Seq 4

Data warehouse → store and analysis of multiple data.

It is a subject oriented collection of data in support of management decession making process.

(management decision make এর জন্য ব্যবহৃত)

Subject oriented Data Warehouse.

↳ organized around major subject such as customer, product, sales.

↳ focus এর মধ্যে decision makers এর দেয়া তথ্য

Data warehouse integrated → constructed by integrating multiple, heterogeneous data sources

(data integration, cleaning are applied)

Data warehouse Time Variant: time horizon (duration) for this is significantly longer than that of operational system.

↳ historical data provide (5-10) years.

Data Warehouse Non-volatile:

Ly means that One data is entered. ઇt is not changed.

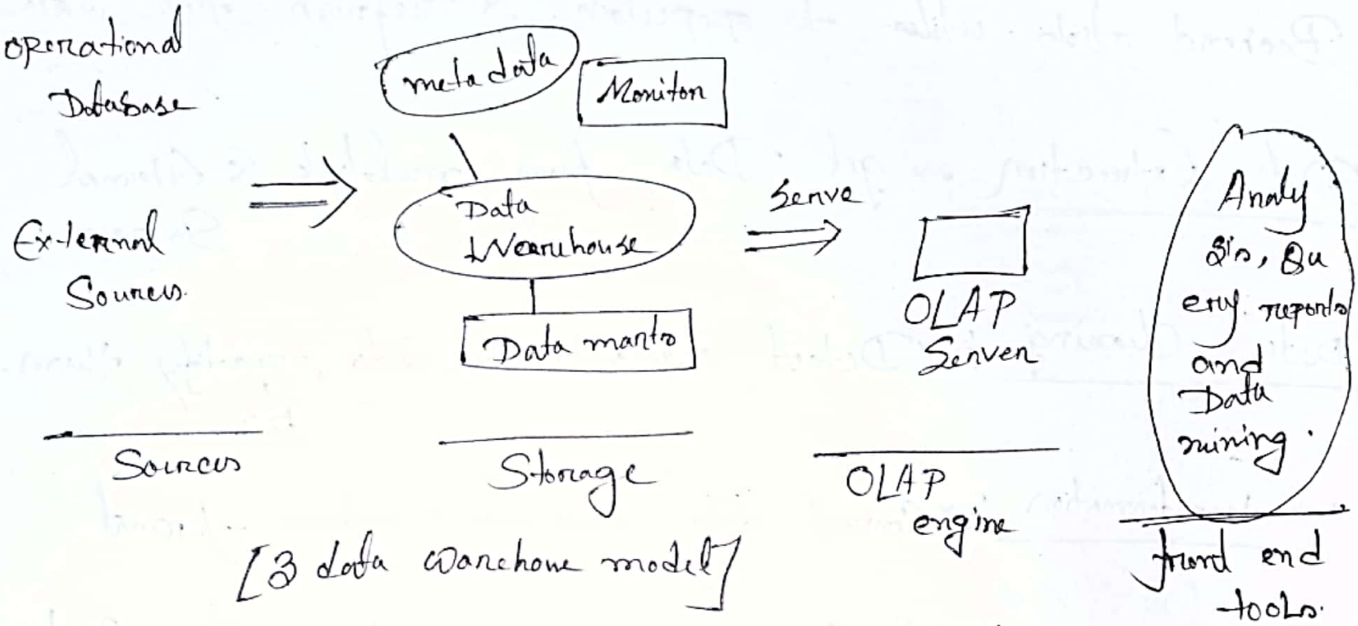
→ ઇt contains integrity, Consistency.

→ Database data daily update થાય છે પણ data Compare decision makes better થાય

OLTP vs OLAP

OLAP	OLTP
① Online analytical processing	① Online Transaction processing
② Consists of historical data	② Consists of only current data
③ Makes use of Data Warehouse	③ It makes uses DBMS.
④ Subject oriented used for Data mining, Decision making.	④ Application oriented used for business tasks.
⑤ It Serves extract info for analysis	⑤ Serves the purpose to insert, del update info from database
⑥ Only read and rarely write operation.	⑥ Both read and write Operation.

Multi tier Data warehousing Architecture



[3 data warehouse model]

Data Mart : Smaller, focused version of a data designed for specific department -> coustomer, item, sales analyze করে

Virtual Data warehouse : Logical data warehouse that provides view of the physically storing.

(budget friendly soln find)

[ঐ Database থেকে virtual Data Warehouse].

Enterprise Data warehouse : provide constant data source for the entire data organization.

→ Integrate from multiple sources

→ Supports enterprise-wide data analysis.

Ex একটি sales company sales, and custom data analyze করে All region এ

Backhand tools

Backend tools utiles to populate & refresh their data.

Data Extraction → get Data from multiple & External Sources.

Data Cleaning :→ Detect errors in the data, rectify them.

" transformation :→ Convert data host → Warehouse format

" Loading : Sort, Summarize, Compute views, check integri

" refresh : Update from the data Warehouse

Meta data Repository

It provides detailed info about the structure, defination and attributes of the data within a system.

