Forward Feature Selection:

It follow the inverse process of the backward elimination process. It means in this technique we donot eliminate the feature instead we will find the feature that can produce the highest increase in the performance of the model. Below steps are performed in this model

i) We start with a single feature only & progressively we will add each feature at a time.

ii) Here we will train the model on each feature seperately.

iii) The feature with the best performance is selected.

iv) The process will be repeated untill we get a significant increase in the performance of the model.

- ## Linear Discreminant Analysis

It is a predictive algorithm model for multi-class classification. It can also be used as dimentionality reduction technique providing a projection of a training dataset that based seperate the example

• Supervised & Un-supervised Learning:

Supervised
are the two technique of machine learning but both the technique is used in different scenario & with different dataset.

Supervised Learning:
It is a ML method in which model are trained labelled data. In supervised learning model need to find the mapping fun to map the input variable (x) with the output variable (y).

$$y = f(x)$$

Supervised learning needs supervision to train the model which is similar as a a student learn things in presence of parents. SL can be used for two type of problem classification & recreation.

Suppouse we have an image of different types of fruit. The task of our supervised learning model is to identify the fruits & classify accordingly. It identify the shape size & color of the fruit.

Un-supervised ML:

It is another ML method in which pattern from the unlabelled data. The goal of USL is to find this structure & pattern from

the input data. It does not need any supervision instead it find pattern from the data by its own. It can be used for two types of problem clustering & association.

Take the above example with the help of suitable ~~example~~ algm then the model will train itself & divide the fruits into different groups according to the most similar features between them.

Supervised Learning	Unsupervised Learning
i) SL algm are trained using labelled data.	i) UL algm are trained using unlabelled data.
ii) SL model take direct feedback to check if it is predicting correct output or not.	ii) It does not take any feedback.
iii) SL model predict the output	iii) UL model find the hidden pattern in the dataset.
iv) In SL input data is provided to the model along with the output.	iv) In UL only input data is provided to the model.

v) It's goal is to train the model so that it can predict the output when it is given new data.	v) It's goal is to find the hidden pattern & useful data from the unknown dataset.
vi) It needs supervision to train the model.	vi) It does not need any supervision to train the model.
vii) It can be categorized in classification & regression preoblem.	vii) It can be categorized in clustering &, association problems.
viii) It can be used for those cases where we Knows the input as well as output.	viii) It use for the cases where we have only the input data & no corresponding output data.
ix) It produce an accurate result.	ix) It may produce less accurate result. as compared to SL
x) It include various algm linear regression, logistic regression, support vector machine & decision tree	x) It includes various algm clustering, KNN,

Decision Tree:

A type of data mining technique, decision tree in data mining built a model for classification of data. The model are built in the form of tree structure & hence belongs to the supervis form of learning.

Structure of a decision tree:

It consist of a root ~~d~~node, branch, & a leaf node. The branch node are the outcome of a tree & the internal node represent the test on an attribute. The leaf node represent a class level.

Working of a decision tree:

a) A decision tree works under the supervised learning approach for the both discrete & continous variable. The dataset is split into subset on the basis of the datasets most significant attribute Identification of the attribute & spitting is done through the algorithm.

b) The structure of the decision tree consist a root node which is significant predictor node. The process of splitting occur from the decision ~~d~~node which are the

sub-node of the tree. The node which do not split further are termed as leaf node.

c) The dataset is divided into homogenous & non overlappling reason following a top down approach.

d) Untill & unless a stop crieteria is reached the decision tree will keep on running.

e) With the building of a decision tree lot of noise outliner are generated. To remove this outliner & noisy data a method of tree pruning is applied.

f) Accuracy of a model is checked on a test set consisting of test tuple & class.

