POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

PAGE NO. .....

Unit I

Supervised learning Algorithm

Analytics is a collection of techniques and tools used for creating value from data. Techniques include concepts such as artificial intelligence (AI), machine learning (ML), and deep learning (DL) algorithms.

AI → Algorithm and systems that exhibit human like intelligence.

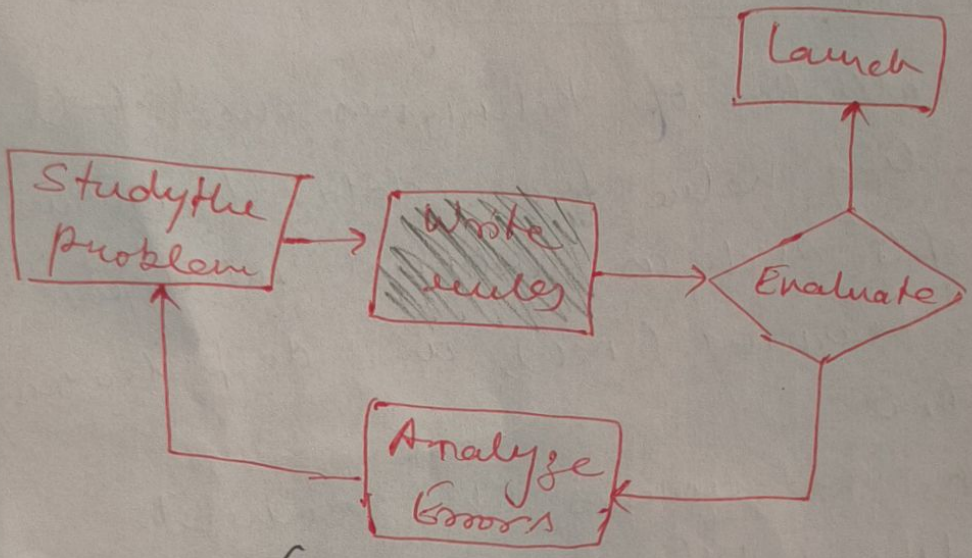
ML → Subset of AI that can learn to perform a task with extracted data and/or models.

DL → Subset of ML that imitate the functioning of human brain to solve problems.

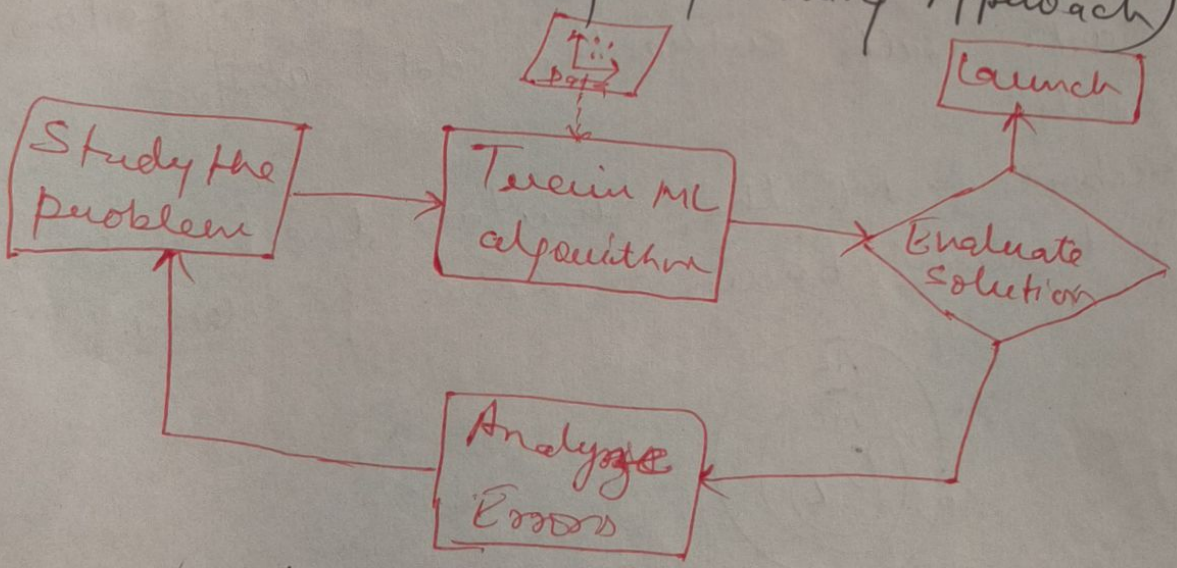
ML is a set of algorithms that have the capability to learn to perform tasks such as prediction and classification effectively using data.

⁂ An algorithm can be called a learning algorithm when it improves on a performance metric while performing a task, for eg:, accuracy of classification such as fraud, customer churn etc.

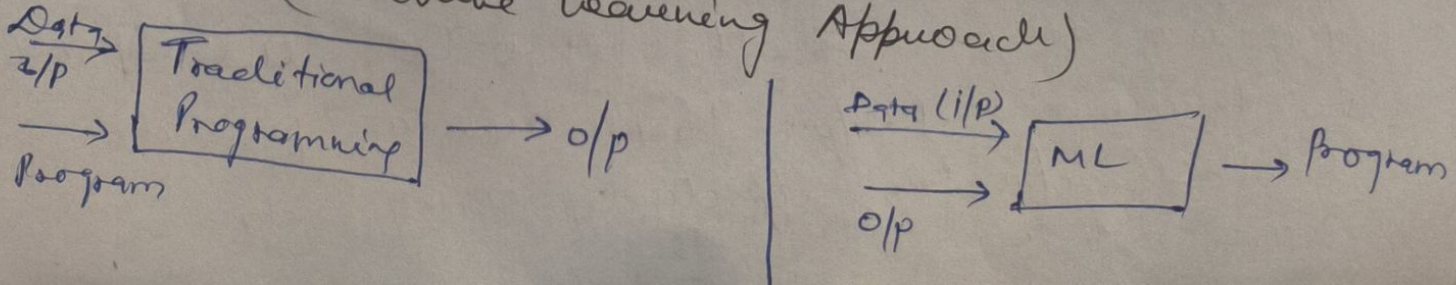
Difference between Traditional programming Approach and Machine learning approach:



(Traditional programming Approach)



(Machine learning Approach)



POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

PAGE NO. .....

→ Advantage of ML :-

1. Easily identifies trends and patterns : eg Amazon
2. No human intervention needed : eg anti-virus sfw
   They learn to filter new threats as they are recognized.
3. Continuous improvement eg: ~~data~~ weather forecasting
4. Handling multi-dimensional and multi variety data

→ Disadvantage of ML :-

1. Data Acquisition
2. Time and Resources
3. Interpretation of Result
4. High error-susceptibility

→ Some important terms :-

↳ Labeled Data : LD has both the i/p and o/p Parameter in a complete machine-readable format. Used in Supervised learning.

↳ Unlabeled Data : is a piece of data that has not been tagged with labels for identifying the characteristics. It is raw data and used in ~~xxxx~~ unsupervised learning.

Machine learning algorithms are classified into four categories:

1. Supervised Machine learning
2. Unsupervised Machine learning
3. Semi Supervised learning
4. Reinforcement learning

Supervised learning

Supervised learning is defined by its use of labeled datasets to train algorithm that to classify data or predict outcomes accurately.

working

Supervised learning uses a training set to teach models to yield the desired o/p. This training dataset includes i/p and correct o/p, which allow the model to learn over time.

The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

→ There are two varieties of Supervised learning algorithms:
- Classification Algorithm
- Regression Algorithm

→ Classification based supervised learning methods identify which category a set of data items belong to.

→ Regression based Supervised learning methods

try to predict output based on input variables.
Examples of Supervised Machine learning algorithms are as follows:

1. K- Nearest Neighbors Algorithm
2. Linear Regression
3. Logistic Regression
4. Support Vector machine
5. Decision Tree
6. Random forest
7. Neural Network

Unsupervised learning Algo

Unsupervised learning is where we have only input data and no corresponding o/p variable. From that data, it discover pattern that help solve for clustering or association problem.