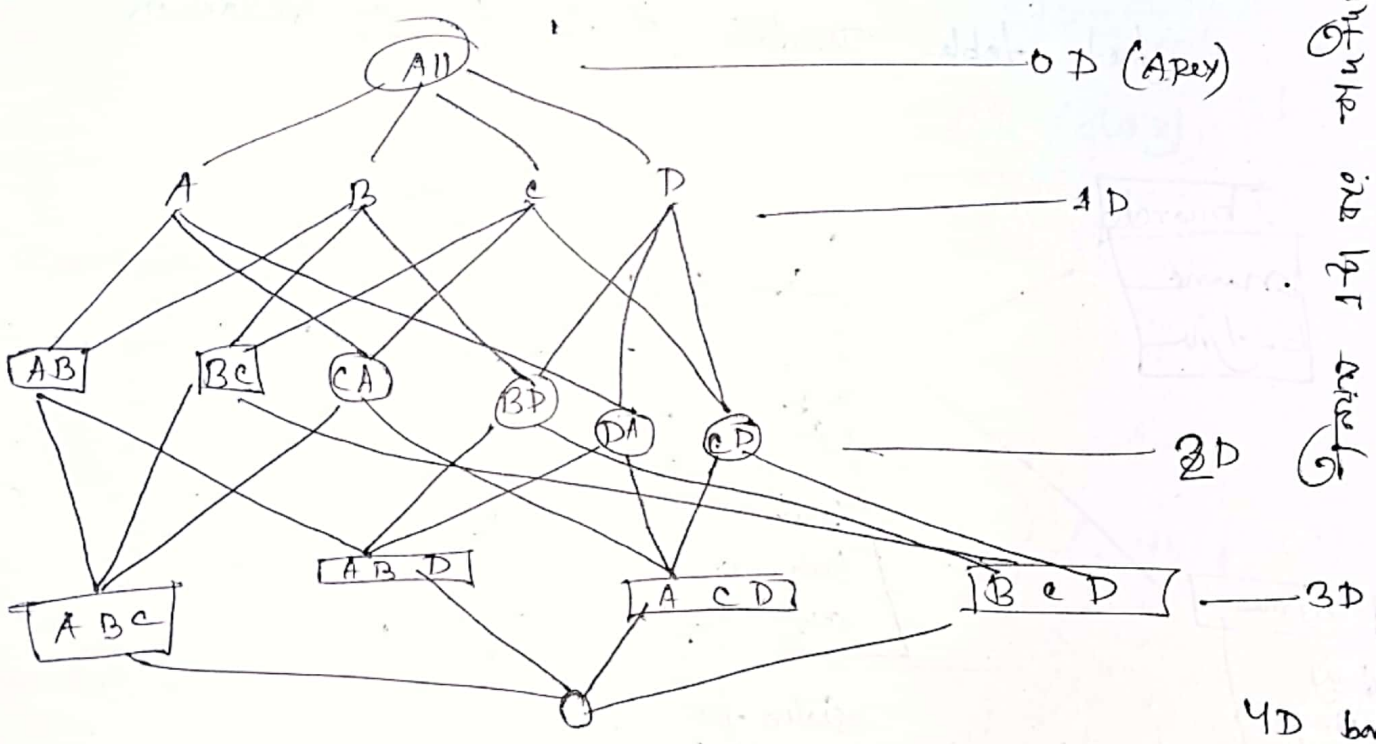
It includes details like data source, types, relationship, understand and manage more effectively.

Data cubes

Multidimensional data model views data in the form of cube

— Literature, an n-d base called a base cuboid

→ Top most 0-D cuboid hold highest level summarization called apex cuboid.



Data cube (Lattice of cuboids)

table

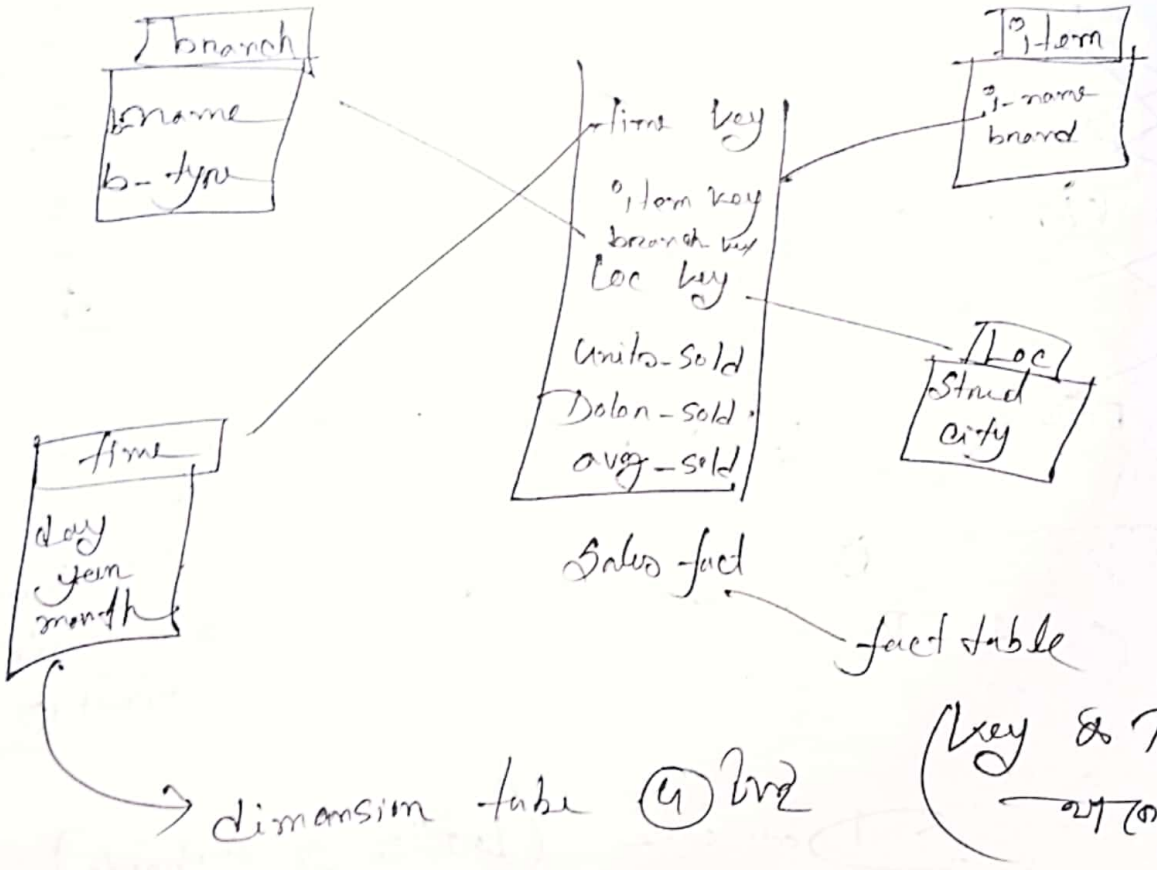
Dimension → item sale, item_name or time (day week)

fact table → contain measure (dollars sold)

