1.1. PLAN

We will first implement the Machine Learning pipeline only (offline, no Flask or database yet):

Load dataset (Obfuscated-MalMem2022.csv).

Data cleaning (drop missing, cap outliers at 99th percentile).

Feature engineering (extract labels, select 7 features for Autoencoder, all 55 features for RF).

Scaling (StandardScaler).

Split data (70% training, 30% testing).

Then:

Train Autoencoder (only on benign samples with 7 features).

Train Random Forest 1 (binary: Benign vs Malware).

Train Random Forest 2 (multi-class: Malware subtypes).

Finally:

Save:

Trained Autoencoder model (autoencoder\_model.h5).

Random Forest classifiers (rf\_binary.pkl and rf\_multiclass.pkl).

Scalers (scaler\_full.pkl for all 55 features, scaler\_ae.pkl for 7 features).

1.2. SETUP

We will create a new folder inside your existing project:

/ml\_models/

Inside it, we will have:

train\_model.py → For training all models.

predict\_pipeline.py → Later for hybrid prediction function (used by Flask).

Saved models will be inside:

/ml\_models/saved\_models/

1.3. FILE STRUCTURE (after this step)

project\_root/

├── node\_modules/

├── public/

├── src/

├── ml\_models/

│ ├── train\_model.py

│ ├── predict\_pipeline.py (later)

│ └── saved\_models/

│ ├── autoencoder\_model.h5

│ ├── rf\_binary.pkl

│ ├── rf\_multiclass.pkl

│ ├── scaler\_full.pkl

│ └── scaler\_ae.pkl

├── package.json

├── README

├── vite.config.js

├── ...

| **Step** | **What Happened** |
| --- | --- |

|  |  |
| --- | --- |
| Load Data | Loaded from CSV into DataFrame |

|  |  |
| --- | --- |
| Clean Data | Dropped NA values, capped outliers at 99th percentile |

|  |  |
| --- | --- |
| Label Separation | Created binary labels (y\_binary) and multi-class malware type (y\_multiclass) |

|  |  |
| --- | --- |
| Feature Selection (AE) | Selected 7 key features for autoencoder |

|  |  |
| --- | --- |
| Scaling | StandardScaler separately for full and AE features |

|  |  |
| --- | --- |
| Save Scalers | Saved both scaler\_full.pkl and scaler\_ae.pkl |

|  |  |
| --- | --- |
| Split Train/Test Sets | 70% train, 30% test |

|  |  |
| --- | --- |
| Prepare AE Data | Extracted benign-only samples for Autoencoder training |