CSP 571: DATA PROCESSING & ANALYSIS

PROJECT – PROPOSAL & OUTLINE

Team Members:

Shanika Kadidal Sundresh (A20585446)

Anusha Venkatesh (A20594323)

Neha Kiran Nayak (A20583245)

1. Project Proposal

Formal Description & Research Goal

The project focuses on **Stock Market Analysis Deploying Machine Learning Pipelines** with a particular emphasis on **NVIDIA (NVDA) stock**. The goal is to **analyze stock market trends, predict future stock prices and assess market volatility** using machine learning techniques. By leveraging historical stock data, the project aims to develop predictive models to assist in investment decision-making.

Research Questions

1. Can machine learning models accurately predict the future prices of NVIDIA stock?
2. What are the key factors influencing stock price fluctuations for NVIDIA?
3. How do different machine learning models compare in their effectiveness for stock price prediction?
4. Can stock market volatility be effectively modeled using historical data and advanced time-series forecasting methods?

Methodology

- **Data Collection**: Utilizing stock data from Kaggle (NVIDIA stock dataset).
- **Data Preprocessing**: Cleaning missing values, normalizing price data, handling outliers and feature engineering.
- **Model Selection :**
    1. Linear Regression
    2. Ridge & Lasso Regression
    3. Decision Tree Regression
    4. Random Forest Regression
    5. Support Vector Regression (SVR)
    6. Long Short - Term Memory (LSTM)

- **Model Training and Evaluation**:
  1. Splitting the dataset into training and testing sets.
  2. Using performance metrics to compare different models to determine the best performer.

Metrics for Evaluation

- **Prediction Accuracy**: Measured using RMSE and MSE.
- **Model Stability**: How well the model performs across different timeframes.
- **Trading Strategy Effectiveness**: Assessing whether predictions can be used for profitable trading strategies.

2. Project Outline

Literature Review & Related Work

The project is inspired by previous research on stock market prediction, including:

1. **Nvidia's stock returns prediction using machine learning** (Chlebus et al., 2021).
2. **Stock market prediction using ML techniques** (Pradip et al., 2018).
3. **Stock market volatility modeling with time series data** (Idrees et al., 2019).
4. **Comparative analysis of ML models for stock forecasting** (Gupta et al., 2025).
5. **Optimizing LSTM for stock prediction** (Yadav et al., 2020).