keys of each of the related dimension tables

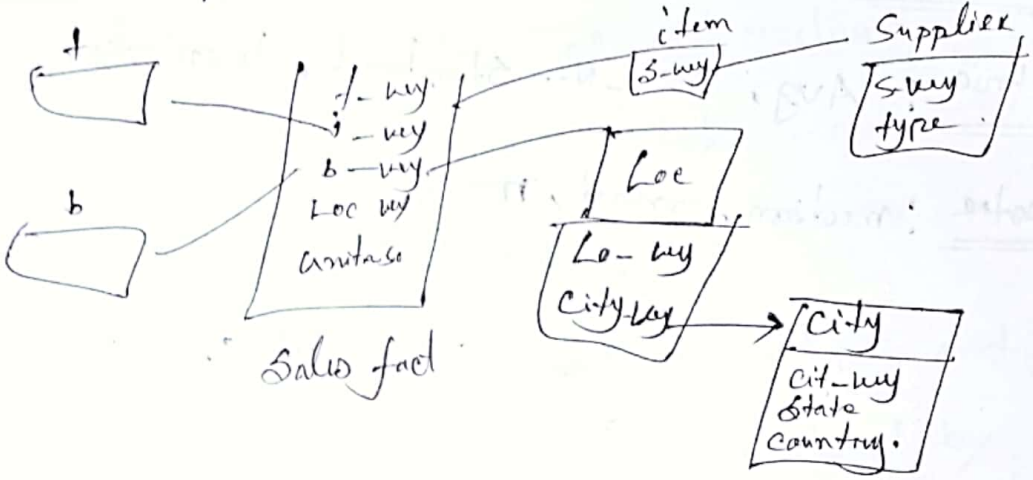
Star Schema

fact table middle of the 5d of dimension



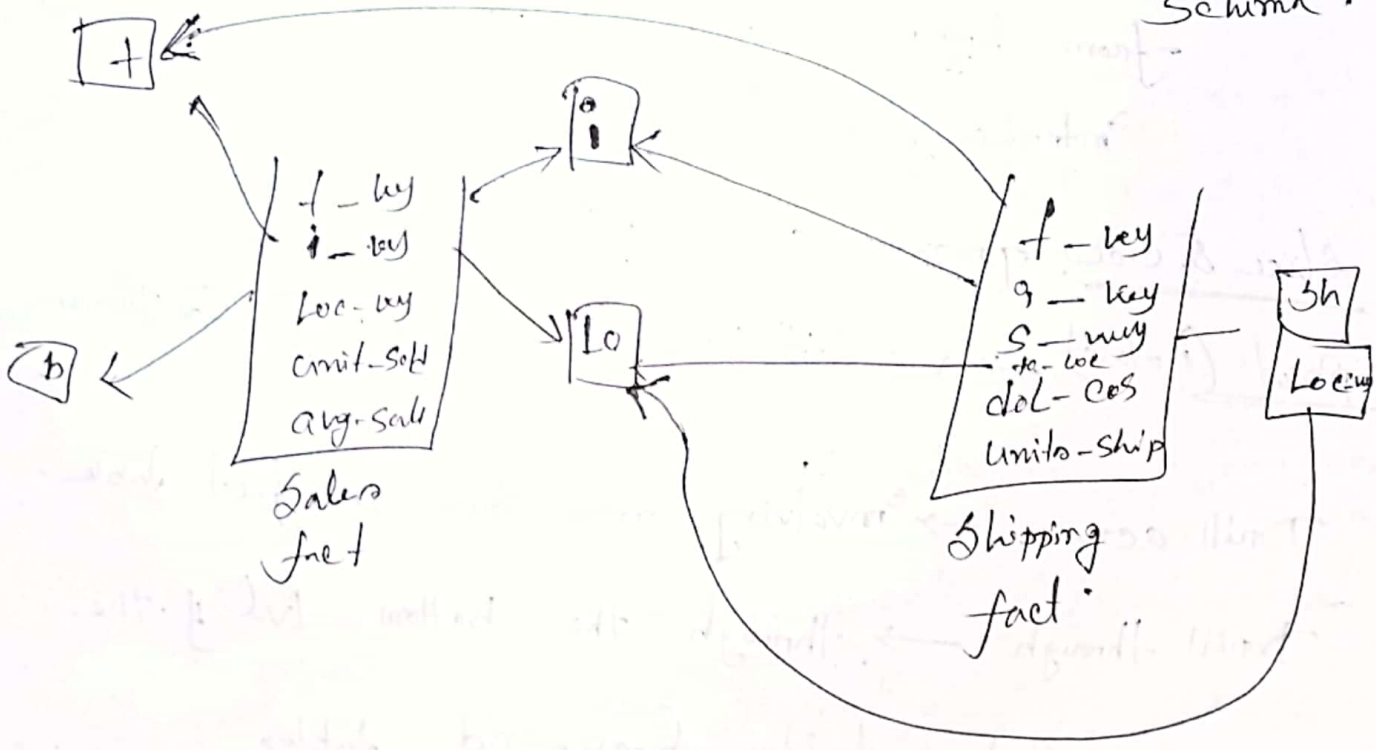
Snow-flake Schema

Star Schema is normalized into a set of smaller dimension tables and shape similar to snowflake.



fact_constellation / galaxy Schema

multiples fact tables share dimension tables. called galaxy Schema.



Data Cube Measure

Distributive: Count, sum, min, max

Algebric: Avg, Min_N_ Standard deviation.

Holistic: median, mod, rank

Typical OLAP Operation

roll up — Summarize data ← dimension reduction.

Drill down (rool down) → reverse of roll up.

- from higher lvl summary to lower lvl summary
- introducing new dimention.

Slice & dice: project and select

pivot (rotate) → visualization, 3d to series of 2D planes

Drill across → involving more than one fact table

Drill through → through the bottom lvl of the cube to its back-end tables

Design of Data Warehouse

4 views regarding the design.

Top Down view: Allows selection of the relevant info necessary for the data warehouse. Info matches current & future needs

Data Source view: The info captured by stored and managed by OP system. Info modeled by CASE and entity relationship model tools

Data Warehouse View: Consists of fact & Dimension table. Represents info that stored in the warehouse.

Business Query View: See the prospective of Data from end user's perspective.

Design process

(*) Top Down, bottom up approaches or a Combination of both.

