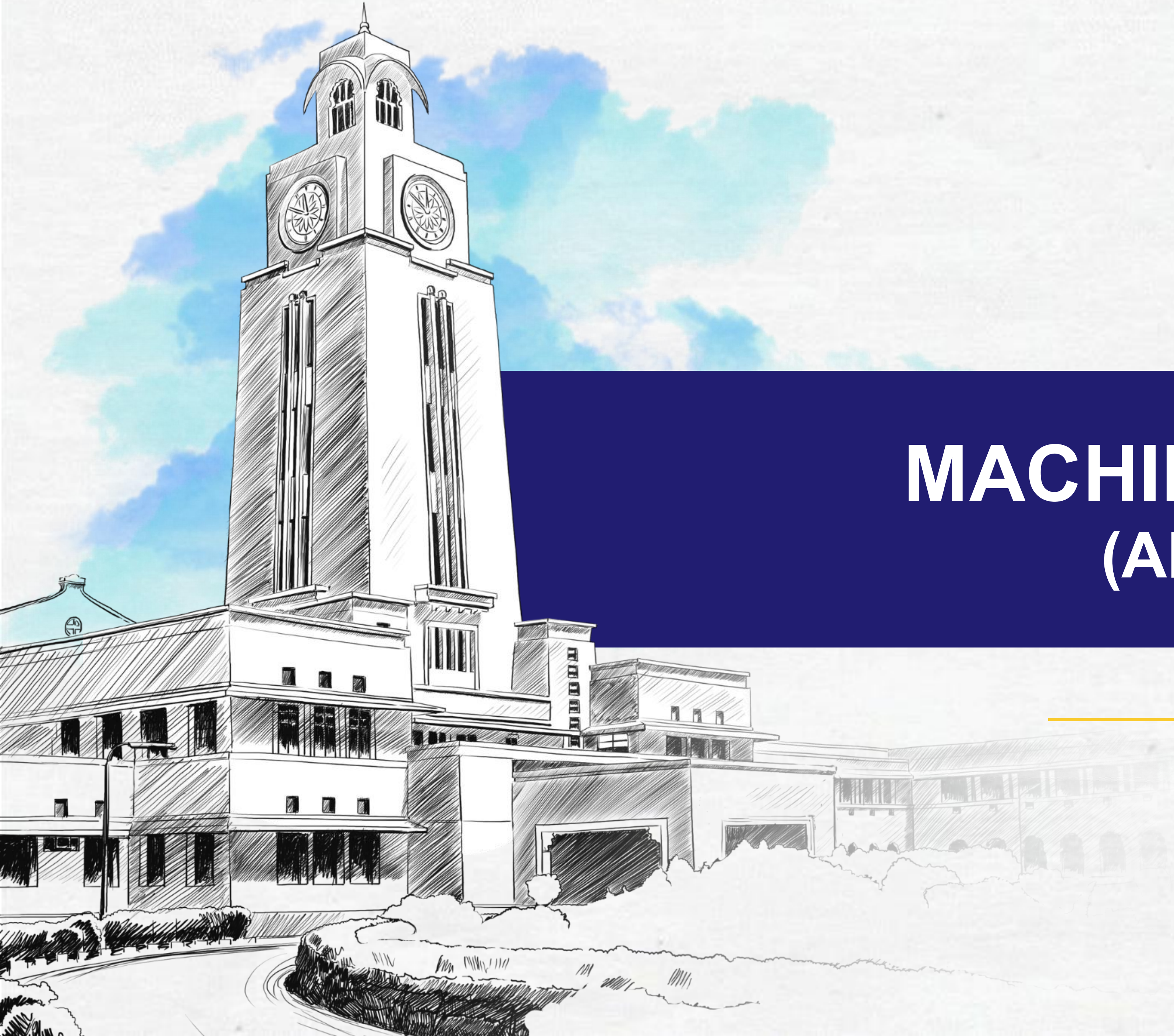

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

MACHINE LEARNING

(AIML ZG565)

Webinar 1
(14th September,2025)

ML Pipeline

A **Machine Learning (ML) pipeline** is a step-by-step process that helps us build a model to make predictions from data. Instead of doing everything manually, we follow an organized workflow to handle data, train a model, test it, and deploy it so it can make decisions automatically.

