Data mining Unit 3 Handwritten

Data Mining And Analytics (SRM Institute of Science and Technology)

Unit - 3

Classification

Classification is a data mining function that assign items in a collection to target categories or classes.

The ~~ta~~ goal of classification is to accurately predict the target class for each case in the data.

Example of cases where the data analysis task is classification :-

(a) A bank officer want to analyze the data in order to know which customer are risky or which are safe.

(b) A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

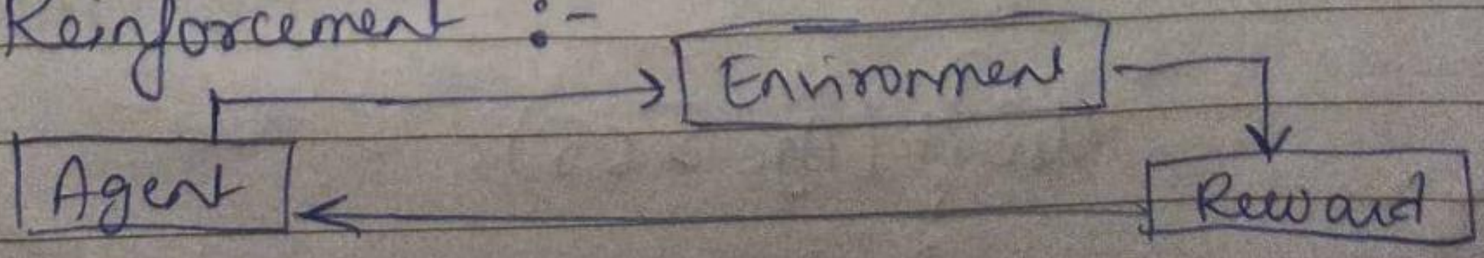
ML (Machine Learning)

→ Supervised Approach :-

X_1 X_2 X_3 (Y) → output is known

→ Unsupervised Approach :- output not known

→ Reinforcement :-

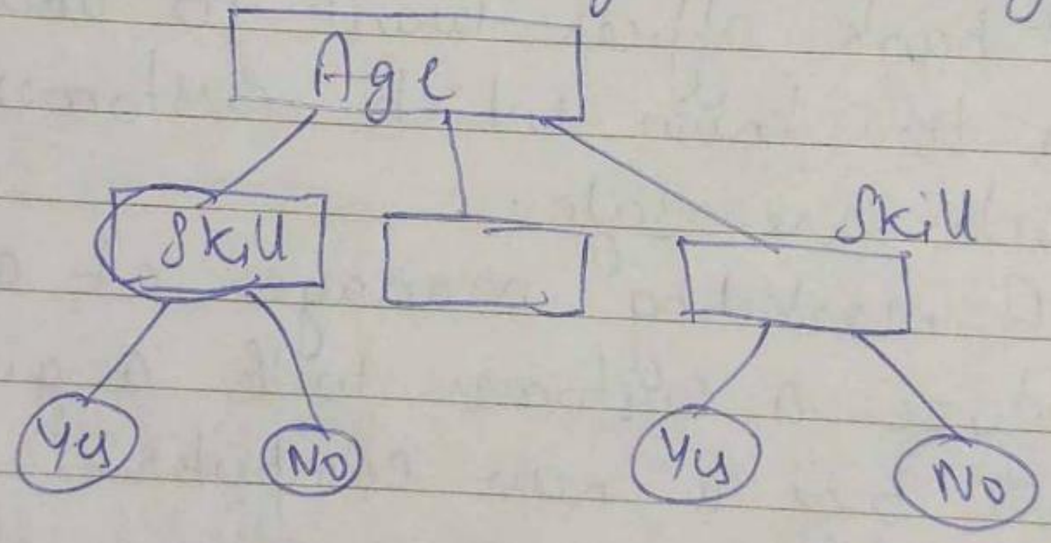


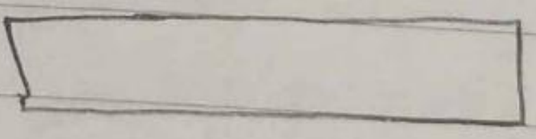
* Classification comes under Supervised approach

In classification, basically what we do is to learn understand the data and then by using known output we can predict the output of unknown data.