- Start with overall design and planning
- Starts with prototype & experiments (rapid)

(*) From DW org point of view:

Waterfall: Structured and systematic analysis at each step before proceeding to the next.

Spiral: Increased fun system, short turn around time

(*) Typical Data Warehouse design process:

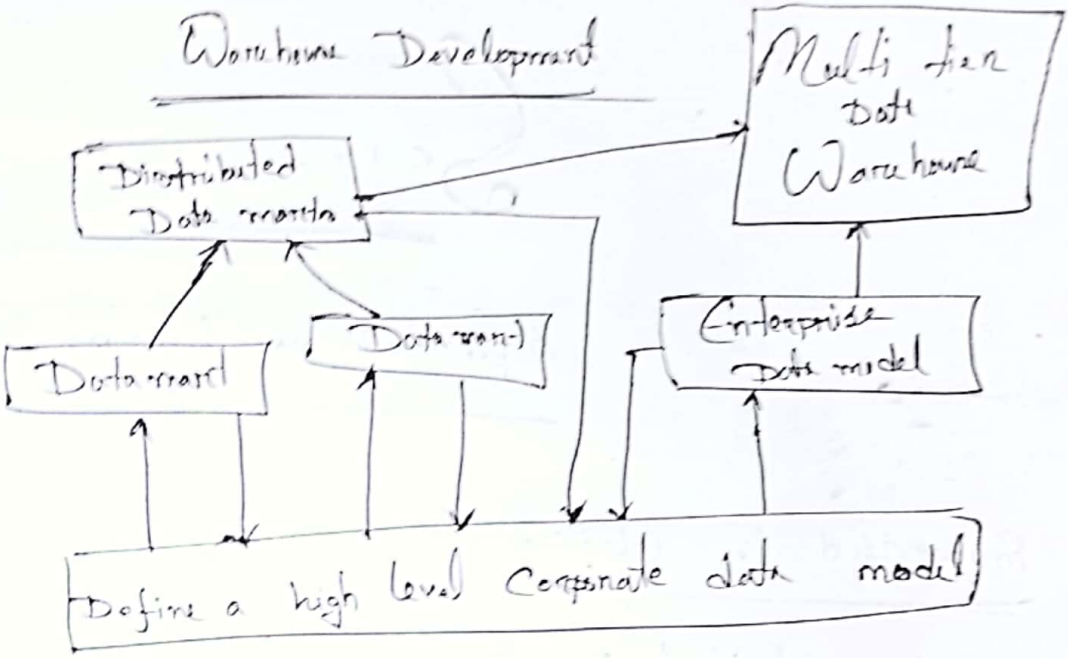
→ Choose a business process to model → sales, invoices

→ Choose a atomic Lvl data of buss. proces → individual transaction, individual daily snapshot.

→ Choose a dimension that will apply to each fact table record → typical dimension are item, customer, supp

→ Choose a measure that will popolate each fact tabl like Dolors sold, units sold.

Warehouse Development



Data Warehouse Usage

3 kinds of Warehouse application.

① Information processing

- Supports querying, basic statistical analysis, reporting tables, charts and graphs.

② Analytical processing

- Multidimensional analysis of Data Warehouse data.
- Support basic OLAP operation, drilling, pivoting

③ Data Mining

- Knowledge discovery from hidden patterns.
- performing classification and prediction presenting the mining results using visualization tools

Seq-5

Supervised vs Unsupervised Learning

Supervised is classification

Model is trained on labeled data.

Goal is predict output for new input.

Ex - Linear Regression, Logistic Regression, SVM, NN

unsupervised (clustring) → groping similler data.

where model works with unlabeled data.

Algorithm tries to find patterns, structures or relationship

Ex → K-mean clustering, Autoencoders.

⟹

* Classification is discrete catagories (yes or no)
* Numeric prediction is continious Numeric value

Classification	Numeric Prediction
→ Accuracy, recall, precision	→ Mean squared, R^2 score
- predict new inputs output	→ Predict unknown or (Loss appeared missing values.

Classification 2 step

① Model Construction (Training)

- Build a model using a labeled training set.
  - * Model represents as classification rules, d-t on mathematical formula.
  - * Data consists of samples with predefined class labels.

② Model usage (Testing) → prediction.

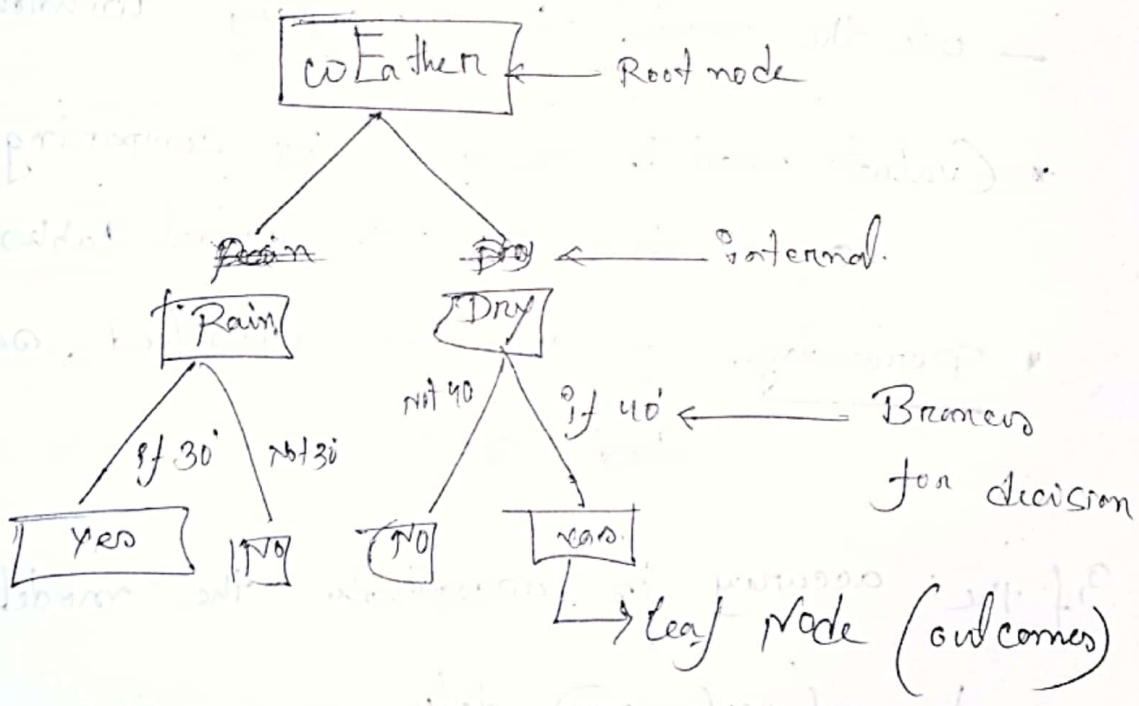
- use the model for classifying unknown data.
  - * Evaluate model's accuracy by comparing prediction on a test set with actual tables.
  - * Percentage of correctly classified samples in the test set.

