24/12/18

UNIT-III

Classification

Table:- vertebrate Dataset.

Name	BodyTemperature	SkinCover	GivesBirth	HasLegs	ClassLabel
Human	warm-blooded	hair	yes	yes	Mamal
Python	cold-blooded	scales	No	No	Reptiles
Salmon	cold-blooded	scales	yes	No	fish
Whale	warm-blooded	hair	No	No	Mammal
bat	warm-blooded	hair	No	yes	Mammal
Pegion	warm-blooded	feathers	No	yes	Birds
cat	warm-blooded	fur	yes	yes	Mammal
turtle	warm-blooded	scales	No	yes	Reptile.

Classification:-

* Given a collection of record (training set)
* Each record is characterized by a tuple (x, y), where x is the attribute set and y is class label.

  x: attribute, predictor, independent variable, input:

  y: class, response, dependent variable, output.
* The above table shows a sample data set use for classifying vertebrate into one of the following categories
  - Mammal
  - Reptile
  - fish
  - Bird
* The attribute set includes properties of vertebrate such as Body temperature, skin cover.

* The class label must be a discrete attribute

* classification is the task of learning a target function 'f' that maps each attribute set 'x' to one of the pre-defined class labels 'y'.

* The target function is also known as classification model, this is used for the following purposes.

1. Descriptive modeling:-

→ A classification model can helps as an explanatory tool to differentiates b/w objects of different class.

2. predictive modeling:-

* A classification model can also be used to predict the class label of unknown records.

* Suppose we are given the following characteristics of a creature known as gila monster.

Name	Body temp	Skin cover	gives birth	Has legs	class label
Gila monster	cold	Scales	No	Yes	?

* By using the data set shown in the table we determine the class to which the creature belongs.

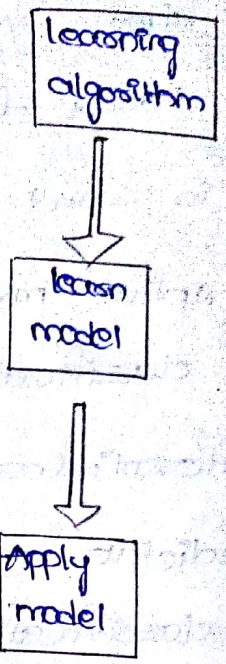
Task	Attribute set, x	class label, y
categorizing email messages	features extracted from email message header and content	spam or non-spam.
Identifying tumor cells.	Features extracted from MRI scans	Malignant or benign cells.
cataloging galaxies	Features extracted from telescope images	Elliptical, spiral or irregular-shaped galaxies.

www.Jntufastupdates.com

General approach to solve a classification problem:-

Tid	attribute 1	attribute 2	attribute 3	class
1.	No	Large	125K	Yes
2.	Yes	Medium	100K	No
3.	Yes	Small	150K	No
4.	No	Large	160K	Yes
5.	Yes	Small	250K	No

Training set.



Tid	attr-1	attr-2	attr-3	class
6	yes	large	100K	?
7	No	small	125K	?
8	No	medium	109K	?
9	yes	large	150K	?
10	yes	medium	120K	?

Test set

* A classification technique is a systematic approach to built classification models from an input data set.
* Examples include decision tree classifiers, rule based classifiers, neural networks, support vector machines, Naive bayes
* Each technic uses a learning algorithm to identify a model that best fits the relationship b/w the attribute set and class label of input data.
* The above figure shows a general approach to solving a classification problem.
* First, a training set consist of records with class labels, this is used to build a classification model, which is applied on test set with unknown class labels.

www.bhtechupdates.com

* Evalution of performance of a model is shown in a table known as confusion matrix is shown -figure.

		Predicted class	
		class=1	class=0
Actual Class.	class=1	f_{11}	f_{10}
	class=0	f_{01}	f_{00}

* Each entry f_{ij} denoted the no.of records from class 'i' predected to be of class 'j'.

