Unit - I

PAGE NO.:
DATE: 19/25

Pattern Recognition

What is pattern ?

-> A pattern means something that repeats or shows a consistent trends in the data.

-> Pattern can exist in many ~~fra~~ forms such as shape, sound, number or behaviour.

-> Pattern can be physically observed in photos, videos or direct observation.

-> Some patterns are not directly visible but can be found using mathematical or satistical algorithm.

What is pattern recognition ?

-> A pattern refers to any ~~from~~ form of structure, trend or repetition found around us.

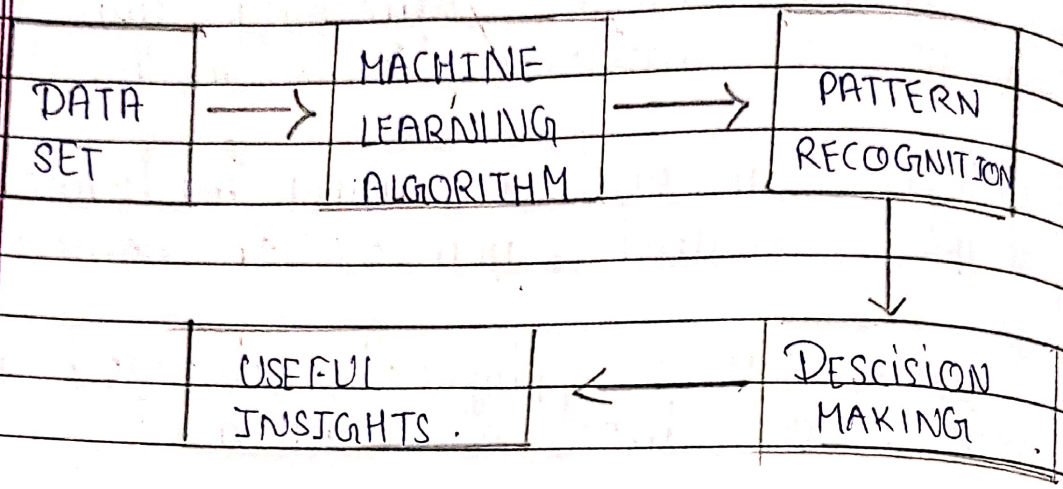
-> Pattern can be seen physically like colors or shapes & they can be observed mathematically by applying algorithms.

-> P.R is a scientific process of identifying & understanding patterns in data. This is usually done with the help of computer program & machine learning.

-> The process involves analyzing incoming data extracting useful features & identifying regularity. is the part of pattern recognition

PAGE NO.:
DATE: 18/05

→ Once patterns are identified they can be applied to the task like prediction & decision-making.



P.R has two main approach :-

i) Explorative : It focuses on detecting unknown patterns within the data.

Descriptive : It categorized the pattern once it is detected.

For an intelligent system the ability to recognise & process patterns is a necessary foundation which helps in ~~desicate~~ decision-making & reasoning.

21/9/20

Goal of Pattern Recognition :

→ Pattern recognition teaching computer to recognise the pattern just like human being.

PAGE NO.:
DATE: 2/9/25

-> A person naturally makes decision by identifying & comparing these patterns.

-> The main goal of pattern recognition is to explain & understand the complex decision-making process.

-> Another important goal is to automate the decision-making without human involvement.

-> P.R simplifies complicated data learn intelligence & automate the decision-making.

A Dataset for Pattern Recognition :-

-> A dataset is a collection of desired data that contains data from the similar field.

-> A good P.R system should be able to recognize familiar patterns quickly & accurately.

-> It should also be capable of recognising and classifying unfamiliar objects.

-> It should also identify shape & object from different angles.

-> It should be identify patterns if objects are hidden or incomplete.

PAGE NO.:
DATE: 2/9/25

A.1 Training & learning in P.R

-> Learning is the process through which a recognition system gets trained & predicts the output correctly.

-> The performance of a P.R system also depends on algorithms used.

-> There are two-~~type~~ process in learning as given below:

a) Training dataset :-

-> About 80% of total dataset is used as training.

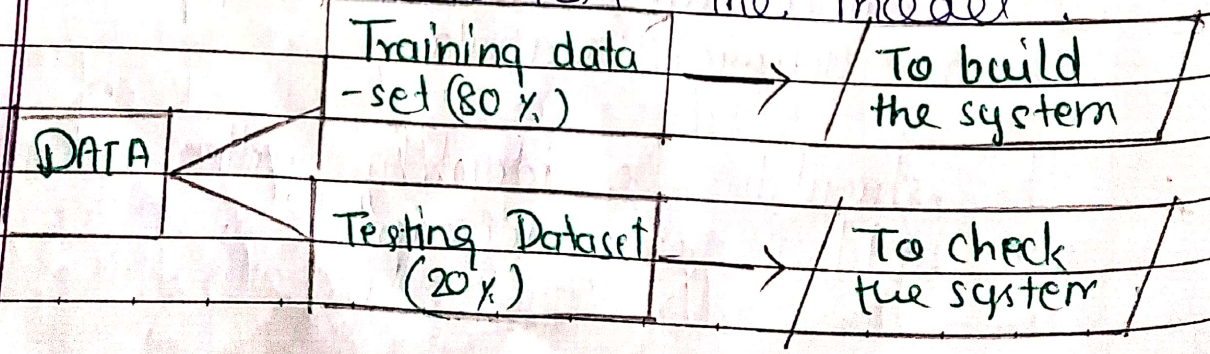
-> Training dataset is used to build & train the model.

-> Different algorithm are applied to extract the data & recognise the pattern.

b) Testing dataset :-

-> It is about 20% of the total dataset

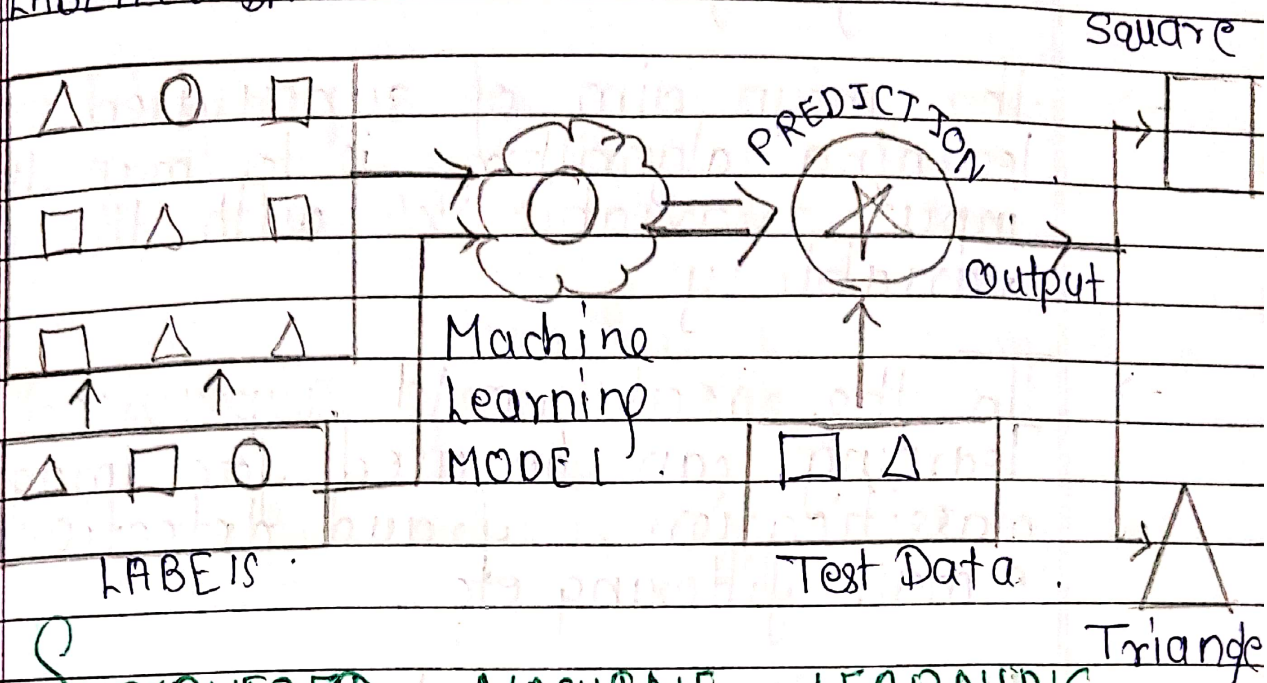
-> It is used to test the model.



SUPERVISED MACHINE LEARNING

PAGE NO.:
DATE: 19/25

Diagram



SUPERVISED MACHINE LEARNING

-> Supervised learning is a type of machine learning in which machines are trained using labelled data. On the basis of labelled training data machine predict the output

-> The labelled data means some input data is already tagged with the correct output

-> The training data provided to the machine work as the supervisor that teaches the maching to predict the output

-> Supervise learning is the process of providing input data as well as

PAGE NO.:
DATE: 4/9/25

correct o/p data by machine learning algorithm.

=> The main aim of supervised learning algorithm is to map the input variable 'x' with the o/p variable 'y'

-> In the real world supervised learning can be used for image classification, fraud detection, e-mail filtering etc.

-> How supervise learning work?

-> In supervise learning models are trained using the labelled dataset where model has knowledge of every data.

-> Once the model training is complete the model is tested using the testing data to predict the output.