Eg: Grouping customers by their corresponding habits.

- Unsupervised learning algo. divided in two categories:
  - → Clustering: A clustering problem is where we want to discover the inherent grouping in the data. Such as based on purchasing behaviour

~~Eg of unsupervised learning~~:

grouping the customers.

→ Association: An Rule based method for finding relationships b/w variables in a given dataset. These method are frequently used for market basket analysis, allowing companies to better understand relationships b/wo different product. Eg Amazon customer's

Eg. of unsupervised learning:

1. K-means for clustering problems.
2. Apriori algorithm for association rule learning problems.

Semi Supervised learning:

Some algorithm deal with partially labeled training data along with a lot of unlabelled data.

eg: Google photos (Photo archive)

Reinforcement Learning: RL is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.

Eg: Google deepmind Chess engine.

POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

PAGE NO. .....

Linear Regression Model:

Linear Regression is most well known algorithm in statistics and machine learning.

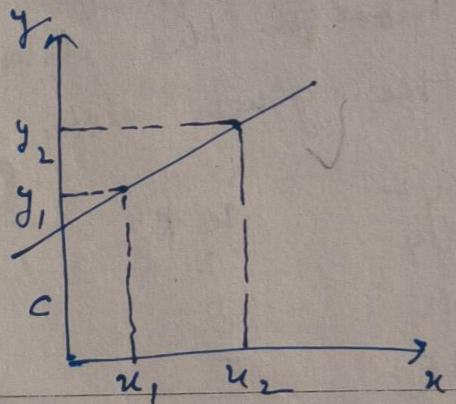
→ Linear regression is a Linear model.

eg. a model that assumes a linear relationship b/w the i/p variables (x) and the single o/p variable (y).

The simplest form of a SLR equation with one dependant and one independant variable is represented by -

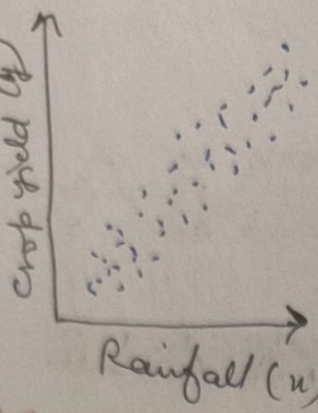
$$y = mx + c$$

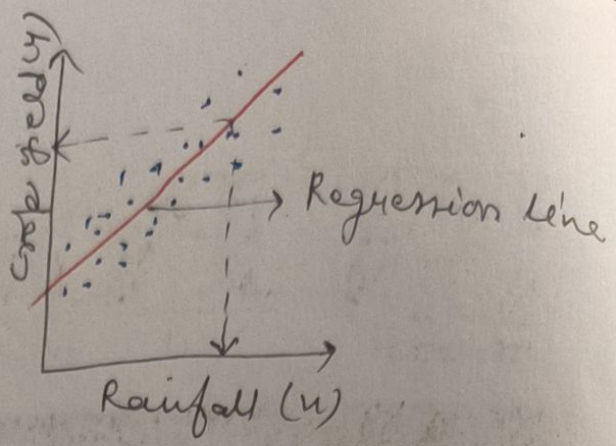
where $$m = \frac{y_2 - y_1}{x_2 - x_1}$$



y - Dependant variable
x - independant variable
m → slope of the line
c - coefficient of the line

For eg: lets plot the crop field based on the amount of rainfall. Here, rainfall is the independant variable and crop field is the dependant variable.

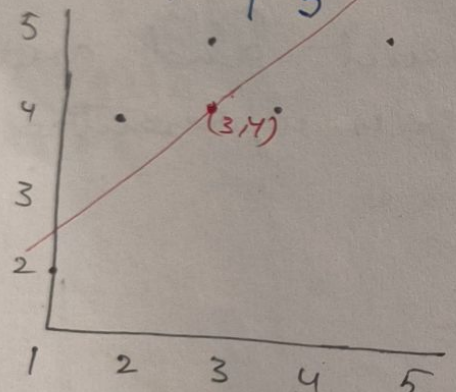




Eg:

Let's consider a sample data set with five rows and find out how to draw the regression line.

x	y
1	2
2	4
3	5
4	4
5	5



Now plot the x and y on the graph. Now we want to draw a regression line in ~~such~~ such a way that fits most of the data point in the given data set.

x	y	x^2	y^2	x*y
1	2	1	4	2
2	4	4	16	8
3	5	9	25	15
4	4	16	16	16
5	5	25	25	25
$\Sigma=15$	$\Sigma=20$	$\Sigma=55$	$\Sigma=86$	$\Sigma=66$

POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

PAGE NO. .....

Linear equation is represented by $y = mx + c$

$$m = \frac{((n * \Sigma(x*y)) - (\Sigma(x) * \Sigma(y))}{((n * \Sigma(x^2)) - (\Sigma(x)^2)}$$

$$m = \frac{((5 * 66) - (15 * 20))}{((5 * 55)) - (225)} = \boxed{0.6}$$

$$c = \frac{((\Sigma(y) * \Sigma(x^2)) - (\Sigma(x) * \Sigma(x*y))}{((n * \Sigma(x^2)) - (\Sigma(x)^2)}$$

$$c = \boxed{2.2}$$

Line equation is $y = mx + c$

y_{pred} values are:

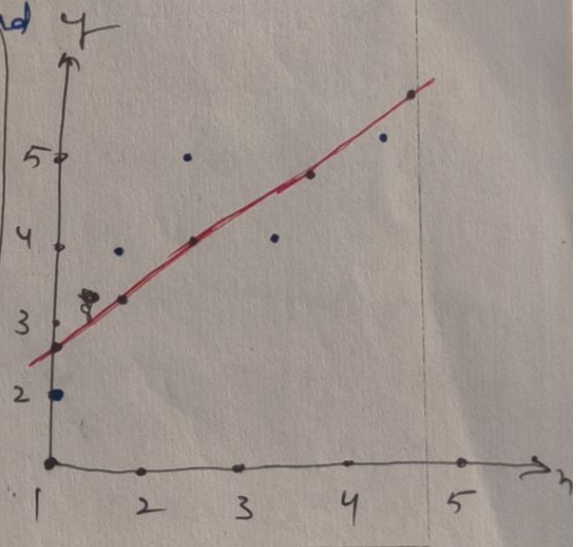
$y = 0.6 * 1 + 2.2 = 2.8$

$y = 0.6 * 2 + 2.2 = 3.4$

$y = 0.6 * 3 + 2.2 = 4$

$y = 0.6 * 4 + 2.2 = 4.6$

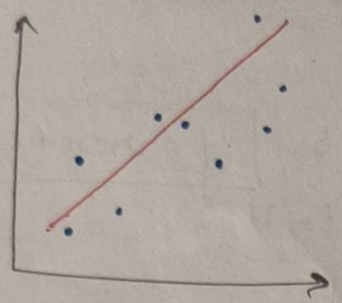
$y = 0.6 * 5 + 2.2 = 5.2$



Blue point represent actual value of y

Black point represent predicted value of y based on the model we created.

distance between actual value and predi
values of y are known as errors or Residual
The best-fit line should have the least su
of squares of these errors, also known as
'e-square'.



We keep the line moving through the data points
to make sure the best fit line has the least
squared distance between the data points and
the regression line.

POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

PAGE NO. .....

Logistic Regression :

Logistic Regression is a mathematical model used in statistics to estimate the probability of an event occuring using some previous data.