* Based on entries in confusion matrix, the total no.of correct predictions made by the model is $(f_{11} + f_{00})$ and the total no.of incorrect predictions is $(f_{10} + f_{01})$

* The performance model can be expressed in terms of its accuracy as follows.

$$\text{Accuracy} = \frac{\text{Number of predictions}}{\text{Total no.of predictions.}}$$

$$= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

* Similarly the performance model can be expressed in terms of it error rate as follows.

$$\text{Error rate} = \frac{\text{Number of incorrect predictions}}{\text{Total Number of predictions}}$$

$$= \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Decission Tree Induction:-

Working of a Decision Tree:-

* Consider a vertebrate classification problem, instead of classifying vertebrates into four spices we assign them two categories. 1) Mammal & 2) Non-Mammal

* suppose a new spice is discovered by scientist, how to we tell whether it is a mammal (or) Non-mammal? one method is to pose a series of questions about the characteristics of spices.

* The first question is whether it is warm (or) cold blooded, if it is warm then it is mammal otherwise non-mammal.

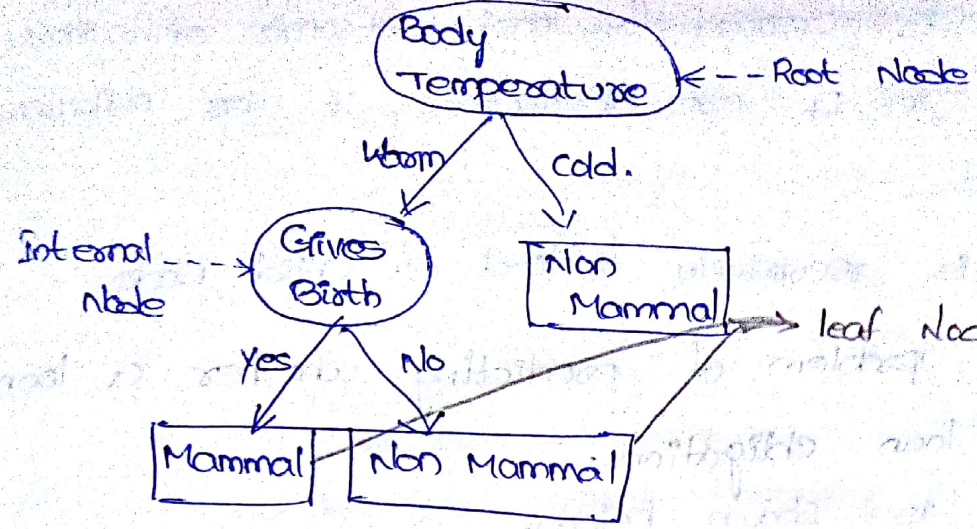
* In the next case we need to ask a question: Do the females of spices gives birth to their young? those that give birth are definitely mammals.

* The following figure shows a decision tree for mammal classification problem, the tree has three nodes 1. root node 2. Internal node 3. leaf node

Root Node:- It has no incoming edges and 0 (or) more outgoing edges.

Internal Node:- It has only one incoming edge and two (or) more outgoing edges.

Leaf Node (or) External Node:- It has exactly one incoming edge and no outgoing edges.



How to build a decision Tree;

* There are many decision trees that can be constructed from a given set of attributes, while some of the trees are more accurate than others.
* Finding the optimal tree by using efficient algorithms, these algorithms use a Greedy strategy that grows a decision tree by making a series of optimum decisions about which attribute to use for partition the data
* one such algorithm is Hunt's algorithm

Many algorithms
HUNT's
CART
ID3, C4.5
SLIQ, SPRINT

** Hunt's Algorithm;

* In this algorithm, a decision tree is grown in a recursive fashion by partitioning the records into successively purer subsets.
* Let D_t be the set of training records that are associated with Node t and $y = \{y_1, y_2, y_3, \ldots, y_c\}$ be class labels.
* The following is a recursive defination of hunts algorithm

Step:1 If all the records in D_t belongs to the same class y_t, then t is a leaf node labeled as y_t.

Step:2 If D_t contains the records belongs to more than one class, an attribute test condition is selected to partition the records into smaller subsets.

* A child node is created for each outcome of test condition and the records in D_t are distributed to the children based on outcomes.

* The algorithm is recursively applied to child node

Ex:- consider the problem of predicting whether a loan applicant will repay his loan obligations.

The example is shown below.