Type of decision tree:

Decision lead to the development of models for classification & regression:

1 Classification: The classification include the building up of model describing important class level. They are applied in the area of machine learning & pattern recognization. Decision tree in machine learning through classification model lead to fraud detection, medical

diagonisis, etc. To step process of a classification model includes.

a) Learning - A classification model based on the training data is built.

b) Classification - Model accuracy is checked & then used for classification of the new data. Class level are in the form of discrete value like yes or no, etc

2. Regression: It is used for the regression analysis of data such that the prediction of numerical attribute. These are also called continous value. Therefore instead of predicting the class level, the regression model predict the continous value.

a) List of algm used in decision tree:

ID3: The whole set of data S is considered as the root node while forming the decision tree. Iteration is then carried out on every attribute & splitting of data into fragments. The algm check & take those attribute which were not taken before the iteration. Splitting data in the ID3 algm is time consuming & is not a ideal

alg^m as it overfit the data.

C4.5, CART, CHAID

Sore Throat	Fever	Swollen Gland	Congestion	Heache	Diagnosis
Yes	Yes	Yes	Yes	Yes	Strep Throat
No	No	No	Yes	Yes	Allergy
Yes	Yes	No	Yes	No	Cold
Yes	No	Yes	No	No	Strep Throat
No	Yes	No	Yes	No	Cold
No	No	No	Yes	No	Allergy
No	No	Yes	No	No	Strep Throat
Yes	No	No	Yes	Yes	Allergy
No	Yes	No	Yes	Yes	Cold
Yes	Yes	No	Yes	Yes	Cold

(1) Information Gain

$$I(P,n) = \frac{-P}{S}\log_2\frac{P}{S} - \frac{n}{S}\log_2\frac{n}{S}$$

$$S = (P+n)$$

(2) Entropy

$$E(A) = \sum_{i=1}^{v} \frac{P_i + n_i}{P+n} (P_i, n_i)$$

③ Grain

$$\boxed{(A) = I(P,n) - E(A)}$$

$$\left\{\log_2 x = \frac{\log_{10} x}{\log_{10} 2}\right\}$$

Sample Space

$\Rightarrow S.T + A + C$

$\downarrow \quad \downarrow \quad \downarrow$

$3 + 3 + 4 = 10$

Inf^n Grain

$$I(ST, A, C) = -\left[\frac{3}{10}\log_2\frac{3}{10} + \frac{3}{10}\log_2\frac{3}{10} + \frac{4}{10}\log_2\frac{4}{10}\right]$$

$$= -[0.3\log_2 0.3 + 0.3\log_2 0.3 + 0.4\log_2 0.4]$$

$$= -\left[0.3 \times \frac{\log_{10} 0.3}{\log_{10} 2} + 0.3 \times \frac{\log_{10} 0.3}{\log_{10} 2} + 0.4 \times \frac{\log_{10} 0.4}{\log_{10} 2}\right]$$

$$= -\left[0.3 \times \frac{(-0.522)}{0.301} + 0.3 \times \frac{(-0.522)}{0.301} + 0.4 \times \frac{(-0397}{0.301}\right.$$

$= 0.6 \times 1.73 + 0.4 \times 1.34$

$= 1.038 + 0.527$

$= 1.56$

Find Splitting Attribute

i) For first, ~~sho~~ sore throat,

	S.T.	A	C	
Yes	2	1	2	= 5
No	1	2	2	= 5

S.T.

$I(ST, A, C) =$

$$I(Yes) = -\left[\frac{2}{5}\log_2\frac{2}{5} + \frac{1}{5}\log_2\frac{1}{5} + \frac{2}{5}\log_2\frac{2}{5}\right]$$

$$= 1.52$$

$$I(NO) = 1.52$$

ii) Entropy (Sore Throat)

$$= \frac{5}{10}^{(P)} * 1.52 + \frac{5}{10}^{(S)} * 1.52$$

$I(Y)$ $I(N)$

$\therefore$ Entropy (S) = 1.52

iii) Gain; $A = I(P,n) - E(A)$

$= 1.56 - 1.52$

$= 0.04$

(d) For Fever,

	ST	A	C	
Yes	1	0	4	= 5
No	2	3	0	= 5

$$I(Yes) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{0}{5}\log_2\frac{0}{5} - \frac{4}{5}\log_2\frac{4}{5}$$

