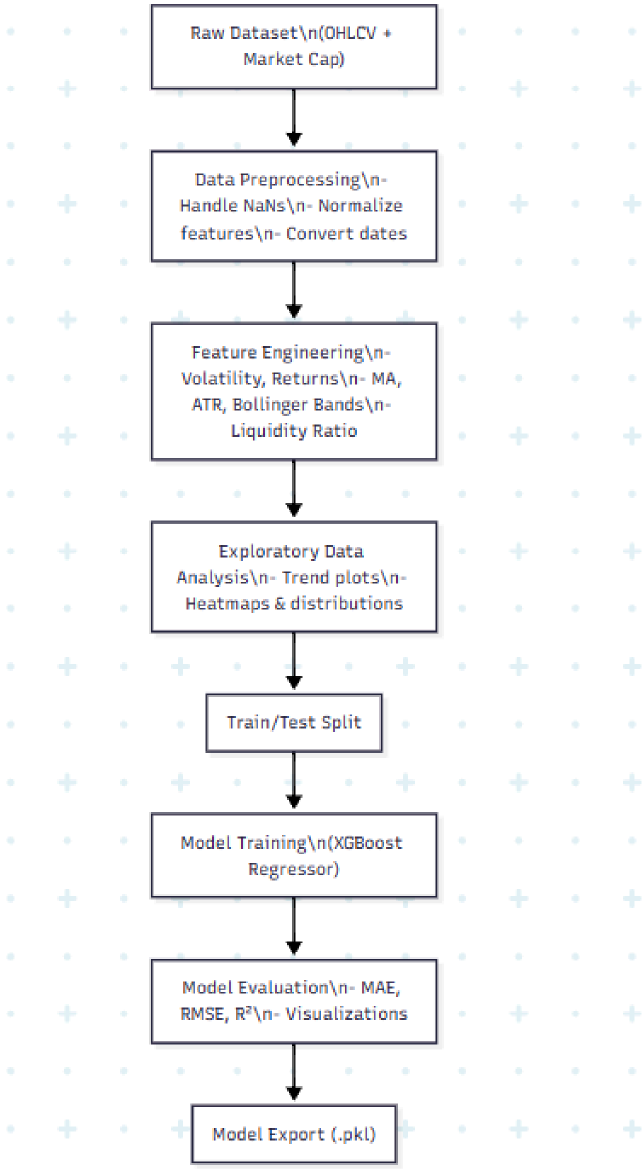
Pipeline Architecture

Objective:

This section outlines the **step-by-step flow of data and processing** used in the Cryptocurrency Volatility Prediction project, from raw dataset ingestion to prediction output using a trained machine learning model.

Pipeline Overview:

The pipeline is composed of several interdependent stages that ensure structured, clean, and feature-rich data is passed into a robust ML model for accurate volatility forecasting.


Raw Dataset\n(OHLCV + Market Cap)
Data Preprocessing\n- Handle NaNs\n- Normalize features\n- Convert dates
Feature Engineering\n- Volatility, Returns\n- MA, ATR, Bollinger Bands\n- Liquidity Ratio
Exploratory Data Analysis\n- Trend plots\n- Heatmaps & distributions
Train/Test Split
Model Training\n(XGBoost Regressor)
Model Evaluation\n- MAE, RMSE, R²\n- Visualizations
Model Export (.pkl)

Pipeline Architecture

Component Descriptions:

1. Raw Dataset

- Daily historical data for over 50 cryptocurrencies
- Includes Date, Open, High, Low, Close, Volume, and Market Cap

2. Data Preprocessing

- Removed unnecessary or irrelevant columns
- Converted date column into datetime format
- Sorted records chronologically
- Handled missing and infinite values using forward-fill strategy
- Normalized numerical values with MinMaxScaler

3. Feature Engineering

New features were derived to enhance predictive power:

Feature	**Description**
volatility	(high - low) / open
volatility_7d	Rolling 7-day average of volatility
Return	Daily percent return
MA_7, MA_14	Moving averages of closing price
Liquidity Ratio	volume / marketCap
bb_bandwidth	Bollinger Band width (price spread)
atr_14	14-day average true range (high - low)

4. Exploratory Data Analysis (EDA)

- Plotted trends for top cryptocurrencies
- Analyzed correlation between numerical features
- Visualized volatility patterns and distributions

5. Train-Test Split

- Used train_test_split from scikit-learn (80/20 split)
- Ensured stratified sampling by cryptocurrency symbol if necessary

Pipeline Architecture

6. Model Training

- Algorithm: XGBRegressor
- Trained using engineered features to predict volatility_7d
- Hyperparameters (n_estimators, max_depth, learning_rate) tuned for best performance

7. Model Evaluation

- Metrics used:
  - MAE: Mean Absolute Error
  - RMSE: Root Mean Squared Error
  - R^2: Coefficient of Determination

- Visual comparisons:
  - Actual vs Predicted Scatter Plot
  - Residuals Histogram
  - Time Series Line Plot for predictions

8. Model Export

- Saved trained model as xgboost_volatility.pkl using joblib
- Can be loaded in external applications (Flask API, Streamlit app) for deployment

Summary:

This pipeline ensures that the data is clean, feature-rich, and optimized for machine learning, enabling accurate forecasting of crypto volatility. It is modular, allowing easy improvements such as additional features, model updates, or real-time deployment.