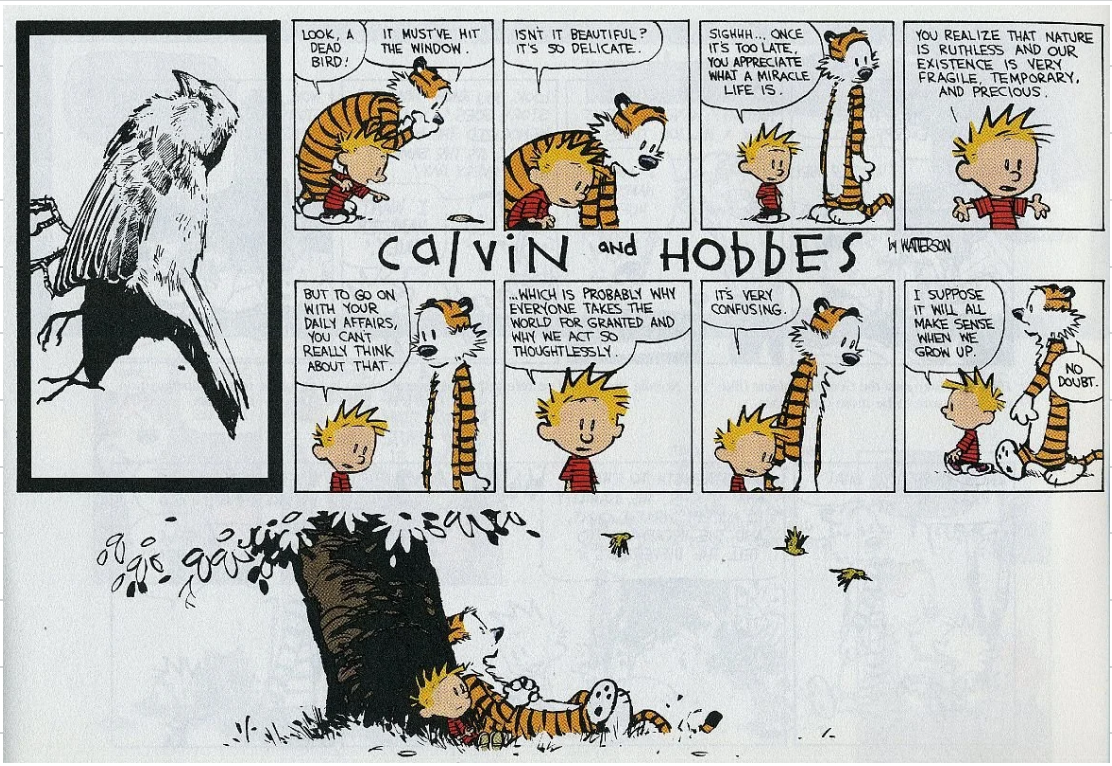
Session-4

Decision Tree - 2

Sep 01, 2025



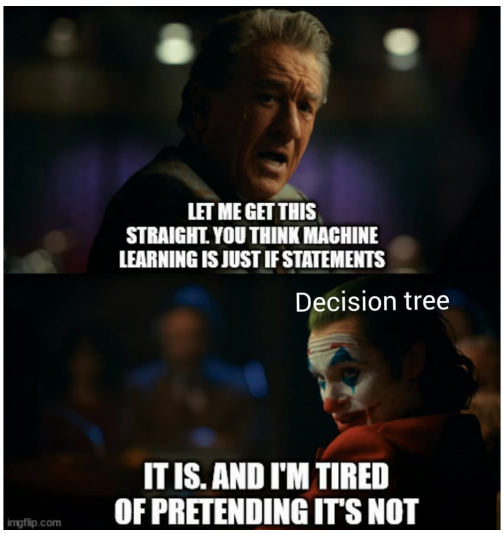
AGENDA

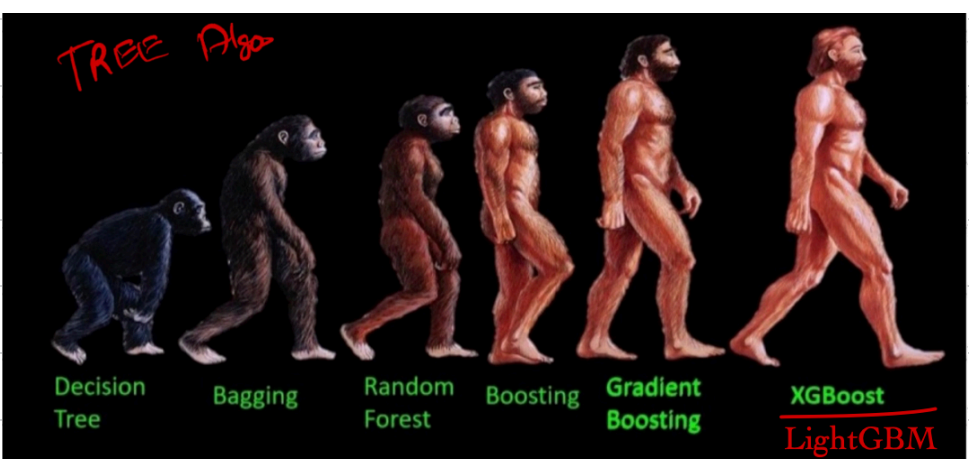
1. Recap for K-Fold cross validate
2. Algorithm for Decision trees
3. Gini Impurity
4. How to split on numerical features
5. Overfit Vs Underfit
6. Hyper - Parameter Tuning
7. Feature Importance
8. Regression using DT

Algorithm for Decision Trees

What an Amoeba can do ?

* sense many cues (food chemicals, toxins, temperature, stiffness),
* integrate those signals,
* change strategy (e.g., wander randomly until it detects a gradient, then move up it),
* keep short-lived "memories" in its biochemistry (previous exposure changes later responses),
* and navigate messy environments by reorganizing its cytoskeleton on the fly.





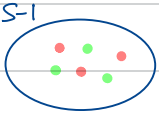
Gini Impurity

$$\text{Entropy} \rightarrow -\sum_{i=1}^{n} P(y_i) * \log(P(y_i))$$

$$GI(Y) = 1 - \sum_{i=1}^{K} P(y_i)^2$$

GI is more sensitive to heterogeneity

S-1

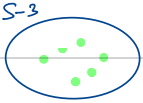


$P(R) = 3/6$