$$= -\left[\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}\right]$$

$$= -\frac{1}{5}\left[\log_2\frac{1}{5} - 4\log_2\frac{4}{5}\right]$$

$$= -\frac{1}{5}\left[\log_2 0.2 - 4\log_2 0.8\right]$$

$$= -\frac{1}{5}\left[\frac{\log_{10}0.2}{\log_{10}2} - 4\cdot\frac{\log_{10}0.8}{\log_{10}2}\right]$$

$$= -\frac{1}{5}\left[\frac{-0.699}{0.30} - 4\cdot\frac{-0.0969}{0.30}\right]$$

$$= -\frac{1}{5}(-2.33 + 4(0.32))$$

$$= -\frac{1}{5}(-1.05)$$

$$= 0.21$$

$$I(No) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= -0.4\log_2 0.4 - 0.6\log_2 0.6$$

$$= -0.4\left(\frac{\log_{10}0.4}{\log_{10}2}\right) - 0.6\left(\frac{\log_{10}0.6}{\log_{10}2}\right)$$

$$= -0.4\left(\frac{-0.3979}{0.30}\right) - 0.6\left(\frac{-0.2218}{0.30}\right)$$

0.525
0.462
0.063

$$I(No) = -0.4(-1.326) - 0.6(-0.739)$$
$$= 0.5304 + 0.4434$$
$$= 0.9738$$

Entropy (Fever)

$$= \frac{5}{10} * 0.71 + \frac{5}{10} * 0.9738$$
$$= 0.355 + 0.4869$$
$$= 0.84$$

Gain, $A = I(P,n) - E(A)$
$= 1.56 - 0.84$
$= 0.72$

⊙ For Swollen Gland,

	ST	A	C	
Yes	3	0	0	= 3
No	0	3	4	= 7

$$I(Yes) = -\frac{3}{3}\log_2\frac{3}{3} = -1 \cdot \frac{\log_{10}1}{\log_{10}2} = -1 \cdot \frac{0}{\cdot\cdot} = 0$$

$$I(No) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$$
$$= -0.428(\log_2 0.428) - 0.571(\log_2 0.574)$$
$$= -0.428\left(\frac{\log_{10}0.428}{\log_{10}2}\right) - 0.574\left(\frac{\log_{10}0.574}{\log_{10}2}\right)$$
$$= -0.428\left(\frac{-0.368}{0.30}\right) - 0.571\left(\frac{-0.243}{0.30}\right) = 0.063$$
$$= 0.973$$

Entropy (Swollen Orland)

$= \frac{3}{10} * 0 + \frac{7}{10} * 0.063$

$= 0.044$ 0.68

Grain (Swollen Orland) $= 0.88$

◎ For Congestion

$I(Yes) = -\left[\frac{1}{8}\log_2\frac{1}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{4}{8}\log_2\frac{4}{8}\right]$

$= 1.36$

$I(No) = -\left[\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2} + \frac{0}{2}\log_2\frac{0}{2}\right]$

$= 0$

Entropy (Congestion) $= \frac{8}{10} \times 1.36 + \frac{2}{10} \times 0$

$= 1.09$

Grain $= 1.56 - 1.09$

$= 0.47$

◎ For Headache –

	ST	A	C	
Yes	1	2	2	= 5
No	2	1	2	= 5

$I(Yes) = \frac{1}{5}\log_2\frac{1}{5} + \frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}$