Data collection → Preprocessing → Feature Engineering → Train-Test Split → Model Training → Evaluation → Deployment → Monitoring

Data Collection: Purpose & Strategy

🎯 Purpose of Data Collection

Collect relevant, high-quality data to solve business problems.
Ensure data reflects real-world scenarios the model will face.

📋 Key Strategies for Data Collection

- **Identify Data Sources**
  - Internal databases (e.g., CRM, ERP, transaction logs).
  - External sources (public datasets, APIs, web scraping).
  - IoT sensors, logs, surveys, third-party data providers.
- **Define Data Requirements**
  - Type of data: Structured (tables), Unstructured (images, text), Time-series, etc.
  - Volume: Enough data to train and validate the model.
  - Frequency: One-time vs. continuous data collection (e.g., streaming data).
  - Label availability: Supervised learning requires labelled data.
- **Data Quality Checks**
  - Check for missing values, inconsistencies, duplicates.
  - Ensure data accuracy, completeness, and reliability.

Data Collection: Purpose & Strategy

- **Data Privacy & Compliance**
  - Ensure compliance with regulations (e.g., GDPR).
  - Anonymize sensitive data when necessary.
  - Obtain consent if required.
- **Automate Data Collection**
  - Build ETL (Extract, Transform, Load) pipelines for continuous data ingestion.
  - Use APIs or automated scripts to collect data regularly.
- **Document Data Collection Process**
  - Log data sources, extraction methods, timestamps, schema.
  - Maintain metadata for traceability.

Exploratory Data Analysis (EDA) Techniques

What is EDA?
Analysing and visualizing data to understand patterns, detect anomalies, and summarize key insights before modelling.

1. Summary Statistics
•**Mean, Median, Mode** → Measures of central tendency.
Example: Mean age = (25 + 30 + 35) ÷ 3 = 30.
•**Standard Deviation (SD)** → Spread of data around the mean.
•**Skewness** → Asymmetry of data distribution.
- Positive skew → Right tail longer (e.g., income data).
- Negative skew → Left tail longer.

2. Data Visualization
•**Histogram** → Visualizes frequency distribution of variables.
•**Box Plot** → Displays Min, Q1, Median, Q3, Max, and outliers.
•**Bar Chart** → Shows categorical data distribution (e.g., customers by country).

3. Normality Check
•Normal distribution → Bell-shaped curve.

Exploratory Data Analysis (EDA) Techniques

4. Relationship Analysis

Scatter Plot → Visualize correlation between two variables.

Example: Age vs. Monthly Spending.

Correlation Matrix → Pearson correlation between features.

5. Outlier Detection (Detailed Explanation)

What is an Outlier?

•An outlier is a data point that differs significantly from other observations in the dataset.

•Can occur due to measurement errors, data entry mistakes, or genuine rare events.

IQR (Interquartile Range) Method

•IQR = Q3 (75th percentile) – Q1 (25th percentile).

•Outliers are data points that lie outside:

- Lower bound → Q1 – 1.5 × IQR
- Upper bound → Q3 + 1.5 × IQR

Exploratory Data Analysis (EDA) Techniques

Example:

- Dataset: [10, 12, 14, 15, 16, 18, 100]
  - Q1 = 12, Q3 = 16 → IQR = 16 – 12 = 4
  - Lower bound = 12 – (1.5 × 4) = 6
  - Upper bound = 16 + (1.5 × 4) = 22
  - 100 is above the upper bound → It is an outlier.

Z-score Method

- Z = (X – Mean) ÷ Standard Deviation (SD).
- If |Z| > 3 → Data point is considered an outlier.

Example:

- Mean = 50, SD = 5
- Data point X = 70 → Z = (70 – 50) ÷ 5 = 4 → Outlier.

Why Detect Outliers?

- Outliers can skew the model training process.
- Removing or handling them improves model accuracy and stability.

Data Pre-processing Techniques

1. Data Cleaning

•**Definition**: Removing errors, inconsistencies, and duplicates from raw data.

•**Why Important**:

Raw data often contains errors like duplicate records, wrong values, or inconsistent formatting which may harm the model.

2. Handling Missing Data

•**Why Does It Happen?**

Data not collected, system errors, or privacy restrictions.

•**Common Methods**:

- **Drop Rows/Columns**:
  - Drop a row if one or two values are missing.
  - Drop a column if >50% values are missing.
- **Mean Imputation**:
  Replace missing numeric values with the column mean.
  Example → Age column: [25, –, 30, 35] → Mean = (25 + 30 + 35)/3 = 30 → Replace missing value with 30.
- **Median Imputation**:
  More robust for skewed data.
  Example → Salaries: [20k, 22k, –, 100k] → Median = 22k → Replace missing value with 22k.