$P(G) = 3/6$

= 1 - ((3/6) ^ 2 + (3/6) ^ 2) =0.5

S-3



$P(R) = 0$

$P(G) = 1$

$= 1 - (0^2 + 1^2)$

$= 0$

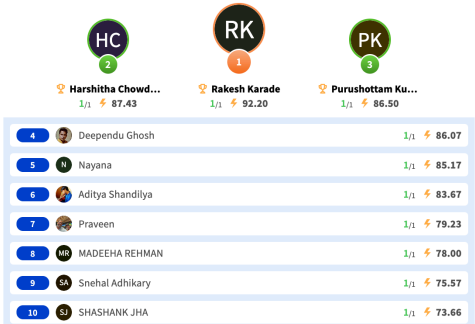
A decision tree model uses Gini impurity as the splitting criterion. If a node has 60 instances of class A and 40 instances of class B, what is the Gini impurity at that node?

0 users have participated

A	0.24	0%
B	0.4	0%
C	0.6	0%
D	0.48	0%

End Quiz Now

Based on all quizzes from the session



(1-(0.6 ** 2) - (0.4) ** 2)

$= (1 - (0.6^2 + 0.4^2))$

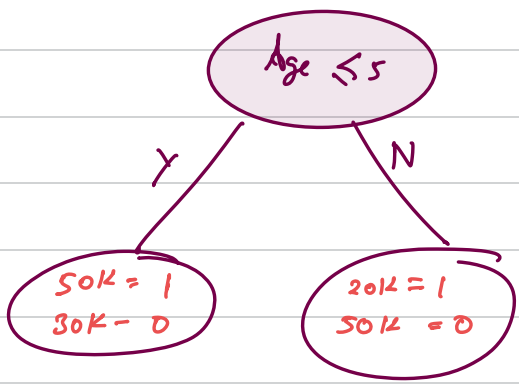
Algorithm of Decision Trees

1. For all the features in the dataset, and every single value in each of the features, compute information gain.
2. Create a split based on the feature/value pair, which has the highest information gain.
3. Repeat 1-2 until you've achieved max-depth.

# of rows = 150k

1 – Sold
0 – Not Sold

Age	Mileage	Sold / Not
5	14	1
7	20	0
⋮	⋮	⋮
2	16	1



$1 - \left(\frac{50}{80}^2 + \frac{30}{80}^2\right)$ 0.47

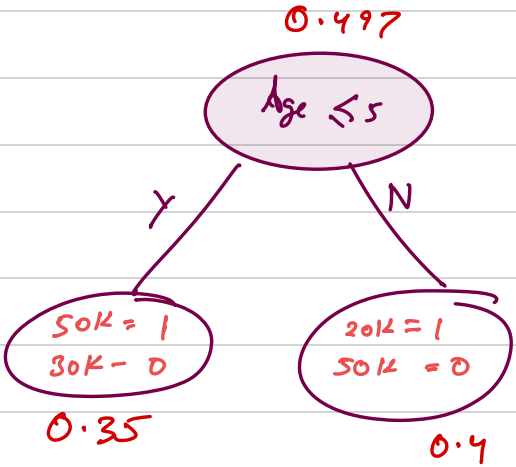
Root node

F_i, T_i – Feature, Value
Node with
highest I.G.

