1/3/23

Date / /
Page No.

DECISION TREE & RANDOM FOREST

Topics-
- → Introduction
- → Entropy & information Gain
- → Process Kaggle Titanic Dataset
- → Implementing Information Gain
- → Implementing Decision Tree
- → Making predictions
- → Decision Tree using Sklearn
- → Random Forest / Ensembles

Entropy & Information Gain

→ Entropy
measures the impurity or uncertainity present in the data.
(measure of randomness)

OR

degree of disorder present in the data.

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Randomness ∝ Entropy

Where,
S = set of all instances
N = number of distinct class values
p_i = event probability

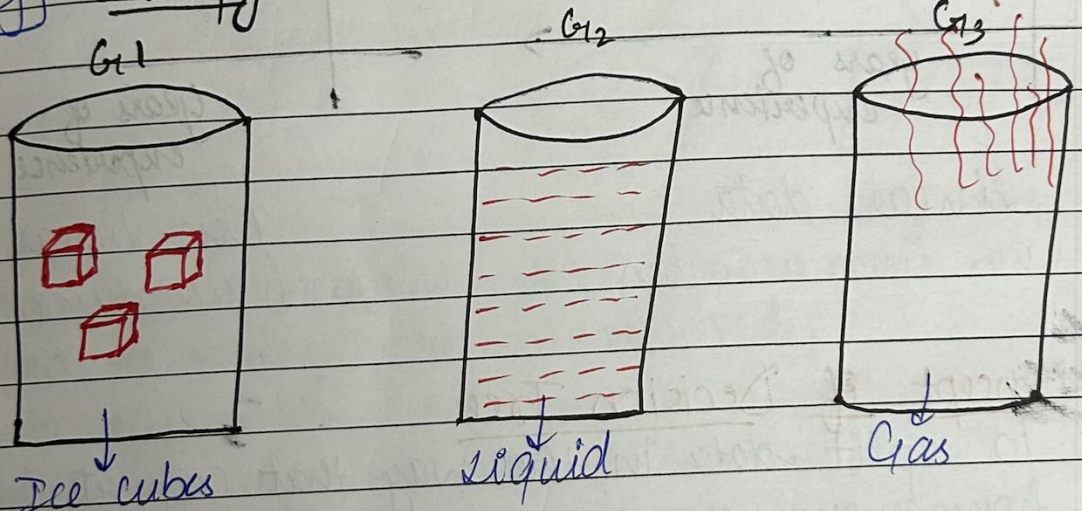
→ Information Gain
It indicates how much information a particular feature / variable gives us about the final outcome

Date / /
Page No.

$$Gain(A,S) = H(S) - \sum |S_i| \cdot H(S_i) = H(S) - H(A,S)$$

where,
$H(S)$ = entropy of whole dataset
$|S_i|$ = no. of instance with i value
$|S|$ = total no. of instance in dataset
V = set of distinct values of A
$H(S_i)$ = entropy of subset of A
$H(A,S)$ = entropy of an attribute A.

Eg) Entropy



⇒ G13 ↑ — has highest entropy

G12 | — medium

G11 ↓ — has lowest entropy

⇒ High entropy (messy data)
Low entropy (clean)