$= 1.52$

$I(No) = 1.52$

Entropy(headache) $= \frac{5}{10} \times 1.52 + \frac{5}{10} \times 1.52$

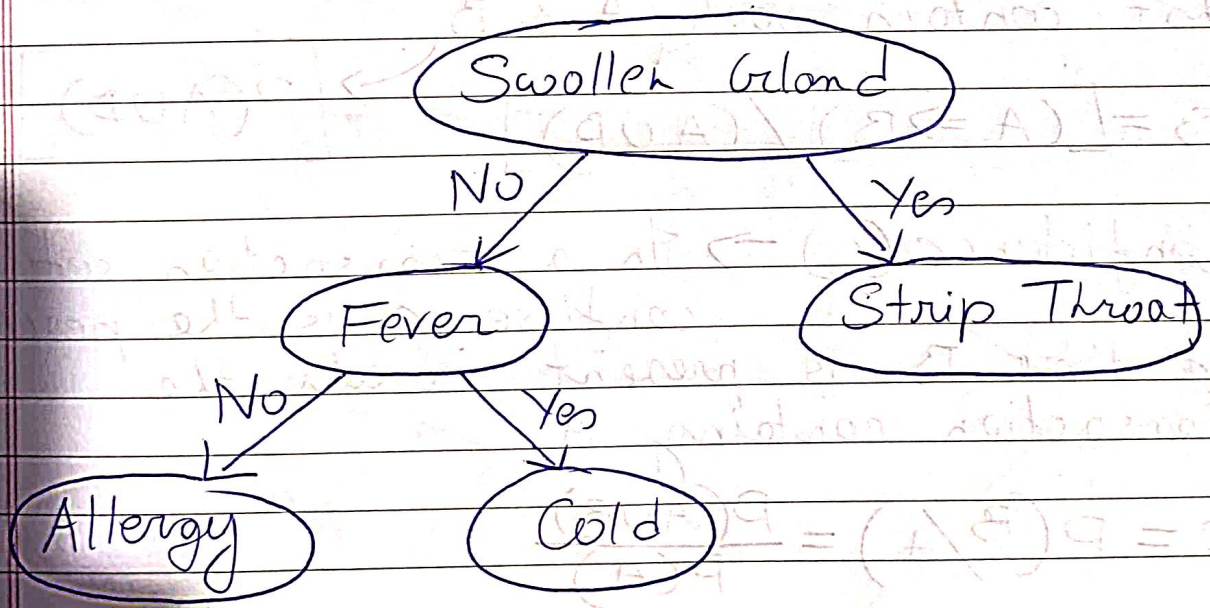
$= 1.52$

Gain(headache) $= 1.56 - 1.52$

$= 0.04$

Split attribute -

Gain ST - 0.04
fever - 0.72
Scr - 0.88
Congestion - 0.47
Headache - 0.04



* Association Rule:

Tid	Data
1	→ Bread, butter, milk
2	→ Bread, butter
3	→ Milk, egg
4	→ Bread, butter, egg.

→ Data set.

i) Set of item in a transaction is called Market Basket.

ii) Mostly used in retail.

iii) If A then B [A → B]

* Support (S) → Percentage of transaction that contain both A & B → [P(AUB)]

$$S = [(A \Rightarrow B) / (A \cup B)]$$

* Confidence (C) → In a transaction set confidence is the percentage of time B is present in all the transaction containg A.