#rows = 150K

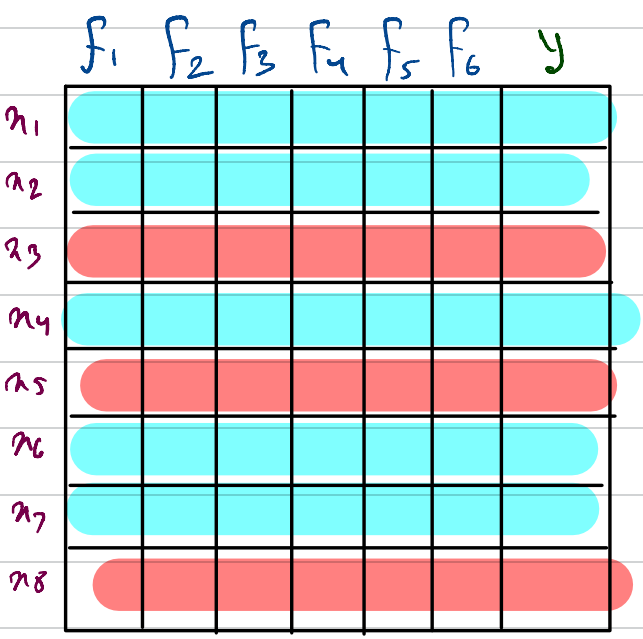
0 — 80K

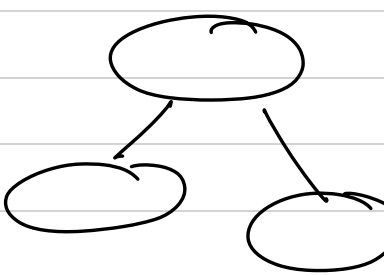
1 — 70K



$0.35 \times \frac{80}{150} + 0.4 \times \frac{70}{150} = 0.373$

$IG = 0.497 - 0.373 = 0.124$





For all the blue-rows, I'll use all the features, and values to find out which feature, value pair has highest information gain, and use that for splitting.

Same for the other leaf node, i.e. the red lines.

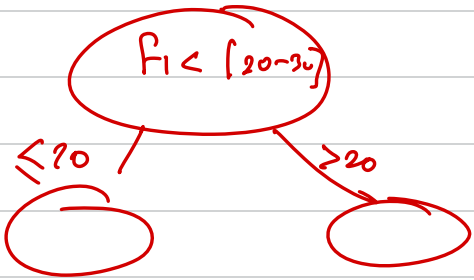
Brute Force Approach:

1. For each numerical feature:
   i. Sort numerical feature in ascending ordering.
   ii. With each unique value in the sorted list, compute IG (Information Gain), store it.
2. Do 1. For all the features

3. Compute Argmax, and find feature, value pair, and do a split, keep doing this until you reach max-depth.

F1 → 0 — 100 (Floating point)

[0 - 10] 01 10'
[10 - 20] 20'
[20 - 30] 30'
:
[90 - 100]

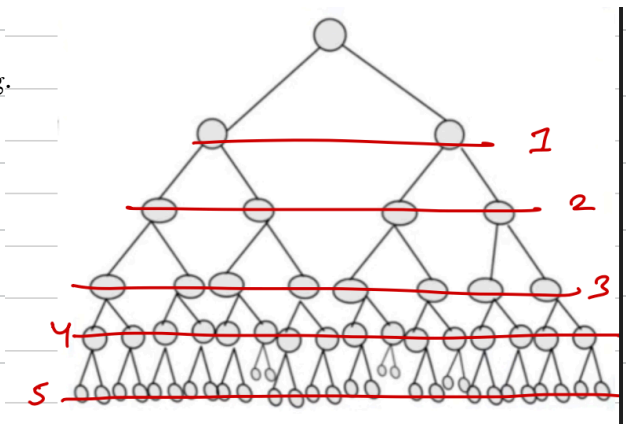


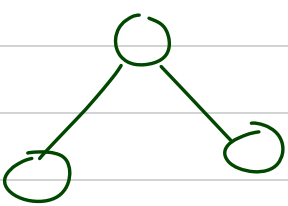
0 – 100

0 – 20 20 – 100

OverFitting and UnderFitting

Max-depth, very high value - overfitting.





underfitting.

prune[2]
/pru:n/

verb

gerund or present participle: **pruning**

trim (a tree, shrub, or bush) by cutting away dead or overgrown branches or stems, especially to encourage growth.