If the accuracy is acceptable the model is used to classify new data.

If test set used for model selection it is called a validation set. (test set)

Decision Tree

It is a flowchart structure where each internal node represents a test on an attribute. each branch represents the outcome of a test, each node represents a class label (classification) or value (regression)



DT Example

Day	Wather	Temp	Humidity	wind	Play
1	S	H	H	W	N
2	S	H	H	S	N
3	C	H	H	W	Y
4	R	M	H	W	Y
5	R	C	N	W	Y
6	R	C	N	S	N
7	C	C	N	S	Y
8	S	M	H	W	N
9	S	C	N	W	Y
10	R	M	N	W	Y
11	S	M	N	S	Y
12	C	M	H	S	Y
13	C	H	N	W	Y
14	R	M	H	S	N

(+) Yes ps
(-) " neg

Step 1 — Calculate Information Gain

Weather

- Entropy of entire dataset

here $y = 9$, $N = 5$, total = 14

$$\text{Entropy} = -J \log_2 J$$

$$S(+9, -5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94.$$

- entropy of all attributes

Sunny entropy $(+2, -3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

Cloudy " $(+4, 0) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$

Rain " $(+3, -2) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97.$

whole হতে অন্যগুলো: $\frac{\text{যা আছে}}{\text{total}}$ দিয়ে

Information gain Weather $=$ Entropy (whole data) $- \frac{5}{14}\text{Ent}(S) - \frac{4}{14}\text{ent}(C) - \frac{5}{14}\text{ent}(R)$

$= \sout{0.94 - (0.97 - 0 - 0.97)}$

$= 0.94 - \frac{5}{14} \times 0.97 - \frac{4}{14} \cdot 0 - \frac{5}{14} \times 0.97$

$= 0.2471.$

② calculate IG for temp

$S\{+9, -5\} = 0.94.$ ← whole Ent (Same always)

Ent of Hot $\{+2, -2\} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0$

" " Mild $\{+4, -2\} = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.91$

" " cold $\{+3, -1\} = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.81$

IG for Temp $=$ whole ent $- \frac{4}{14}E(H) - \frac{6}{14}E(M) - \frac{4}{14}E(C)$

$= 0.94 - \frac{4}{14} \times 1.0 - \frac{6}{14} \times 0.91 - \frac{4}{14} \times 0.81$

$= 0.029.$

③ Same for Humanidity

whole $\epsilon = 0.94$.

ⓔ High $(+3,-4) = 0.98$

ⓔ Normal $(+6,-1) = 0.59$

Ig for Humanidity = ϵ of whole $- \frac{7}{14}\epsilon(H) - \frac{7}{14}\epsilon(N)$

$= 0.15$

④ wind

whole $\epsilon = 0.94$

ⓔ Strong $(+3,-3) = 1.0$

ϵ ~~Normal~~ weak $(+6,-2) = 0.81$

Ig of wind = whole $\epsilon - \frac{6}{14} - \frac{8}{14} * \epsilon(\omega)$

$= 0.94 - \frac{6}{14} \times 1.0 - \frac{8}{14} \times 0.81$

$= 0.0478$.

Gain (S, weather) = 0.246 ⟶ ↑ so root node.

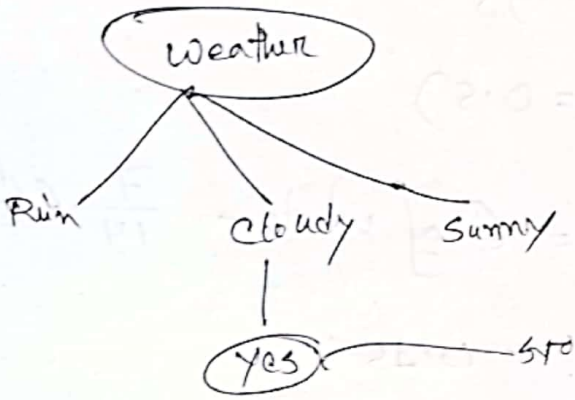
" (S, Temp) = 0.029

" (S, Humanidity) = 0.15

" (S, wind) = 0.0478

So weather node root.

cloudy, Rain, Sunny



so cloudy yes leaf node find

So Rain and Sunny এগুলো Ig ck.

Sunny. (weather Root so বাকী 3 টা)

Rain Ig

Temperature

Ent for Sunny $(+2,-3) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.97$

Ent (H) $(+0,-2) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log\frac{2}{2} = 0$

" (M) $(+1,-1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1.$

" Cool $(+1,-0) = -\frac{1}{1}\log\frac{1}{1} - \frac{1}{1}\log\frac{1}{1} = 0.$

Ig of Temp = whole E - $\frac{2}{5}$ Ent (H) - $\frac{2}{5}$ E(M) - $\frac{1}{5}$ E(C)