-> Supervise learning is not suitable for the complex task.

-> Steps involved in (S.L)

Step1: First, determine the type of training dataset.

Step2: Collect or gather the labelled training data.

PAGE NO.:
DATE: 4/9/25

step 3: Split the dataset into training, testing & validation.

step 4: Determine the suitable algorithm for the model.
Eg: Support Vector Machine

step 5: Execute the algorithm on the training dataset & generate the model.

step 6: Test the model accuracy by giving test data.

Types of Supervise Machine learning Algorithm

i) Regression.
ii) Classification.

i) Regression: Regression alg. are used if there is relationship b/w i/p & o/p variable.

Types of regression:
a) Linear regression.
b) Non-linear regression.
c) Polynomial regression.
d) Regression tree.

PAGE NO.:
DATE: 4/9/25

27. Classification : Classification alg. are used if the o/p data has categories.
Like : yes/no, male/female, true/false, etc.

Types of Classification :
a) Random forest Algorithm.
b) Decision tree algorithm.
c) Logistic regression.
d) Support Vector Machine.

UNSUPERVISED MACHINE LEARNING

→ As we know that supervised machine learning is a process in which we train model using the labelled data.

→ But, in unsupervised machine learning no labelled dataset is used & model & model try finds the hidden pattern in the data.

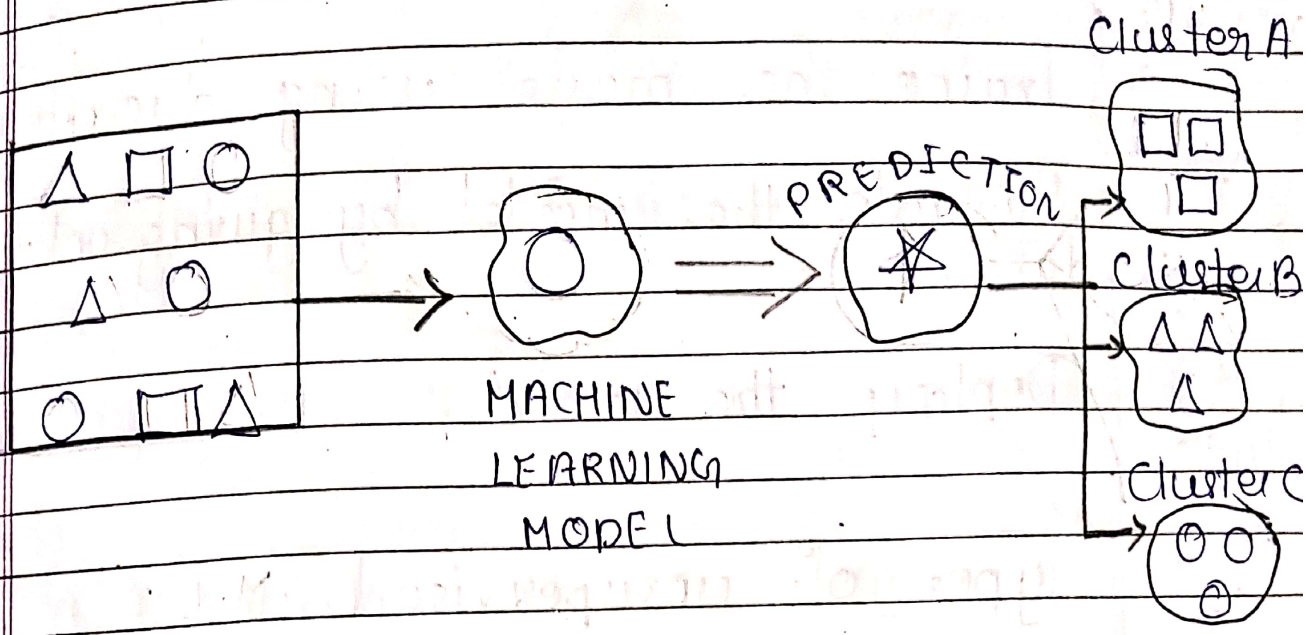
→ Unsupervised M.L work as a human brain in which model trained itself from learning new things.

→ The algorithm learn to automatically group similar th items & find the idden patterns.

Model like karo.
Test a M.L program
train the comp. & tr detect two algo

PAGE NO.:
DATE: 8/9/25

-> This algorithm can be used for task like image segmentation, clustering & association



Why use unsupervised learning?

-> Unsupervised learning is helpful for finding the useful insights from the data.

-> Unsupervised learning is similar to the human learning & thinking capacity.

-> Unsupervised learning work on unlabelled data & uncategorized data.

-> It is also important to find the hidden pattern b/w i/o data.

PAGE NO.:
DATE: 8/9/2

How unsupervised works.

1 -> Collect the data in form of dataset or like live photo, video, text.

2 -> Train the model using algorithms.

3 -> Evaluate the model by giving unlabelled data.

4) Deploy the model.

Types of unsupervised M.L:

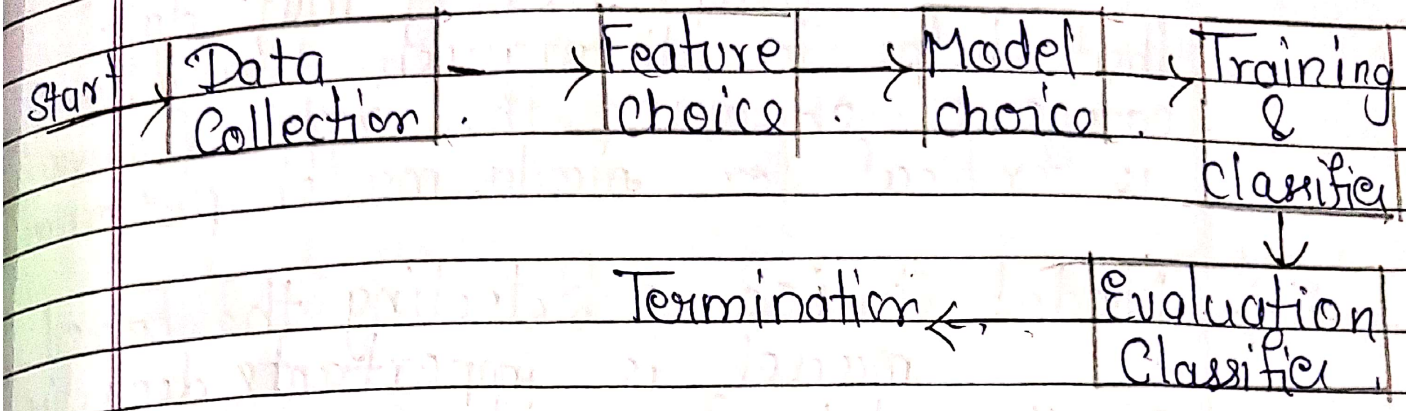
i) Clustering: Clustering is the method of grouping the object into similiar cluster in which all similiar objects are in one cluster & not similiar object in another cluster.

ii) Association: An association is an unsupervised M.L method which is used to find the relationship b/w the input variabe 'x' & with output variable 'y'.

Unsupervised learning Algorithms:

1) K-means Algorithm
2) KNN. (K-Nearest Neighbour) Algorithm
3) Hierarichal Algorithm.
4) PCA (Principale componenentanalysis.
5) NN (Neural Networks),

Activities For Designing Pattern Recognition system :



-> Designing a pattern recognition system evolves a sequence well-define activities to ensure accurate and reliable recognition.

-> It is a step-by-step process from data collection to deployment.

-> Each activities & phase ensures that system learn meaningful patterns.

-> Below points are essential for building the excellent pattern recognition.

i) Data Collection : The first step is gathering relevant data for the problem. The data could include image, audio, video & text or sensor reading.

PAGE NO.:
DATE: 9/9/

ii) Feature Choice (Feature Selection) : Features are the measurable properties of that data that helps to distinguish b/w different categories. choosing the right features is critical for model performance

iii) Model Choice : Selecting the correct model is important depending on the data & problem. Model can be linear, non-linear, neural based.

iv) Training & Classifier : Training involves fitting the model to the collected data. The model learns the relationship b/w i/p features & o/p categories. which includes hyper parameter tuning & optimization

v) Evaluation Classifier : After the training system is tested on unseen data to check the accuracy.

PAGE NO.:
DATE 9/9/23

DECISION BOUNDARY (D.B)

-> D.B is a fundamental concept in M.L & pattern recognition.

-> It is used to separate the data into different classes base on their feature

-> It is # a boundary or surface that separates different classes or categories.

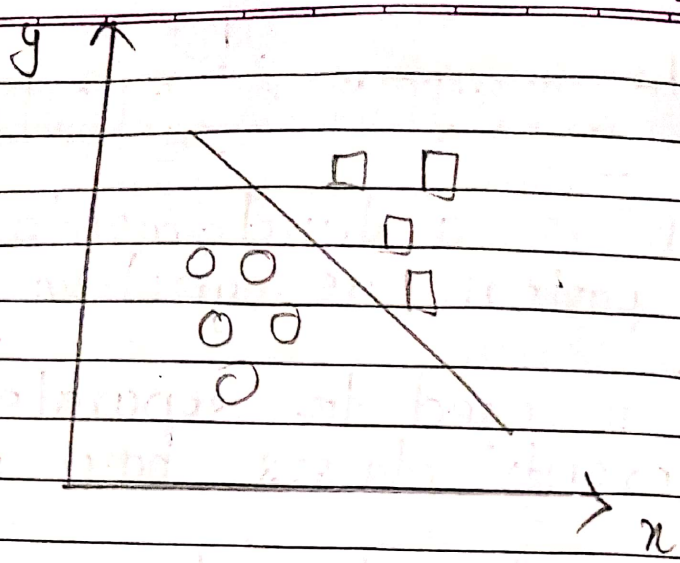
-> D.B are generated using different methods like maximum likelihood, support vector machine, decision tree, neural network.

