Decision Trees

09 January 2024 07:53

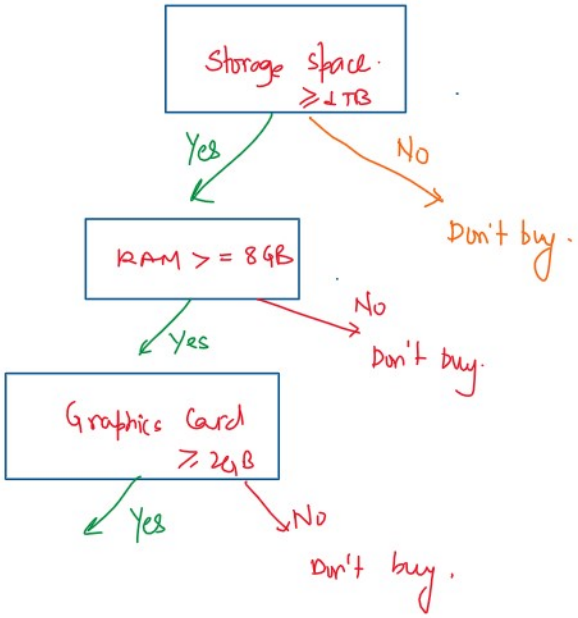
Intuition behind decision Trees

"Decision trees are everywhere".

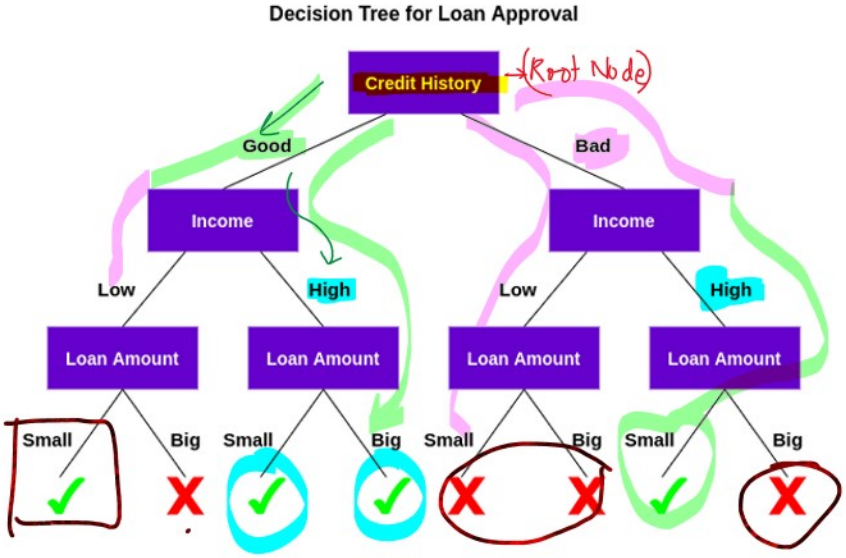
Person looking to buy a laptop

Storage → ??
1 TB

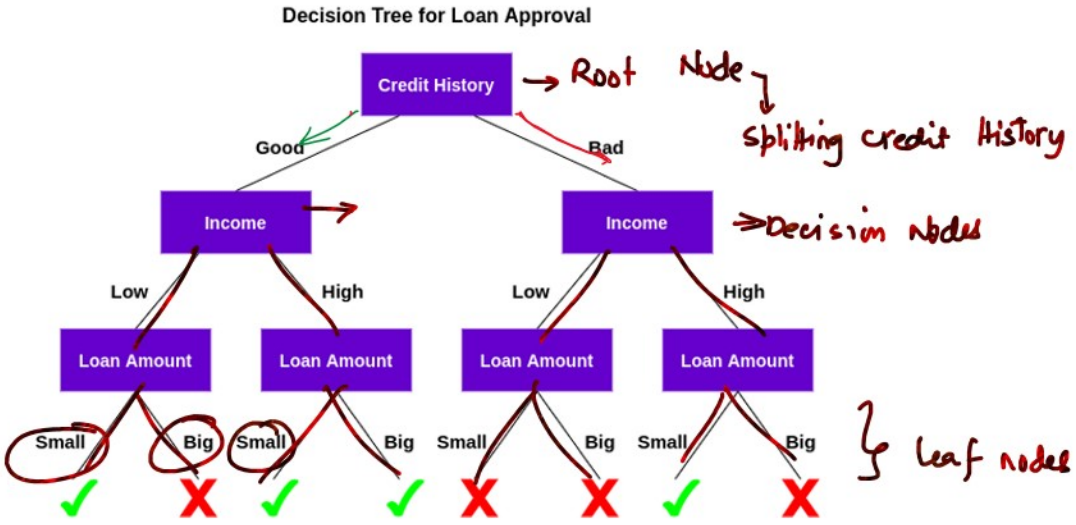
Display → 22.
LED ?
OLED ?