"now is the time to prune roses"

Similar: cut back | trim | thin | thin out | pinch back | crop | clip

- cut away (a branch or stem) from a tree, shrub, etc.
  "**prune back** the branches"

Similar: cut off | lop (off) | chop off | hack off | clip | snip (off)

- reduce the extent of (something) by removing superfluous or unwanted parts.
  "the workforce was pruned"

What is pruning in decision trees?

0 users have participated

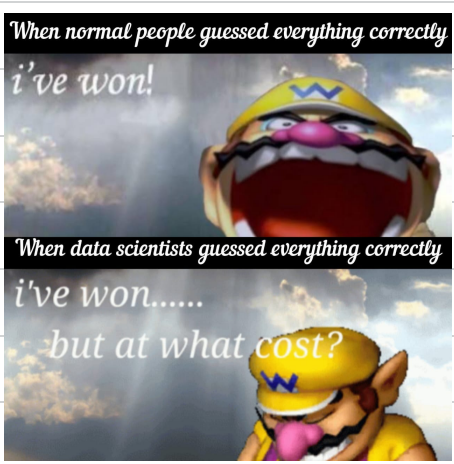
A	The process of removing outliers	0%
B	The process of handling missing values	0%
C ✓	The process removing unnecessary branches	0%
D	process of transforming categorical variables	0%

End Quiz Now

Based on all quizzes from the session

2	RK 1	HC 3
Deependu Ghosh 2/2 181.57	Rakesh Karade 2/2 185.30	Harshitha Chowd... 2/2 178.20

4	Aditya Shandilya	2/2 175.87
5	Purushottam Kumar	2/2 174.30
6	Snehal Adhikary	2/2 169.23
7	SHASHANK JHA	2/2 168.50
8	Kiran Hebasur	2/2 162.23
9	Praveen	2/2 162.23
10	MADEEHA REHMAN	2/2 161.73



What does each of these hyperparameters do?

- **Criterion**:
  - Defines which impurity measure to use.
  - By default, it is set to Gini Impurity.
- **max_depth**:
  - Defines maximium depth upto which a tree will grow.
  - By default, it is **None** i.e.
  - it'll grow until all leaves are pure.
- **min_samples_split**:
  - Defines the minimum no. of datapoints required to split further.
  - Helps control depth which therefore prevents **overfit**.
  - By default, it is 2.
- **min_samples_leaf**:
  - Defines the minimum no. of samples a leaf node can have.
- **max_leaf_nodes**:
  - Defines the maximum no. of leaf nodes a tree can have.
- **max_features**:
  - Selects features to be used while deciding the best split.
  - By default, it considers all the features for split.
- **class_weight**:
  - Assign weights to different classes during training.
  - Helps in dealing with imbalanced data.

How the performance of a DT will be affected if we tend to increase maximum number of pure leaf nodes?

0 users have participated

A	Model will underfit	0%
B	Model will be optimized	0%
✓ C	Model will overfit	0%
D	Can't be sure	0%

End Quiz Now

Based on all quizzes from the session

Which of the given DTs will be affected by an outlier?

4 options

Active Duration(Most preferred: 30 seconds)

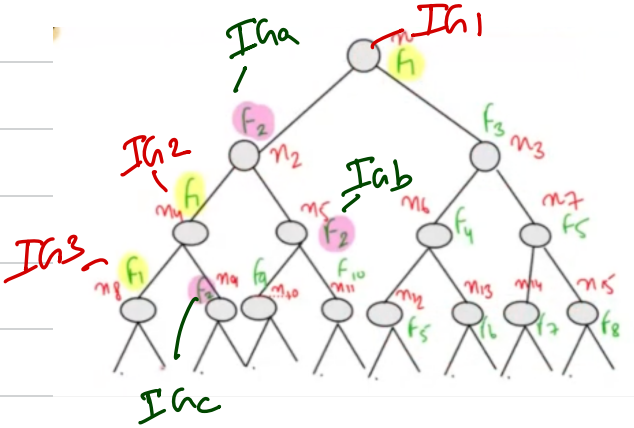
Appears for 60 Secs

A	DT with depth=50
B	DT with depth=5
C	Both of them
D	None of them

Feature Importance

$Imp(F_1) =$

$$\frac{n}{N} \times Ig_1 + \frac{n_4}{N} \times Ig_2 + \frac{n_8}{N} \times Ig_3$$



$$Imp(F_2) = \frac{n_2}{N} \times Ig_a + \frac{n_5}{N} \times Ig_b + \frac{n_9}{N} \times Ig_c$$