-> A D.B is a line or curve that divides the data into two or more categories

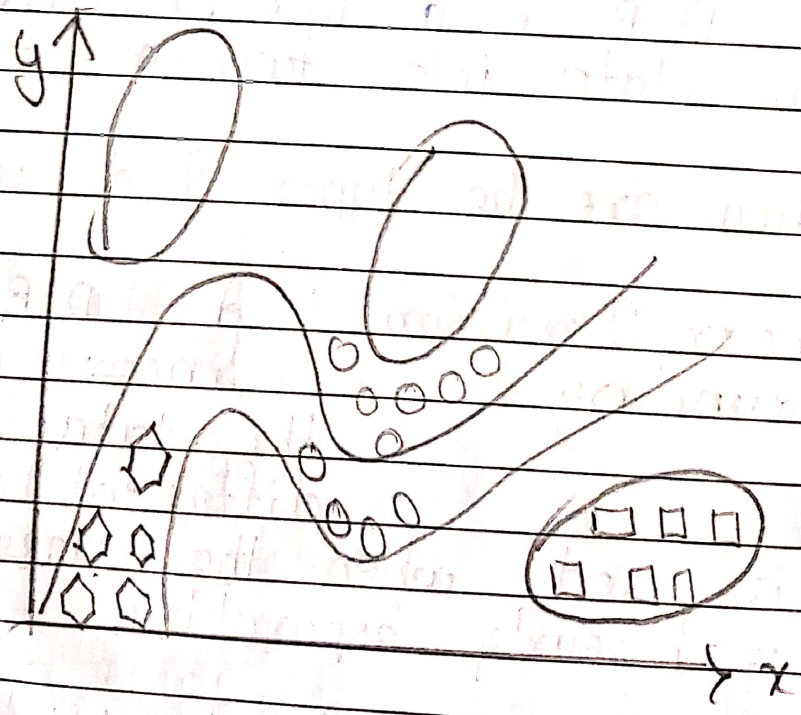
-> Below are the types of decision boundary

1) Linear Decision Boundary : A L.D.B is a straight linear that separates the data into two different classes. It is used when the classification problem is linearly separable.

The equation of straight line is $y = mx + b$ where, m = slope of the line.



ii). Non-linear Decision boundary :
A Non-linear D.B is a curved line that separates the data into two or more classes. Non-linear D.B can take different forms such as circle, ellipse or parabola.



PAGE NO.:
DATE: 4/9/25

Challenges of D.S

i) Overfitting
ii) Underfitting
iii) class imbalance.

4/9/25

i) Overfitting : If is overfitting in machine learning when a model learns the training data too well including noise and random fluctuation.

-> It happens when a model learn not only the actual patterns but also the random noise in the training data.

-> Small data sets are more likely to have the overfitting problem because of limited data.

-> The prevention method of overfitting is: cross-validation, early stopping the training connecting more data & training collecting more data & ~~preve~~ pruning.

ii) Underfitting : It happens when a machine learning model is too simple to capture the pattern in the data.

-> It perform poorly on both training & testing data.

-> Fixing the underfitting uses more complex model train, the model for longer time and add relevant feature.

iii) Class Imbalance : It occure when number of samples in different classes of a data set is unequal

-> This happen mostly in classification problem specially in binary classification

-> Model predict the high accuracy classes & failed to detect the minority class

iv) Noise : It refers to the random irrelevant, unpredictable variance ~~variance~~ in the data

-> It is basically disturbance or error in the data caused by human mistake measurement error or natural randomness

DECISION REGIONS :

-> The boundary that distinguish a class from another in a classification problem is known as decision region.

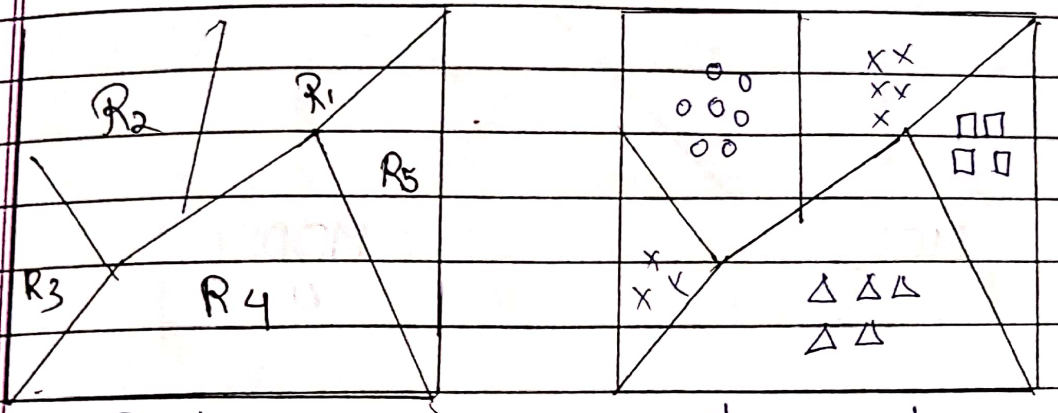
-) A DR is a boundary or surface that divides the input space into region each of which is associated with a distinct class or category

-) Each patter represented ~~for~~ as a point a particular set of ordinate which are define by the value

PAGE NO.:
DATE: 11/9/25

-> Pattern shape are change to decision region which is marked by the patternspace.

-> All of the pattern within a usable decision regicon which belong to the same class.



Regions. classes in regions.

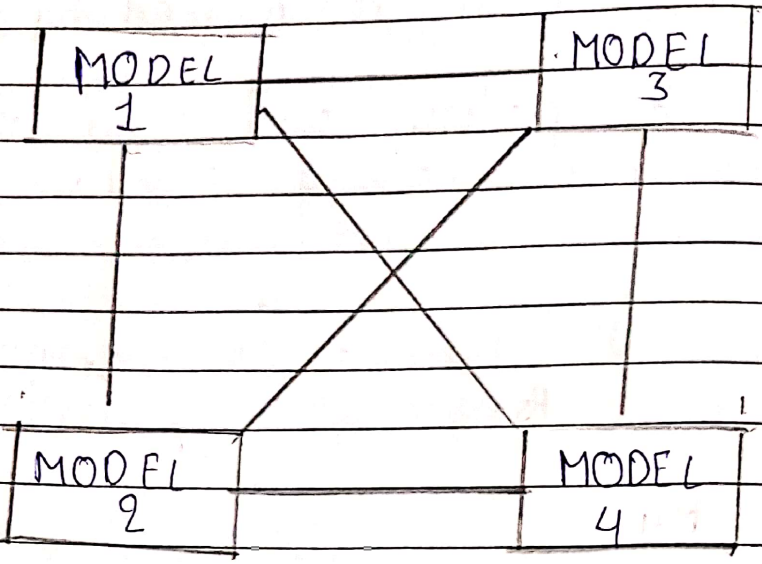
Types of Decision Region

i) Linear decision region. (LD boundary ka ma
ii) Non-linear decision region (NLD boundary ka ma

Algorithm of Decision Region

i). LDA (linear discriminant Analysis).
ii). Logistic Regretion.
iii). Support Vector machine.
iv). Decision tree.
v) Neural Network.

Metric Space

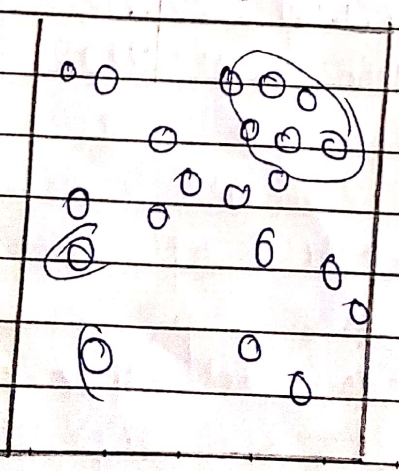


Distance Matrix D

D ↓

.	1	2	3	4	---	. . .
1	0	δ_{12}	δ_{13}	δ_{14}	-	-
2	δ_{21}	0	δ_{23}	δ_{24}	-	-
3	δ_{31}	δ_{32}	0	δ_{34}	-	-
4	δ_{41}	δ_{42}	δ_{43}	0	-	-
	--	--	--	-	-	-
	--	--	--	-	-	-

2D Projection of MDS Plots



PAGE NO.:
DATE: 6/9/24

-> In mathematics a matrix space is a set where a distance (called a matrix) is defined b/w elements of set.

-> Matrix space method has been employed in various application such as internet search engine, image classification, etc.

-> The distance matrix defines the matrix space.

-> The matrix space can be visualized using multi-dimensional scaling (MDS) whose o/p is shown in the last diagram.