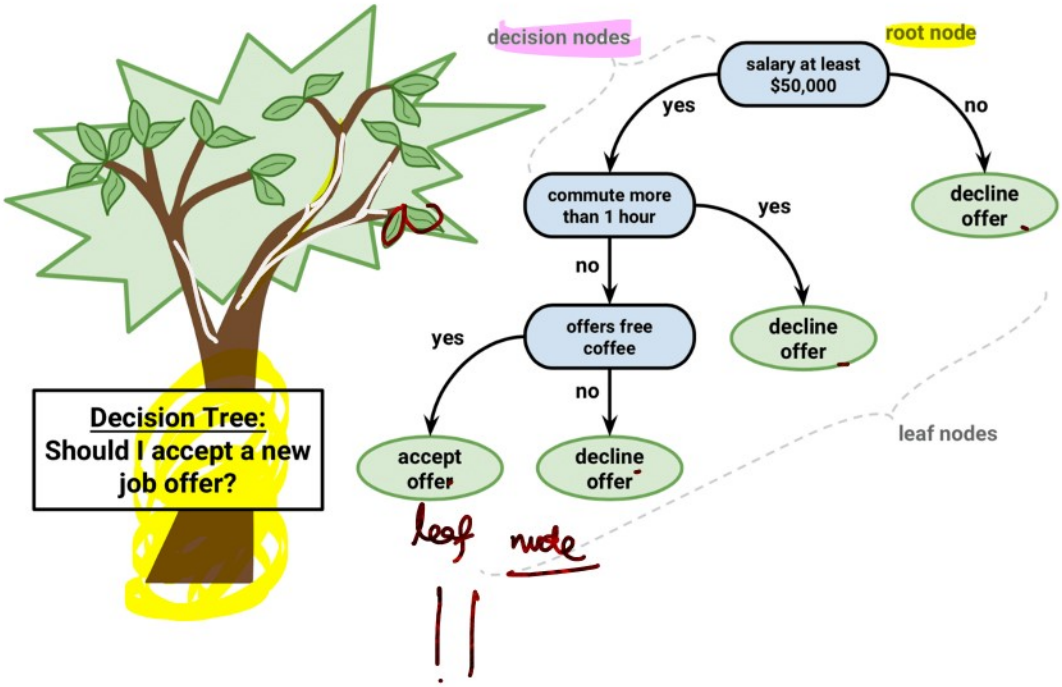


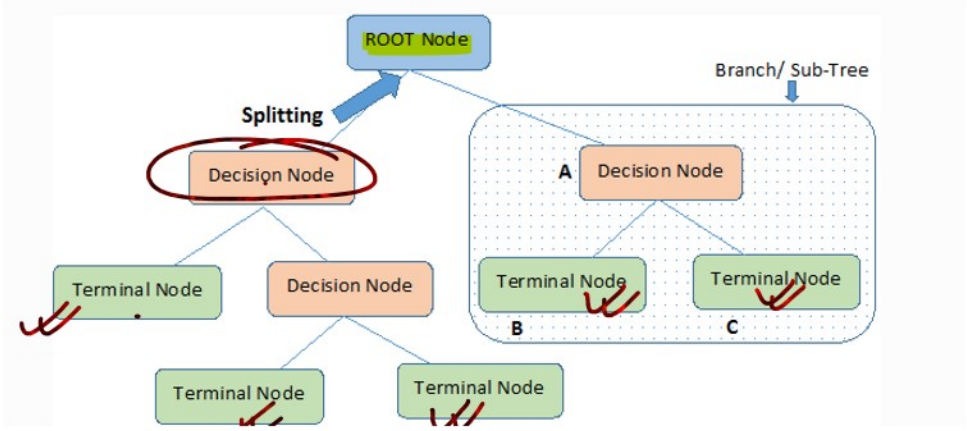
loan approval Decision Trees

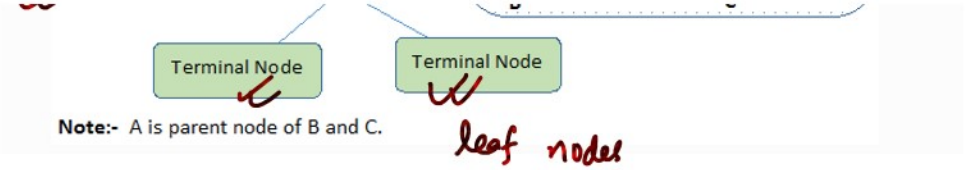


Decision Trees Terminologies



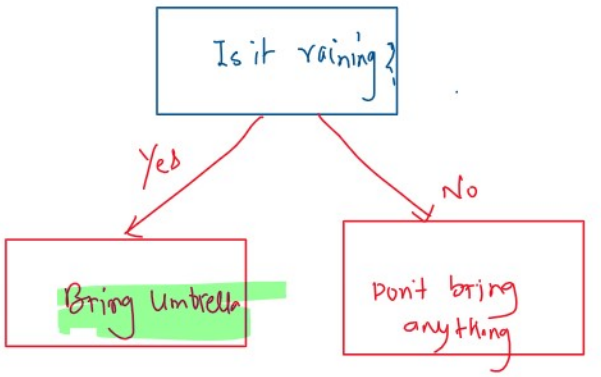


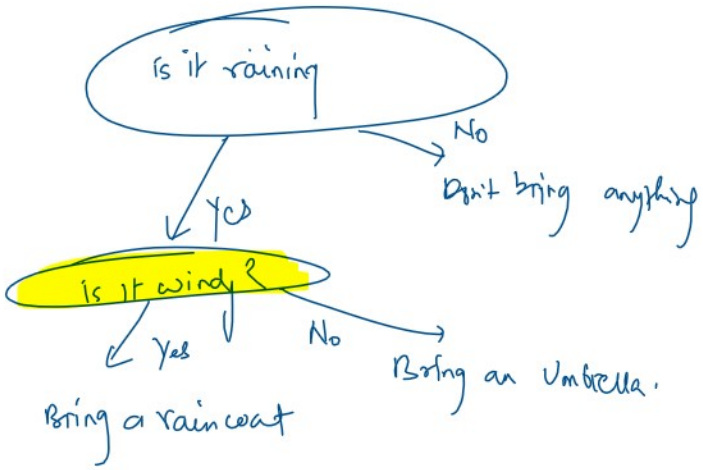




Note:- A is parent node of B and C.

leaf nodes





ROOT NODE :

It represents entire population or sample and this further gets divided into two or more homogeneous sets.