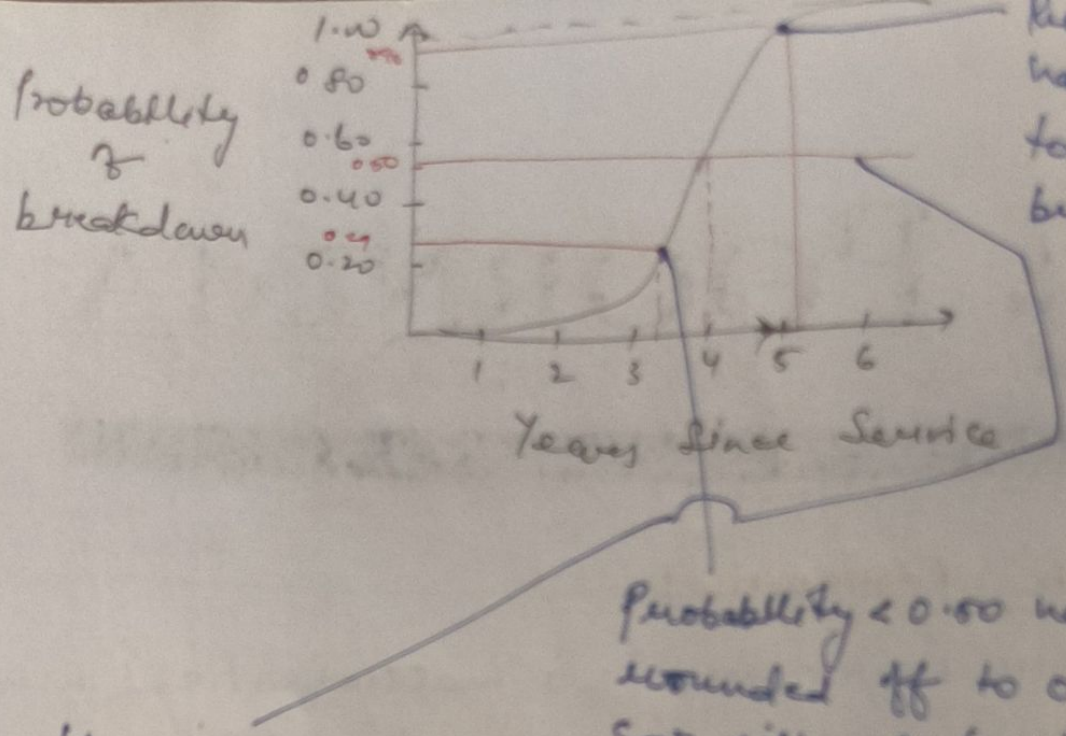
Logistic Regression works with binary data, where the event either happens (1) or doesn't happen (0).

Eg: Employee got a promotion or not.

→ Logistic Regression is used to describe data and the relationship between one dependent variable and one or more independant variable. The independ variable can be nominal, ordinal or of interval type.

→ Logistic Regression name is derived from logistic function and this function is also called sigmoid function. Value of logistic function lies between 0 and 1.

Eg:- The following is an example of a logistic function we can use to find the probability of a vehicle breaking down, depending on how many years it has been since it was service last.



Here, the threshold value 0.50 indicates that the car is more likely to breakdown after 4 years of usage

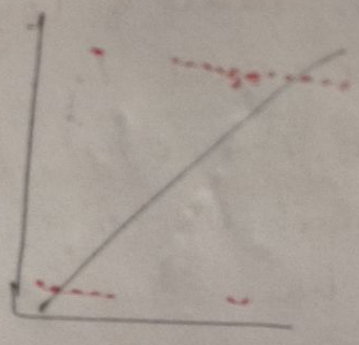
→ This curve is called Sigmoid curve or S-curve, which mathematically represented like

$$\sigma(y) = 1/(1+\exp(-y))$$

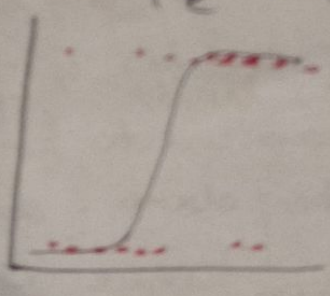
→ The Sigmoid function is also called squashing function as its domain, is the set of all real numbers and ~~values~~ range is (0,1).

→ If the input is either a very large negative number or a very large positive number, the o/p is always between 0 and 1.

$y = mx+b$



$$y = \frac{1}{1+e^{-(mx+b)}}$$



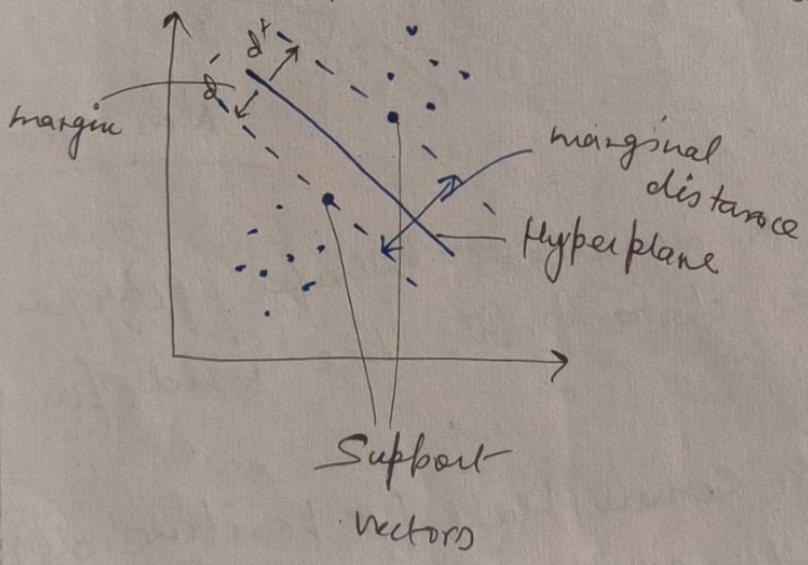
PAGE NO. .....

Support Vector Machine:

Objective of SVM algorithm is to find a linear surface (hyper plane) in N dimensional space such that it distinctly classified the data points.

→ A linear surface is a line in 2D, plane in 3D, hyperplane in higher dimensions.

→ Our objective is to find a higher plane such that it has maximum margin distance b/w the data points of the classes.



1. Support vector
2. Hyperplane
3. Marginal Distance
4. Linear Separable
5. Non Linear Separable

→ The higher the marginal distance, the more generalized our model.

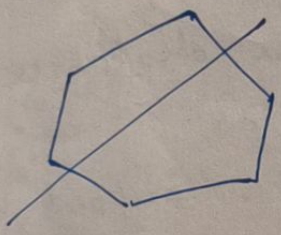
→ The data points through which lines/hyp
(π^+) and (π^-) pass through are called as Support vector. Hence, the algorithm is called as Support vector Machine.

→ Decision Boundary (Hyperplane)

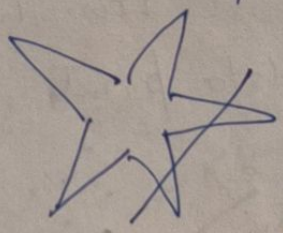
→ We always choose that hyperplane which is having the largest margin (max width margin) Maximal margin hyperplane should be selected which decrease error rate and increase accuracy.

→ Geometric Intitution of SVM:-

Convex polygon: A shape is said to be a convex polygon if a line drawn on that shape intercept only in 2 points.



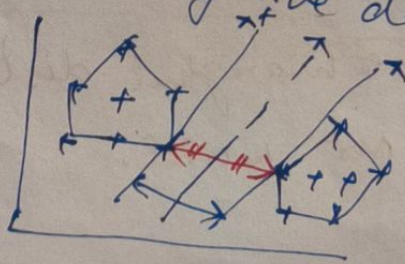
Convex.



Non-convex

Convex Hull: A smallest shape/polygon where all the data points lie inside/on the polygon.

Step 1: Create a convex Hull for positive data points and negative data points.

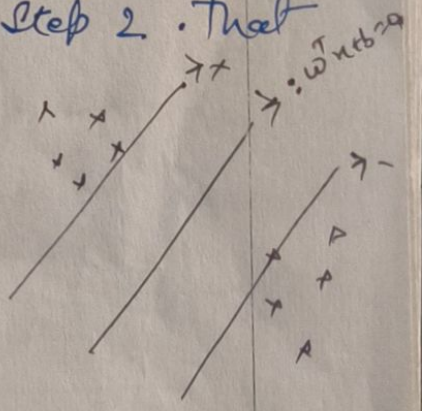


POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

PAGE NO.:

Step 2: Draw a shortest line that connect 1 data point of positive class and 1 data point of negative class (nearest point of each class)

Step 3: Draw a line/Hyperplane that bisects the shortest line drawn in Step 2. That line is Hyperplane.