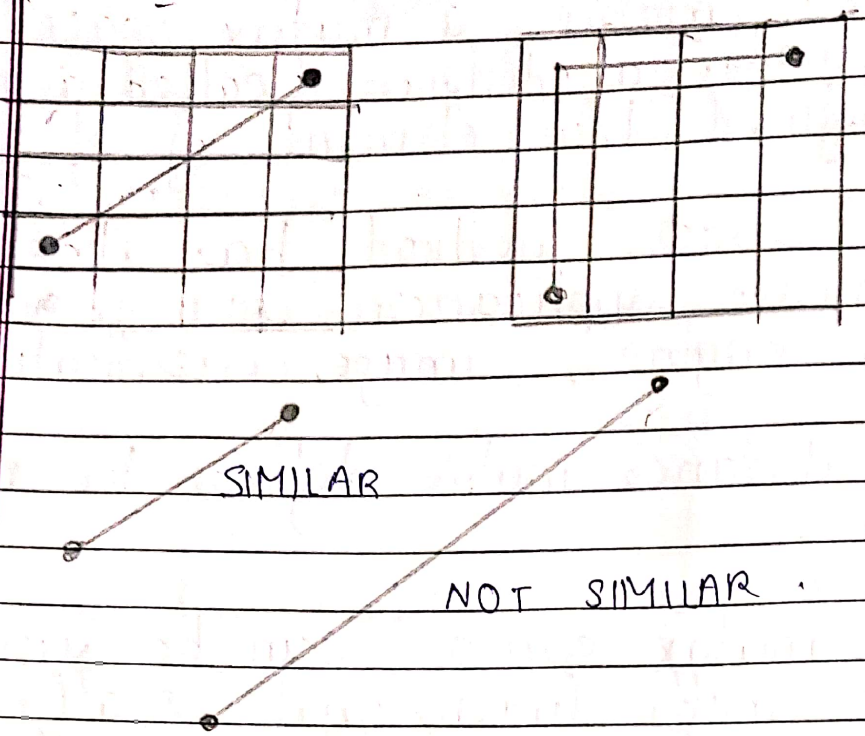
-> The MDS takes the distance matrix & translate the model into different points in a Euclidean distance.

-> The location of the point in the EUCLIDEAN space are optimize using the distance matrix.

-> The MDS plot is a visual & diagonic tool similiar to the PCA algorithm (Principle Component Algorithm.

PAGE NO.:
DATE: 16/9/20

Distance Metrics



→ Clustering is an important part of data cleaning used in the field of artificial intelligence, deep learning & data science.

→ Distance matrics basicly deals with find the distance b/w two data points & determining if they can clustered together

→ Distance matrics is a key-part of several machine learning algorithms

→ These can be used in both supervise & unsupervised learning.

→ An effective distance matrics improve the performance of machine learning model & check whether the model is used for classification & clustering.

PAGE NO.:
DATE: 6/9/25

→ We can calculate the distance b/w two points & then define the similarity b/w them.

→ Types of Distance matrices ~~or~~ in ML are
i) EUCLIDEAN DISTANCE
ii) MANHATTAN.
iii) MINKOWSKI
iv) HAMMING.

PAGE NO.:
DATE: 19/1/25

UNIT - II

CLASSIFICATION

Classification In M.L :

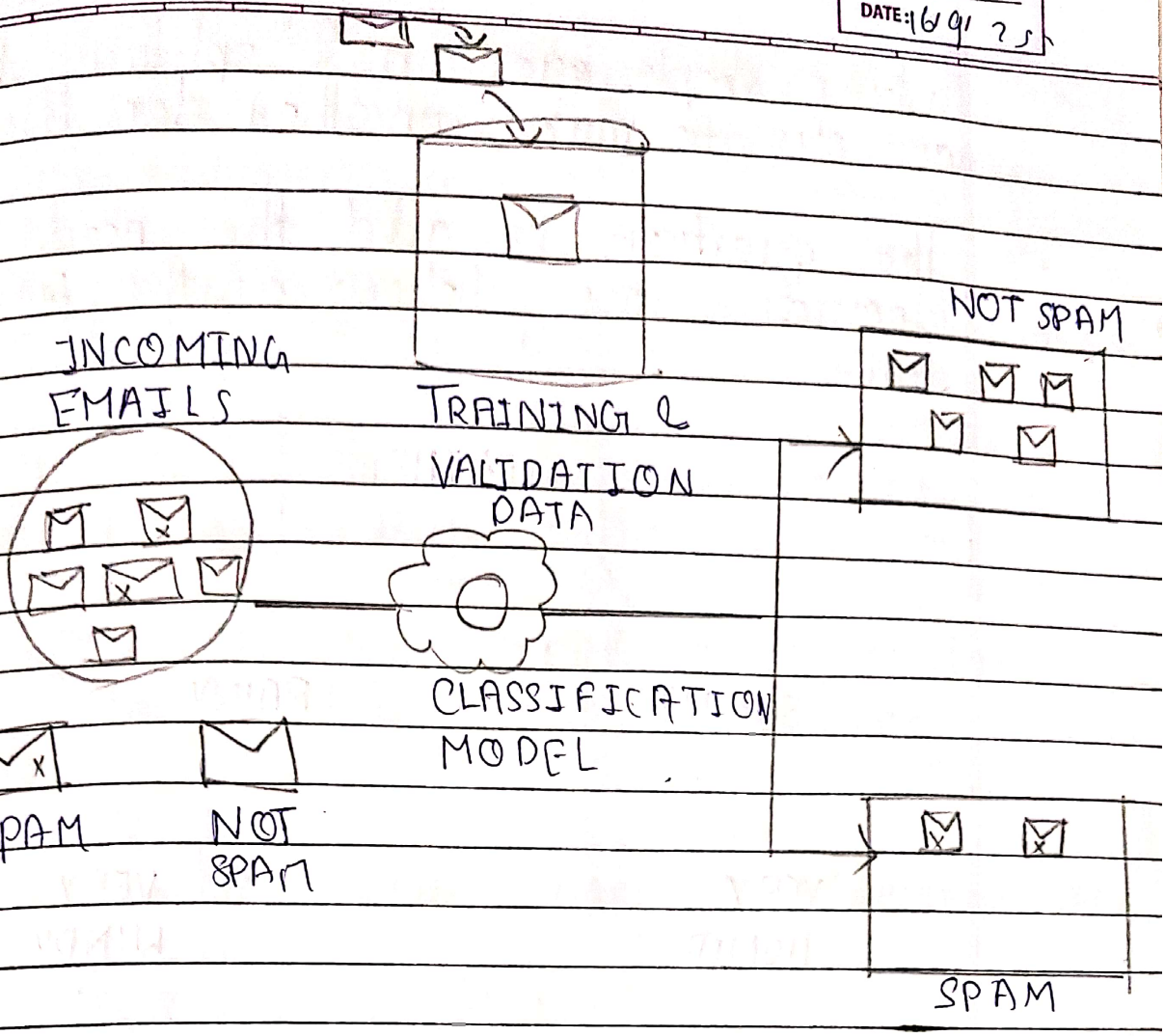
Classification is defined as the process of recognition, understanding and grouping of objects into pre-defined categories.

With the help of pre-categorized training dataset, classification in M.L has wide range of algorithm to classify the dataset.

Classification algorithms used in machine learning utilize the i/p training data for the purpose of predicting that data & fall into one predefine category.

One most common eg. of classification application is filtering the email into spam or not spam categories.

PAGE NO.:
DATE: 16/9/25



Types of Classification:

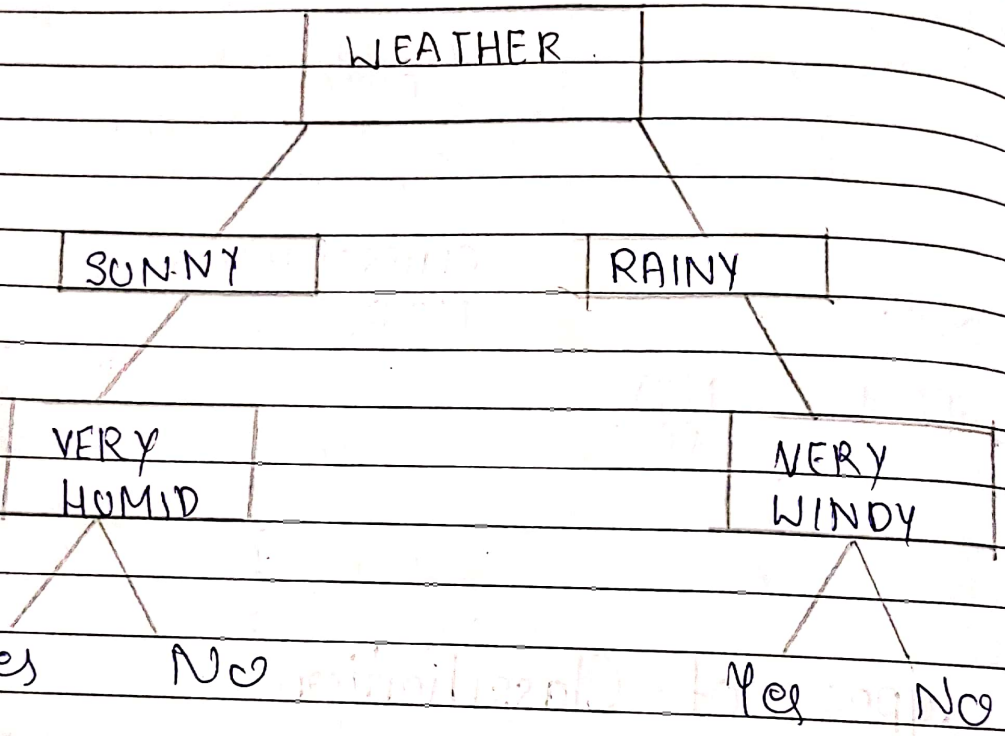
1) DESCISION TREE :
2) NAIVE BAYES
3) LOGISTIC REGRESSION
4) SUPPORT VECTOR MACHINE
5) RANDOM FOREST
6) K- NEAREST NEIGHBOUR

1) DECISION TREE : A decision tree is an algorithm that is used to visually represents decision making.

