER) is better in terms of finding high accuracy rules for the small classes?