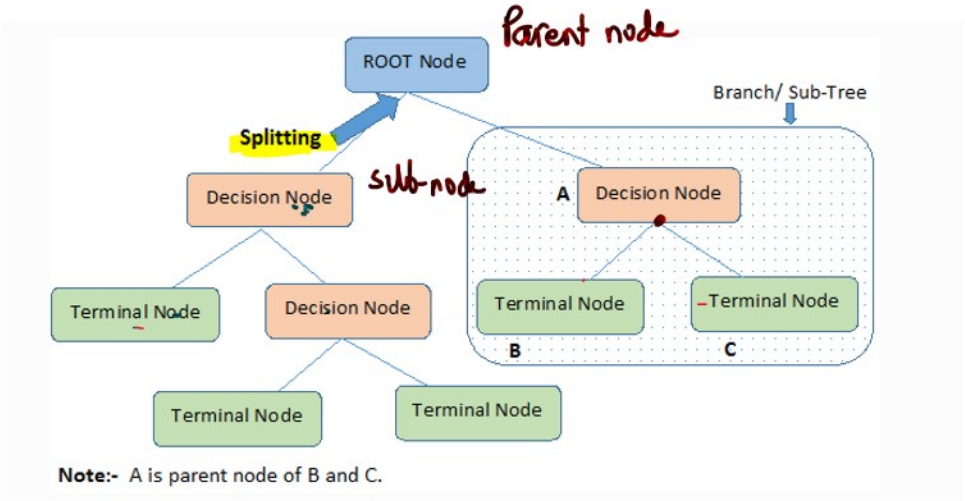
SPILITTING

- Is a process of dividing a node into two or more subnodes.

DECISION NODE

- When a sub-node splits into further sub-nodes, then it

is called decision node.



LEAF / TERMINAL NODE

- Nodes which do not split is called leaf or terminal node.

BRANCH / SUB-TREE

- a sub section of entire tree is called branch or sub-tree

PRUNING :

- When we remove sub-nodes of a decision tree, this process is called pruning.

Advantages

1. Easy to Understand :

- Decision Tree o/p is very easy to understand even for people from non-analytics background.
- doesn't require any statistical knowledge to read and interpret the decision trees
- its graphical representation is really intuitive, helping users validate their hypotheses.

2. EDA: Exploratory Data Analysis

- Decision tree is one of the fastest tool for

- Decision tree is one of the fastest tool for identifying the most influential variables and relationship among the variables.

3. Data type

it can handle both numerical and categorical variable.

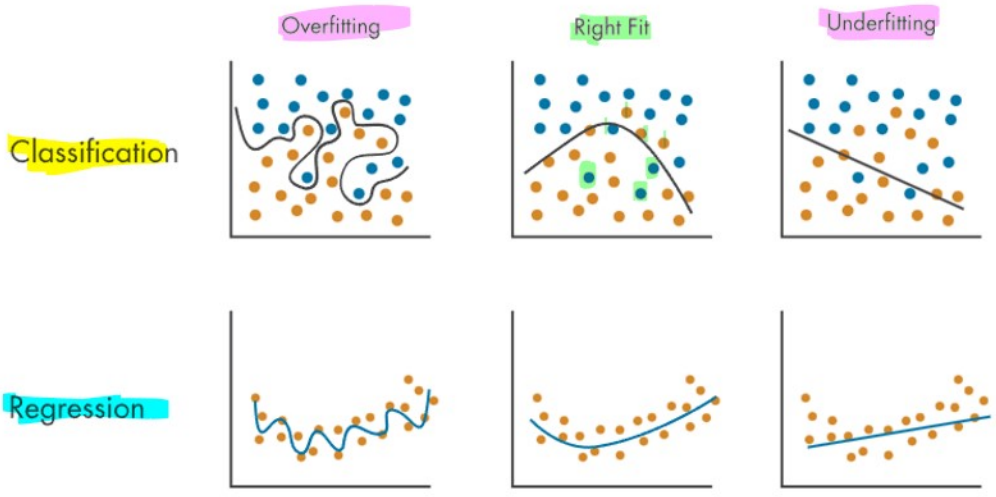
Disadvantages

1. overfitting

Overfitting is one of the most practical difficultly for decision tree models.

overfitting and Underfitting

Overfitting : it occurs when a model learns the training data too well (Too much), capturing noise or random fluctuations as well that may not be representative of the true underlying patterns in the data.