$= 0.57$

Humidity

(E) whole Sunny = 0.97

E (H) = 0

E (N) = 0

Ig = 0.97 - 0 - 0 = 0.97

wind

E whole sunny = 0.97

E (S) = 1

E (W) = 0.918

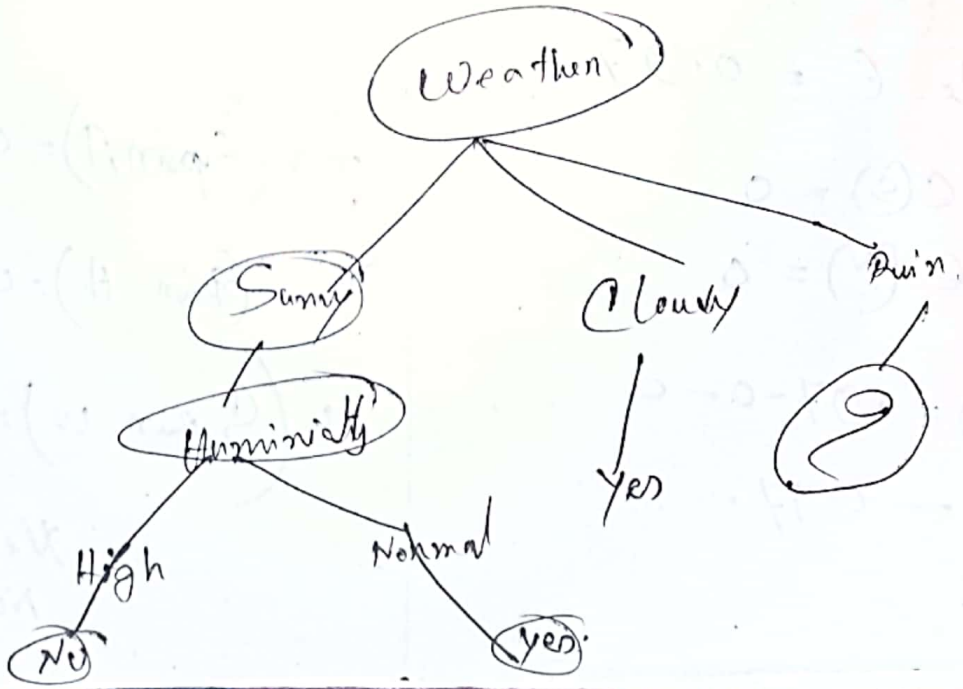
Ig (W) = 0.97 - 1 * $\frac{2}{5}$ - $\frac{3}{5}$ * 0.918

= 0.019

Gr (Sunny, T) = 0.57

Gr (Sunny, H) = 0.97 → to next

Gr (Sunny, W) = 0.019



For Rain

(E) Temp $= (+3-2) = 0.97$

$E(H) = 0$

$E(M) = 0.198$

$E(C) = 1.0$

Ig of Tem $= 0.97 - \frac{0}{5}.0 - \frac{3}{5}.0918 - \frac{2}{5}.1$

$= 0.019$

Humminidity

whole $E = 0.97$

$E(H) = 1$

$E(N) = 0.918$

Ig (Hum) $= 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0.918$

$= 0.019$

Coind

cohole $E = 0.97$

$E(S) = 0$

$E(M) = 0$

$Ig(W) = 0.97 - 0 - 0$

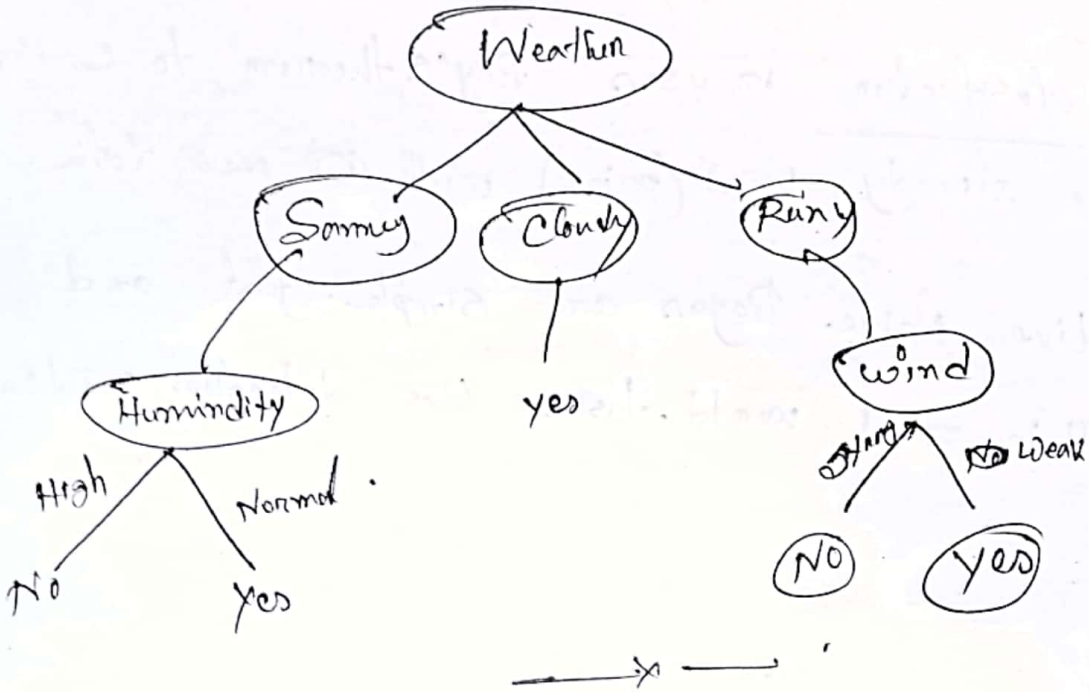
$= 0.97.$

gain (Sprain, T) $= 0.019$

" (Rain H) $= 0.019$

" (Sprain W) $= 0.97$

↑ Next Node wind



Measure impurity

$$\text{Entropy} = \sum_j -P_j \log_2 P_j$$

$$\text{Gini Index} = 1 - \sum P_j^2$$

$$\text{Classification Error} = 1 - \max\{P_j\}$$

Bayesian classification is uses Bay's theorem to combine what we already know (prior) with new data.

Methods like Naive Bayes are simple, fast, and work well in real world tasks like detection, medical diagnosis.

Bayes theorum

formulla = $P(H|X) = \dfrac{P(X|H)\,P(H)}{P(X)}$

$= P(X|H) \times P(H) / P(X)$

let X is unknown.

Naive → variables are independent

perform classification tasks.