Equation of ~~Hyplane~~ Hyperplane:

Let $\pi: \quad w^T x + b = 0$

$\pi^+: \quad w^T x_1 + b = 1$

$\pi^-: \quad w^T x_2 + b = -1$

$w^T(x_2 - x_1) = 2$

$\frac{w^T}{\|w\|}(x_2 - x_1) = \frac{2}{\|w\|}$ optimization function

margin (d) $= \frac{2}{\|w\|}$

w^T → Transpose of slope values

→ we can remove w^T with norm of w^T $\|w\|$

→ In w^T, some direction involved

→ magnitude of w off.

optimize the under best condition such that y_i

→ we need to maximize Optimize function.

$y_i = 1: \quad w^T x + b \geq 1$

$y_i = -1: \quad w^T x + b \leq -1$

$\Rightarrow y_i * w^T x_i + b_i \geq 1$ (for all x_i)

If y_i is not $>$ or $= 1$ then there is mis classification

PAGE NO. ..... (67)

Decision Tree in Machine Learning:

In decision Tree each branch of the tree represents a possible action/decision occurrence or reaction. It is utilized for both regression and classification task.

Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree.

Nest if else ⟹ Decision Tree

→ Important terms related to Decision Tree:

Entropy: Entropy is the measure of randomness or unpredictability in the dataset.
(Initially the dataset is having higher entropy).

→ Entropy is a metric used to train the decision tree. These metric measures the quality of split.

Information Entropy for a dataset with C classes are denoted as:

$$E = -\sum_{i}^{C} p_i \log_2 p_i$$

where p_i is the probability of randomly picking an element of class i.

Eg: Consider a dataset with 1 blue, 2 gree, 3 red. Then

$$E = -(p_b \log_2 p_b + p_g \log_2 p_g + p_r \log_2 p_r)$$

$$p_b = \frac{1}{6}, \quad p_g = \frac{2}{6}, \quad p_r = \frac{3}{6}$$

$$E = -\left(\frac{1}{6}\log_2\left(\frac{1}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right)$$

$$E = 1.46$$

Eg: Consider 3 blue only in a dataset

$$E = -\frac{3}{3}\log_2\frac{3}{3}$$

$$= -1\log_2 1$$

$$E = 0$$

Information Gain: how can we quantify the quality of a split?

IG = [How the features are ~~selected~~ selected]

Information gain is the measure of decrease in Entropy after the dataset is split. IG is calculated for a split by subtracting weighted entropies of each branch from the original entropies.

Information Gain = how much Entropy we removed.

→ Purity → Pure split

→ Entropy

→ Gini ~~coefficient~~ impurity $\left[1 - \sum_{i=1}^{n} (p)^2\right]$

→ How the features are selected

→ Information Gain

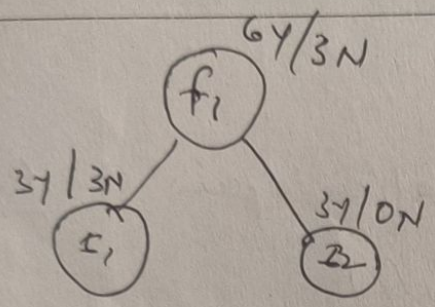
Higher IG = more entropy removed

IG is the measure of decrease in entropy after the dataset is split.

[leaf node carries the classification labels that we are trying to predict.

The best-split is choosen by maximizing the IG.

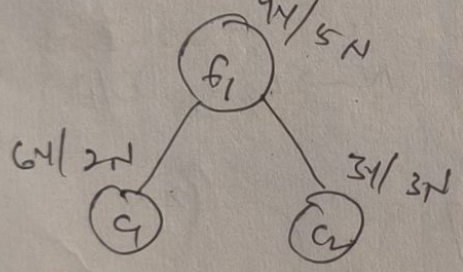
PAGE NO. .....

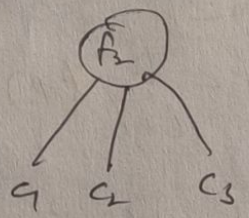


If Entropy = 0 then
Pure split

Entropy = 1 then
Impure split
further division

→ which feature to take to split ??



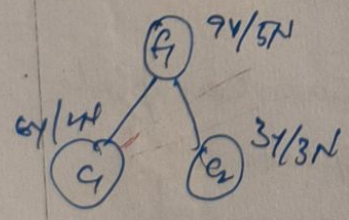


Information Gain :

$$\text{Gain}(S, f_1) = H(S) - \sum_{v \in Val} \frac{|S_v|}{|S|} H(S_v)$$

$|S_v|$ → on particular node total sample
$|S|$ → total sample

$H(S_{v_1}) \rightarrow C_1$
$H(S_{v_2}) \rightarrow C_2$



$H(S)$ is Entropy of root node.

$$H(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$H(S) = 0.94$$

$$H(S_{VC_1}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$$

$$H(S_{VC_1}) = 0.81$$

$$H(S_{VC_2}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$

$$H(S_{VC_2}) = 1$$

$$\text{Gain}(S, f_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

$$= \underline{0.049}$$

⇒ Suppose for feature 2 Information Gain is 0.051

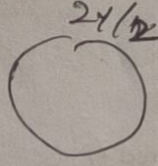
Gain (S, f_2) >> Gain (S, f_1)

↘ use feature 2, we will start the split.

Highest IG value feature is used for splitting

→ Gini Impurity

$$GI = 1 - \sum_{i=1}^{N} (P)^2 \qquad N \rightarrow \text{no. of o/p} \begin{cases} \text{Yes} \\ \text{NO} \end{cases}$$



$$= 1 - [(P_+)^2 + (P_-)^2]$$

$$= 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$$

$$= 1 - \frac{1}{2}$$

$$= 0.5$$ ⟹ It means complete impure split.

→ Gini calculates fast as in comparison to Entropy. Entropy use log function and Gini use simple math.

PAGE NO. .....

Naive Bayes: Naive Bayes works on the principle of conditional probability as given by the Bayes theorem.

Bayes theorem gives the conditional probability of an event A given the another event B has occured.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Where:

P(A) = The probability of A occuring
P(B) = The probability of B occuring
P(A,B) = The probability of A given B
P(B,A) = The probability of B given A

Eg:- Tossing two coins:

Sample space = {HH, HT, TH, TT}

P (first coin being tails when the second coin is head) = ?

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A) * 1/2}{1/2}$$

$$= \frac{\frac{1}{4} * \frac{1}{2}}{\frac{1}{2}} = 0.25$$

Application → Sentiment Analysis
Face Recognition
Weather Prediction
Medical Diagnosis

$$P(Y \mid x_1, x_2 \ldots x_n) = \frac{P(x_1 \mid y) * P(x_2 \mid Y) * P(x_3 \mid y) \ldots P(x_n \mid y) * P(y)}{P(x_1) * P(x_2) * P(x_3) \ldots P(x_n)}$$

Sanjit $P(N \mid \ldots)$

Person	COVID (x_1) (Yes/NO)	FLU (x_2) (Yes/NO)		FEVER (YES/NO)
1	Yes	Yes	NO	Yes
2	NO	Yes	Yes	Yes
3	Yes	NO	Yes	Yes
4	NO	NO	NO	NO
5	Yes	NO	NO	Yes
6	NO	NO	NO	Yes
7	Yes	NO	NO	Yes
8	Yes	Yes	NO	NO
9	~~Yes~~ NO		Yes	Yes
10	~~Yes~~ NO	~~Yes~~	Yes	NO

Given Person (flu, COVID)

$P(\text{yes} \mid \text{flu}, \text{covid}) = P(\text{flu} \mid \text{yes}) * P(\text{covid} \mid \text{yes}) * P(\text{yes})$

$P(\text{NO} \mid \text{flu}, \text{covid}) = P(\text{flu} \mid \text{NO}) * P(\text{covid} \mid \text{NO}) * P(\text{NO})$

Soln

Step 1: prior probability:

$P(\text{fever} = \text{yes}) = 7/10$

$P(\text{fever} = \text{NO}) = 3/10$

Step 2: Conditional probability:

	Yes	NO
covid	4/7	2/3
FLU	3/7	2/3

fever

(Yes) $= \frac{4}{7} \times \frac{3}{7} \times \frac{7}{10}$

$= \frac{12}{17}$

$= .17$

(No) $= \frac{2}{3} \times \frac{2}{3} \times \frac{3}{10}$

$= \frac{4}{30}$

$= .13$

Probability value of High for Yes.
So for Given problem. value is Yes
↳ fever.