-> A decision three can be made by

asking simple question & splitting the answers to make another decision.

→ The question is add the node & decisions are below at the least leaves.



From the above figure depending on the weather condition like ~~sum~~ sunny & rainy we can play cricket.

→ A decision tree is a supervise M.L alg. used to predict o/p of the target variable.

→ Supervise learning is used as there is labelled data to make the prediction.

PAGE NO.:
DATE: 29/12/25

→ Advantage of decision tree :

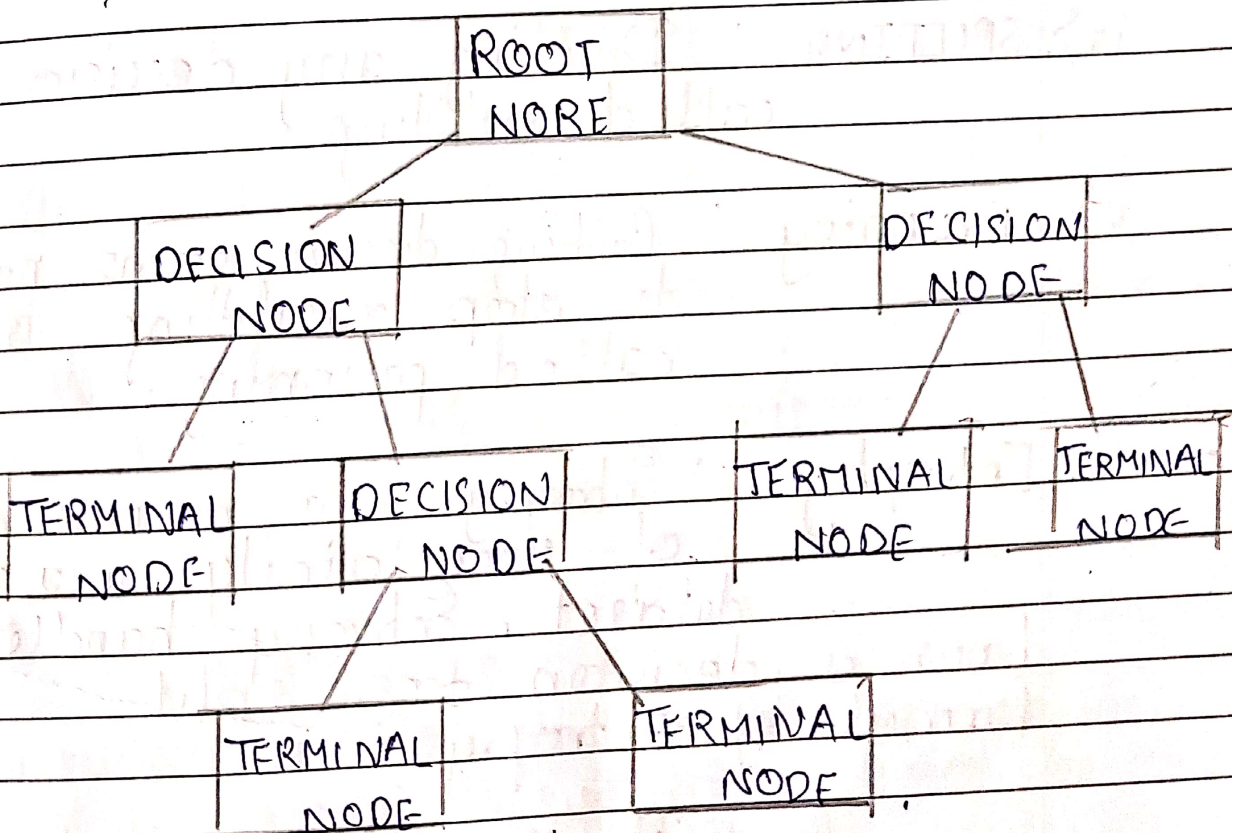
a) Decision tree are simple to understand, interpret & visualize

b) They can effectively handle both numerical & category wise data

c) They can determine worst case & best case easily.

d) Decision tree require little data preparation & data normalization

Important term used in Decision Tree:



PAGE NO.:
DATE: 22/9/25

i) ROOT NODE : ROOT NODE is always the top node of the decision tree. It represents the entire population & can be further divided into different set.

ii) DECISION NODE [Test] : Decision nodes are the nodes on which test is performed & it is further spilted into different sub-nodes or atleast two branches.

iii) TERMINAL NODE or [LEAF NODE] [RESULT] A leaf node in a decision tree carries the final result.

iv) SPLITTING : Dividing any descision is called spliting

v) Prunning : Cutting-down some nodes to stop overfitting is called prunning.

vi) Entrophy : Entropy is a measurement of uncertainity in a given dataset, Entropy handles how a decision tree split formula of Entropy is :

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

PAGE NO.:
DATE: 19/1/23

where,
P = Probability.

Eg: Play Golf

Yes	NO
9	5

14.

Using formula:

Entropy (5, 9)

Entropy $\left(\frac{5}{14}, \frac{9}{14}\right)$

Entropy (0.36, 0.64)

$= -(0.36 \log_2 0.36) - (0.64 \log_2 0.64)$

$= (0.36 \times (-1.4739)) - (0.64 - (-0.6439))$

$= -(-0.530604) - (-0.4129096)$

$= 0.530604 + 0.4129096$

$= 0.94$.

2) NAIVE BAYES:

→ Naive Bayes alg. is the supervised M.L algorithm which is based on the 'bayse theorem for solving the classification problem.

PAGE NO.:
DATE 22/9/25

-> It is mainly used in text classification that includes high dimensional training dataset.

-> It is the most simple & fast machine learning model to make the prediction quickly.

-> It is a probability classifier which means it predicts on the basis of probability of an object.

-> Some popular eg. of naive bayes algorithms are: sentiment Analysis & spam filtering.

-> NAIVE : It is called NAIVE because it assume that the occurence of a certain feature is independent from the other features.

Eg: If a fruit is in red color, sweet ~~test~~ taste & spherical shape the fruit is identified as apple.

-> BAYES : BAYES Algorithm is also known as BAYES Rules which is used to determine the probability of the hypothesis with prior knowledge.

PAGE NO.:
DATE: 21/9/25

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

where,

$P(A|B)$ - POSTERIOR PROBABILITY.
$P(B|A)$ = LIKELIHOOD PROBABILITY
$P(A)$ = PRIOR PROBABILITY.
$P(B)$ = MARGINAL PROBABILITY.

Working of NAIVE BAYES CLASSIFIER

0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Rainy Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes.

Dataset of weather

Weather	Yes	No	Total	← Frequency Table of weather Dataset
Overcast	5	0		
Rainy	2	2		
Sunny	3	2		
	10	4	= 14	

PAGE NO.:
DATE: 23/9/25

→ Suppose we have a dataset of weather condition & corresponding target variable play, so with the help of Naive Bayes classifier we have to decide whether to play cricket or not in a particular weather condition below are the to solve the problem.

a) Convert the given dataset into frequency table.
b) Generate the likelihood table.
c) Now, apply the Bayes theorem to calculate the probability.

Likelihood Table :-

Weather	NO	YES	
Overcast	0	5	5/14 = 0.35
Raing	2	2	4/14 = 0.29
Sunny	2	3	5/14 = 0.35.
Total	4	10	= 14

4/14 = 0.29 10/14 = 0.71

Applying Bayes Theorem

$$P(\text{Sunny}/\text{yes}) = \frac{P(\text{Sunny}/\text{yes}) \times P(\text{Yes})}{P(\text{Sunny})}$$

PAGE NO.:
DATE: 28/9/21

$P(\text{Sunny} | \text{yes}) = \frac{3}{10} = 0.3$

$P(\text{Sunny}) = 0.35$

$P(\text{Yes}) = 0.71$

So, by formula : $\frac{0.3 \times 0.71}{0.35}$

$= 0.60$

$P(\text{Sunny} | \text{No}) = \frac{P(\text{Sunny} / \text{No}) \times P(\text{No})}{P(\text{Sunny})}$

$P(\text{Sunny} | \text{No}) = \frac{2}{4} = 0.5$

$P(\text{Sunny}) = 0.35$

$P(\text{No}) = 0.29$

So, by formula : $\frac{0.5 \times 0.29}{0.35} \Rightarrow 0.41$

$P(\text{Sun} | \text{yes}) > P(\text{No})$

So, we can play Cricket.

Advantages of Naives Bayes Classifica:

→ It can be used for binary classifica and multiclane classification..

PAGE NO.:
DATE: 23/9/25

-> It is one of the most fast ML algorithm.

-> It is most popular for text classification.

Disadvantage of Naive Bayes :-

-> Sometimes it cannot learn the relationship b/w features.

3).

