Assignment - II
Data Mining & Warehouse

— Abhishek Roy
Roll - 58 Department of CSE (3rd yr)

MAR 2, 2023

Q.

department	status	age	salary	count
Sales	Senior	31 — 35	46K — 50K	30
Sales	Junior	26 — 30	26K — 30K	40
Systems	Junior	31 — 35	31K — 35K	40
Systems	Junior	21 — 25	46K — 50K	20
Systems	Senior	31 — 35	66K — 70K	5
Systems	Junior	26 — 30	46K — 50K	3
Systems	Senior	41 — 45	66K — 70K	3
marketing	Senior	36 — 40	46K — 50K	10
marketing	Junior	31 — 35	41K — 45K	4
Secretary	Senior	46 — 50	36K — 40K	4
Secretary	Junior	26 — 30	26K — 30K	6

a] How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple?

Ans → The basic decision tree algo should be modified as follows to take into consideration the count of each generalized data tuple —

- The count of each tuple must be integrated into the calculation of the attribute selection measure (such as information gain)
- Take the count into consideration to determine the most common class among the tuples.

b] Use your algo to construct a decision tree from given data.

Ans First we calculate gini (Gini)-Index for entire dataset

Total - count = 150
Sales - count = 110
System - count = 28
marketing - count = 14
secretary - count = 10

→ Gini-Total = [1 − ((sales-count/total-count)×2 + (system-count/total-count)×2 + (marketing-count/total-count)×2 + (secretary-count/total-count))]

Gini-total = [1 − (0.7332 + 0.1872 + 0.0932 + 0.0672)]

Gini-Index for each = **0.612**

→ now calculate attribute →

- Department
  → sales: senior = 0
  junior = $[1 - ((\frac{40}{70}) \times 2 + (\frac{30}{70}) \times 2] =$ **0.489**

→ System : Senior = $[1-((\frac{5}{8})\times 2+(\frac{3}{8})\times 2)]$ = **0.469**

Junior = $[1-((\frac{23}{28})\times 2+(\frac{5}{28})\times 2)]$ = **0.408**

→ Marketing : senior = 0 Junior = 0

→ secretary : senior = **0.375** Junior = **0.5**

→ Status : Senior → Sales = 0.469, Systems = 0.375, Marketing = 0.5, Secretary = 0

Junior → Sales = 0.489, Systems = 0.408, Marketing = 0, Secretary = 0.5

→ Age : (21–25) = 0

(26–30) → Sales = 0.489, Systems = 0.408, Marketing = 0, Secretary = 0.5

(31–35) → Sales = 0.489, Systems = 0.469, Marketing = 0, Secretary = 0

(36–40) = 0

(41–45) = 0

(46–50) → Secretary = 0.375

→ Salary : (26K–30K) → Sales = 0.489, Systems = 0, Marketing = 0, Secretary = 0.5

(31K–35K) → Sales = 0.489, System = 0.469, Marketing = 0, Secretary = 0

(36K–40K) → Secretary = 0.375

(41K–45K) → Marketing = 0

(46K–50K) → Sales = 0.489, System = 0.469, Marketing = 0, Secretary = 0

→ Attribute of lowest Gini-Index is Department with value = **0.373**

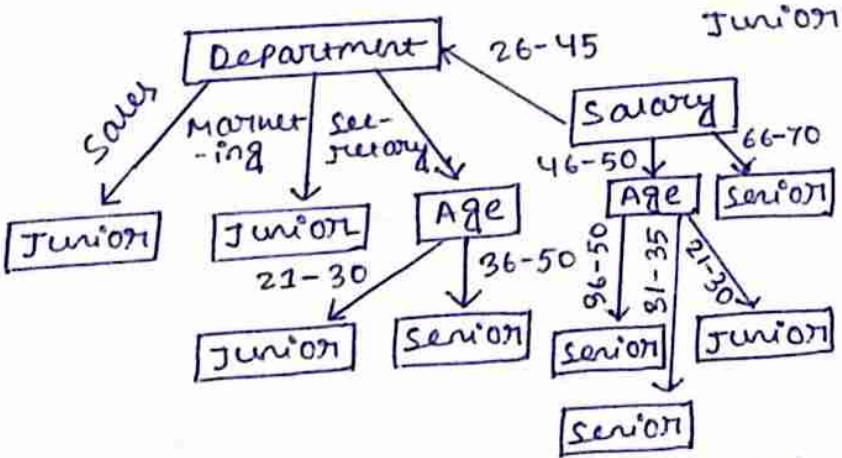
Split dataset with based on department attribute → Sales : status →

Senior = 0

Junior = $\left[1-\left[(\frac{40}{70})\times 2+(\frac{30}{70})\times 2\right]\right]$

$= [1-(1.14+0.85)]$

$= 0.997$



← — — — [Decision-tree]

C] Given a data-tuple having the values "systems", "26-30" & "46-50K" for the attributes department / age / salary respectively what would be the naive Bayesian classification?

→ Given a data tuples with the values — "System" "Junior" "26---30" for the attribute department status / age respectively what would a naive Bayesian classification for salary tuple be →

$P(X|Senior) = 0$

$P(X|Junior) = 0.018$

– Thus a naive Bayesian classification predicts Junior (Ans)