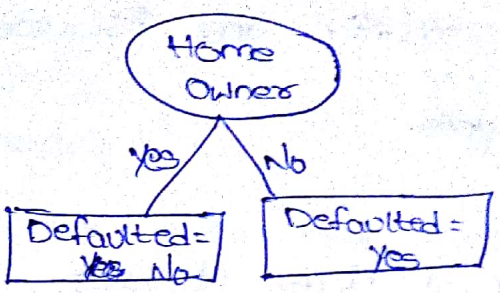
TID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	single	125 K	No
2	No	Married	100 K	No
3	No	single	70 K	No
4	Yes	Married	120 K	No
5	No	Diverse	95 K	Yes
6	No	Married	60 K	No
7	Yes	Diverse	220 K	No
8	No	single	85 K	Yes
9	No	Married	75 K	No
10	No	single	90 K	Yes

* Each record contains the personal information of a borrower along with a class label indicating whether a borrower has defaulted on loan payments.

* The initial tree for classification problem contains a single node with class label Defaulted = No which makes that most of the borrowers successfully repaid their loans.

Defaulted = No

* The records are divided into smaller subsets based on the outcomes of Home Owner test condition as shown below.

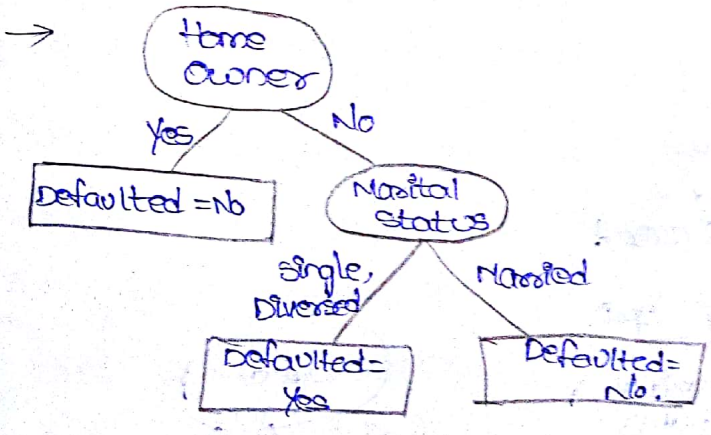


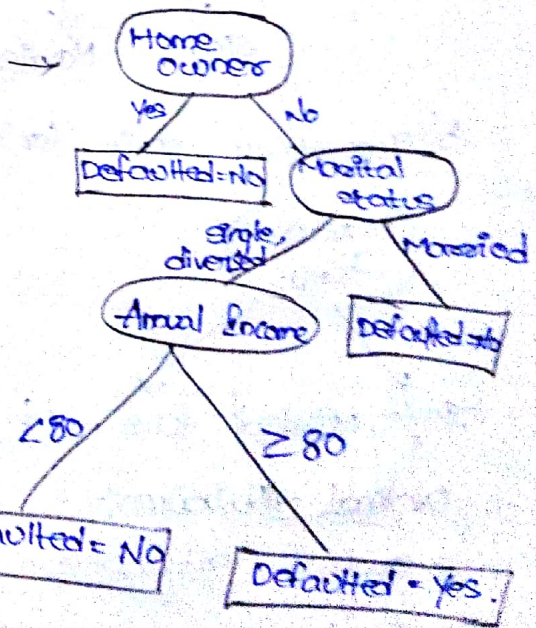
* Hunt's algorithm is then applied recursively to each child of the root node.

* From the training set, all borrowers who are home owners successfully repaid their loans.

* The left child of the root is a leaf node, for the right child we need to continue recursive step of Hunt's algorithm, until all the records belongs to same class.

* The resulting trees as shown below.





* Hunt's algorithm will work if every combination of attribute values is present in the training data and each combination has a unique class label.

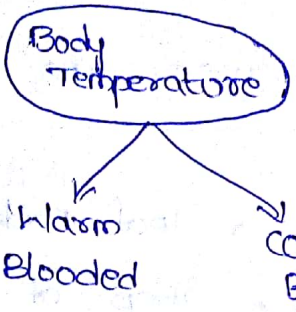
Methods for expressing attribute Test Condition:-

Decision tree induction algorithms must provide a method for expressing an attribute test condition and its corresponding outcomes for different attribute types.

