**End-to-End Machine Learning**

**EDA (Exploratory Data Analysis):**

EDA is the process of examining and analyzing data to gain insights, discover pattern, and understand the characteristics of the data.

It help understand the dataset and identify the issue that could affect model performance downstream.

Contain some steps like Proportion of missing value

**Understanding the data**

Summary of the features -> df.info

Show non null entries and feature type -> df.info

Class imbalance: impact the performance of the machine learning model, it cause model to always predict the majority class -> df.value\_counts( )

Missing values: df.null

Outliers: Are the data points that are different from the other points, it will cause measurement error, data entry error -> use boxplot, or IQR

**Data Preparation**

Involves identifying dara cleaning steps derived from EDA

Steps:

1. Null/ missing values

Drop missing values

For columns -> df.drop()

For rows df.dropna()

Dealing with null and empty values

Imputataion

1. Fill values with subtitles
2. fill with mean or median/ use constant or previous value

Advanced imputation:

1. K-nearest neighbour
2. SMOTE

**Feature Engineering and Selection**

Process of creating features that enhance the performance of ML models

- Modify preexisting features

- Design new features

**Normalization**

- Scale numeric feature to 0 or 1

**Standardization**

- scale data to have mean = 0, vairance =1

**Model Traning**

Occam’s Razor

**Model Options**

Logistic Regression: find decision boundary between classes

Decision Tree: find simple rules to classify data

Suppport Vector Mahcine: find plan to separate classes

Random Forest: Combine multiple decision trees

Deep learning model:

1. Neural Networks
2. Convolutional Neural Networks
3. Generative pretrained transformers
4. nearest neighbour (KNN)

XGBoost

**Logging experiments(MLFLOW)**

**Model Evaluation and Validation**

Accuracy

Confution matrix

Balance accuracy

Cross validation