CS 6350/DS 4350: HW1 – Q2 – Decision Trees on the University (Scholarship) Dataset

Samir Abdelrahman

Yujin Song, Shubham Sanjay Sawant

Fall 2025

Grading Rubric

Question	Points	Scoring rules	Graded Points
1-(a)	8	Total Boolean functions – 2 pts Reasoning for total functions – 2 pts Consistent functions – 2 pts Reasoning for consistent functions – 2 pts	
1-(b)	6	Final entropy result – 2 pts Reasoning process – 4 pts	
1-(c)	10	Every feature result – 1 pt (total 4 pts) Reasoning process of GPA/Extracurricular – 4 pts Reasoning process of Recommendation/Year – 2 pts	
1-(d)	4	The root result – 2 pts Reasoning process – 2 pts	
1-(e)	12	A textual description or a simple diagram – 6 pts Reasoning process – 6 pts	
1-(f)	2	Result – 1 pt Reasoning process – 1 pt	
1-(g)	8	Every feature result (0.5 pts) and reasoning process (1 pt) – 1.5 pts each (total 6 pts) The root result – 1 pt Compare with entropy and Gini – 1 pt	
Total:	**50**		

Problem 1

A university wants to decide whether to offer a scholarship to an applicant. The decision depends on the following four features, analogous to classic decision-tree exercises (e.g., the weather example):

- GPA (Low, Medium, High)

- Extracurricular (None, Some, Strong)

- Recommendation (Weak, Strong)

- Year (Junior, Senior)

You are given **10 past applications** (Table 1). For each application, the four feature values and the final scholarship decision are recorded.

Table 1: Training data for the scholarship prediction problem.

GPA	Extracurricular	Recommendation	Year	Scholarship
High	Strong	Strong	Senior	+
Low	None	Weak	Junior	-
Medium	Some	Strong	Senior	+
Low	Strong	Weak	Junior	-
High	Some	Weak	Senior	+
Medium	None	Weak	Senior	-
High	None	Strong	Junior	+
Medium	Strong	Strong	Junior	+
Low	Some	Strong	Senior	-
Medium	Strong	Weak	Senior	+

(a) [8 pts] How many Boolean functions are possible over the input space? How many are consistent with the given dataset (10 labeled rows)? Show your reasoning.

(b) [6 pts] Compute the entropy of the labels in Table 1. Show your counts and the formula.

(c) [10 pts] Compute the **information gain (using entropy)** for each feature (GPA, Extracurricular, Recommendation, Year). Report the results in Table 2.

Table 2: Information gain (entropy) for each feature.

Feature	Information Gain
GPA	
Extracurricular	
Recommendation	
Year	

(d) [4 pts] Based on Table 2, which attribute would **ID3** choose as the root? Justify briefly.

(e) [12 pts] Construct a decision tree using your chosen root that correctly classifies the training data. A textual description or a simple diagram is acceptable.

(f) [2 pts] Test your tree on the **3 new applications** in Table 3 and report the accuracy.

Table 3: Test data for the scholarship prediction problem.

GPA	Extracurricular	Recommendation	Year	Scholarship
High	None	Weak	Senior	+
Low	Strong	Strong	Junior	-
Medium	Some	Weak	Senior	-

(g) [8 pts] **Gini impurity as a split criterion(within the 50 points).**
$Gini(Z) = 1 - \sum p_i^2$, where p_i is the fraction of class i in set Z. Define

$$Gain_{\text{Gini}}(A) = Gini(Y) - \sum_v P(A = v) \cdot Gini(Y \mid A = v).$$

Compute $Gain_{\text{Gini}}$ for each feature and state which attribute Gini would choose as the root. Do entropy and Gini agree?