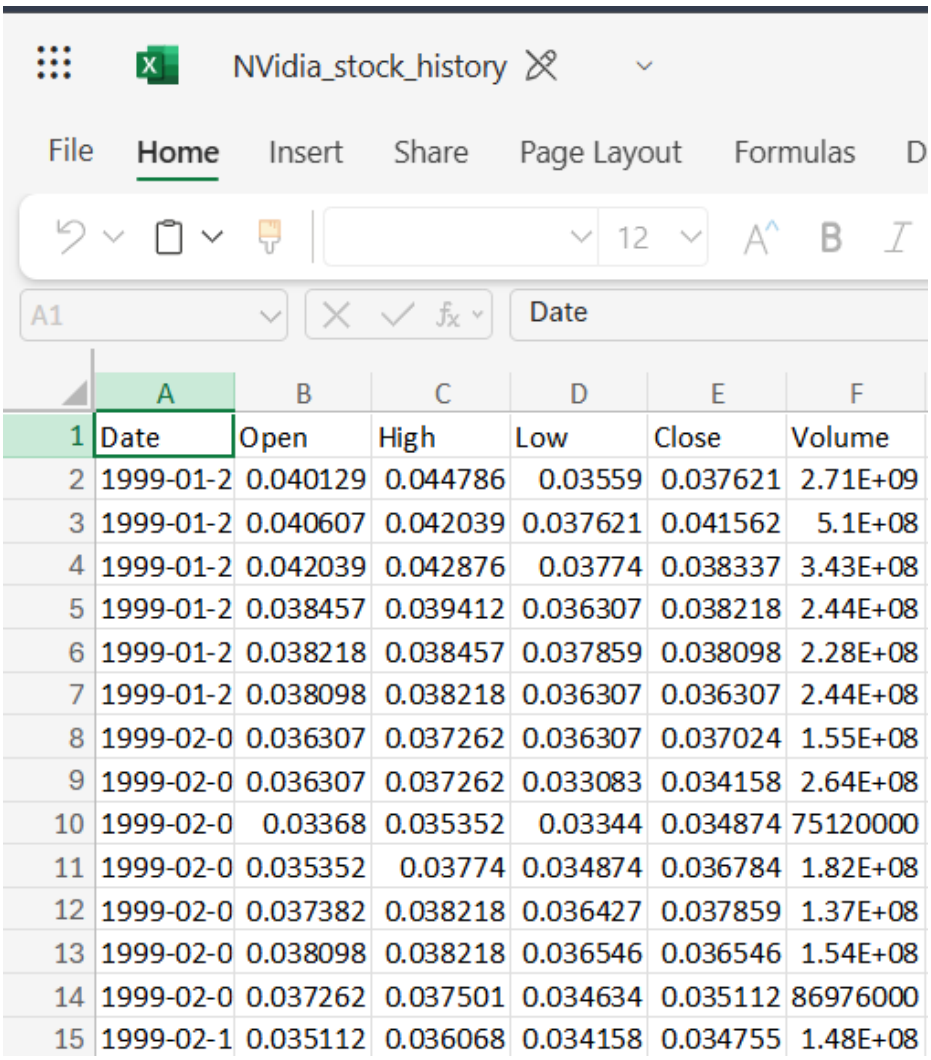
Data Sources & Description

- **Dataset Source**: Kaggle (NVIDIA stock data from 1999–2024).
- **Features**: The dataset covers NVIDIA stock prices from **1999 to 2024** and has 6442 entries.

  Columns & Description

  1. **Date** → Date of the stock entry (converted to DateTime format for time-series analysis).
  2. **Open** → Opening stock price.
  3. **High** → Highest stock price of the day.
  4. **Low** → Lowest stock price of the day.
  5. **Close** → Closing stock price (**primary target for prediction**).
  6. **Volume** → Number of shares traded (can indicate market activity and volatility).

  Target Variable : Close Price → The final stock price of the day, commonly used for forecasting future trends.



	A	B	C	D	E	F
1	Date	Open	High	Low	Close	Volume
2	1999-01-2	0.040129	0.044786	0.03559	0.037621	2.71E+09
3	1999-01-2	0.040607	0.042039	0.037621	0.041562	5.1E+08
4	1999-01-2	0.042039	0.042876	0.03774	0.038337	3.43E+08
5	1999-01-2	0.038457	0.039412	0.036307	0.038218	2.44E+08
6	1999-01-2	0.038218	0.038457	0.037859	0.038098	2.28E+08
7	1999-01-2	0.038098	0.038218	0.036307	0.036307	2.44E+08
8	1999-02-0	0.036307	0.037262	0.036307	0.037024	1.55E+08
9	1999-02-0	0.036307	0.037262	0.033083	0.034158	2.64E+08
10	1999-02-0	0.03368	0.035352	0.03344	0.034874	75120000
11	1999-02-0	0.035352	0.03774	0.034874	0.036784	1.82E+08
12	1999-02-0	0.037382	0.038218	0.036427	0.037859	1.37E+08
13	1999-02-0	0.038098	0.038218	0.036546	0.036546	1.54E+08
14	1999-02-0	0.037262	0.037501	0.034634	0.035112	86976000
15	1999-02-1	0.035112	0.036068	0.034158	0.034755	1.48E+08

- **Data Challenges**:

1. Missing data handling.
2. Data normalization for better ML model performance.
3. Possible feature engineering for additional predictive power.

Data Processing & Pipeline

- **Data Cleaning**: Handling missing values, correcting anomalies.
- **Feature Engineering**: Creating indicators like Moving Averages, Bollinger Bands,and RSI.
- **Data Transformation**: Normalization and scaling for machine learning algorithms.
- **Outlier Detection**: Removing extreme values that may skew predictions.

Data Stylized Facts

- Distribution analysis of stock prices.
- Identifying correlations between different market indicators.
- Detecting seasonality and trends in NVIDIA stock.

Model Selection

- **Feature Selection**: Identifying the most relevant indicators for predicting stock price movements.
- **Training Models**: Linear Regression, Ridge & Lasso Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression (SVR), LSTM.

Software Packages & Tools

- **Programming:** R (R-Studio)
- **ML Libraries:** scikit-learn, TensorFlow/Keras
- **Data Source:** Kaggle
- **Visualization Tools**: Matplotlib, Seaborn.