PAGE NO.

K nearest neighbors (KNN) :- It uses the past data for learning and produces prediction.

KNN classifies a data point based on how its neighbors are classified. Stores all available cases that are being provided while training and then classifies new cases based on a similarity measures.

→ The process of choosing K value is termed as parameter tuning and it is very important aspect that is going to determine the accuracy of model.

↳ value of K can be chosen by applying following methods:

- Hit and trial
- Sqrt(n), where n stands for total number of data samples in dataset.
- Odd value of K is selected to avoid confusion between two classes of data.

↳ When do we choose KNN - then

- Data-set should be properly labelled.
- KNN cannot be used for huge data. It is lazy learner.
- KNN provides higher accuracy for small dataset.

How does KNN works:

Euclidean distance of two point (A, B)

$$= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Eg:-

Weight	Height	Body type	ED
62	182	Obese	13
55	170	Obese	2
56	174	Non-Obese	4.1
57	173	Obese	3
65	172	Obese	8.2
58	169	Obese	1.4
57	167	Non-Obese	6.7
69	176	Obese	13.4
64	173	Obese	7.6

K=3

Testing Sample:-

57	170	?

~~label~~ ~~By Euclidean Distance~~

Higher votes are for obese.

KNN Classifier in Machine Learning: KNN is supervised machine learning algorithm that can be used to solve the classification and regression problems.

-The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement, **non-parametric**, **lazy** learning, **supervised** machine learning algorithm that can be used to solve both classification and regression problems using feature similarity.

K- Nearest Neighbors is a

- **Non parametric** as it does not make an assumption about the underlying data distribution pattern

- **Lazy algorithm** as KNN does not have a training step. All data points will be used only at the time of prediction. With no training step, prediction step is costly.

- **Supervised machine learning algorithm** as target variable is known

- Used for both **Classification and Regression**

- Uses **feature similarity/nearest neighbors** to predict the cluster that the new point will fall into.

- KNN algorithm assumes that similar things exist in close proximity. (means near to each other). Therefore, KNN captures the idea of similarity (distance, proximity or closeness) with some mathematics

(i.e. calculating the distance of a new data point with nearest point).

- KNN classified a data point based on how its neighbor's are classified.

KNN algorithms steps:

1. Compute a distance value between the item to be predicted and every item in the training dataset.

2. Pick the k closest data points (the items with the k lowest distances)

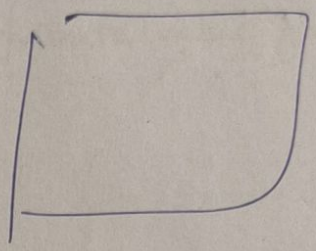
3. Get the labels/values of the selected k neighbors

- If regression, return the mean of the 'k' neighbors
- If classification, return the mode of the 'k' neighbors.

~~Gangan~~ (9) [~~Gapan Tenpid~~]

Decision Tree (Numerical Example)

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



4 attributes
1 target

Step: 1 Find the overall entropy for the dataset denoted by S and then find the entropy for each attribute value for each and specific features.

$$Values(Outlook) = Sunny, Overcast, Rain$$

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$

$S_{Sunny} \leftarrow [2+, 3-]$ $\quad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$

$S_{Overcast} \leftarrow [4+, 0-]$ $\quad Entropy(S_{Overcast}) = -\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4} = 0$

$S_{Rain} \leftarrow [3+, 2-]$ $\quad Entropy(S_{Rain}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$

$$Gain(S, Outlook) = Entropy(S) - \sum_{v \in (Sunny, Overcast, Rain)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Outlook) = Entropy(S) - \frac{5}{14}Entropy(S_{Sunny}) - \frac{4}{14}Entropy(S_{Overcast}) - \frac{5}{14}Entropy(S_{Rain})$$

$$Gain(S, Outlook) = 0.94 - \frac{5}{14}0.971 - \frac{4}{14}0 - \frac{5}{14}0.971 = 0.2464$$

Attribute (Temp)

$Values(Temp) = Hot, Mild, Cool$

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$

$S_{Hot} \leftarrow [2+, 2-]$ $\quad Entropy(S_{Hot}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0$

$S_{Mild} \leftarrow [4+, 2-]$ $\quad Entropy(S_{Mild}) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.9183$

$S_{Cool} \leftarrow [3+, 1-]$ $\quad Entropy(S_{Cool}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$

$$Gain(S, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$Gain(S, Temp)$

$$= Entropy(S) - \frac{4}{14}Entropy(S_{Hot}) - \frac{6}{14}Entropy(S_{Mild}) - \frac{4}{14}Entropy(S_{Cool})$$

$$= 0.94 - \frac{4}{14} \times 1.0 -$$

$$= .0289$$

$$\Rightarrow 0.04$$

Attribute (Humidity)

$Values\,(Humidity) = High, Normal$

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$

$S_{High} \leftarrow [3+, 4-]$ $\quad Entropy(S_{High}) = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7} = 0.9852$

$S_{Normal} \leftarrow [6+, 1-]$ $\quad Entropy(S_{Normal}) = -\frac{6}{7}log_2\frac{6}{7} - \frac{1}{7}log_2\frac{1}{7} = 0.5916$

$$Gain\,(S, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$Gain(S, Humidity)$

$$= Entropy(S) - \frac{7}{14}Entropy(S_{High}) - \frac{7}{14}Entropy(S_{Normal})$$

$$Gain(S, Humidity) = 0.94 - \frac{7}{14}0.9852 - \frac{7}{14}0.5916 = 0.1516$$

Attribute (Wind)

$Values(Wind) = Strong, Weak$

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$

$S_{Strong} \leftarrow [3+, 3-]$ $\quad Entropy(S_{Strong}) = 1.0$

$S_{Weak} \leftarrow [6+, 2-]$ $\quad Entropy(S_{Weak}) = -\frac{6}{8}log_2\frac{6}{8} - \frac{2}{8}log_2\frac{2}{8} = 0.8113$

$$Gain(S, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Wind) = Entropy(S) - \frac{6}{14}Entropy(S_{Strong}) - \frac{8}{14}Entropy(S_{Weak})$$

$$Gain(S, Wind) = 0.94 - \frac{6}{14}1.0 - \frac{8}{14}0.8113 = 0.0478$$

0.0478

Now Information gain for all attributes is:

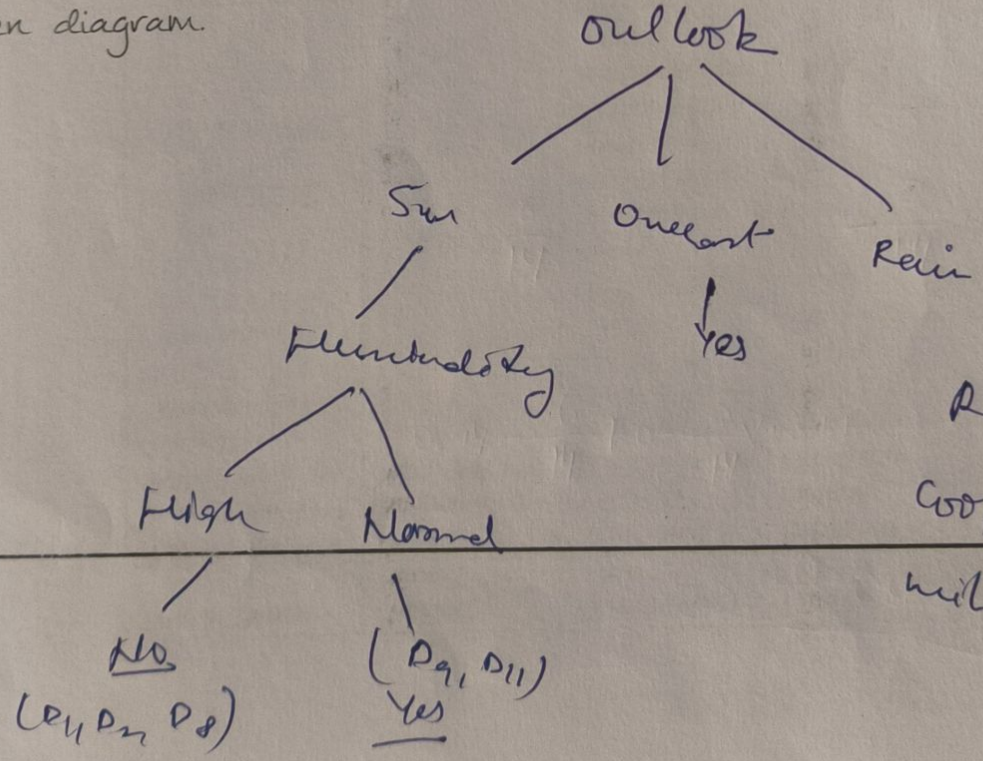
Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

$$Gain(S_{Rain}, Temp) = 0.0192$$

$$Gain(S_{Rain}, Humidity) = 0.0192$$

$$Gain(S_{Rain}, Wind) = 0.97$$

Now choose wind as next attribute since it is having the highest gain. So updated and final decision tree look like this in the given diagram.



Rain = 0.97

Cool = 1.0

mild = .9183

Now Information gain for all attributes is:

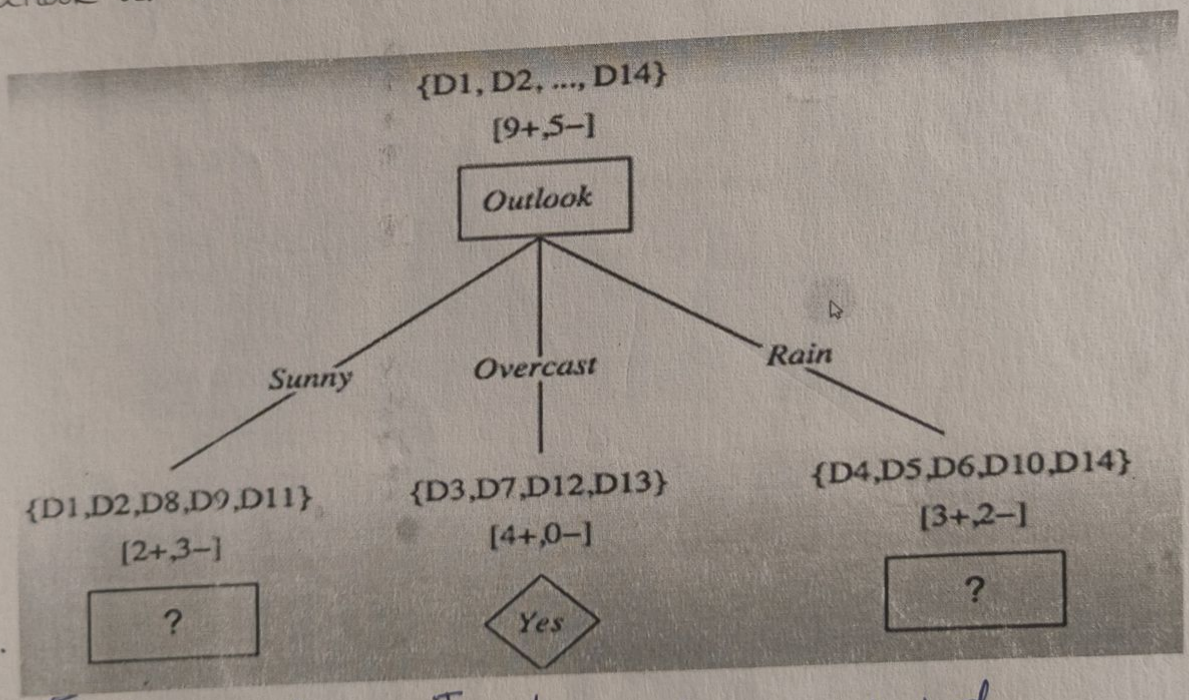
$Gain(S, Outlook) = 0.2464$

$Gain(S, Temp) = 0.0289$

$Gain(S, Humidity) = 0.1516$

$Gain(S, Wind) = 0.0478$

So we choose Outlook variable as an root node. After choosing Outlook as an root node then the Decision tree look like this:



Tem, Hum, Wind

0.97

~~Tem, Hum, Wind~~

0.97

Wind

0.97

Page | 6

Hot = 0
Mild = 1
Cool = 0.0
= 0.570

High = 0.0
Normal = 0
=
= 0.97

Strong = 1.0
Weak = .9183
= .0192

POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

PAGE NO. .....

Random Forest Algorithm:

Random forest is constructed using multiple decision trees and final decision is obtained by the majority votes of the decision trees.

↳ If the problem is of classification type, decision tree algorithm takes the majority votes.

↳ If the problem is of regression type, decision tree algorithm takes the means of decision.

Random forest or random decision forest is an ensemble learning method for classification, regression.

✓ [By this overfitting problem will be resolved].

✓ In classification gives more accurate result in comparison to regression.

For eg:-

Mango type	Source	Sweetness (on scale of 5)	Diameter (cm)
A	M	4.5	7
A	M	4.5	6.8
D	UP	3.8	9
D	UP	3.6	7.9
K	M	3.7	6
K	M	4	6.5

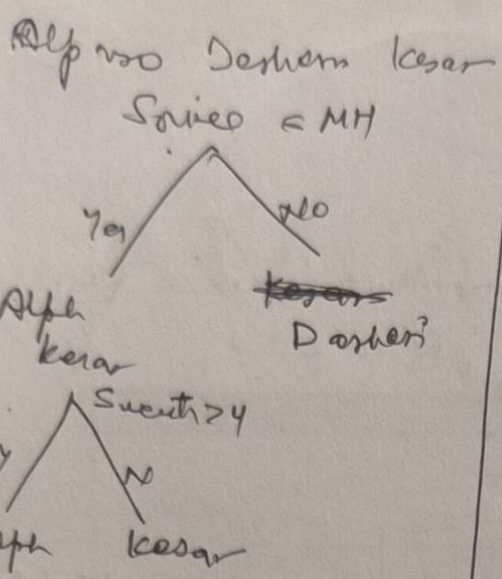
MT	Source	Sweetness
A	Malur	4.5
D	UP	3.8
K	M	3.7
K	M	4

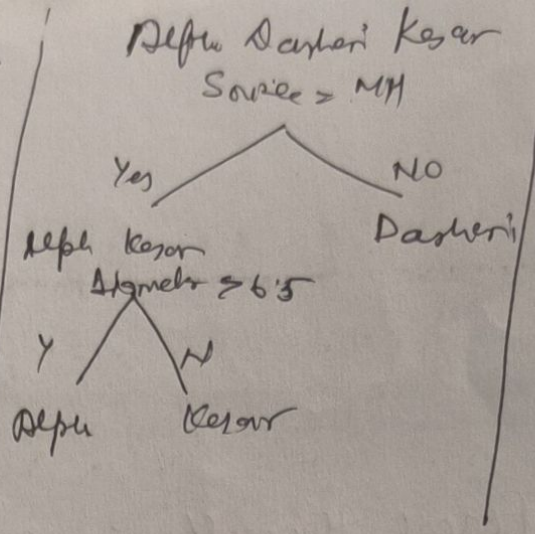
MT	Source	Diameter
A	Malur	7
D	UP	7.9
K	M	8.5
K	M	6.5

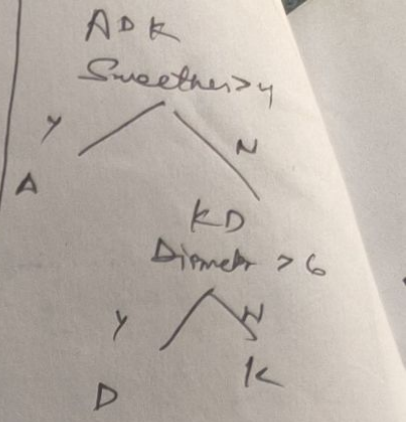
MT	Sweetness	Diameter
A	4.5	6.5
A	4.5	7
K	3.7	6
D	3.8	9

→ The process of selecting features for bootstrap sample is known as Feature Selection.

