1.1 Introduction of ML models

Machine Learning:

ML is a subset of AI

Enables computers to learn from data

Make predictions or decisions without being explicitly programmed.

What is a Machine Learning Model?

- ML model is a mathematical representation.
- Trained on data
- To identify patterns and relationships
- Allowing to make predictions on unseen data.
- Training: The model learns from existing data.
- Testing: The model's performance is evaluated on new (unseen) data.
- Prediction: The trained model makes predictions on future or unknown inputs.

Types of Machine Learning:

1. Supervised Learning - labeled data

2. Unsupervised Learning - unlabeled data

3. Reinforcement Learning - interacting with an environment (reward/punishment)

4. Semi-supervised learning - both supervised and unsupervised

1.2 Training a model for Supervised learning

Uses input features (X) and target output (y)

To train a model to predict outcomes for new data.

1. Collect Data
2. Split Data
3. Choose a Model
4. Train the Model
5. Evaluate the Model
6. Tune Hyperparameters

1.3 Features: Understanding Data, Feature Extraction & Engineering

Feature: is individual measurable property or characteristic of dataset.

Example: In a housing dataset → area, bedrooms, location, price.

Understanding Data: Before feature engg. understand the data through:

Data types (numerical, categorical, text, etc.)

Missing values

Outliers

Correlations and distributions

Feature Extraction

Converting raw data (like text, images, or audio) into usable numerical features.

Text → TF-IDF vectors or word embeddings

Image → Pixel values, color histograms

Audio → MFCCs (Mel-frequency cepstral coefficients)

Feature Engineering

Creating new input features from existing data to improve model performance.

Example: total_rooms / households in a housing dataset gives average rooms per household.

1.4 Feature Engineering on Different Data Types

A. Numerical Data

Handling Missing Values – mean/median imputation

Binning – group continuous values into bins

Polynomial Features – add power or interaction terms

Log Transformation – reduce skewness in data

Example: log(income) helps normalize skewed income data.

B. Categorical Data

1. Label Encoding – assign integer values to categories

Example: {Low=0, Medium=1, High=2}

2. One-Hot Encoding – create binary columns for each category

Example: City → [Delhi, Mumbai, Kolkata]

3. Target Encoding – replace category with mean of target variable

C. Text Data

Convert text into numerical features that capture meaning.

Techniques:

1. Bag of Words (BoW): counts word occurrences.

2. TF-IDF (Term Frequency–Inverse Document Frequency): gives weight to important words.

3. Word Embeddings: convert words into dense vectors (Word2Vec, GloVe, BERT).

4. Feature Cleaning:

- Remove stopwords, punctuation

- Lowercasing

- Lemmatization or stemming

1.5 Feature Scaling & Feature Selection

Feature Scaling

Ensures that features are on the same scale

Standardization (Z-score) : Mean = 0, Std = 1

Min-Max Scaling: Scales to [0, 1]

Robust Scaling : Uses median and IQR (handles outliers better)

Feature Selection

Reducing the number of input variables to avoid overfitting and improve performance.

Methods:

1. **Filter Methods: -** Correlation, Mutual Information

2. **Wrapper Methods: -** Recursive Feature Elimination (RFE)

3. **Embedded Methods:-**

Feature importance from tree-based models (Random Forest, XGBoost)