$$C = P(B/A) = \frac{P(A \cup B)}{P(A)}$$

FP Tree :- Algorithm

Tid	items
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

Given, minimum support(s) = 30%
Total transaction = 8
Now,
30% of 8 = 3

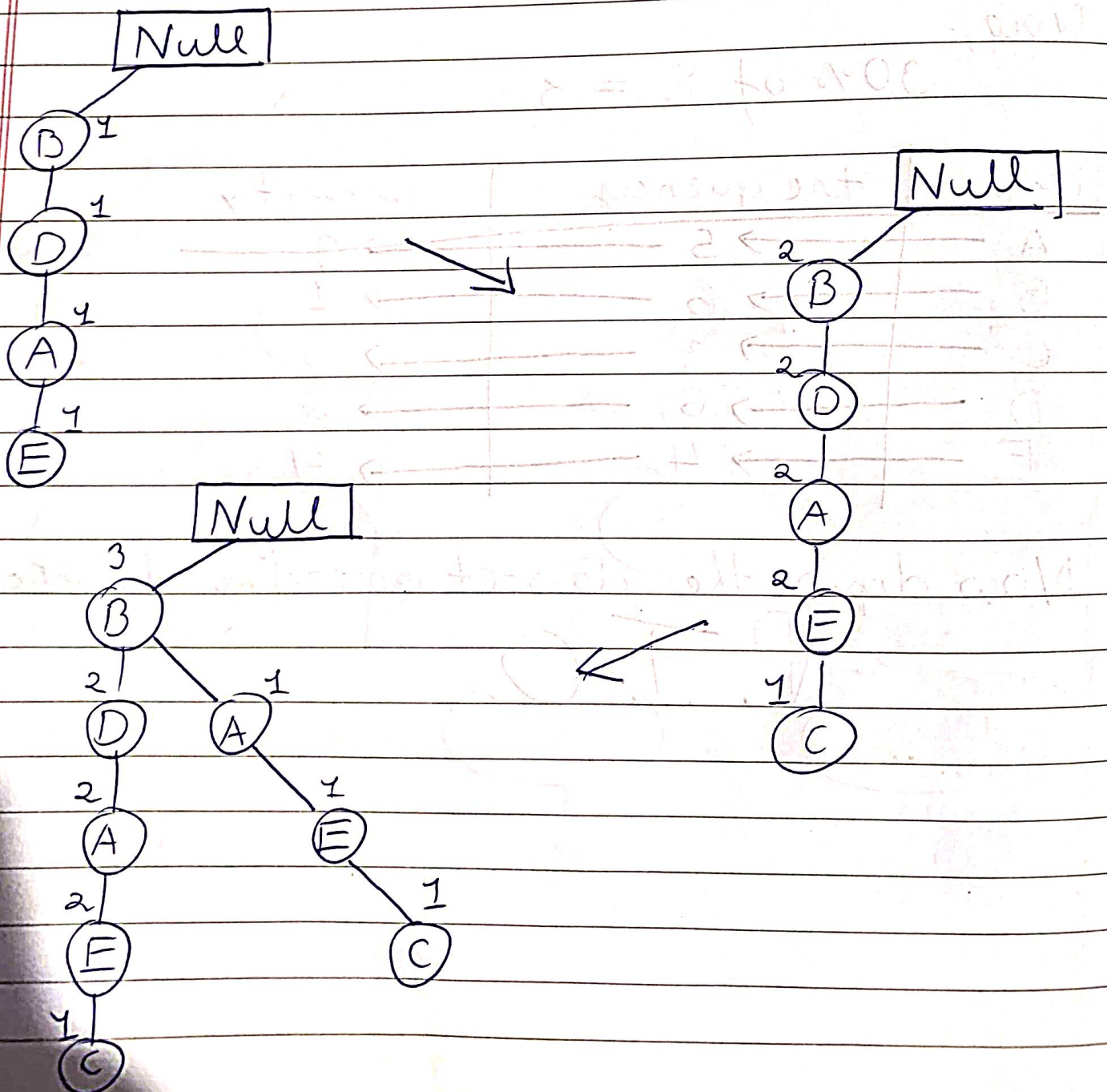
item	frequency	priority
A	5	3
B	6	1
C	3	5
D	6	2
E	4	4

