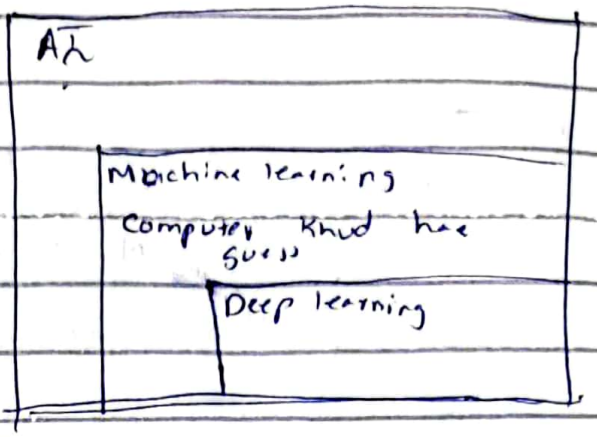
Session 1-2.

① Machine learning

"Field of study that gives computer the ability to learn without being explicitly programmed"



a) Start with data → E (experience)

b) perform some sort of task → T (task) (Tasks is called supervised learning, classification or inductive learning)

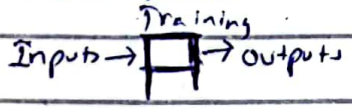
c) Validation test → Performance measure P

Traditional

ML

ML

- Supervised (• Classification, • Regression → numerical)
  - learn explicitly
  - Output clear
  - Predicts future
  - Direct feedback

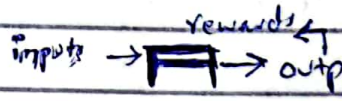


  - Medical Diagnosis
  - labeled data
  - 2 steps - learn (training)
    - Testing
  - Decision trees, Linear Regression, KNN, SVM, Neural networks (black box)

- Unsupervised (• detection of anomalies, • Association, • clustering, • Dimensionality Reduction)
  - machine learn
  - do not predict

  - Fast Students new
  - unlabeled data.
  - K-means clustering
  - Apriori Algo

- Reinforcement
  - Approach to AI
  - learning from +ve or -ve reinforcement
  - maximize reward



  - Reward based learning
  - netflix, youtube recommendations.

Machine learning Process

gather data → clean data for homogeneity → Model building (select right ML algo) → Gaining insights → Data visualization

Accuracy:- (Classification problem)

$$\text{Accuracy} = \frac{\text{No. of correct classification}}{\text{Total no of test cases}}$$

Classification:- (2 things)

- We predict y labels (classes) for input u.
- Yes or No (fraud or not fraud)
- $y = f(u)$ → input
  - output
  - Classification function

Regression:- Numerical data (1 thing)

Supervised

- feedback
- Used for prediction
- Known number of classes
- less complex than →

Examples

- Text recognition (training no rqmt)
- Spam detection
- Face detection
- OCR
- Fraud detection.

Unsupervised.

no feedback
Used for analysis
Unknown numbg of classes.

Examples

Noise removal from datasets!
Recommandation engines
Customer Behaviour prediction.

- Different Matric :-

Confusion Matrix

Accuracy

Precision

Auc (Area under the curve)

MAPE (Mean absolute % error)

MAE (Mean absolute error)

MSE (Mean Squared error)

F1 Score

	Actual	
Predicted	TP	FP
	FN	TN

We can not apply unbalance data to accuracy.

- **Accuracy :-** $\frac{TP + TN}{Total}$ (TP, TN) • accuracy_score

- **Precision :-** $\frac{TP}{TP + FP}$ (TP, FP) (How many did we match)

Sb kay sb predicted Cheezon ma

- **Recall or Sensitivity :-** $\frac{TP}{TP + FN}$ (TP, FN) (How many we miss)

If every person has cancer, recall is 100%.

- If focus on minimize FN, Recall → 100% without precision too bad

- If on " FP, Precision → 100%.

- **Specificity :-** (Opposite to recall)

(FP / TN)

$$\frac{TN}{TN+FP}$$

- **F1 Score :-**

Single score that represent Precision and recall.

F1 Score = Harmonic Mean (Precision, recall)

$$F_1 \text{ Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **AUC :-** (Scale - invariant) (0 - 1 range) (If 100% correct AUC 1 hoti hai)

a) True - Positive rate (TPR)

$$TPR = \frac{TP}{TP+FN}$$

b) False - Positive rate (FPR)

$$FPR = \frac{FP}{FP+TN}$$

- **MAE :-** (Avg diff b/w orignal and predicted values)

$$= \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)$$

- **MSE :-**

$$= \frac{1}{n} \sum_{i} (y_i - \hat{y}_i)^2$$

- **MAPE :-**

$$= \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Decision Tree (if one attribute 0, answer is zero)

- Calculate entropy and gain of all the columns
- From these columns whose gain is more it means that impurity is less and we select that as a root node.

- **Step 1:-** Calculate entropy of whole dataset.

$$-\frac{C(P)}{C(T)} \times \log_2 \frac{C(P)}{C(T)} - \frac{C(N)}{C(T)} \times \log_2 \frac{C(N)}{C(T)}$$

Step 2:- Entropy of all attributes of that column.

→ Same formula.

Step 3:- gain = Entropy(whole data) − $\frac{1^{st} \text{ attribute total values}}{\text{Total}}$

1st
Entropy of attribute −

Step 4:- Maximum gain jis ka ans ko root node.

Root node ⇒ splitting attribute.

Attribute Selection Measures (ASM) / Splitting rules.

Popular are ⇒
- Information gain (decrease in entropy)
- Gain ratio
- Gini index

Data imbalancing ⇒ Red blck] Jo mam na samjhaya tha

Performance metrices

① Data Preparation → feature engineering (features nikaltay hain) → data modeling → Performance measure.

② Confusion matrix (used in classification) (like ek banday ko cancer machine Yes. dusray ko cancer mi, No, lakin agr apa ulte ho jae. Yes ko No aur No ko yes. Ab ais table ko confusion matrix)

Truly Positive.

Predicted (rows) \ Actual (columns)	positive	negetive	
Positive	TP 2	FP 3	⇒ 5
negetive	FN 1	TN 3	⇒ 4
	⇓ 3	⇓ 6	

Falsy negitive (Agr false likha hwa hy tou agla word galat ho ga) (e.g FN: galat N hy)

Big cases minimize (Cancer)

Small cases minimize (email spam)

(We have to overcome FP and FN according to the situations)

Accuracy

TP	FP
FN	TN

(Accuracy achi nahi hy, data imbalancing)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision:

$$\frac{TP}{TP + FP}$$