08-11-2025

Agenda:

- → Decision tree Regressor
- → Ensemble method
  - → Bagging & Boosting
- → Bootstrap dataset
- → Random forest
- → code implementation (DT-classifier)

Decision Tree classifier & Regressor

	Classification	Regression
output	class label	Numeric label
Impurity measure	entropy / gini	variance / MSE
leaf prediction	Majority class	Mean of values
split criteria	Information gain	variance reduction

Dataset

ID	Square feet	price (₹ in lakhs)
1	800	50
2	850	52
3	900	54
4	1500	90
5	1600	95
6	1700	100

Step 1: Compute parent variance (price)

$$\mu(\text{price}) = 73.5 \qquad \text{var} = \frac{1}{n}\sum (y_i - \mu)^2$$

$$= \frac{1}{6}\sum\left((52-73.5)^2 + (\cdots)\right)$$

$$= \frac{2831.5}{6}$$

$$= 471.9$$

Step 2: For each feature → Compute below

→ we have only one feature:

→ we pick the "square feet" column:

→ sort values:

800, 850, 900, 1500, 1600, 1700

→ split point (midpoint between value):

825, 875, 1200, 1550, 1650

→ we test & compute variance Reduction (VR)

split at 1200 :

square feet ≥ 1200

← left right →

800	50
850	52
900	54

μ(left) → 52

var(left) → 8/3 → 2.67

1500	90
1600	95
1700	100

μ(right) → 95

var(right) → 50/3

→ 16.67

weighted variance :

$$\text{var}_{\text{split}} = \frac{3}{6} \times 2.67 + \frac{3}{6} \times 16.67$$

$$= 1.335 + 8.335 = \underline{9.67}$$

471.9 → 9.67

variance Reduction: 471.9 − 9.67

= 462.23

Sqft < 1200 ?

Yes ↙ ↘ NO

800	50
850	52
900	54

Step 1 :

midpoint : 825, 875

split 825 :

↙ ↘

800, 50 | 850, 52 ; 900, 54

↓ | ↓

var(left) | var(right)

↓ | ↓

0 | 1

w. variance $= \frac{1}{3} \times 0 + \frac{2}{3} \times 1$

$= \frac{2}{3} = 0.67$

V.R $= 2.67 - 0.67$

$= 2.0$

1500	90
1600	95
1700	100

↓

midpoint : 1550, 1650

split : 1550

↙ ↘

1500, 90 | 1600, 95 ; 1700, 100

↓ | ↓

var(left) | var(right)

↓ | ↓

0 | 6.25

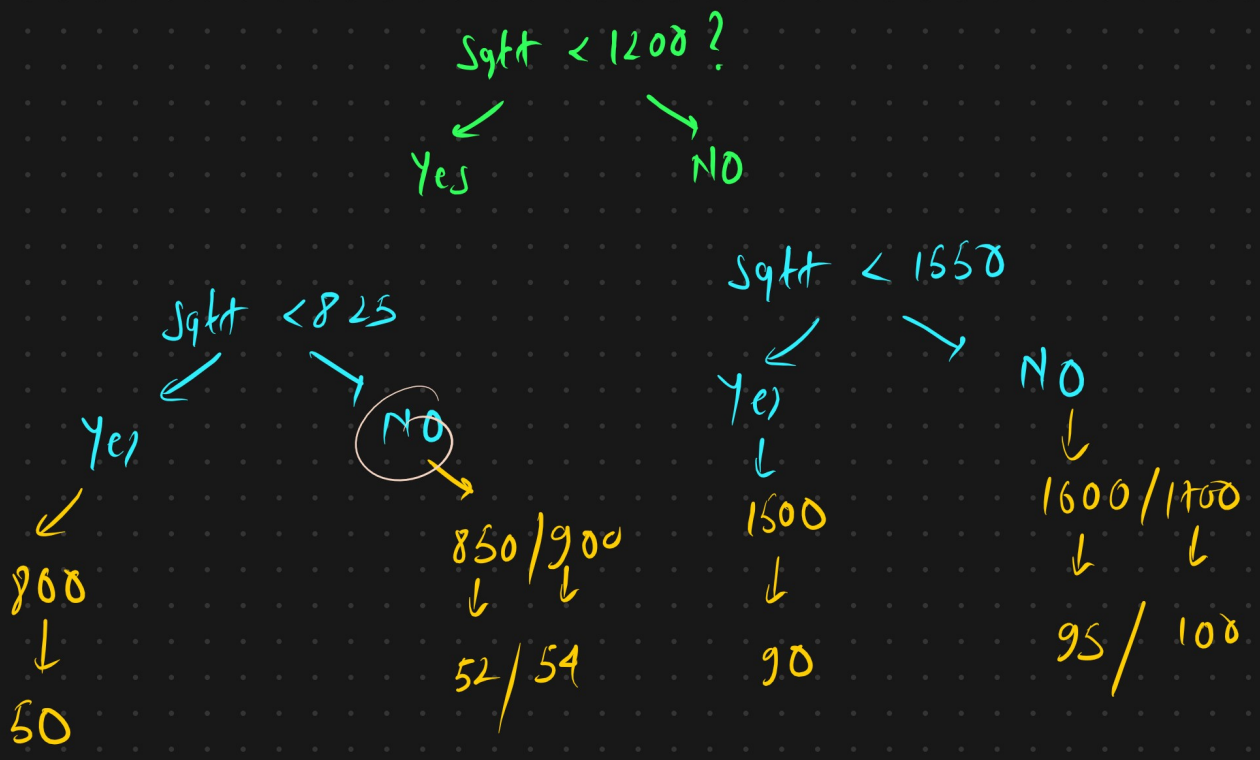
w. variance :

