Week 1 — Disease Prediction Using Patient Data

Objective

Learn basic ML workflow to predict heart disease using the UCI Cleveland dataset.

Dataset

- Source: UCI ML Repository (Cleveland subset, processed version)
- Size: 303 rows × 14 columns
- Target: `target` (0–4) → binarized to `target_bin` (0 = healthy, 1 = disease)

Preprocessing

- Missing values handled (median for numerics, mode for `ca`/`thal`)
- Features scaled to [0, 1] with MinMaxScaler
- Final feature set: 13 columns

Exploratory Data Analysis (EDA)

- Class balance: ~54% healthy, ~46% disease
- Correlation heatmap + feature histograms to understand relationships

Models

Model	Accuracy
Logistic Regression	0.8525
Random Forest	0.9016

Selected: Random Forest (higher accuracy)

Outcome

- Random Forest selected as the better model by accuracy.
- Balanced dataset → accuracy is a fair metric, but for medical tasks, precision/recall/F1 are also important.