→ The process of creating rows for subsets is known as Random Sampling.







Feature Selection in Random Forest

→ For classification, by default the feature selection is taken as:

$$= \sqrt{\text{Total number of features in data}}$$

→ For Regression, by default the feature selection is taken as:

$$= \frac{\text{total number of feature in Dataset}}{3}$$

Ensembling means Aggregating the results of decision trees and taking the majority votes in classification.
and mean in case of regression.

New Sample :-

Source	Sweetness	Diameter	
Maharashtra	3.9	6.4	⇒ class of maya = ?
①	②	③	
↓	↓	↓	
Kesar	~~Alph~~ Kesar	D	⇒ class is Kesar.

Advantages:

Low variance
Reduced overfitting
No need for Normalization
High accuracy
High Scalability

Versatility
Handles missing data

Random Forest Algorithm: A random forest is actually just a bunch of decision trees bundles together (that's why it is called as forest).

--Random forest is a **Supervised Machine Learning Algorithm** that is **used widely in Classification and Regression problems**. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Consider the following dataset-

TRAINING DATASET

COLOR	DIAMETER	LABEL
RED	3	APPLE
YELLOW	3	LEMON
PURPLE	1	GRAPES
RED	3	APPLE
YELLOW	3	LEMON
PURPLE	1	GRAPES

CONDITIONS

COLOR== PURPLE?

DIAMETER=3

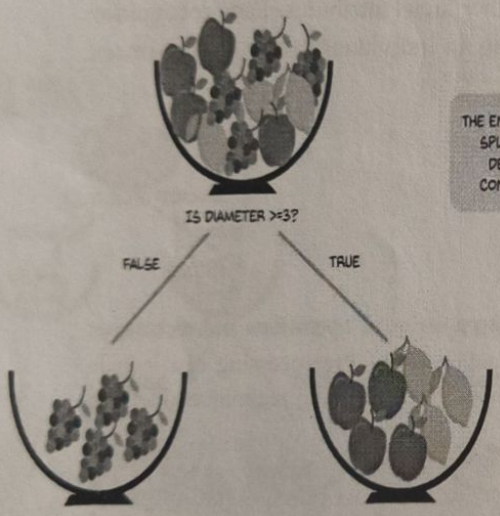
COLOR== YELLOW?

COLOR== RED?