Signs of overfitting

- model performs exceptionally well on the training data.
- However, it fails to generalize to new, unseen data (performs poorly on the test / validation set)

Reasons for overfitting

- Using a highly complex model with too many parameters
- having insufficient data to support the complexity of the model
- training the model for a lot number of epochs.
  - one epoch means in ML, one complete pass of the training dataset through the algorithm.

Error	Overfitting	Right Fit	Underfitting
Training	Low	Low	High
Test	High	Low	High

Mitigation

- Use simpler models imply less no. of features
- regularization techniques to penalize overly complex models.
- increase the amount of training data

Underfitting

- Underfitting occurs when a model is too simple to capture the underlying patterns in the training data, resulting in poor performance on both the training / validata data

Signs of underfitting

- the model performs poorly on training data
- it also performs [illegible] on the test [illegible]

- the model performs poorly on training data
- it also performs poorly on the test/validation set
- model is very simple, with almost no features

Causes for underfitting

- Using a model too simple with too few parameters
- Insufficient training or not allowing the model to learn enough during training.

Mitigation

- increasing model complexity by adding more parameters
- use a more sophisticated model.
- ensure sufficient training time and training data.

Key takeaway

* Idea is to achieve a balance between overfitting and underfitting which is crucial for building models that generalizes well to new, unseen data.

* Techniques such as cross-validation, regularization and hyperparameter tuning are often employed to strike the right balance.

Where to split?

- Decision trees use algorithms to decide to split a node in two or more subnp
- creation of sub-nodes increases the homogeneity of resultant sub-nodes.
- Gini algorithm

- Gini algorithm

Ex: We want to group the students based on target variable (playing cricket or not)

Variable #1 Gender

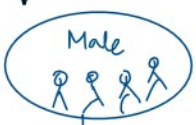
N = 30

Play cricket = 15 (50%)

Calculate Gini for sub-nodes using formula: sum of squares of probability for success and failure

p → success → plays cricket.

q → failure → doesn't play cricket.



students = 20

Play cricket = 13

$\left(\frac{13}{20} \times 100\right) = 65\%$.

students = 10

Play cricket = 2

$\left(\frac{2}{10} \times 100\right) = 20\%$.

$(p^2 + q^2)$.

p q

Gini for sub-node female → $(0.2 * 0.2) + (0.8 * 0.8)$

$= 0.04 + 0.64$

$= 0.68$

Gini for sub-node male

$= (0.65 * 0.65) + (0.35 * 0.35)$

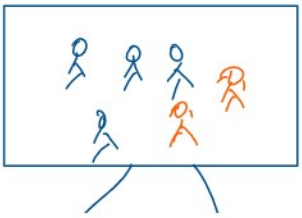
$= 0.4225 + 0.1225$

$= 0.5450 = 0.55$.

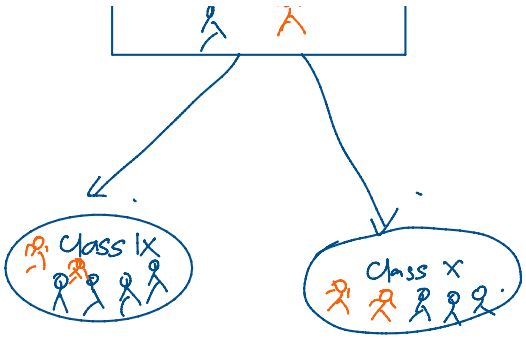
Calculate weighted Gini for split at Gender $= \frac{10}{30} \times 0.68 + \frac{20}{30} \times 0.55$

$= 0.59$

Variable #2 class



N = 30



Students = 14 Students = 16

Play cricket = 6 (43%) Play cricket = 9 (56%)

Gini for sub-node class IX = $(0.43 * 0.43 + 0.57 * 0.57) = 0.51$

Gini for sub-node class X = $(0.56 * 0.56 + 0.44 * 0.44) = 0.51$

Weighted Gini for split on class = $\frac{14}{30} \times 0.51 + \frac{16}{30} \times 0.51$

$= 0.51$