Decision Tree Induction

Decision Tree is a technique used in classification.
It will not be always a binary tree.



for attribute → 

for result →

Ye basically if else if pe kam kar rha h.

```
If (Age < 20)
{
}
else if (Age > 60)
{
}
else {}
```

Now, whenever we choose different attributes on top decision tree will be different.

So, we need to find which attribute will be well suited for top. This process is known as Attribute Selection.

For this we measure Entropy.

$$\text{Entropy} = -\sum P_i \log_2(P_i) \text{ bits}$$

Q.

S.No	Age	Income	Class
1.	Youth	H	No
2.	Youth	H	No
3.	Middle age	H	Yes
4.	Senior	M	Yes
5.	Senior	L	Yes
6.	Senior	L	No
7.	Middle age	L	Yes
8.	Youth	M	No
9.	Youth	L	Yes
10.	Senior	M	Yes
11.	Youth	M	Yes
12.	Middle age	M	Yes
13.	Middle age	H	Yes
14	Senior	M	No

No. of classes = 2.

Step 1 → Count No. of Yes & No.

Yes = 9

No = 5

$$P(Yes) = \frac{9}{14} \qquad P(No) = \frac{5}{14}$$

Step 2 → find information in data (Entropy)

$$info(D) = -\sum P_i \log_2(P_i)$$

$$= -\left(\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \left(\frac{5}{14} \times \log_2\left(\frac{5}{14}\right)\right)\right)$$

$$= -(-0.4098 + (-0.5305))$$

$$= 0.940 \text{ bit}$$

Step 3 :- find info of age :-

	Yes	No	Sum
Youth	2	3	5
Middle Age	4	0	4
Senior	3	2	5

$$info(age) = -\left[\frac{5}{14} \times \left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)\right.$$

$$+ \frac{4}{14} \times \left(\frac{4}{4} \times \log_2\left(\frac{4}{4}\right) + 0 \times \log_2 0\right)$$

$$\left. + \frac{5}{14} \times \left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)\right]$$

$$= -\left[-0.347 + 0 + -0.347\right]$$

$$= 0.694 \text{ bits}$$

Gain = info(D) − info(Age)

$$= 0.940 - 0.694$$

$$= 0.246 \text{ bits}$$

Step 4:- find info of income :-

	Yes	No	Sum
H	2	2	4
M	3	1	4
L	4	2	6

$$\text{info(income)} = -\left[\frac{4}{14} \times \left(\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) + \frac{4}{14} \times \left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] + \frac{6}{14} \times \left(\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right)\right.$$