$\frac{1}{3} \times 0 + \frac{2}{3} \times 6.25$

$= 4.16$

VR : 16.67 − 4.16

: 12.51



min_samples_leaf = 1

min_samples_split = 2

A node must have atleast 2 samples to even consider a split

⚠ Disadvantages

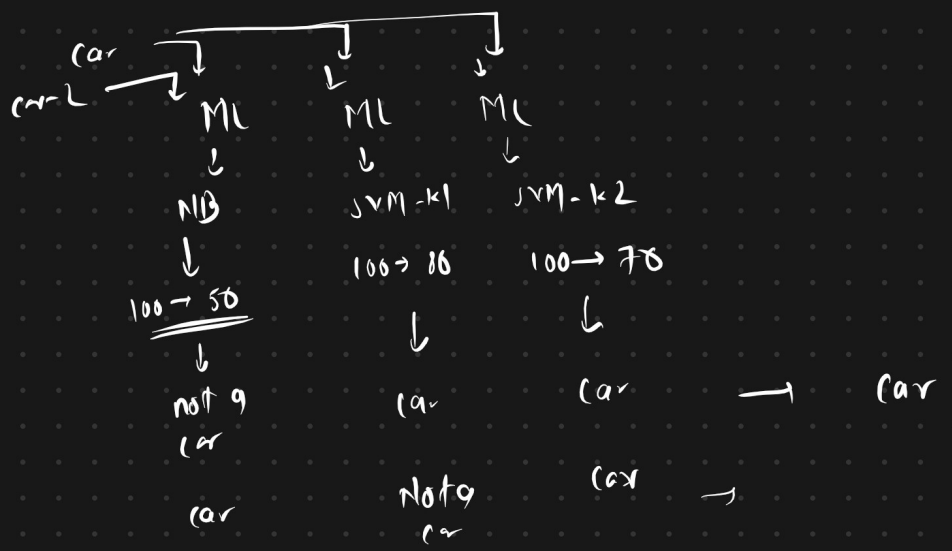
❌ Overfits easily (deep trees)
❌ Unstable (small data changes → big tree changes)
❌ Not smooth — predictions are piecewise constant

Ensemble method / learning / technique:

Car detection / classification

Image ← ML → CAR
→ CAR AT Night / Evening
→ particular color of car
→ Model of car

} Not able to detect

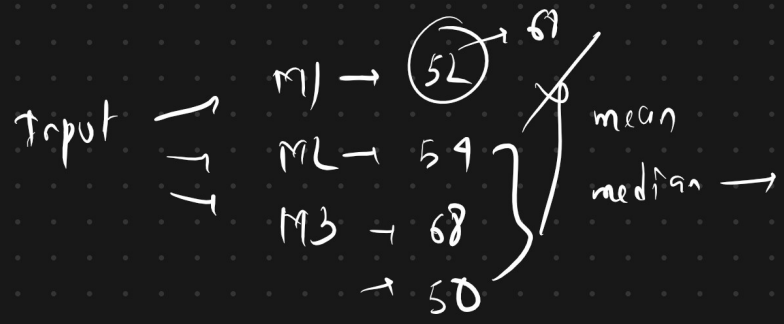


predict → value of a property → continuous value.