⇒ Low entropy - high information

* Entropy $\propto \frac{1}{\text{Information Gain}}$

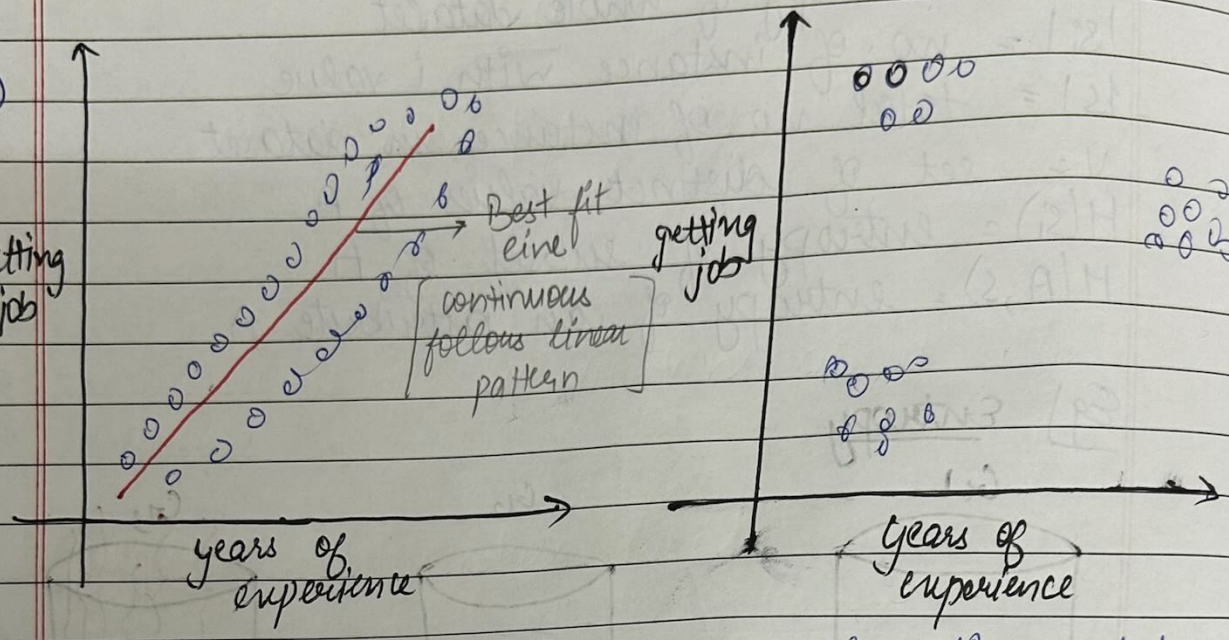
Gini Index & Entropy → Algos.

Date / /
Page No.