→ feature data gula independent থাকে বলে তাই কোনো decision নিতে সাহায্যে Naive বলে

Naive bayes → classify করা হয়

Mal Naive Bayes

Days	outlook	Temp	Humidity	wind	Play
D_1	S	H	H	S	N
D_2	S	H	H	W	N
D_3	O	H	H	W	Y
D_4	R	M	H	W	Y
D_5	R	C	N	S	N
D_6	R	C	N	S	Y
D_7	O	C	N	W	Y
D_8	S	M	H	W	N
D_9	S	C	N	W	Y
D_{10}	R	M	N	S	Y
D_{11}	S	M	N	S	Y
D_{12}	O	M	H	W	Y
D_{13}	O	H	N	W	Y
D_{14}	R	M	H	S	N

O=S, T=C, H=H, W=S then P=? — Q?

probability play? yes? No?

Ans

$$play(yes) = \frac{9}{14} = .64$$

$$play(NO) = \frac{5}{14} = 0.36$$

Outlook	Yes	NO
	9	5
Sunny →	$\frac{2}{9}$	$\frac{3}{5}$
overcast →	$\frac{4}{9}$	$\frac{0}{5}$
rain →	$\frac{3}{9}$	$\frac{2}{5}$

Temp	yes	NO
Hot	$\frac{2}{9}$	$\frac{2}{5}$
cold	$\frac{4}{9}$	$\frac{2}{5}$
cool	$\frac{3}{9}$	$\frac{1}{5}$

Humidity	Y	N
High	3/9	4/5
Non	6/9	1/5

Wind

	Y	N
Strg	3/9	3/5
wea	6/9	2/5

$P(yes|X) = P(yes) \cdot P(Sunny|yes) \cdot P(cool|yes)$

$P(high|yes) \cdot P(strong|yes)$

$= \frac{9}{14} \cdot 2/9 \cdot 3/9 \cdot 3/9 \cdot 3/9$

$= 0.053$

$P(NO|x) = P(no) \, P(Sunny|NO) \, P(cool|NO) \, P(high|NO)$

$P(strog|NO)$

$= 0.206.$

here

$P(yes|x) < P(NO|x)$

So NO Doesn't play.

(*)

Age	income	credit	class
H	f.	L	Y
L	f	H	N
L	E.	H	N
L	E	H	N
H	f	H	Y

predict for [A = Low, Income = fair, Cr = Low] using Naive Baysin. may use Laplacian correction to abaid computing probability value of zero.

$P(Yes) = \frac{2}{5}$ $P(No) = \frac{3}{5}$

Age

	Y	N
H	2/2	0/3
L	0/2	3/3

income

	Y	N
f	2/2	1/3
E	0/2	2/3

Credit

	Y	N
H	1/2	0/3
L	1/2	0/3

Before Correction.

yes

$P(Age=L \mid yes) = 0/2$

$P(inc=fa \mid yes) = 2/2$

$P(cr=low \mid ye) = 1/2$

No

$P(Age=L \mid No) = 3/3$

$P(inc=fai \mid No) = 1/3$

$P(cr=low \mid No) = 0/3$

Abiod Computing 0 used.

Laplacian Correction

যেগুলো তে 2টা করে decision So $k=2$

and add (1) Count (উপরে)

After

for yes

$$P(Age = Low \mid yes) = \frac{0+1}{2+2} = \frac{1}{4}$$

$$P(ine = fa \mid yes) = \frac{2+1}{2+2} = \frac{3}{4}$$

$$P(Cr = Low \mid yes) = \frac{1+1}{2+2} = \frac{2}{4}$$

for NO

$$P(Age = Low \mid No) = \frac{3+1}{3+2} = \frac{4}{5}$$

$$P(in = fa \mid No) = \frac{1+1}{3+2} = \frac{2}{5}$$

$$P(Cr = Low \mid No) = \frac{0+1}{3+2} = \frac{1}{5}$$

So Abiod 0 Succesful

Now posterior probabilities

Yes

$$P(yes|x) = P(yes)\ P(A=low^{yes})\ P(in=f^{yes})\ P(c=low^{yes})$$

$$= 0.\frac{2}{5} \times \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{2}{4}.$$

$$= 0.0375$$

for=NO

$$P(NO|x) = P(NO) \cdot P(A=low)\ P(in=f)\ P(c=low)$$

$$= 0.6 \times 0.8 \times 0.2 \times 0.6$$

$$= 0.384$$

here $P(NO|x) > P(yes|x)$

Ans (NO)

Gaussian Naive Bayes

Person	Hight	weight	foot size
M	6	180	12
M	5.92	190	11
M	5.58	170	12
M	5.92	165	10
f	5.00	100	6
f	5.50	150	8
f	5.42	130	7
f	5.75	150	9

New Instance

H	W	ft
6	130	8

Ans

here $P(M) = \frac{4}{8} = 0.5$ | $female = \frac{4}{8} = 0.5$

male

mean (hight) $\mu = \frac{6.00 + 5.92 + 5.58 + 5.92}{4} = 5.855$

variance $\sigma^2(A) = \frac{(6-5.85)^2 + (5.92-5.85)^2 + (5.58-5.85)^2 + (5.92-5.85)^2}{4-1}$

$= 0.0350$

$\frac{\sum x_i - \mu}{n-1}$

Same way (c) [illegible] mean, Variance find out.

female weight

$$\text{mean} = \frac{180+190+170+165}{4} = 176.25$$

$$\text{Variance} = \frac{(180-176.25)^2 + (190-176.25)^2 + (170-176.25)^2 + (165-176.25)^2}{4-1}$$

$$= 122.92.$$

male foot size

$$\text{mean (FS)} = \frac{12+11+12+10}{4}$$

$$= 11.25$$

$$\text{Variance} = \frac{(12-11.25)^2 + (11-11.25)^2 + (12-11.25)^2 + (10-11.25)^2}{4-1}$$

$$= 0.91667$$

Female

Mean (H) = 5.417

Variance (H) = 0.097.

Mean (W) = 132.5

Variance (W) = 558.333

mean (FS) = 7.5

Variance (FS) = 1.6667.

