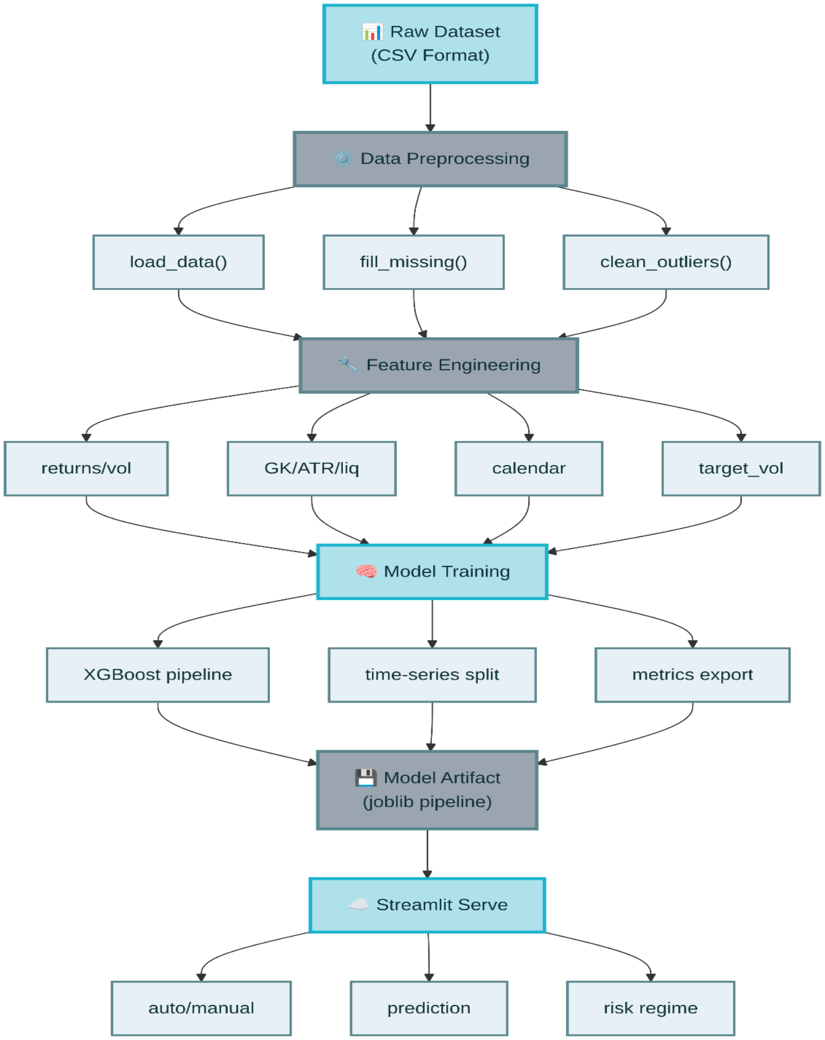
Cryptocurrency Volatility Prediction Pipeline Architecture



Overview:

This document outlines the end-to-end machine learning pipeline developed to forecast next-day cryptocurrency volatility using engineered features and an XGBoost model. The pipeline is modular, reproducible, and designed for time-series data.

Pipeline Stages:

1. Raw Dataset (CSV Format)

- Daily OHLCV and market cap data for multiple cryptocurrencies.
- Stored in standardized CSV format for ingestion.

2. Data Preprocessing

- `load_data`: Loads and normalizes schema across assets.
- `fill_missing`: Applies forward/backward fill per symbol.
- `clean_outliers`: Winsorizes extreme values to reduce noise.

3. Feature Engineering

- **Returns & Volatility:** Log returns, rolling volatility.
- **GK, ATR, Liquidity:** Garman–Klass, Average True Range, volume/cap ratios.
- **Calendar Features:** Day-of-week, month, weekend flag.
- **Target Variable:** Next-day GK volatility.

4. Model Training

- **Pipeline:** Combines preprocessing and XGBoost regressor.
- **Split:** Time-series aware train/val/test separation.
- **Metrics:** RMSE, MAE, R^2 exported for evaluation.

5. Model Artifact

- Trained pipeline saved using `joblib`.
- Stored in `models/xgb_volatility.joblib` for deployment.

6. Streamlit Serve

- **Auto Mode:** Uses latest features from dataset.
- **Manual Mode:** Accepts user-defined feature inputs.
- **Output:** Predicted GK volatility and risk regime classification.

Design Principles

- Modular and reusable code structure.
- Time-series aware modeling and evaluation.
- Interactive deployment via Streamlit for real-time forecasting.