NAME - Himanshu.

Roll no. - 23/4017

Exam Rollno. - 23019582065.

SEM - 5th.

YEAR - 3rd.

Course - B.Sc. Computer Science.

(a) Gini index for the overall collection.

→ The Gini index is:

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2$$

- Count of C0 : 10 (IDs 1-10).
- Count of C1 : 10 (IDs 11-20).

$$P(C0) = \frac{10}{20} = 0.5, \quad P(C1) = 0.5$$

$$Gini(D) = 1-(0.5^2+0.5^2) = 1-(0.25+0.25) = 0.5$$ Ans.

(b) Gini index for Customer ID

→ Each customer ID is unique (20 distinct values). That means each split contains 1 record only, so the purity of each split = 0 (because only one class per. record). → Gini (Customer ID) = 0.

→ weighted average Gini = 0.0. Ans..

(c) Gini index for Gender.

- Males (IDs 1-6, 11-14) → 10 customers.
- Class C0: 6 (IDs 1-6).
- Class C1: 4 (IDs 11-14).
- Gini(M) = 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 1-(0.36+0.16) = 0.48].
- Females (IDs 7-10, 15-20) → 10 customers.
- class C0: 4 (IDs 7-10).
- Class C1: 6 (IDs 15-20).
- Gini (F) = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 1 - (0.16 + 0.36) = 0.48].

- Weighted average.

$$\text{Gini (Gender)} = \frac{10}{20}(0.48) + \frac{10}{20}(0.48) = 0.48 \ \underline{\text{Ans}}$$

d.) Gini index for Car type (multiway split).

values: Family, Sports, Luxury.

- family (IDs 1, 11, 12, 13) → 4 customers.
  - class C0: 1 (ID 1).
  - Class C1: 3 (IDs 11-13).
  - Gini (family) = 1-\left (\frac{1}{4}\right)^2 = 1-(0.0625 + (0.5625) = 0.375].

- Sports (IDs 2-9) → 8 customers.
  - class C0: 8 (all).
  - class C1: 0
  - Gini (Sports) = 0]

- Luxury (IDs 10, 14-20) → 8 customers.
  - Class C0: 1 (ID 10).
  - Class C1: 7 (IDs 14-20).
  - Gini (Luxury) = 1 - \left(\frac{1}{8}\right)^2 = 1 - (0.0156 + 0.7656) = 0.2188.

weighted average:

$$\text{Gini (Car Type)} = \frac{4}{20}(0.375) + \frac{8}{20}(0) + \frac{8}{20}(0.2188).$$

$$= 0.075 + 0 + 0.0875 = 0.1625.$$

{Gini (Car Type) = 0.163 Approx.}.

(c) Gini index for Shirt Size (multiway split).

Values: Small, Medium, Large, Extra Large.

- Small (IDs 1, 7, 8, 15, 16) → 5 customers.
  - Class C0: 3 (IDs 1, 7, 8).
  - Class C1: 2 (IDs 15, 16)
  - Gini (Small) = $1 - (3/5)^2 - (2/5)^2 = 1 - (0.36 + 0.16) = 0.48$].
- Medium (IDs 2, 3, 9, 13, 17, 18, 19) → 7 customers.
  - Class C0: 3 (IDs 2, 3, 9).
  - Class C1: 4 (IDs 13, 17, 18, 19).
  - Gini (Medium) = $1 - (3/7)^2 - (4/7)^2 = 1 - (0.184 + 0.327)$. $= 0.4898$].
- Large (IDs 4, 10, 11, 20) → 4 customers.
  - Class C0: 2 (IDs 4, 10).
  - Class C1: 2 (IDs 11, 20).
  - Gini (Large) = $1 - (0.5^2 + 0.5^2) = 0.5$]
- Extra Large. (IDs 5, 6, 12, 14) → 4 customers.
  - Class C0: 2 (IDs 5, 6).
  - Class C1: 2 (IDs 12, 14).
  - Gini (Extra Large) = 0.5].

→ Weighted average:

$$\text{Gini (Shirt Size)} = \frac{5}{20}(0.48) + \frac{7}{20}(0.4898) + \frac{4}{20}(0.5).$$

$$+ \frac{4}{20}(0.5) = 0.12 + 0.1714 + 0.1 + 0.1 = 0.4914.$$

→ Gini (Shirt Size) = 0.491 Ans..

f). which attribute is better ?

- Customer ID → 0.0 (but useless, see part g).
- Gender → 0.48.
- Car type → 0.163 (best split).
- Shirt Size → 0.491.

→ Car type is the better attribute.

g). why not customer ID ?

Even though customer ID gives the lowest Gini (0.0), it overfits because:

- Each ID is unique (no generalization).
- Spilitting on customer ID memorizes training data without learning patterns.
- A decision tree using ID wont classify new customers correctly.

→ Customer ID should not be used since it has no. predictive power.