$$= -\left[-0.2857 + -0.3936 + -0.3936\right]$$

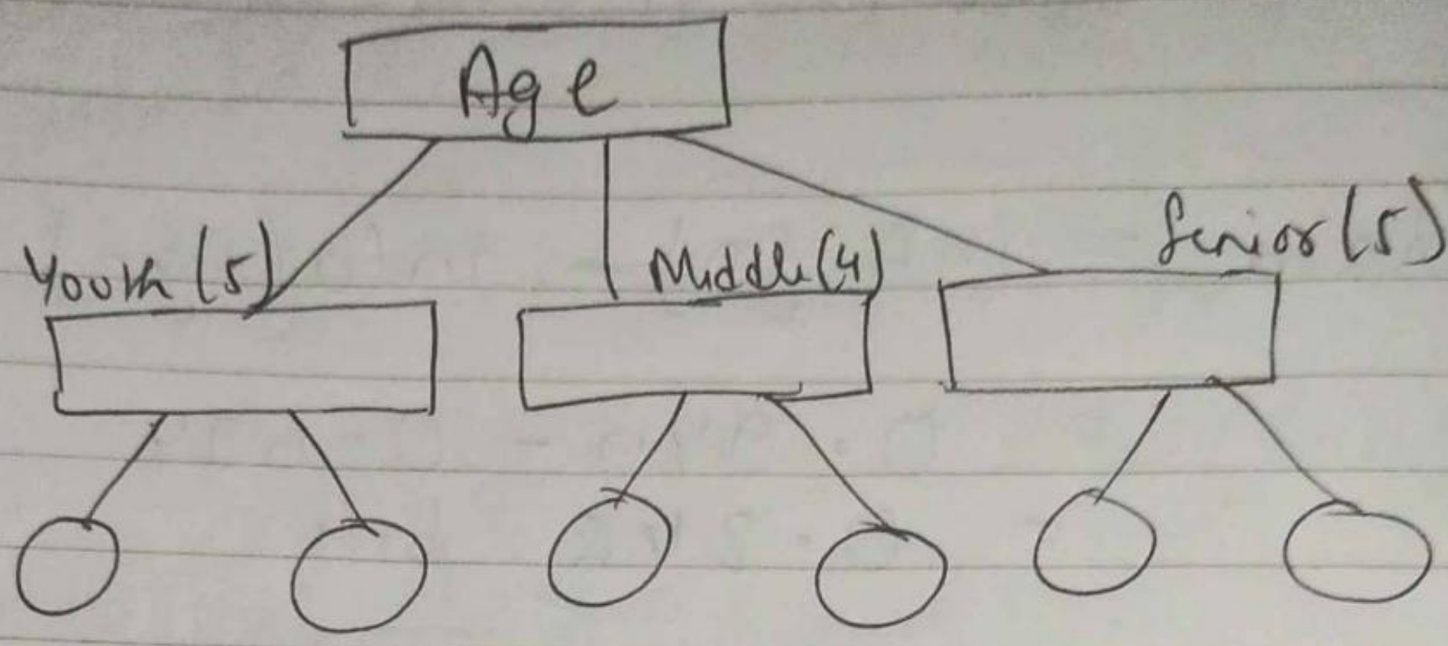
$$= 0.911 \text{ bits}$$

Gain = info(D) − info(income)

$$= 0.940 - 0.911$$

$$= 0.029 \text{ bits}$$

Gain of Age is higher than Gain of income. So it will come on top.



Drawback :- 1) Bias :- it is favouring the attributes which has more no. of partition.

Name of this Algo is ID3 :

Algorithm :-

Generate decision-tree (D, Attribute, Selection method)

1) Create a Node N
2) If D contain the tuples belonging to same class.
   return N with label of class
3) If attribute list is empty then
   return N as leaf node with majority in D.
4) Apply attribute selection criteria to find the best split.
5) Attribute list = Attribute list - best selected attribute

6. For every j in the split
If D_j is empty.
attach a leaf node with majority of class as labels.
Else
return Node Create decision_tree(D_j)

7. Return N.

* $Split\_info_A(D) = -\sum_{i=1}^{v} \left|\frac{d_i}{D}\right| \log_2\left(\frac{D_i}{D}\right)$.

Split_info of Info(B)

$= -\left(\frac{4}{14}\log_2\frac{4}{14} + \frac{6}{14}\log_2\frac{6}{14} + \frac{4}{14}\log_2\frac{4}{14}\right)$.

$= 1.55$ bits

Gain Ratio $= \frac{Info(D)}{SplitInfo_A(D)}$

$= \frac{0.029}{1.55} = 0.018$

* Select that whose Gain ratio is high.

Gini Index :-

Gini Index measures ~~that~~ the impurity of data.

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

$$P(Yes) = \frac{9}{14}$$

$$P(No) = \frac{5}{14}$$

$$Gini(D) = 1 - \left[\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right]$$

$$= 0.46$$

$$= 46\% \text{ impurity.}$$

$$Gini_A(D) = \sum_{i=1}^{v} \frac{|D_i|}{|D|} * Gini(D).$$

* Gini is applied when split is Binary.

So, these 3 were the 3 techniques for designing Decision tree.

* Now the Question arises till what height we should do it.

We just need to draw till a certain level. We will not draw after that level.