linear regression (5 feature) → ML → 96% → 4%

linear regression (8 feature) → ML → 96% → 4%

SVM-k2 (5 fea.) → ML → 97% → 4%

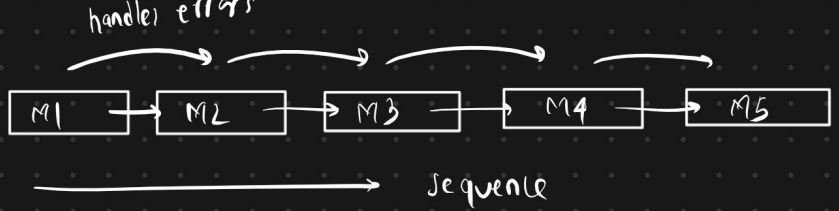


Technique:

Bagging

Aggregate → majority
→ average
→ median

Boosting



Bootstrap :

→ Overfitting

X	y
1	200
2	300
3	400
4	500
5	600
6	700

→ Dataset → DT →

B.T=2

1	200
2	300
2	300
4	500
5	600
1	200

→ set 1 → DT → Tree → 3/6

3, 400
6, 700
(OOB data)

→ generalized

1	200
2	300
3	400
5	600
3	400
6	700

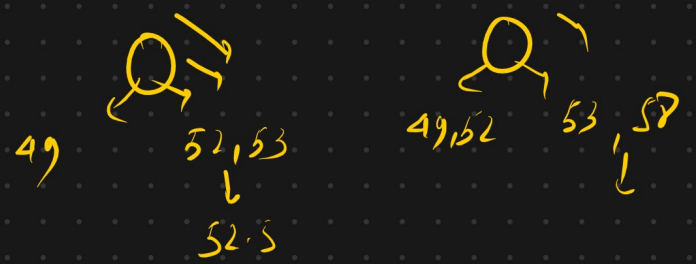
→ set 2 → DT → tree → 4

4, 500
(out of bag data)

when duplicated are allowed : sampling with replacement

BT : 100 → DT → DT

or



BT.Dataset → DT → DT

RF(=100, 200, 300

```
RF( X_train, y_train, n_estimator = 100 )
    1   → B.Data-1   → OOB_data_1   ←
DT  2
    3   → B.Data_3  →
    100 → B.Data_100 → OOB_data_100 ←
```

test the model on OOB data

RF is a self-validated model.

Random forest step by step

max_features = 2
: each split consider a random subset of 2 features.

ID	Sqft	Bedroom	Age(year)	price(lakhs)
A	800	2	10	60
B	900	2	8	70
C	1000	2	7	80
D	1500	3	3	90
E	1600	3	2	95
F	1700	4	1	100

n_estimators = 3, max_features = 2
: each split in a tree will randomly consider 2 of the 3 features.

Step 1: Create Bootstrap samples.

Tree 1 Bootstrap dataset

B(900, 70)
C(1000, 80)
A(800, 60)
B(900, 70) ← repeated
E(1600, 95)
F(1700, 100)

Tree1. OOB: D(1500, 90)

Tree 2 Bootstrap data:
- Data
- OOB

Tree 3 Bootstrap data:
- Data
- OOB

Step 2: for each tree:
- Random feature selection [Sqft, Bedroom, Age]
- Sqft, Age (2 feature)
  - calc parent variance
  - sort data & find mid point.
  - for each mid point:
    - split from mid point
    - Calculate variance of right & left side
    - calc weight variance
    - calc variance reduction
  - find the best mid point using the highest variance reduction.

→ Final prediction

D.T → $f_i(x) \rightarrow \hat{y}$

$$\hat{y}_{RF}(x) = \frac{1}{3}\sum_{i=1}^{3} f_i(x)$$

Sqft = 1200
Bedroom = 3
Age = 4

$f_1(x)$: 97.5
$f_2(x)$: 93.75
$f_3(x)$: 90

$$\hat{y}_{RF}(x) = \frac{1}{3}(97.5 + 93.75 + 90)$$

$$= \underline{93.75}$$