Qn) LOGISTIC REGRESSION ALGORITHM :

-> Logistic regression is one of the most popular ML algorithm under supervised learning technique for predicting the category based dependent variable using a given set of independent variable.

-> The outcome of the logistic regression is a categorical or discret value like yes/No, 0/1, true/false, it gives the value b/w these.

-> Logistic regression is much similar to the linear regression in which logistic regression is used to solve classification problem.

-> In this algorithm we have S-shaped logistic function called S-curve which predict maximum & minimum value.

PAGE NO.:
DATE: 23/9/25

Logistic function (Sigmoid):

=> The sigmoid function is a mathematical function used to map the predicted value to the probabilities.

-> It maps any real value into another value within the range 0 & 1.

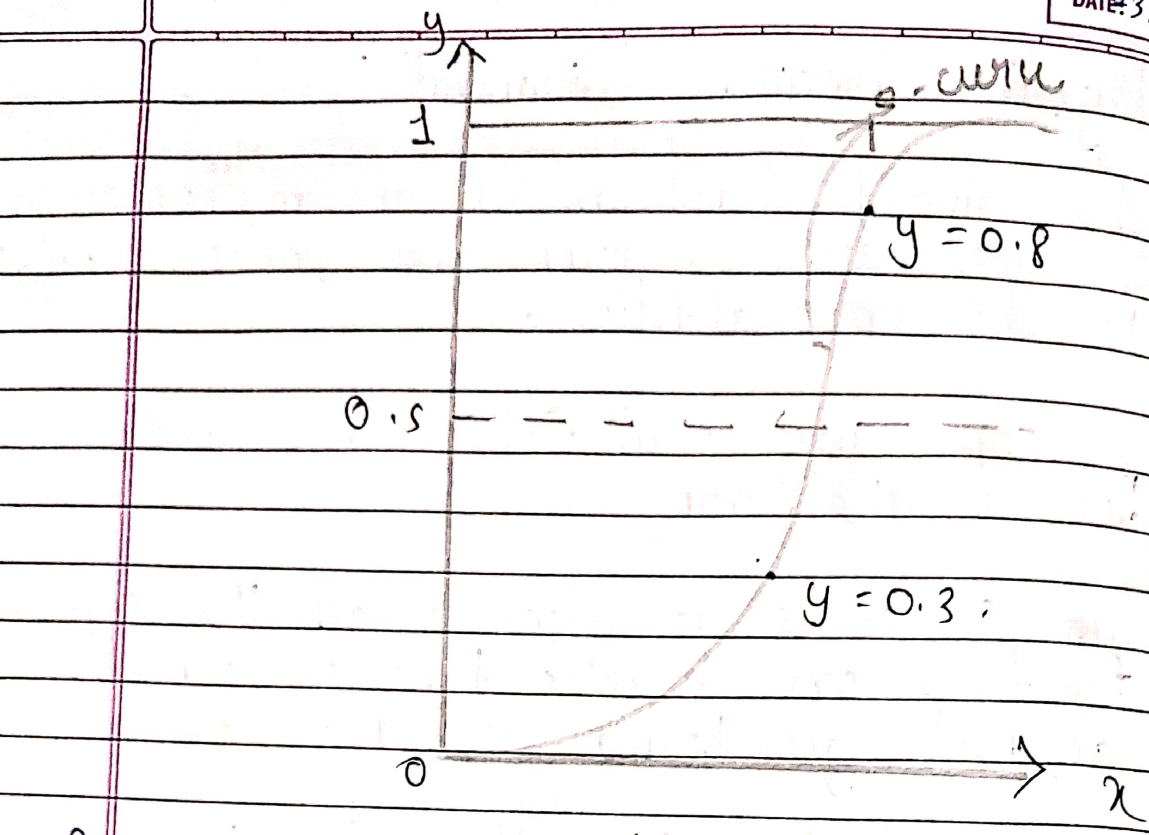
-> The s-curve is also called the sigmoid function or logistic function which cannot go beyond 0 & 1.

Types of Logistic Regression:-

i) Binomial: There can be only two possible types of dependent variable - 0/1, true/false, pass/fail.

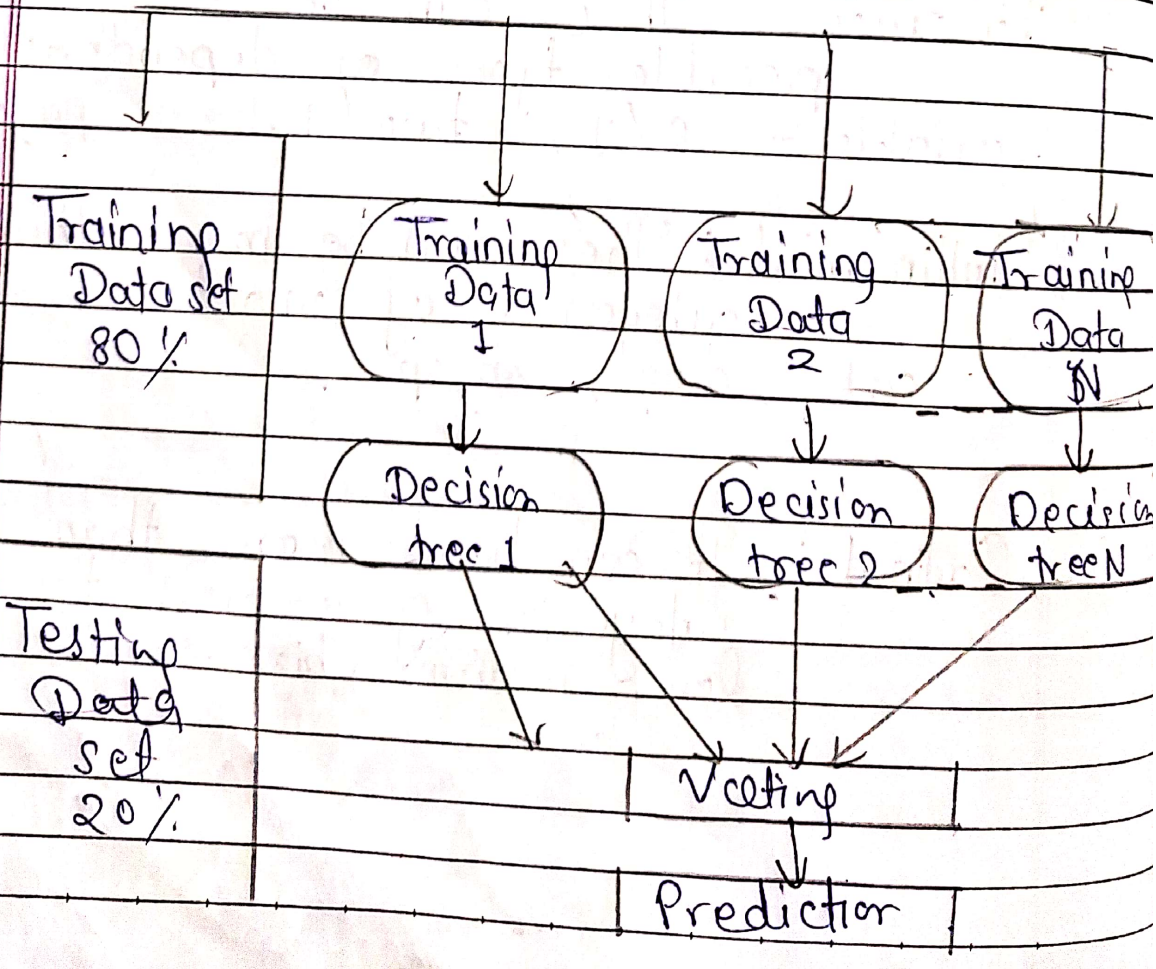
ii) Multinomial: There can be more than 3 categories of unordered type like cat, dogs, sheep.

iii) Ordinal: It can have more than 3 categories of ordered type like low, medium high.



IMP

RANDOM FOREST ALGORITHM



PAGE NO.: _____
DATE: 1/10/23

-> Random forest is the popular M.L algo. that belongs to the supervise learning technique.

-> It can be used for both classification and regression problem.

-> It is based on the ensemble learning technique which means two or more algorithms can be combine

-> Random forest is a classifier that contains a no. of decision trees on various subset of the given dataset, & takes the average to improve the predictivity & accuracy of the dataset.

-> Instead of depending on one decision tree, the random forest takes prediction from different decision tree based on the majority votes of prediction, it predicts the final output

-> The greater no. of decision tree in the forest leads to the higher accuracy

-> How Random forest Algorithm work:

Step 1: Select random 'k' data points from the training set.

PAGE NO.:
DATE: 11/10/25

Step 2: Build the decision tree associated with the selected data point.