Binary Attributes:-

* The test condition for a binary attributes generate two potential outcomes, as shown below.

eg:-

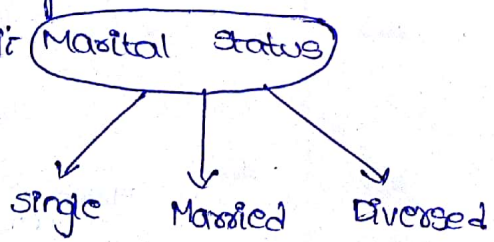


Nominal Attributes:-

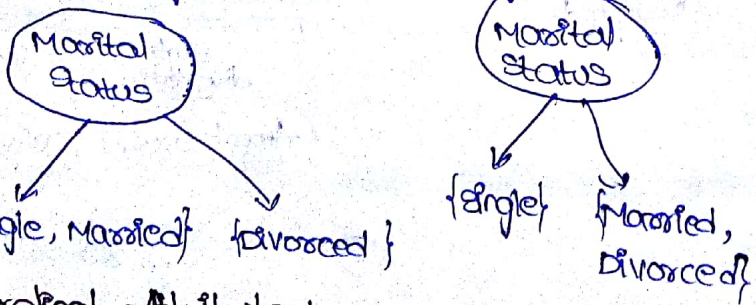
A nominal attribute can have many values, its test condition can be expressed in two ways.

a) Multi-way split

eg:-



b) Two-way split (or) Binary split

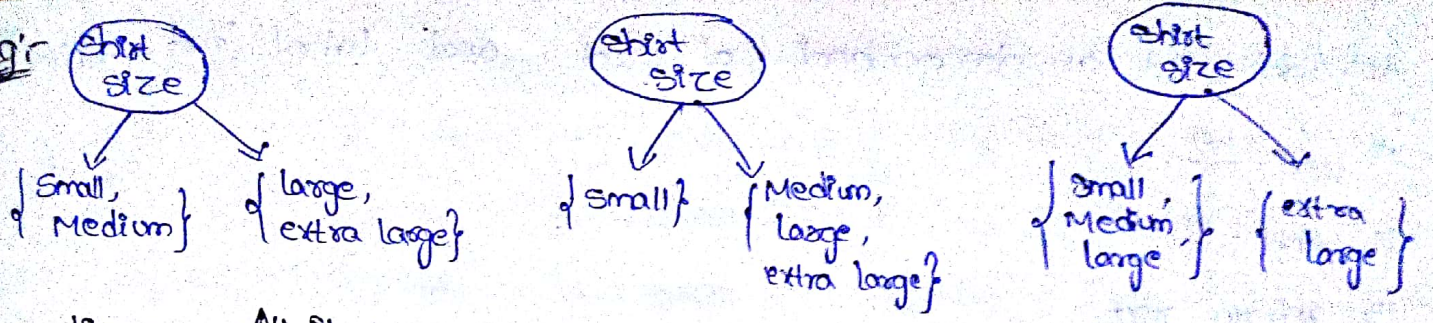


Ordinal Attributes:-

→ ordinal attributes can produce binary or multi-way splits

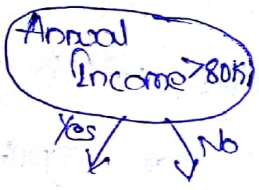
→ ordinal attribute values can be grouped as long as the grouping does not violate, the ordered property of attribute values.

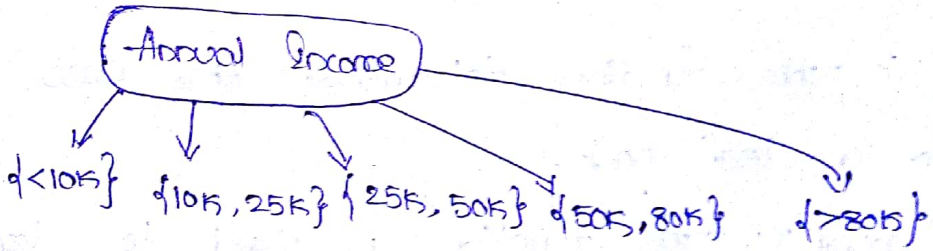
eg:-



continuous ~~Attributes~~:-

For these attributes, the test condition can be expressed as a comparison test $(A < V)$ (or) $(A >= V)$ with binary outcomes, (or) a range query with outcomes of the form $V_i \leq A \leq V_{i+1}$ for $i = 1, 2, \ldots K$.





31/12/2018

Algorithm for Decision tree induction:-

TreeGrowth (E, F) → attribute set; training records data set

1. If stopping_cond(E,F) = true then
2. leaf = createNode( )
3. leaf.label = classify(E)
4. return leaf
5. else
6. root = createNode( )
7. root.test_cond = find_best_split(E,F)
8. let V = {v/v is a possible outcome of root.test_cond}
9. for each v∈V do
10. E_v = {e/root.test_cond(e)=V and e∈E}
11. child = TreeGrowth(E_v, F)

12. add child as descendend of root and label (root → child) = v.