Data Pre-processing Techniques

- **Mode Imputation**:
  For categorical features.
  Example → Gender: [Male, –, Female, Male] → Mode = Male → Replace missing with 'Male'.
- **KNN Imputation**:
  Missing value filled based on nearest neighbors' values.
- **Regression Imputation**:
  Predict missing value using regression models trained on other features.

3. Data Transformation

- **Why**: Convert data into suitable formats for modeling.
- **Examples**:
  - Dates → Separate Year, Month, Day.
    Example → "2025-09-14" → Year = 2025, Month = 9, Day = 14.
  - Text to Numbers → Convert text labels to numbers using encoding.

Data Pre-processing Techniques

4. Normalization

•**Purpose**:

Scale numeric features to a fixed range (typically [0, 1]) to prevent features with large scales dominating the model.

•**Formula**:

X_norm = (X – X_min) ÷ (X_max – X_min)

•**Example**:

Age ranges from 18 to 65 →

X_norm for Age = (X – 18) ÷ (65 – 18).

5. Standardization

•**Purpose**:

Transform data to have mean = 0 and SD = 1 → useful when features follow Gaussian distribution.

•**Formula**:

Z = (X – Mean) ÷ SD

•**Example**:

Salary data → After standardization, transformed to normal distribution.

Data Pre-processing Techniques

6. Encoding Categorical Data

•**Why Needed**:
ML models require numeric input.

•**Techniques**:

- **One-Hot Encoding**:
  Example → Gender column:
  Male → [1, 0], Female → [0, 1].
- **Label Encoding**:
  Male → 0, Female → 1.

7. Feature Scaling

•Ensures features have the same scale so no one feature dominates others.

•Example:

- Height in cm and Weight in kg are scaled to [0, 1] or standardized.

Why Data Pre-processing Is Critical

Removes noise and irrelevant variations.
Provides clean, consistent input to ML models.
Ensures better model accuracy, stability, and faster convergence.

Feature Engineering

What is Feature Engineering?

The process of creating new meaningful features or transforming existing ones to improve machine learning model performance.

Why Feature Engineering Matters

- Helps the model understand the data better.
- Improves predictive power by creating relevant features.
- Converts raw data into actionable input.

Feature Engineering

Technique	Description & Example
1. Feature Creation	Create new features from existing data. Example → From Date → Create features: Year, Month, Weekday, Is Holiday (Boolean).
2. Feature Transformation	Apply mathematical transformations to reduce skewness or make features more meaningful. Example → Apply Log(Salary) to reduce skewness.
3. Binning (Discretization)	Convert continuous variables into categories. Example → Age → Age Group: [0–18], [19–35], [36–60], [60+].
4. Polynomial Features	Create combinations of features to capture non-linear relationships. Example → If features are Age and Years_of_Experience → Create new feature: Age × Years_of_Experience. This helps model patterns like "older employees with more experience are more productive."
5. Encoding Categorical Features	Convert categorical data into numeric form. Example → Country → One-Hot Encoding: India → [1,0,0], USA → [0,1,0], UK → [0,0,1].
6. Feature Scaling	Scale features to a common range so no feature dominates others. Example → Heights in cm (150–200 cm) and Weights in kg (40–100 kg) → Both scaled to [0,1] range so they are comparable.
7. Feature Selection	Select most important features based on correlation, mutual information, or model-based importance. Example → Drop irrelevant columns like 'Customer ID' or low-variance features.

Next steps

Stage	What We Do	Techniques / Algorithms	Business Impact
Model Building	Apply ML algorithms based on problem type	**Regression:** Linear, Random Forest **Classification:** Logistic, Random Forest, XGBoost **Clustering:** K-Means, DBSCAN	Predict trends, classify risk/segments, uncover patterns
Evaluation & Validation	Assess model performance	**Regression:** R^2, RMSE, MAE **Classification:** Accuracy, Precision, Recall, F1-score **Clustering:** Silhouette Score	Ensure model is reliable and actionable
Hyperparameter Tuning	Optimize model for best performance	Grid Search, Random Search, Cross-Validation	Improve accuracy, enable better business decisions
Deployment / Insights	Use model outputs for business actions	Dashboards, Reports, Alerts	Data-driven decision making, proactive strategies