Gaussian distribution

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(\text{Male} \mid \text{Hig}) = \frac{1}{\sqrt{2 \cdot 3.14 \times 0.035}} \times e^{-\frac{(6 - 5.85)^2}{2 \times 0.035}}$$

(6 — instance value)

$$= 1.578$$

Same way ∞

$$P(\text{Male} \mid w) = 5.98 \times e^{-6}$$

e ঘাত ঋণ (10^1) ধরে

$$P(\text{male} \mid fs) = 1.311 \times 10^{-3}$$

Female

$P(F \mid H) = 2.23e^{-1}$ | $P(F \mid w) = 1.67e^{-2}$ | $P(F \mid FS) = 2.86e^{-1}$

Posterior male = $\frac{P(M) * P(H \mid M) * P(w \mid M) * P(Fs \mid M}{\text{evidence} \rightarrow \text{unknown}}$

= হর বাদে = $6.1984e^{-9}$

value হর বাদে $e^{-6} \cdot e^{-3}$

$= e^{-9}$ ধরলে

Posterior female = $\frac{P(f) \cdot P(f/(H)) \, P(w/f) \; P(fs/f)}{\text{evidence}}$ — এই unknown.

$= 5.37 e^{-4}$.

here $P(f) > P(M)$

So test sample is classified as female.

Naive Bayesian classification known as class conditional independence

Because

Independence feature : easy decision নেওয়া যায় independent হলে

Simpler math : assumption as simpler math calculation.

fast prediction : allows quick and easy probability.

efficiency : one feature doesn't change another feature make it more efficient.

* easy implement

[Adv আকারে এগুলা লিখতে দিলে হবে)

Model Evaluation & Selection

Evaluation → How can we measure accuracy.

Methods for estimating accuracy:

- holdout method.
- Cross validation
- Bootstrap.

Confusion matrix ← represent accuracy of model

positive tuples (P)	Negative tuples (N)
main category yes হলে buys pc = yes (does buy)	There are all other categories buys pc = no (doesn't buy pc)

4-hour

True positive (TP): positive tuples → identified as positive

T → Buy pc = yes

P → He does buy pc

True Negative (TN): Negative tuples → correctly Negative

T → buy pc = no

N → he don't buy pc

False positive: Neg tupels ক positive বলেন দুটো (FP)

buy pc = yes, actualy don't buy.

False Negetive: positive tupes দুটো Neg-tive বলেন (FN)

buys pc = NO — actual buy pc.

Example

predicted

Actual class	+C	−C
+C	TP	FN
−C	FP	TN

Ac \ pred	−C	+C
−C	TN	FP
+C	FN	TP

Formalla for math

accuracy — $\frac{TP + TN}{P + N}$

error rate — $\frac{FP + FN}{P + N}$ ‖ (1 − accuracy)

Sensitivity, Tp, recall — $\frac{TP}{P}$

Specifeity, TN rate — $\frac{TN}{N}$

Precession — $\frac{TP}{TP + FP}$

F1 score — $\frac{2 \times \text{precession} \times \text{recall}}{\text{precession} + \text{recall}}$

(*) Math

	Predicted 0 (-)	Predicted 1 (+)	
Actual 0 (-)	90	10	= 100
Actual 1 (+)	10	0	= 10
	100	10	= 110 (P+N)

find accuracy, error rate, precession, recall

Ans

here.

$TP = 90$, $TN = 0$, $FP = 10$, $FN = 10$

যদি 0(+) 1(-) হয়

যদি 0 = (-), 1 = (+) হয়

$TN = 0$, $FP = 10$, $FN = 10$, $TP = 0$

math Solved by this.

$$Accuracy = \frac{TN + TP}{110} = \frac{90+0}{110} = 0.81 = 81\%$$

$$error\ rate = \frac{FP + FN}{P+N} = \frac{10+10}{110} = 0.18 = 18.1\%$$

(1 - accuracy) করলে হবে

$$Precession = \frac{TP}{TP + FP} = \frac{0}{0+10} = 0$$

$$Recall = \frac{TP}{TP + FN} = \frac{0}{0+10} = 0$$

Observations

- যদি precision, recall 0 হয় then → model fails to correctly identify any positive instance.
- accuracy high হলে → model correctly classifies a good proction of the overall data.
- যদি precision high হয় then → model's strong ability to minimize false positive
- recall high হলে → models capturn all actual buyers
  - low হলে → model missed many true positive cases.
- F1 Score high হলে → indicates Balanced performance.
  - low হলে → poor blance precision & recall needs improvement
- Accuracy high but recall low হলে তিন mis leading হয়েছে – Midical Section এটা ঘটবে

Class imbalance problem

when number of events in each class of dataset is not equal. Ex - fraud, spam detection

Dataset imbalance:

when,

- unequal cls distribution.
- if minority cls less than (10-20)% of the total data then dataset is imbalanced.

Measure the performance ~~for imbalanced Algorithm~~:

1. confusion matrix: TN, TP, FN, FP
   - TN: correct neg predict (N-N)
   - TP: correct pos predict (P-P)
   - FN: incorrect neg predict (P-N)
   - FP: incorrect pos predict (N-P)
2. precession: Measure accuracy of positive predictions.

$$\frac{TP}{TP+FP}$$

3. Recall: Measure the ability to identify actual positives

$$\frac{TP}{TP+FN}$$

F1 score , Specificity, → $\frac{TN}{TN + FP}$

(ability of predict true negative.

ROC : plots the TP rate against the FP rate at the various threshold settings.

AUC : Measure overall performance higher is better

Evaluting Classifier Accuracy.

Holdout & Cross validation Methods

Holdout 2 steps

Training sets → model construction

Test set → accuracy estimation.

Random Sampling

- coherce the data split multiple times
- final accuracy is the average all accuracies from this splits.

Cross validation : if k = 10 then split into 10, 9 parts are -train, 1 are -testing.

Leave one out cross validation : each point test other is training

Stratified cross " : distribution similer to the orginal set. (balanced evaluation)

(14)

Roc Curve

It is a graph showing the performance of a classification model at all classification thresholds.

True positive Rate (TPR) vs False positive Rate (FPR)

Auc (Area under Curve)

measure accuracy.

Closer to 1 → Highly accurate model.

" " 0.5 → poor model almost random prediction.