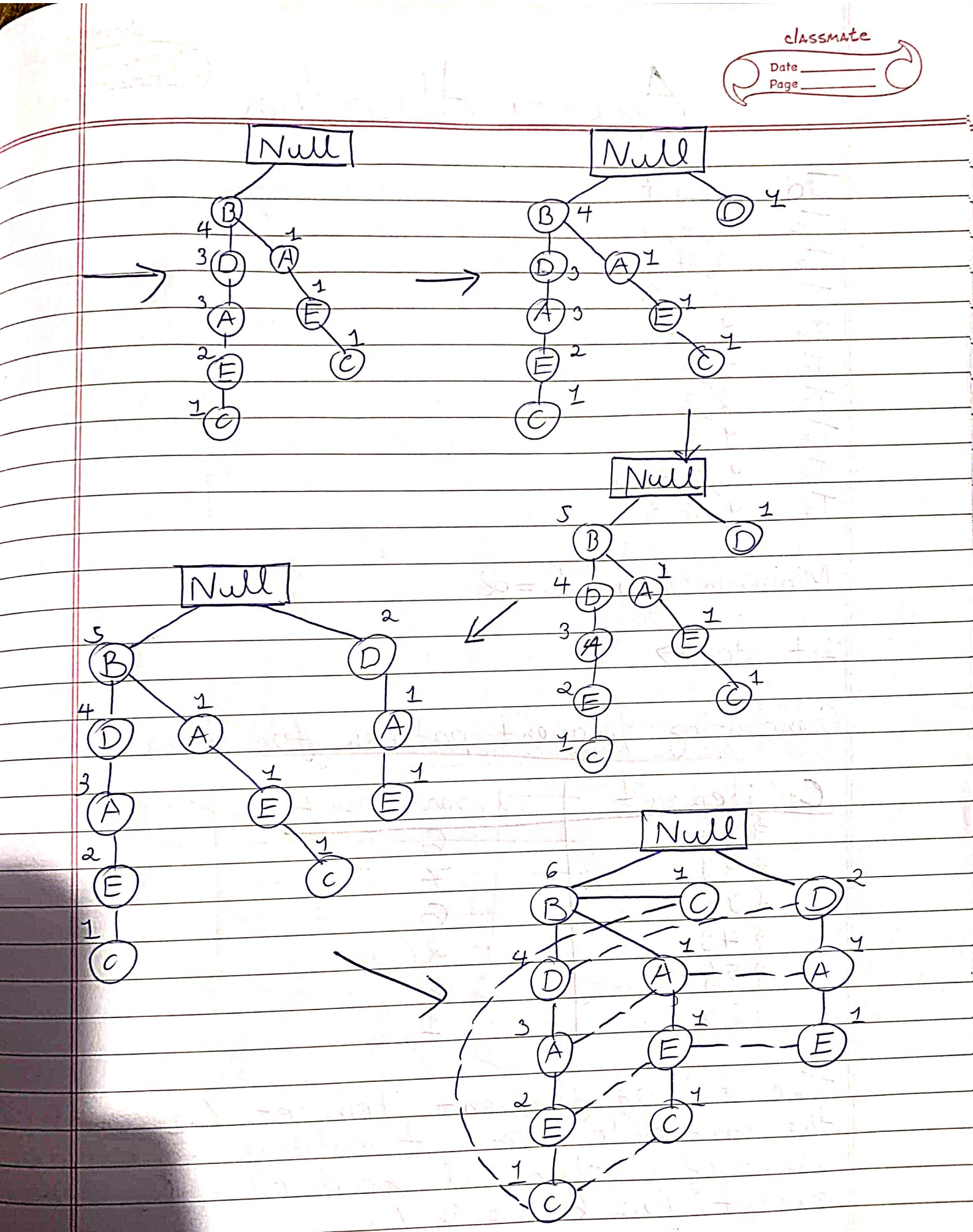
Now draw the dataset according to the priority-

P. T. O.

Tid	element
1	B, D, A, E
2	B, D, A, E, C
3	D, A, E, C
4	B, D, A
5	D
6	B, D
7	D, A, E
8	B, C

In FP growth tree, the root is always null




Null
B
D
A
E
C
Null
B
D
A
E
C
Null
B
D
A
E
C
Null
B
D
A
E
C
Null
B
C
D
A
A
E
E

Find F.P ~~graph~~ growth tree from following data set -

Tid	items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

Sol^{n}:

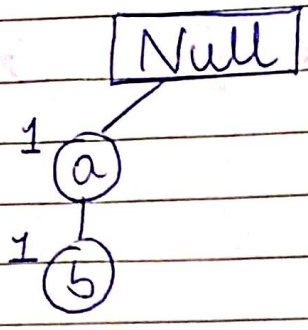
Item	frequency	Priority
a	8	1
b	7	2
c	6	3
d	5	4
e	3	5

Now draw the dataset according to the priority.

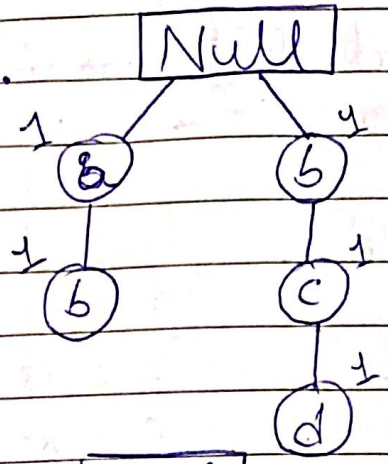
It would same.