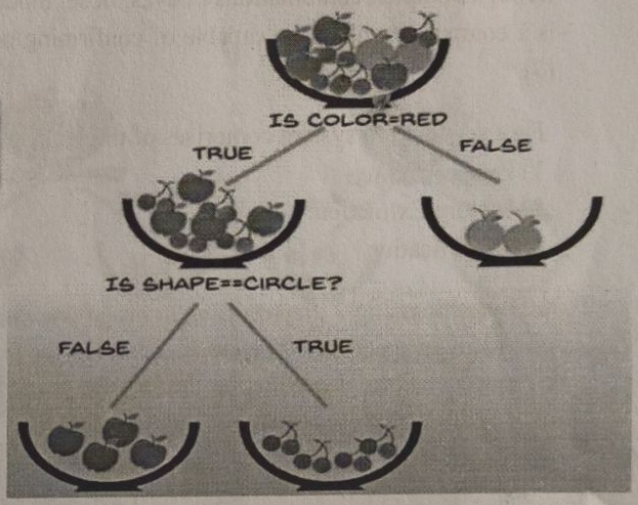
DIAMETER=1

Decision Tree-1

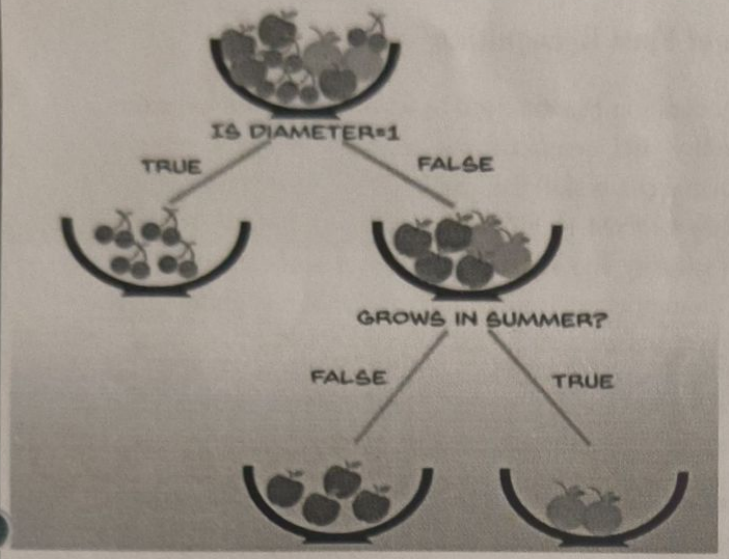
Decision Tree-2



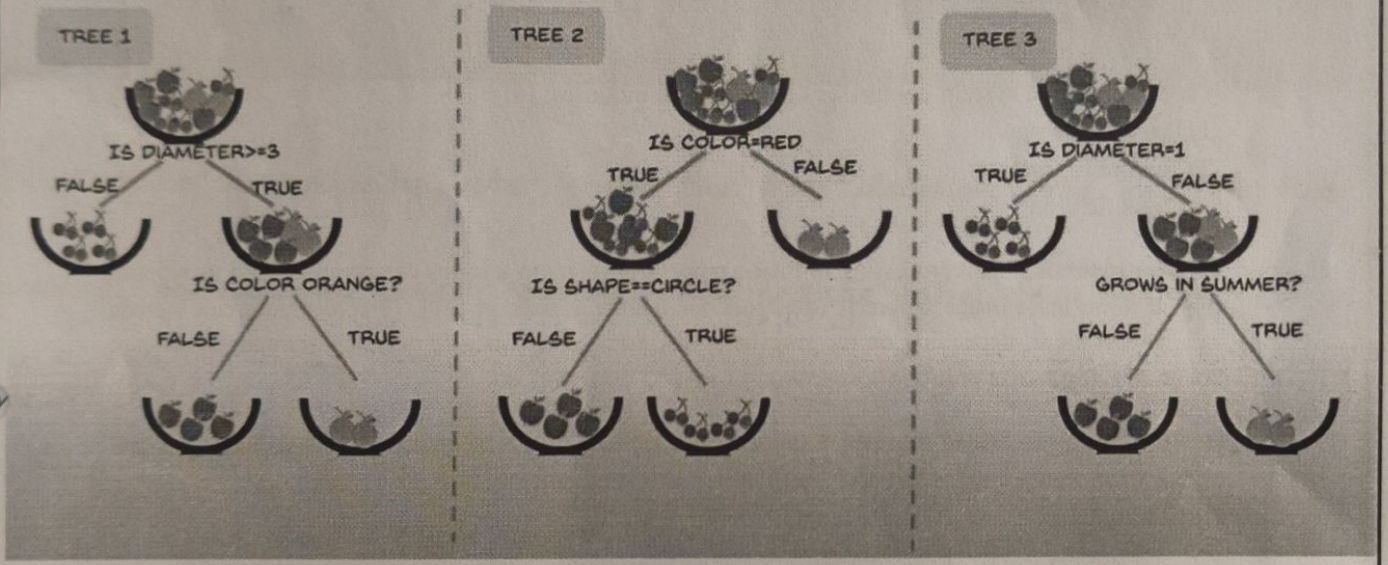
THE ENTROPY AFTER SPLITTING HAS DECREASED CONSIDERABLY



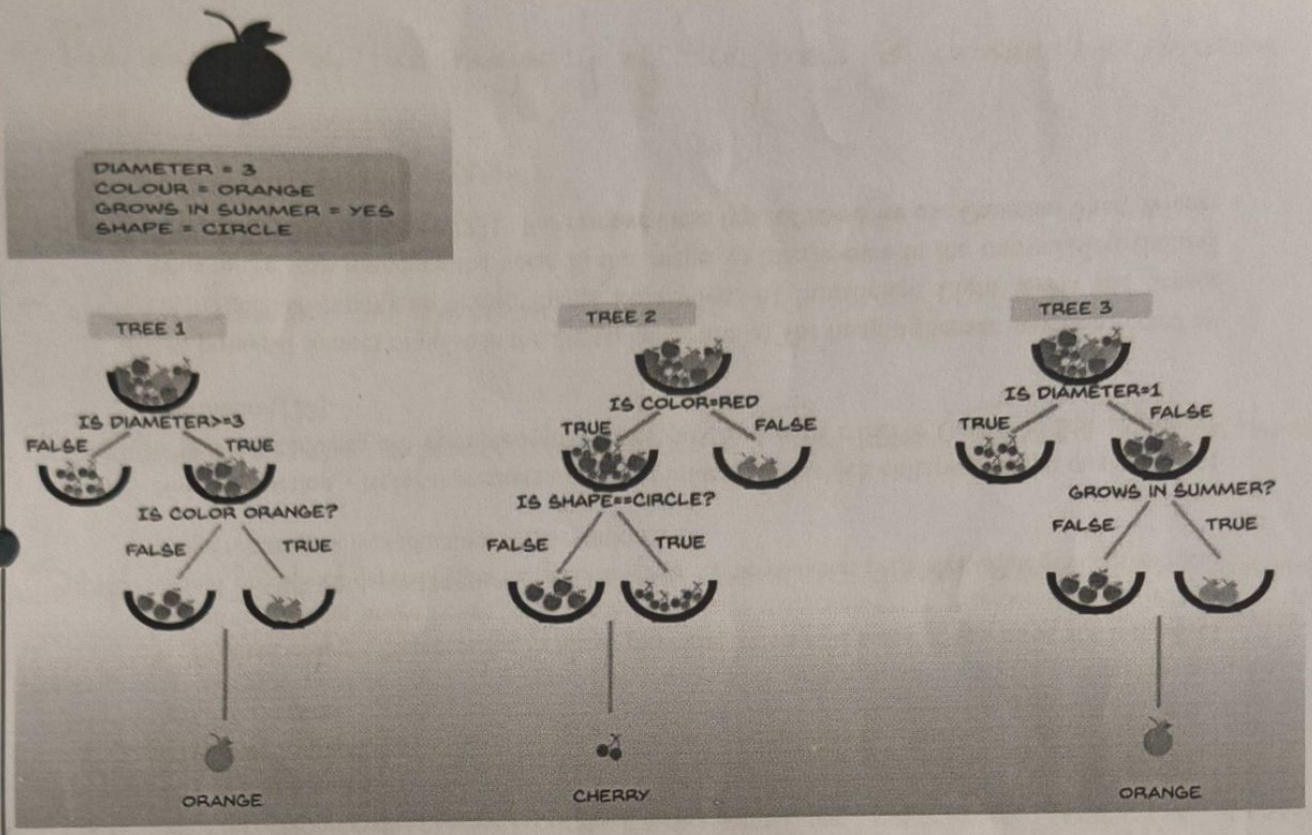
Decision Tree-3



Now all decision trees are as-



Let's take a new fruit and find the name of the fruit.



Now take a majority vote, since the two-decision tree predict the orange, so random forest predict the orange label to classify the new fruit.

--So a random forest or Random decision forest is a method that operates by constructing multiple decision tree during training phase.

--The decision of the majority of the trees is chosen by random forest as the final decision.

Random Forest Algorithm Steps:

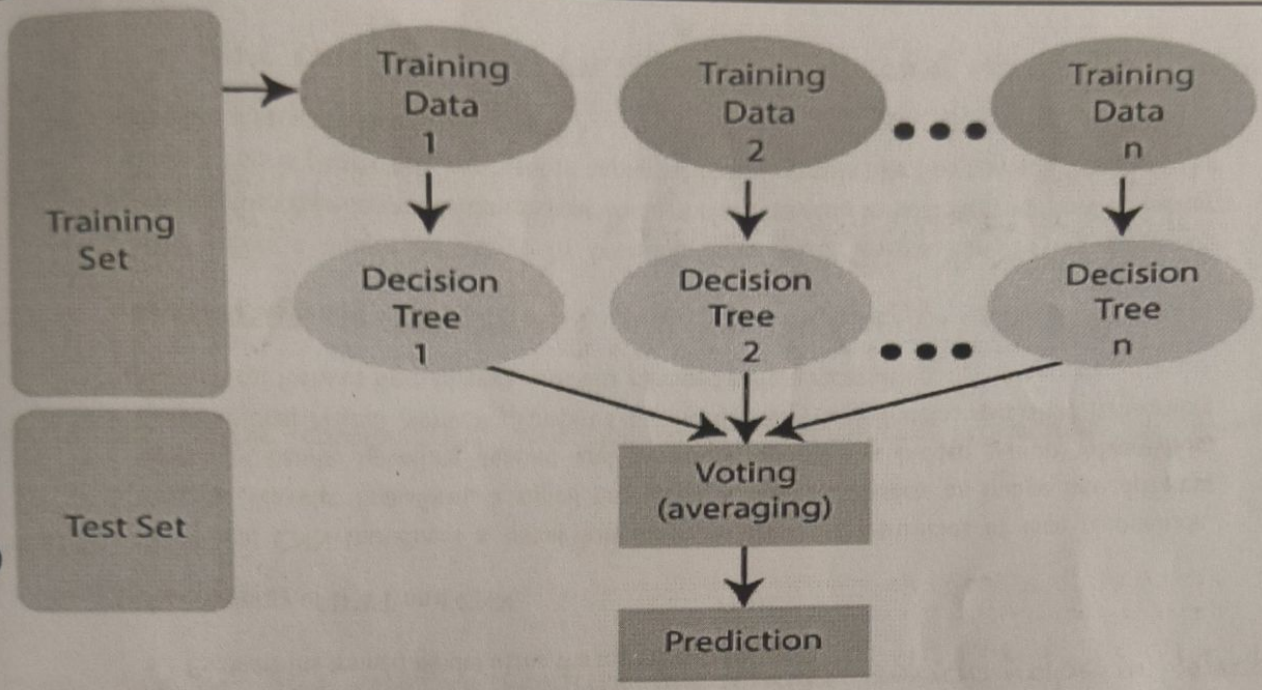
The following steps explain the working Random Forest Algorithm:

Step 1: Select random samples from a given data or training set.

Step 2: This algorithm will construct a decision tree for every training data.

Step 3: Voting will take place by averaging the decision tree.

Step 4: Finally, select the most voted prediction result as the final prediction result.



Advantages:

- Random forests is considered as a highly accurate and robust method because of number of decision trees participating in the process.

- It doesnot suffers from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.

-The algorithm can be used for classification and regression problems.

-Random forests can also handle missing values. There are two ways to handle these. Using median values to replace continuous variables, and computing the proximity-weighted average of missing values.

-you can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

Disadvantages:

-Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then performing voting on it. This whole process is time consuming.

--The model is difficult to interpret as compared to a decision tree, where you can easily make a decision by following the path in the tree.

--Requires much computational power as well as resources as it builds numerous trees to combine their outputs.

--Requires much training time as it combines a lot of decision trees to determine the class.