13. end for

14. end if

15. return root.

The input to this algorithm consists of training records 'E' and attributes set 'F'.

CreateNode( ):- this function extends the decision tree by creating a new node.

find_best_split( ):- This function determines which attribute should be selected as the test cond for splitting training records.

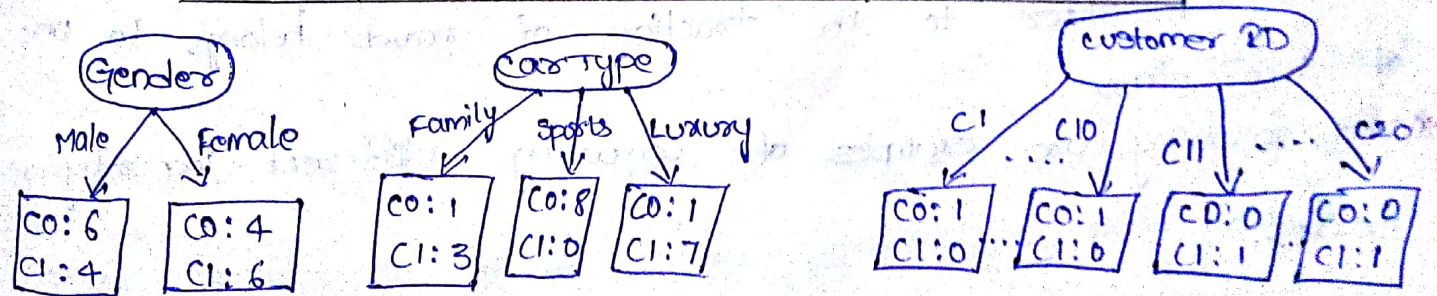
classify( ):- this function determines the class label to be assigned to a leaf node.

stopping_cond( ):- this function is used to terminate the tree growing process by testing whether all records have either the same class label. (or) same attribute values.

2|11|17

Measures for selecting the Best split :-

CID	Gender	car Type	class
1	Male	Family	C0
2	M	Sports	C0
3	M	sports	C0
4	M	sports	C0
5	M	Sports	C0
6	M	sports	C0
7	Female	Sports	C0
8	F	sports	C0
9	F	Sports	C0
10	F	Luxury	C0
11	M	Family	C1
12	M	Family	C1
13	M	Family	C1
14	M	luxury	C1
15	F	luxury	C1
16	F	Luxury	C1
17	F	Luxury	C1
18	F	Luxury	C1
19	F	Luxury	C1
20	F	Luxury	C1



* There are many measures that can be used to determine the best way to split the records, these measures are defined in terms of class distribution of records before and after splitting.

www.Intufastupdates.com

* let $p(i/t)$ denotes the fraction of records belongs to class 'i' at a given node 't'.

* In a two-class problem, the class distribution at any node can be written as (p_0, p_1) where $p_1 = 1 - p_0$

* consider the test condition shown in above figure: the class distribution before splitting is (0.5, 0.5), because there are an equal no.of records from each class.

* If we split the data using Gender attribute, then class distribution of child nodes are (0.6, 0.4) & (0.4, 0.6).

* The measures developed for selecting the best split are based on the degree of impurity of child nodes.

* Examples of impurity measures includes:

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i/t) \log_2 p(i/t)$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i/t)]^2$$

$$\text{Classification Error} = 1 - \max_i [p(i/t)]$$

where

c is no.of classes

p refers to the fraction of records belongs to one of two classes

* The following are examples of computing different impurity measures.

Node N_1	Count
Class=0	0
Class=1	6

$$\text{Gini} = 1 - (0/6)^2 - (6/6)^2 = 0$$

$$\text{Entropy} = -(0/6)\log_2(0/6) - (6/6)\log_2(6/6) = 0$$

$$\text{Error} = 1 - \max_i (0/6, 6/6) = 1 - 1 = 0$$

Node N2	Count
Class = 0	1
Class = 1	5

$$\text{Gini} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.278$$

$$\text{Entropy} = -\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) = 0.65$$

$$\text{Error} = 1 - \text{Max}\left(\frac{1}{5}, \frac{5}{6}\right) = 0.167$$

Node N3	Count
class = 0	3
class = 1	3

$$\text{Gini} = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

$$\text{Entropy} = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 1$$

$$\text{Error} = 1 - \text{Max}\left(\frac{3}{6}, \frac{3}{6}\right) = 0.5$$

* the above example illustrate the consistency among different impurity measures, based on these calculations node N_1 has lowest impurity value followed by N_2 & N_3.