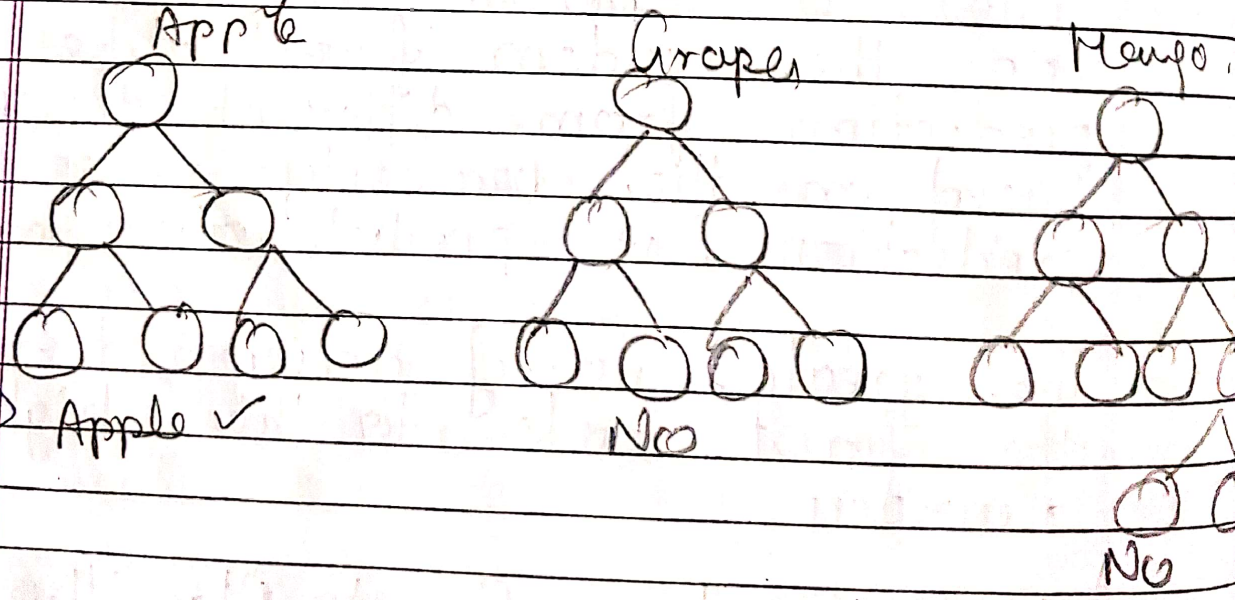
Step 3: Choose the no. of 'n' for decision tree that you want to build.

Step 4: Repeat step 1 & 2 until the final 'n' no. is found.

Step 5: For new datapoints, find the predictions of each decision decision tree & assign the new data point to the category that wins.

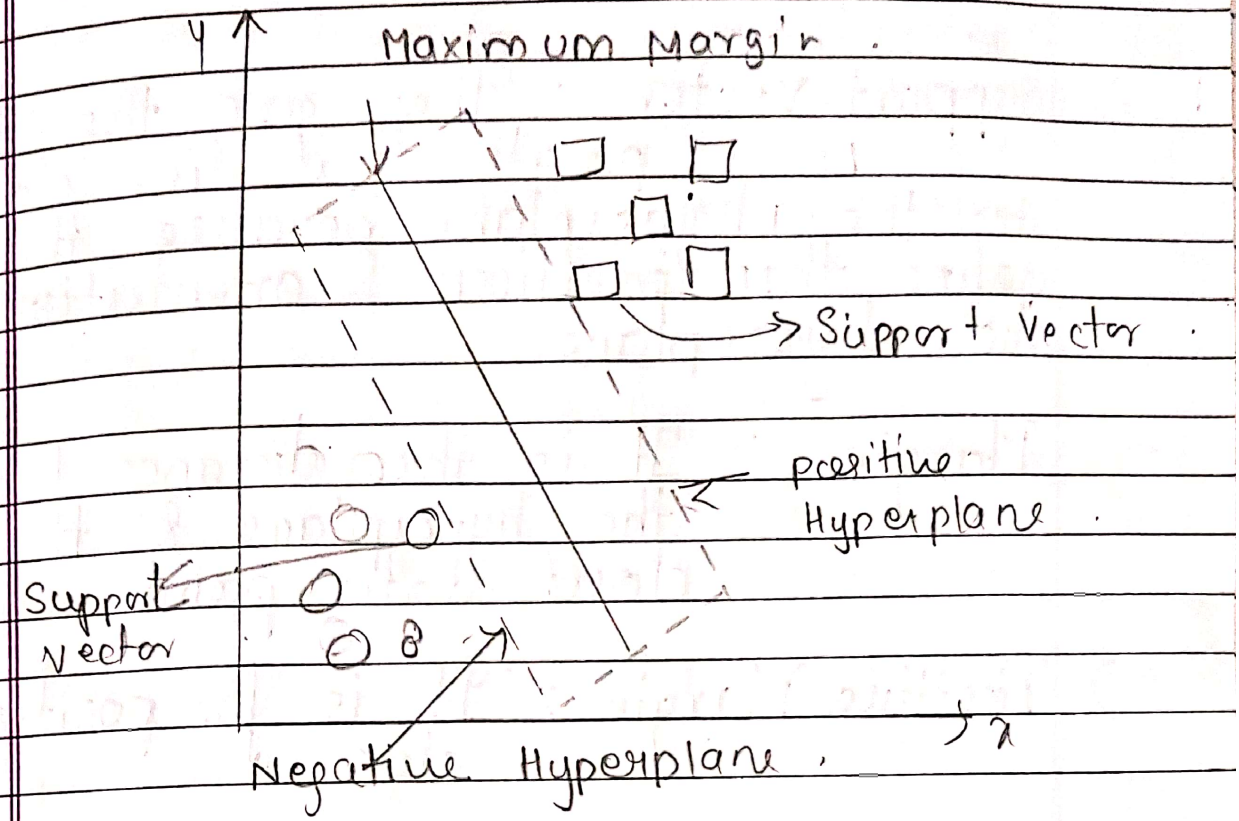
Eg.

Fruit Basket : In this there are some fruits we have to find that.



PAGE NO.:
DATE: 11/10/25

SUPPORT VECTOR MACHINE. (SVM)



→ SVM is a supervised M.L algorithm used for classification & repression problems.

→ The main goal is to find the best boundary [hyperplane] that separates simple classes of data point in the end 'n' dimensional space.

→ Hyperplane is a decision boundary that separates data into different classes
- in 2D its a line,
in 3D it a plane
& in 'n' dimensional it is hyperplane.

PAGE NO.:
DATE: 1/01/24

→ It splits the data like positive & negative, happy & sad, spam & not spam

→ Support vector : They are the data points that lies closer to the hyperplane because they define the position & orientation of the hyperplane

→ Margin : It is the distance b/w the hyperplane & the closet data point

→ Positive Margin : It is for positive class

→ Negative Margin : It is for negative classes

→ The ideal SVM algorithm tries to maximize this margin that provides maximum separation b/w two classes

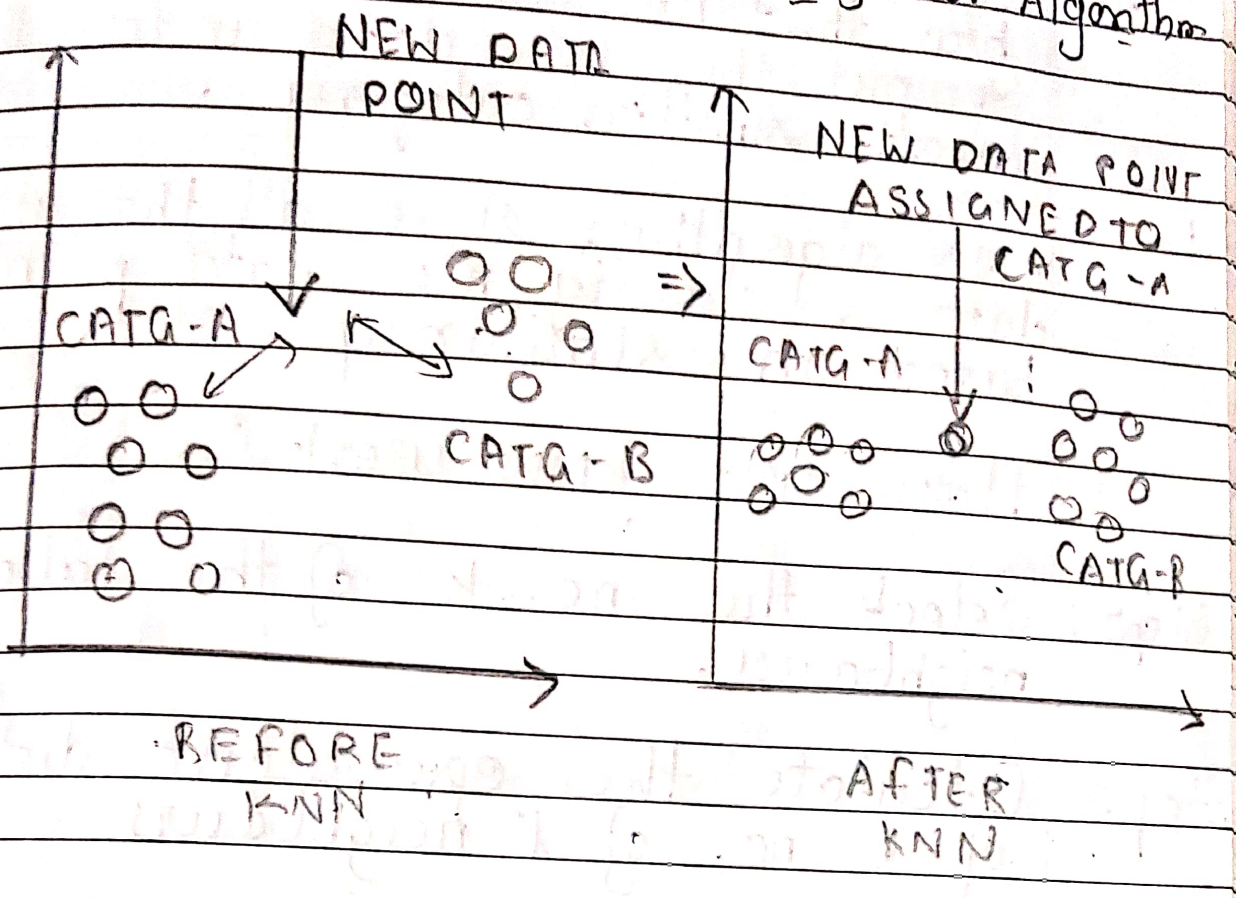
→ Hard Margin in SVM means both classes are perfectly separable.

