Student Name: Yuvaraj P.G

Register Number: 620123106126

Institution: AVS Engineering College

Department: Electronics & Communication Engineering

Date of Submission: 08.05.2025

Github Repository Link: GitHub Link

---

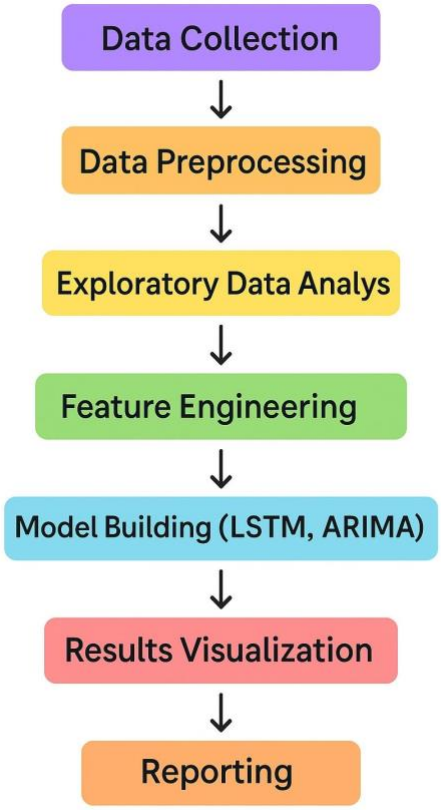
1. Problem Statement

Building on Phase-1, we refine the problem to a regression task: forecasting daily closing prices of selected stocks using historical time-series data. Accurate price prediction enables better investment Decisions and automated trading strategies.

2. Project Objectives

- [Update the project goals now that you're entering Achieve high forecasting accuracy (target RMSE < 1% of price range).
- Compare LSTM and ARIMA models for time-series forecasting.
- Validate models via rolling-window cross-validation.
- Prepare a production-ready codebase in a GitHub repository.

3. Flowchart of the Project Workflow



4. Data Description

- *Source: Yahoo Finance API (yfinance library).*
- *Type: Time-series data (structured).*
- *Records: ~5 years of daily OHLCV data (~1250 records per stock).*
- *Static dataset: snapshot at model training time.*
- *Target variable: Closing price.*

5. Data Preprocessing

- Handle missing trading days by forward-fill imputation.
- Remove duplicate entries.
- Detect outliers via IQR method and cap them.

- Convert date column to datetime type and set as index.
- No categorical encoding required.
- Scale features using MinMaxScaler.

6. Exploratory Data Analysis (EDA)

- Univariate Analysis: Histograms of returns, boxplots of volume.
- Bivariate Analysis: Correlation heatmap of technical indicators vs close price.
- Insights: Identified seasonality patterns and volatility clusters.

7. Feature Engineering

- Added rolling mean and std features (window=7, 30 days).
- Created technical indicators: RSI, MACD.
- Extracted date features: day of week, month.
- Optional PCA for dimensionality reduction on indicator set.

8. Model Building

- Models: ARIMA(p,d,q) and LSTM neural network.
- Train-test split: last 20% of data for testing.
- Rolling-window cross-validation (window=60 days).
- Evaluation metrics: RMSE, MAE, MAPE.

9. Visualization of Results & Model Insights

- Plot of actual vs predicted closing prices.
- Residual error distribution plot.
- Feature importance for ARIMA parameters and LSTM attention weights.
- Comparative bar chart of model RMSE scores.

10. Tools and Technologies Used

- Python 3.9
- Jupyter Notebook / VS Code
- Libraries: pandas, numpy, matplotlib, scikit-learn, statsmodels, TensorFlow/Keras, yfinance
- Visualization: matplotlib, seaborn

11. Team Members and Contributions

- P.G Yuvaraj - Model development (LSTM, ARIMA), GitHub setup
- M. Ramana - Data preprocessing, EDA
- M. Poornachadran - Feature engineering, reporting