Project Title: Cracking the Market Code – AI Driven Tesla Stock Price Prediction

Phase-2 Submission

Student Name: VINISHYAMALA.P

Register Number: 712523104067

Institution: PPG INSTITUTE OF TECHNOLOGY

Department: BE.COMPUTER SCIENCE & ENGINEERING

Date of Submission: 29.04.2025

Github Repository Link:
https://github.com/Vini123vini/NM_DS_vinishyamala_stock-prediction

---

1. Problem Statement

Predicting stock prices is a challenging problem due to market volatility and external influences. The objective is to use historical Tesla stock data to predict the closing price of the stock on future trading days. This will assist traders and analysts in making data-informed investment decisions.

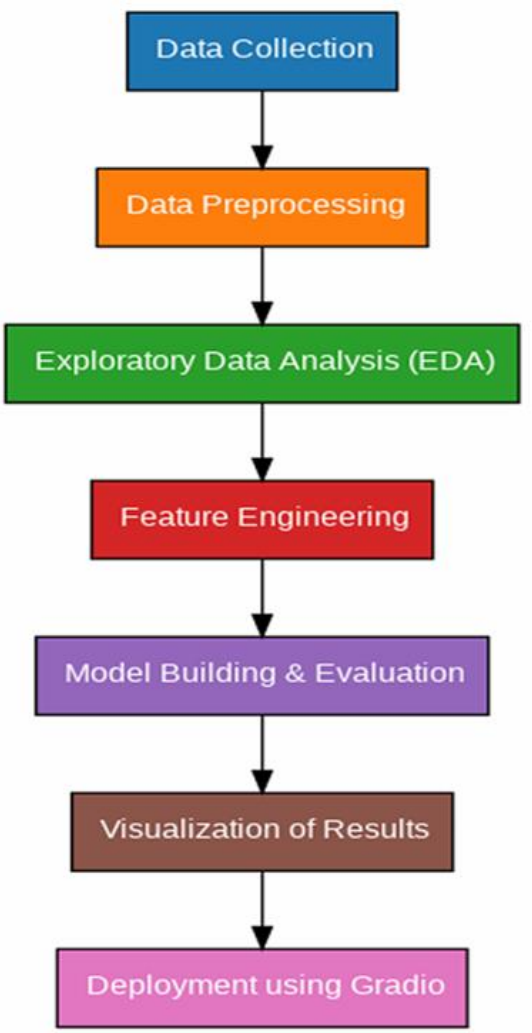
This is a **supervised regression problem**, where the target variable is the **closing price** (a continuous value).

Significance:
Accurate stock price prediction has real-world relevance in algorithmic trading, portfolio management, risk assessment, and financial planning.

2. Project Objectives

- Build a machine learning model to predict Tesla's stock closing prices.
- Identify and use historical features like Open, High, Low, and Volume to train the model.
- Analyze which variables most affect stock price.
- Ensure the model is reliable, generalizable, and interpretable.
- Compare baseline Linear Regression with Random Forest for better accuracy.
- Update the objective to include **lag features** and **technical indicators** after EDA.

3. Flowchart of the Project Workflow



4. Data Description

- **Dataset Name:** Ferrari and Tesla Share Prices (2015–2023)
- **Source:** Kaggle
- **Type:** Structured time-series data
- **Period:** October 2015 to February 2024
- **Records & Features:** Daily stock data with Date, Open, High, Low, Close, Adj Close, and Volume
- **Target Variable:** Close (Closing Stock Price)
- **Focus:** Tesla (TSLA) stock performance
- **Datasetlink:**https://www.kaggle.com/datasets/kapturovalexander/ferrari-and-tesla-share-prices-2015-2023

5. Data Preprocessing

- Converted all monetary values (e.g., $273.36) to float.
- Renamed columns for consistency.
- Converted the Date column to datetime format and sorted
- No missing or duplicate values detected.
- Created additional features:
- features (Previous_Close)
- Percentage change ((Close - Open)/Open)
- Moving Averages (5-day, 10-day)

6. Exploratory Data Analysis (EDA)

- **Univariate Analysis:**
    - Histogram and boxplots of Close Price and Volume showed volatility and some outliers.
- **Bivariate & Multivariate Analysis:**
    - Strong correlation observed between Open, High, Low, and Close.
    - Weak correlation between Volume and price.
    - Time-based line plots showed uptrends and market fluctuations.
- **Key Insights:**
    - Historical prices are strong indicators of future price.
    - Volume has less predictive power but may signal volatility.
    - Lag features and moving averages can help improve model performance

7. Feature Engineering

- Created lag variables: Previous_Close, Previous_Open, etc.
- Derived moving average features: MA_5, MA_10
- Added percentage change: (Close - Open)/Open
- Converted Date to weekday/month to capture seasonal effects.
- Removed redundant features (e.g., Close/Last renamed as Close)

8. Model Building

Algorithms Used:

- **Linear Regression** – simple and interpretable.
- **Random Forest Regressor** – handles non-linearity and feature interactions.

Train-Test Split:

- 80% training and 20% testing
- Data split in chronological order to respect time-series nature

Evaluation Metrics:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 Score

Results:

- Random Forest outperformed Linear Regression in terms of RMSE and R^2.
- Linear Regression provided a good baseline.

9. Visualization of Results & Model Insights

- **Feature Importance** plot showed that Previous_Close, High, and Low were top predictors.
- **Actual vs Predicted Price** line plot for both models.
- **Residual Plot** to analyze prediction error trends.
- **Correlation Matrix** to validate feature relationships.

10. Tools and Technologies Used

- **Programming Language** -Python
- **Environment** -Google Colab / VS Code
- **Libraries** - pandas, numpy, scikit-learn, matplotlib, seaborn
- **Model Algorithms** -Linear Regression, Random Forest Regressor
- **Visualization Tools** -matplotlib, seaborn

11. Team Members and Contributions

VISHNURAJ. N	Collected and cleaned the Tesla stock dataset, handled missing values, formatted date columns, and ensured data consistency for modeling.
VISHNU. M	Performed Exploratory Data Analysis (EDA), created visualizations (line plots, correlation heatmaps), and derived insights for model relevance.
VINISHYAMALA.P	Built and trained the Linear Regression and Random Forest models, conducted model comparison, and handled feature importance interpretation.
ROSHINI. A	Evaluated the performance of both models using MAE, RMSE, and R^2 score; created plots for residuals and feature importance to assess model behavior.
RAGAVI. K	Designed the Streamlit web app for interactive prediction, allowing users to upload Tesla stock data and view forecasted results visually.