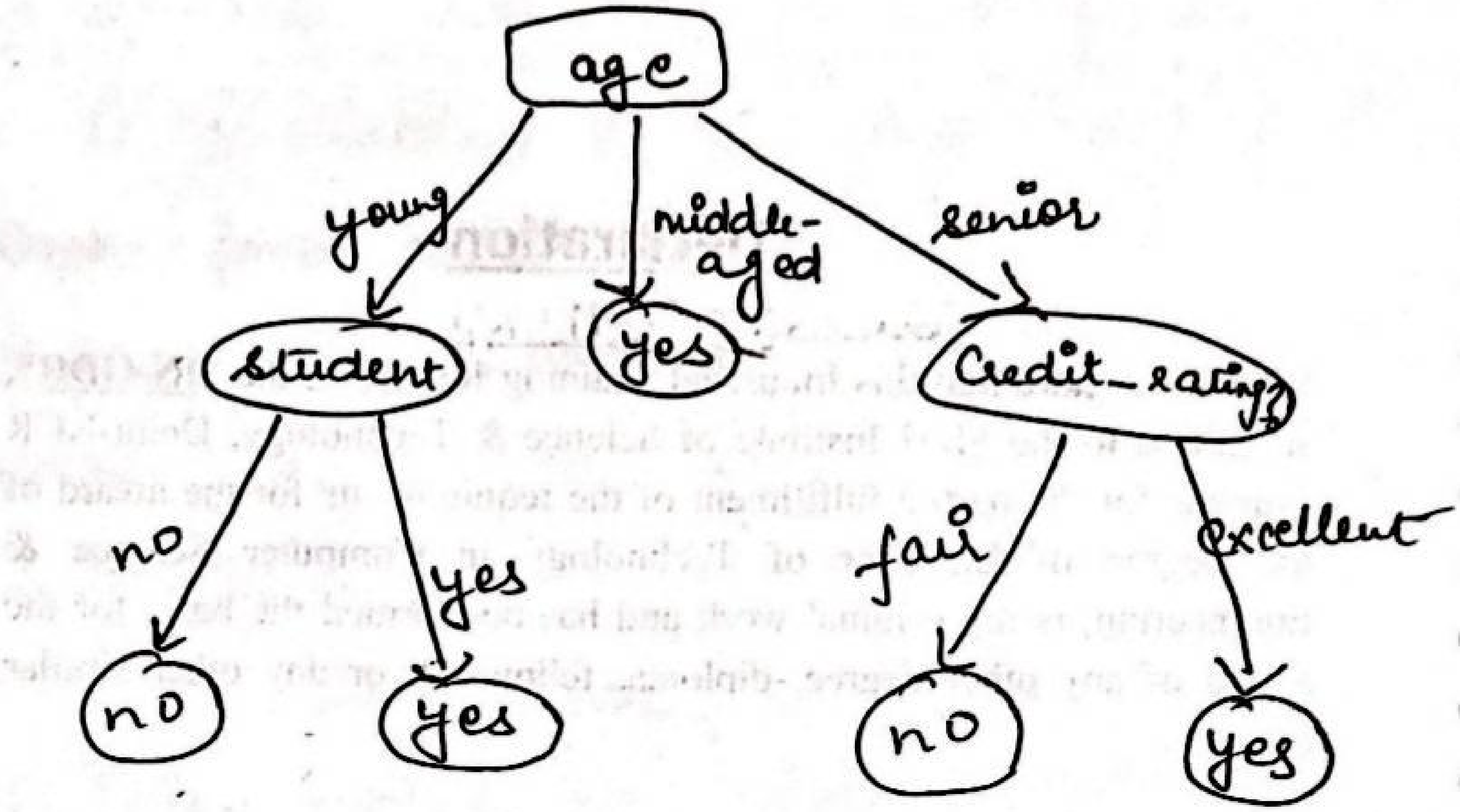
2 Ways to decide height

(i) Pre pouring
(ii) Post pouring.

Decision Tree induction

It is a method of learning the decision trees from the training sets. The dataset is broken down into smaller subsets and is present in the form of nodes of a trees. The tree structure has a root node, internal nodes or decision nodes, leaf nodes and branches. The root node is the topmost node.

→ Each internal node denotes a test of an attribute, each branch ~~nodes~~ denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.



Decision tree induction algorithm

developed a decision tree algorithm known as ID3 (Iterative Dichotomiser) later, he presented C4.5, which was the successor of ID3. ID3 and C4.5 adopt a greedy approach. In this algo, there is no backtracking; the trees are constructed in a top-down recursive divide and conquer manner.

Tree Pruning

Tree Pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Tree - Pruning Approaches

There are two approaches to prune a tree -

Pre-pruning → The tree is pruned by halting its construction early.

Post-pruning → This approach removes a sub-tree from a fully grown tree.

Key factors

Entropy :-

It refers to a common way to measure impurity. In the decision trees, it measures the randomness or impurity in data sets.

Information Gain :-

Information gain refers to the decline in entropy after the dataset is split. It is also called entropy reduction. Building a decision tree is all about discovering attributes that return the highest data gain.