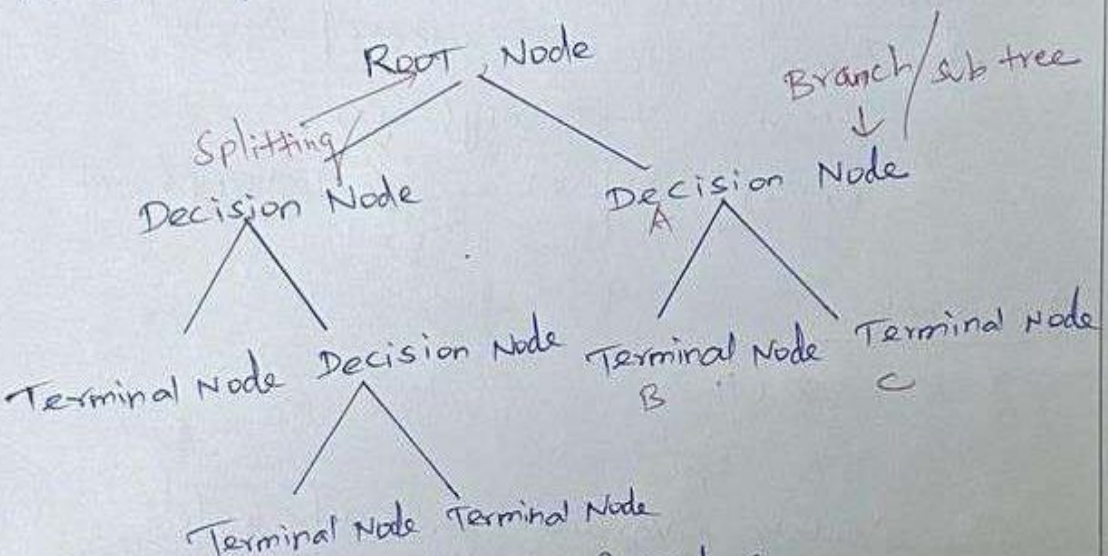
DECISION TREE BASED METHODS FOR CLASSIFICATION

Terminology Related to Decision Trees

1. Root Node
2. Splitting
3. Decision Node
4. Leaf / Terminal Node
5. Pruning
6. Branch
7. Parent and child Node



A is Parent node of B and C

Algorithm used in Decision Trees

- ID3 → [Extension of D_3]
- C4.5 → [Successor of ID_3]
- CART → [Classification And Regression Tree]
- CHAID → [chi-square Automatic Interaction Detection]
- MARS → Multivirate Adoptive Regression Splines

ID3 Algorithm — [Entropy (H), Information gain (IG)]

* Entropy is a measure of randomness in the information being processed. Flipping a coin is an example of an action that provides information that is random

$$E(S) = \sum_{i=1}^{C} -P_i \log_2 P_i$$

play golf

Yes 9 | No 5 → = Entropy (5,9)

$$= E(T,X) = \sum_{c \in X} P(c)\, E(c)$$

* Information gain

T → Current State X → selected Attribute.

= Entropy (T) — Entropy (T, x)

= Entropy (before) — $\sum_{j=1}^{k}$ Entropy (j, after)

* Variance

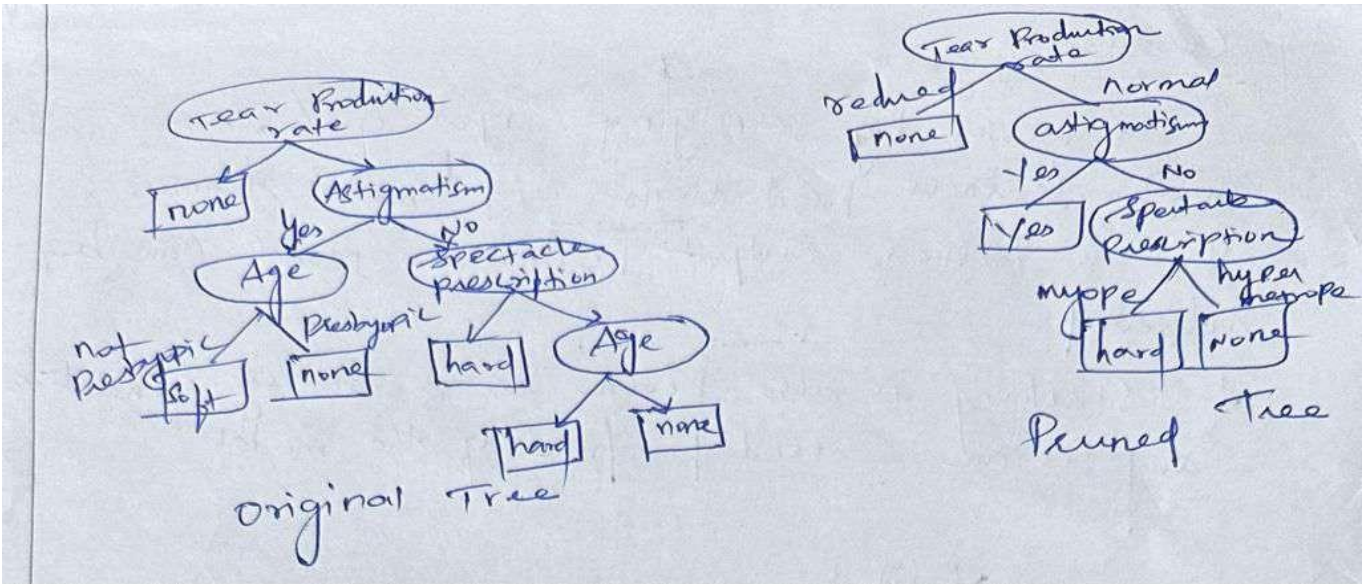
$$\frac{\sum (x - \bar{x})^2}{n}$$

① Calculate Variance for each node

② Calculate Variance for each split as the weighted average.

Pruning

Segregating the actual training set into 2 sets: training data set, D and Validation data set


Tear Production rate
none
Astigmatism
Yes
No
Age
Spectacle prescription
Not Presbyopic
Presbyopic
Soft
none
hard
Age
hard
none
Original Tree
Tear Production rate
reduced
Normal
none
astigmatism
Yes
No
Yes
Spectacle Prescription
myope
hypermetrope
hard
none
Pruned
Tree