Tid	items
1	
2	
3	
4	
5	
6	
7	
8	

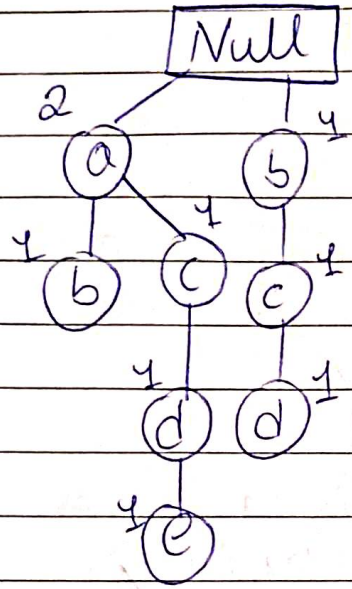
Step 1



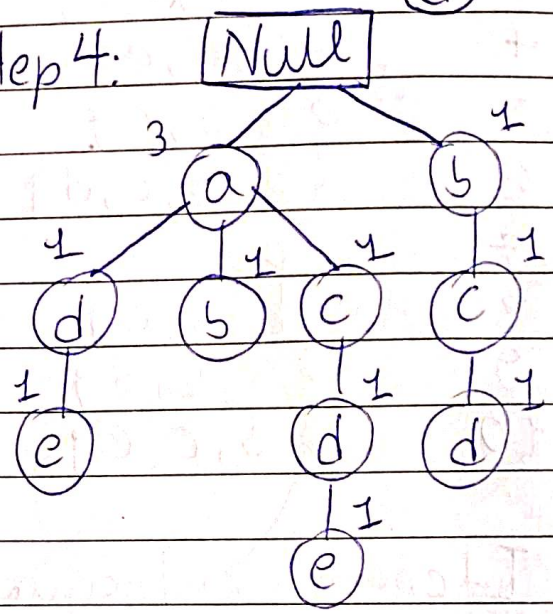
Step 2.



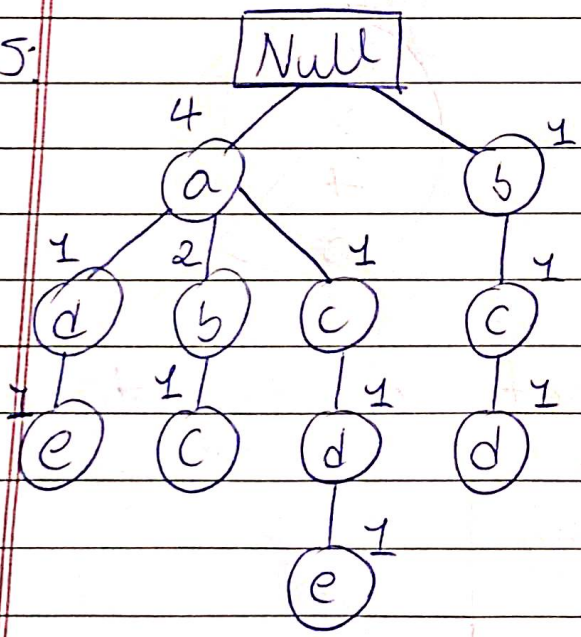
Step 3:



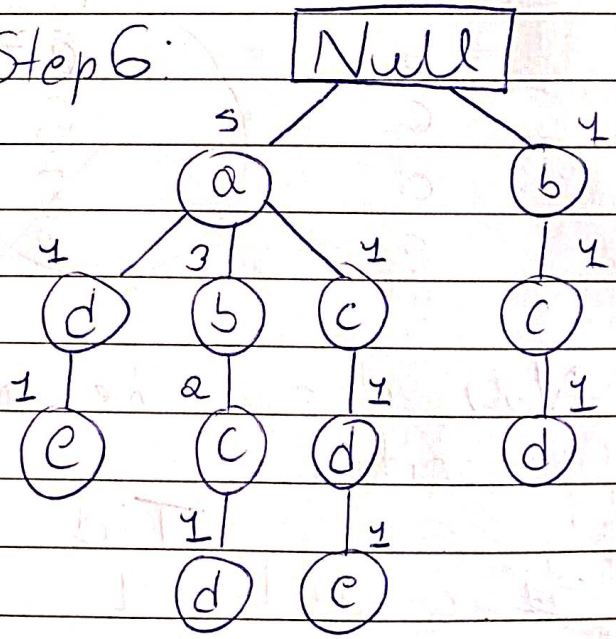
Step 4:



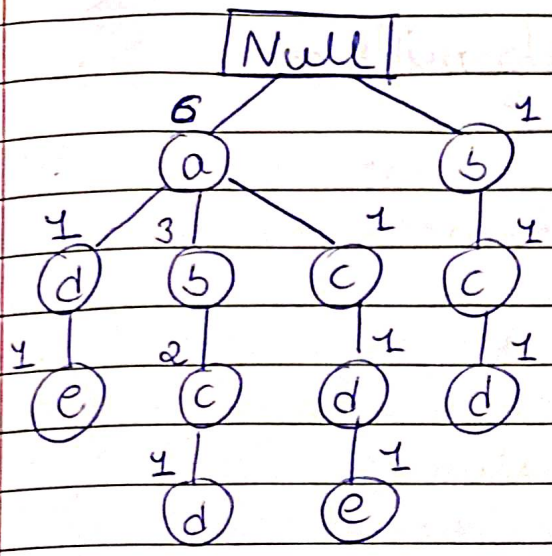
Step 5:



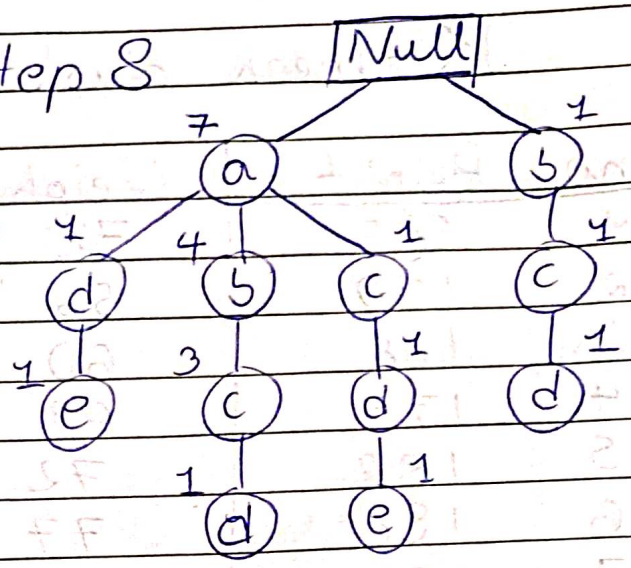
Step 6:



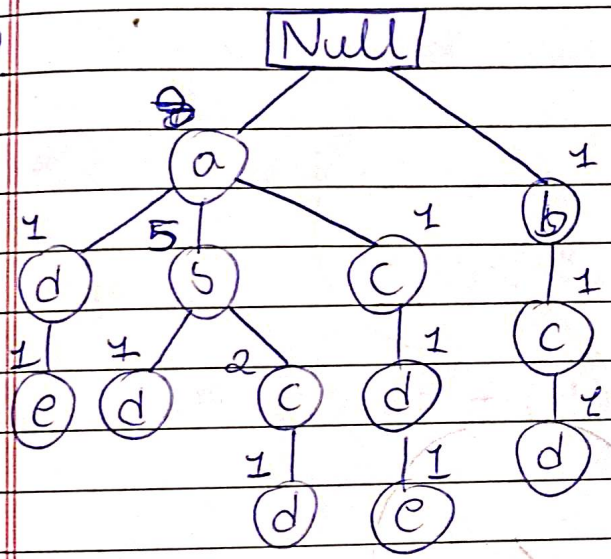
Step 7:



Step 8



Step 9



Step 10