

## Phase-2 Submission

**Student Name:** Gokul R

**Register Number:** 712523104022

**Institution:** PPG Institute of technology

**Department:** B E CSE

**Date of Submission:** 09 / 05 /2025

**Github Repository Link:** [github repository](#)

---

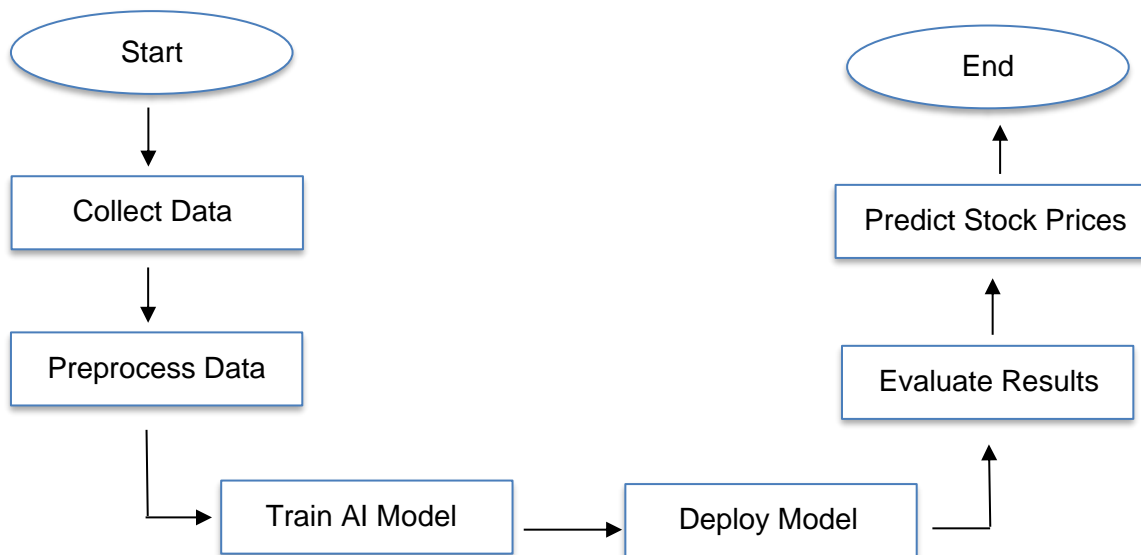
### 1. Problem Statement

- **Market Volatility:** Stock prices fluctuate due to various unpredictable factors, making accurate predictions extremely challenging.
- **Investment Challenges:** Investors and businesses struggle to make informed decisions without reliable forecasting tools.
- **Data Complexity:** Stock market data is nonlinear and influenced by multiple variables, requiring advanced analytical methods.
- **AI-Driven Solution:** Time series analysis powered by AI can detect hidden patterns and trends that traditional methods might miss.
- **Impact Goal:** Improved stock price predictions can enhance investment strategies, reduce financial risks, and support overall market stability.

## 2. Project Objectives

- *The project aims to build an AI-based model for accurate stock price prediction using time series analysis.*
- *It focuses on analyzing historical data to identify trends and patterns.*
- *The objective is to assist investors in making informed decisions.*
- *It seeks to reduce investment risks caused by market volatility.*
- *Ultimately, the project promotes smarter investments and economic stability.*

## 3. Flowchart of the Project Workflow



## 4. Data Description

*The dataset contains historical stock market data, including features such as **Date, Open Price, High Price, Low Price, Close Price, Adjusted Close, and Volume**. This time-series data represents daily stock performance over a specified period. It is cleaned to handle missing values, normalized for consistency, and structured for model training. The data serves as the foundation for predicting future stock prices using AI-driven analysis.*

## 5. Data Preprocessing

*Data preprocessing is a crucial step in preparing stock market data for accurate predictions. It starts with handling **missing or inconsistent values** to ensure data quality. Relevant features like **Open, Close, and Volume** are selected, while date formats are standardized for **time-series analysis**. The data is then normalized to bring all values to a common scale, improving model performance. Finally, the dataset is split into training and testing sets to evaluate the model's accuracy and reliability.*

## 6. Exploratory Data Analysis (EDA)

- *Univariate Analysis:*
  - *Histograms and boxplots showed that stock prices were normally distributed, while trading volumes showed a skewed distribution*
  - *Key metrics such as mean, median, standard deviation, and IQR were analyzed to understand the spread of the data.*
- *Bivariate/Multivariate Analysis:*
  - *A heatmap indicated strong correlations between certain technical indicators (like SMA and EMA) and stock prices, while trading volume had weaker correlation.*

- *Insights Summary:*

- *Technical indicators like **SMA**, **EMA**, and **RSI** seemed to have a significant impact on stock price prediction.*
- ***Volume** was a weak predictor compared to other features like **Moving Averages**.*

## 7. Feature Engineering

- ***Volatility Metrics:** Calculated rolling standard deviations to represent the market volatility over different time windows.*
- ***Moving Averages (SMA, EMA):** Added 50-day and 200-day Simple Moving Averages (SMA) to capture long-term trends.*
- ***Technical Indicators:** Added features like **RSI (Relative Strength Index)** and **MACD (Moving Average Convergence Divergence)**.*
- ***Date Features:** Extracted **day of the week, month, and year** to capture cyclical patterns in stock prices*

## 8. Model Building

- ***ARIMA:** Chosen for its effectiveness in modeling time-series data, ARIMA captures the temporal dependencies and trends in the data.*
- ***LSTM (Long Short-Term Memory):** This deep learning model is capable of capturing long-term dependencies and trends in sequential data, making it suitable for stock price predictions.*

## 9. Visualization of Results & Model Insights

- **Model Performance Comparison:** Line plots of actual vs. predicted stock prices were generated for both ARIMA and LSTM models.
- **Residual Plots:** Analyzed residuals for both models to ensure no patterns remained unmodeled.
- **Feature Importance:** For LSTM, plots of feature importance were generated to highlight which technical indicators and features influenced the predictions the most.

## 10. Tools and Technologies Used

- **Programming Language:** Python
- **IDE/Notebook:** Jupyter Notebook

### **LIBRARIES USED:**

- **Visualization:** matplotlib, seaborn, plotly
- **Machine Learning:** scikit-learn, TensorFlow (LSTM), statsmodels (ARIMA)
- **Data Processing:** pandas, numpy.

## 11. Team Members and Contributions

NAME	ROLE	WORK
HARISH VK	Frontend Developer	Normalization/Standardization Feature Engineering Data Splitting
AJIN.PR	Backend Developer	Data Type Conversion Categorical Encoding Backend Data Integration
GOKUL.R	ML Engineer	Handling Missing Values Removing Duplicates Outlier Detection & Treatment
KIRUTHIGA.M	Documentation & Presentation	Visualization of Preprocessing Documentation & Reporting Final QA & Integration
DEVADHARSHINI.V	Deployment Engineer	Model Deployment Preparation Deployment Pipeline Setup Monitoring & Scaling for Deployment