* Decision Tree is used for regression & classification

⇒ Decision Tree is used for non-linear data

⇒

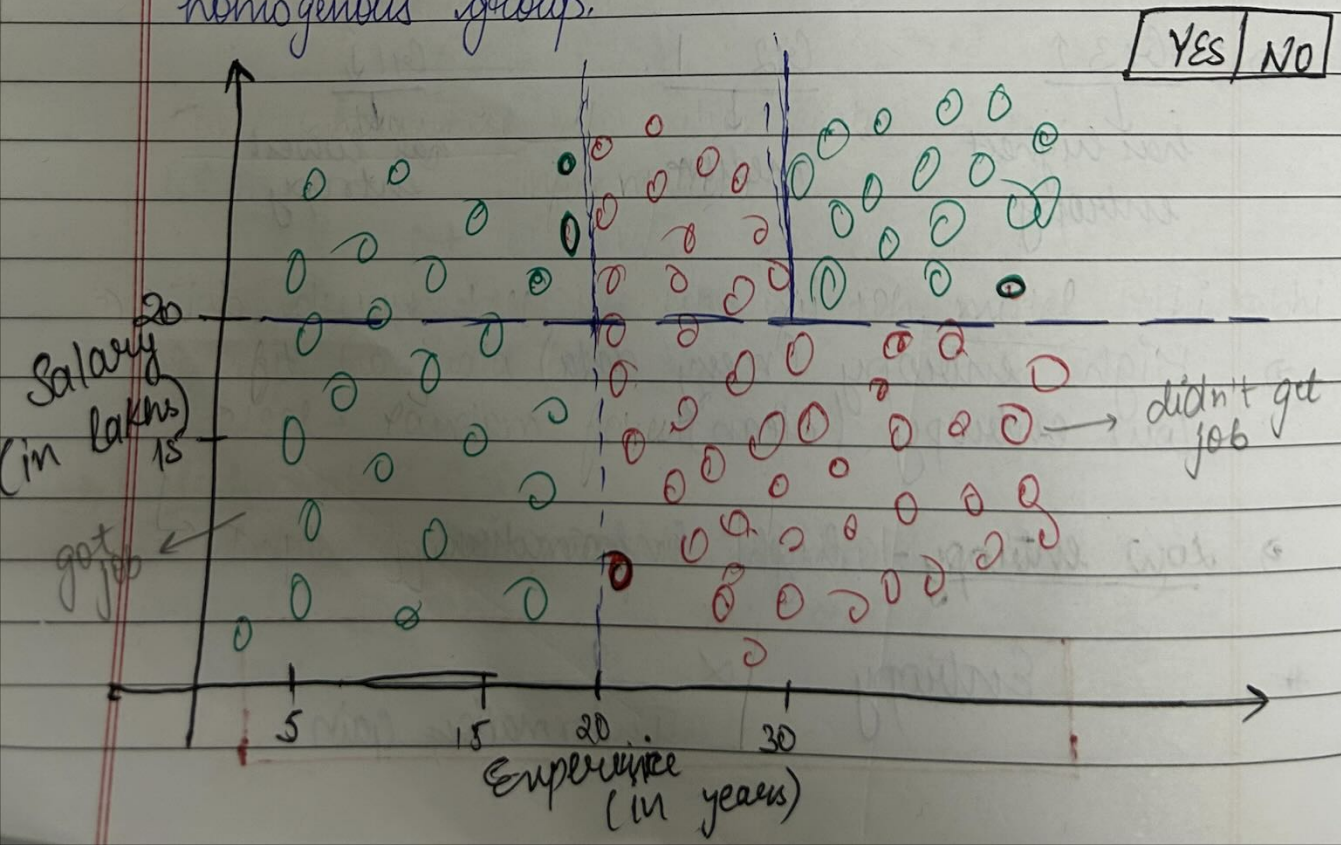


Linear data
(uses linear regression)

Non-Linear data
(uses decision-tree)

Concept of Decision Tree.

To split data in the way that converts into homogenous group.

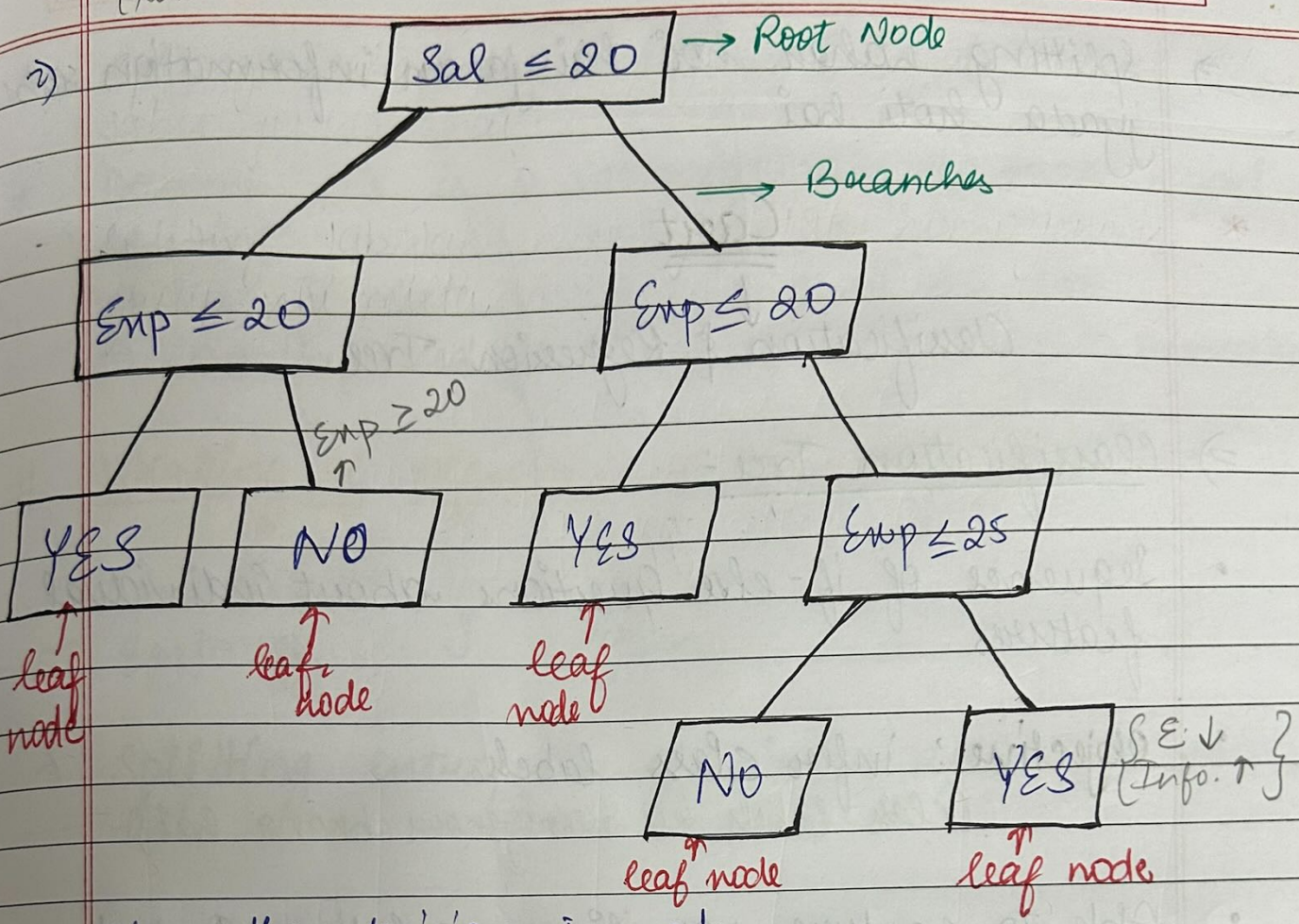


Non-linear Algo

SVM (Kernel Trick)

Decision Tree

Date / /
Page No.