→ Soft Margin mean both classes are overlapping

→ SVM alg. is used in face recognition, sentiment analysis, speech recognition, image classification.

PAGE NO.:
DATE: 3/10/25

(KNN) K - Nearest Neighbor Algorithm



→ KNN is one of the simplest M.L algorithm based supervised ~~lan~~ learning technique.

→ It can be used for both classification & regression problems.

→ KNN is a non-parametric algorithm which means it does not make any assumption on the underlying data.

→ It is also called lazy-learner algorithm because it does not learn from training dataset immediately, instead it stores the dataset & used it at the time of classification.

PAGE NO.:
DATE: 13/10/23

→ KNN algorithm assume the similarity b/w the new data & available data & put the new data into the most similar category.

→ KNN algorithm store all the available data & classify new data points based on similarity.

→ How does KNN work?

Step1: Select the no. k of the labo neighbour.

Step2: Calculate the equivalent distance of 'k' no. of K neighbours.

Step3: Take the k nearest k neighbours as per the calculated equivalent distance.

Step4: Among these k-neighbours count the no. of data points in each category.

Step5: Assign the new data points to the category that has maximum neighbours.

Advantage:

→ It is simple to implement, can handle noisy training data & can work with large dataset.

PAGE NO.:
DATE: 3/10/21

Disadvantage

-> If value of k is large then some times model become complex & distance calculations becomes difficult

DIFFERENT APPROACHES TO PROTOTYPE SELECTION IN P.R

Step 1: Understand your project goal:

a) Identify what you want to achieve.

b) like prediction, classification, recognition, etc.

b) Choose the algorithm according to the need like:

-> Supervise for prediction based on labelled data.

-> Unsupervise for clustering

-> Reinforcement for learning from the past actions & classification for image recognition.

Step 2: - Analyze your data:

a) Check your datatype & quality like raw data, biased data, structured or unstructured data.

b) Ensures sufficient cleaning data for supervise learning.

c) If data is limited or unstructured prefer unsupervised learning.

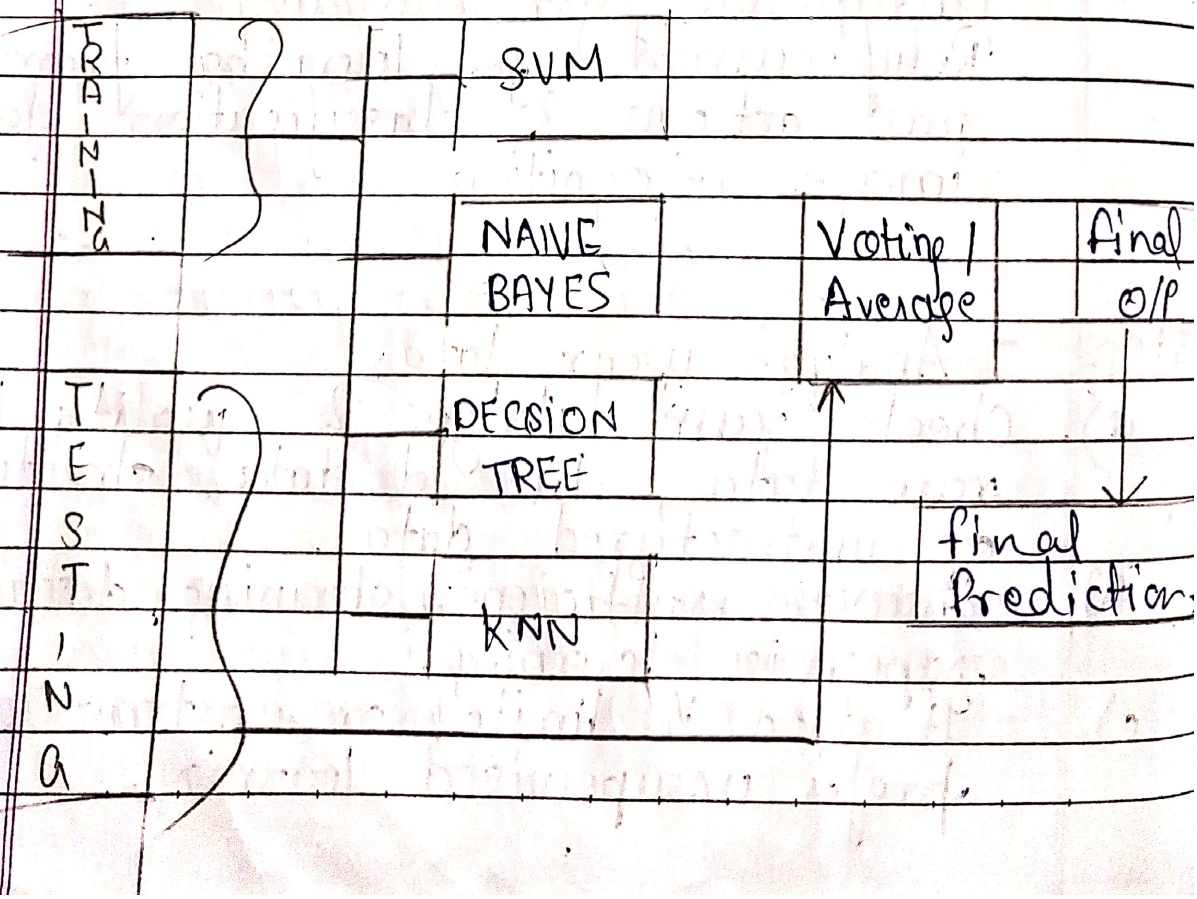
PAGE NO.:
DATE: 13/10/2

Step 3: Evaluate speed & training time.
a). Decide priority b/w speed & accuracy.
b). If you have more & better data training will be longer & result will be better.

Step 4: Check linearity of the data:
a). Use linear algorithm like linear regression, SVM algorithm for simple linearly separable data.

b). for complex patterns or high dimensional data use non-linear algorithms such as random forest, neural network.

COMBINATION OF CLASSIFIER ENSEMBLE METH



PAGE NO.:
DATE: 13/10/25

-> An ensemble of classifier means using a group or combination of different classifier whose individual o/p are combined to make a final decision.

-> Instead of depending on one model we take the opinion of many model & merge their result.

-> There are several region of combining of different classifier such as:

a) The training data may not be enough to create one perfect model

b) Sometimes, single algorithm cannot find the best possible solution.

c) Sometime, the dataset is very large & cannot handled by single classifier.

-> O/p of each classifier is merged & voting is done to make the complex decision.

-> In simple words ensemble of the classifier is a committee where each model give its vote & together they make smater & more reliable decision.

PAGE NO:
DATE 13/10/25

NORMALIZATION

→ It is one of the most frequently used data preparation technique which helps us to change the value of numeric value of numeric column in the dataset to use a common scale.

→ It is used whenever the attribute of the dataset has different ranges

→ It help to enhance the performance & reliability of the model.

→ It is not necessary for all the dataset, but it is required only when features of m.l model have different ranges.

→ Normalization is a transformation technique that improves the accuracy of the model.

→ It is used when we don't know feature distribution exactly.

→ It is useful in scaling technique such as kNN & artificial Neural Network (ANN).

PAGE NO.:
DATE: 13/10/23

Formula :

$$x_n = \frac{(x - x_{min})}{x_{max} - x_{min}},$$

where,

x_n = Value of normalization.
x_{max} = Maximum value of feature.
x_{min} = Minimum value of feature.

Standardization

x' -

It is a scaling technique also known as Z- ~~Z normalization~~ Z-score normalization in which values are centered around the mean with a unit standard deviation.

Mathematical we can calculate the standardization by subtracting the feature value from the mean dividing it by standard deviation.

This technique is helpful for various m.l algorithm like KNN, k-mean clustering & PCA (Priciple Compoment Analysis).

This is used when we are exactly aware of the feature distribution & your data follows Gaussian distribution.

PAGE NO.:
DATE: 4/10/25

-> It is also useful when data has variable dimensions & techniques such as linear - regression & logistic regression

-> Formula :

$$x' = \frac{x - \mu}{\sigma}$$

μ = mean of feature values

σ = Standard deviation feature value.

Normalization	Standardization
-> This technique use minimum & maximum value for scaling of model.	-> This technique use mean & standard deviation for scaling of model.
-> It is useful when features are of different scale	-> It is useful when the mean of variable is set to 0 & standard deviation set to 1.

Normalisation	Standardization
Scale value ranges b/w [0,1] [-1,1].	Scale values are not restricted to a specific range.
-> It get affected by outlier.	-> Not affected by outlier.
-> It is also called scaling normaliza -tion.	-> It is also called Z-score normalizatin.
-> It is useful when feature distribution is unknown.	-> It is useful when feature distribution is normal.