* To determine how well a test condition performs, we need to compare the degree of impurity of parent node (before splitting) with the degree of impurity of child nodes (after spliting)

* The larger their difference, the better test condition.

* The gain Δ is a criteria that can be used to determine the goodness of a split

$$\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(V_j)}{N} \cdot I(V_j)$$

* Where $I(\cdot)$ is the impurity measure of a given node, N is total no.of records at parent node, k is no.of attribute values, $N(V_j)$ is no.of records associated with child node.

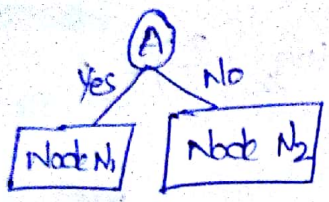
* Finally, when entropy is used as impurity measure, the difference in entropy is known as InformationGain, Δ_{info}

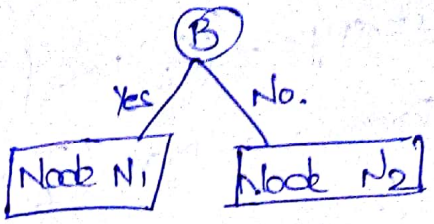
Splitting of Binary Attributes:-

	Parent
C0	6
C1	6
Gini = 0.5	

$1-\left(\frac{6}{12}\right)^2-\left(\frac{6}{12}\right)^2$

$= 1 - 0.5 = 0.5$





	N_1	N_2
C0	4	2
C1	3	3

	N_1	N_2
C0	1	5
C1	4	2

N_1

$1-\left(\frac{4}{7}\right)^2-\left(\frac{3}{7}\right)^2 = 1-0.326-0.183$

$= 0.4898$

N_2

$1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2$

$= 1-0.16-0.36$

$= 0.48$

$N_1 \rightarrow 1-\left(\frac{1}{5}\right)^2-\left(\frac{4}{5}\right)^2 = 1-0.04-0.64 = 0.32$

$N_2 = 1-\left(\frac{5}{7}\right)^2-\left(\frac{2}{7}\right)^2 = 1-0.510-0.081$

$= 0.408$

* If attribute A is choosen to split data, GINI index for node $N_1 = 0.4898$ and for node N_2 is 0.48
* the weighted average of GINI Index for attribute A is

$\frac{7}{12}(0.4898) + \frac{5}{12}(0.48)$

$= 0.486$

* Similarly, we can show that the average of GINI Index for attribute B is 0.375

$= \frac{5}{12}(0.328) + \frac{7}{12}(0.408)$

$= 0.375$

* for attribute B have a smaller GINI index poefered over A.

Splitting of Nominal attributes:-

* A nominal attribute can produce either binary (or) multiway splits as shown below

Multi-way split

	Car Type		
	Family	Sports	Luxury
C0	1	8	1
C1	3	0	7
Gini	0.163		

Two-way split

	Car Type	
	{Sports, Luxury}	{Family}
C0	9	1
C1	7	3
Gini	0.468	

	Car Type	
	{Sports}	{family, luxury}
C0	8	2
C1	0	10
Gini	0.167	

* We prefer multi-way splitting compared with Binary split.

Continuous	Annual Income ≤ 80	Income ≥ 80
Yes	0	3
No	3	4

Gain Ratio:-

* A splitting criteria is known as Gain Ratio, is used to determine the goodness of a split
* This criteria is defined as follows

$$\text{Gain Ratio} = \frac{\Delta \text{Info}}{\text{SplitInfo}}$$

* where $\text{SplitInfo} = -\sum_{i=1}^{k} P(v_i) \log_2 P(v_i)$

where k is total no. of splits

id	A	B	C	Class label
1.	T	F	T	+
2	T	F	F	+
3	T	T	T	−
4	T	T	F	+
5	F	F	T	−
6	F	F	F	+
7	F	T	T	−
8	F	T	F	−

A

+	3	1
−	1	3

B

+	1	3
−	3	1

C

+	1	3
−	3	1

$N_1 = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$

$N_2 = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$