Now the decision is made.

Ques) How to verify the decision tree?

⇒ No. of regions = No. of leaf nodes then the decision tree is right

Ques) Predict if the person gets the job or not, if the person has salary = 15, exp = 30.

⇒ Salary = 15
Exp = 30

Output = NO

⇒ Decision tree makes different boundaries.

⇒ Splitting tab tak hoti hai jab tak homogenous groups nhi milta.

⇒ Splitting wahan hoti hai jahan information sabse jyada hoti hai

* **Cart**

↓

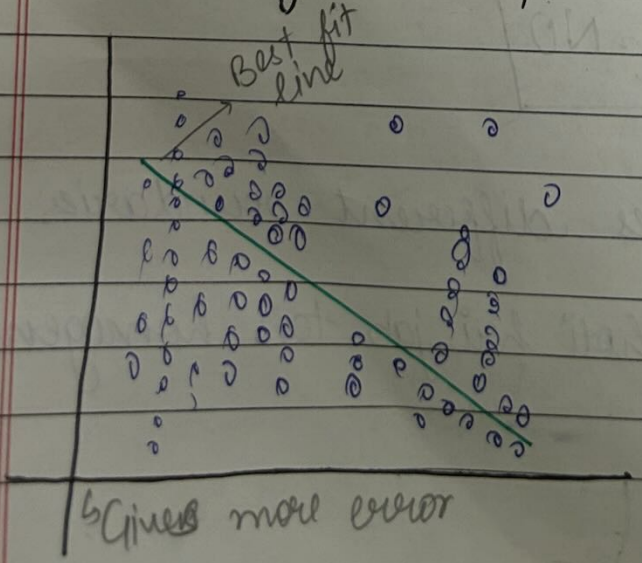
Classification & Regression Tree

⇒ **Classification Tree -**

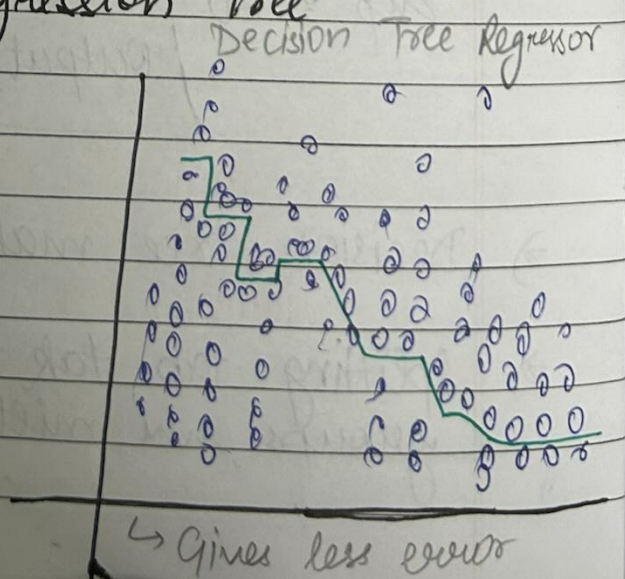
- Sequence of if-else questions about individual features
- Objective : infer class labels
  (class labels ko homogenous bnana hai)
- Able to capture non-linear relationships b/w features & labels
- Don't require feature scaling
  (Eg - standarization, normalization - etc.)

Regression Tree

Linear Regression v/s Regression Tree



↳ Gives more error



↳ Gives less error

Date / /
Page No.

* leaf node is the answer

* Decision Tree is a datastructure to make model of data.
  → Uses if else at every node of the tree
  → can be used for both classification & regression.

Working of Algorithm of Decision Tree

→ Gini Index } Information gain
→ Entropy

→ Splitting entropy or gini index apne aap karta hai.

⇒ Splitting is done by

Gini Index, Entropy, ID3

decide where to split & where not to

* Random Forest → Combination of Decision Trees.