Final Report

Group 13

Andrew doss

Hussien Hussien

Phoebe fana

Sandy beshay

Abstract:

The stock market's volatility necessitates accurate prediction of stock prices to inform investment decisions effectively. With the stock exchange generating vast volumes of data daily, compounded by intraday fluctuations, analyzing these changes and forecasting future prices and market trends for specific company stocks has become increasingly challenging. In response to this critical need, our project aims to predict stock market movements, specifically whether they will increase or decrease, by employing various machine learning models. We compare the performance of these models to identify the one with the highest accuracy, offering valuable insights for investors and analysts to make well-informed decisions.

Introduction:

The stock market stands as a cornerstone of global finance, where investors seek opportunities to capitalize on asset growth and achieve their financial objectives. However, the inherent unpredictability of the market presents a formidable challenge in anticipating stock price movements accurately. As the daily process on the stock exchange generates an overwhelming amount of data, the need for sophisticated analysis and predictive models has never been greater.

Our project aims to address the significant requirement for predicting stock market movements by leveraging the potential of machine learning. Recognizing the diverse nature of stock market dynamics, we employ multiple machine learning models to gain comprehensive insights into their effectiveness. These models include but are not limited to Support Vector Machines (SVM), Random Forest, Gradient Boosting, Neural Networks, and K-Nearest Neighbors (KNN).

The primary objective of this paper is to outline our project's methodology and approach to comparing the performance of various machine learning models. We detail the data preprocessing steps, feature engineering techniques, and model training procedures to ensure the models receive meaningful input for prediction.

Moreover, we present the experimental results and evaluation metrics to assess each model's accuracy in predicting whether the stock market will increase or decrease. We compare the models based on metrics such as accuracy and precision to identify the best-performing model.

Our findings shed light on the efficacy of different machine learning models in predicting stock market trends and offer valuable insights into their strengths and limitations. The comparison allows us to determine the model with the highest accuracy, providing investors and analysts with a robust tool to make sound investment decisions.

system architecture:

1. Data Collection and Preprocessing:

- Data Sources: The system takes a real time data of Microsoft Corporation's stock performance from 1986 to 2023, The code essentially employs the Yahoo Finance API (Application Programming Interface) to fetch historical stock price data for the Microsoft (MSFT) stock.

- Data Preprocessing: we remove two columns which not important, we predict new column “tomorrow” to help us to find the target, then we find the target, finally we remove data that before 1990.

Date	Open	High	Low	Close	Volume	Tomorrow	Target
1990-01-02 00:00:00-05:00	-0.974987	-0.957942	-0.987966	-0.957942	17.786466	-0.952323	1
1990-01-03 00:00:00-05:00	-0.949526	-0.941182	-0.960763	-0.952323	18.549728	-0.923335	1
1990-01-04 00:00:00-05:00	-0.952323	-0.921976	-0.957942	-0.923335	18.649733	-0.948130	0
1990-01-05 00:00:00-05:00	-0.927425	-0.921976	-0.949526	-0.948130	18.057792	-0.932906	1
1990-01-08 00:00:00-05:00	-0.949526	-0.932906	-0.960763	-0.932906	17.892750	-0.935657	0
...	...	...	...	...	...	...	...
2023-07-11 00:00:00-04:00	5.802300	5.807722	5.789960	5.806550	17.100107	5.820676	1
2023-07-12 00:00:00-04:00	5.818895	5.833787	5.816129	5.820676	17.216551	5.836739	1
2023-07-13 00:00:00-04:00	5.827651	5.839886	5.826059	5.836739	16.839208	5.844240	1
2023-07-14 00:00:00-04:00	5.851024	5.862011	5.841542	5.844240	17.158450	5.845658	1
2023-07-17 00:00:00-04:00	5.845513	5.849296	5.835395	5.845658	16.829274	5.884686	1

8449 rows × 7 columns

2. Machine Learning Model Selection and Training:

- Model Selection: The system employs multiple machine learning models such as Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Tree, Ada boosting, Gradient Boosting, Naïve bayes, Logistic regression, and LSTM.

- Model Training: Historical data is split into training and testing sets. The selected models are trained on the training data using various algorithms and hyperparameters to optimize their performance.

3. Model Evaluation and Comparison:

- Evaluation Metrics: The system uses evaluation metrics like accuracy and precision

- Model Comparison: The performance of different models is compared to identify the one with the highest accuracy in predicting stock market movements (KNN) .

4. Voting Classifier:

- The Voting Classifier is introduced as an ensemble learning technique that combines the predictions of multiple individual machine learning models.

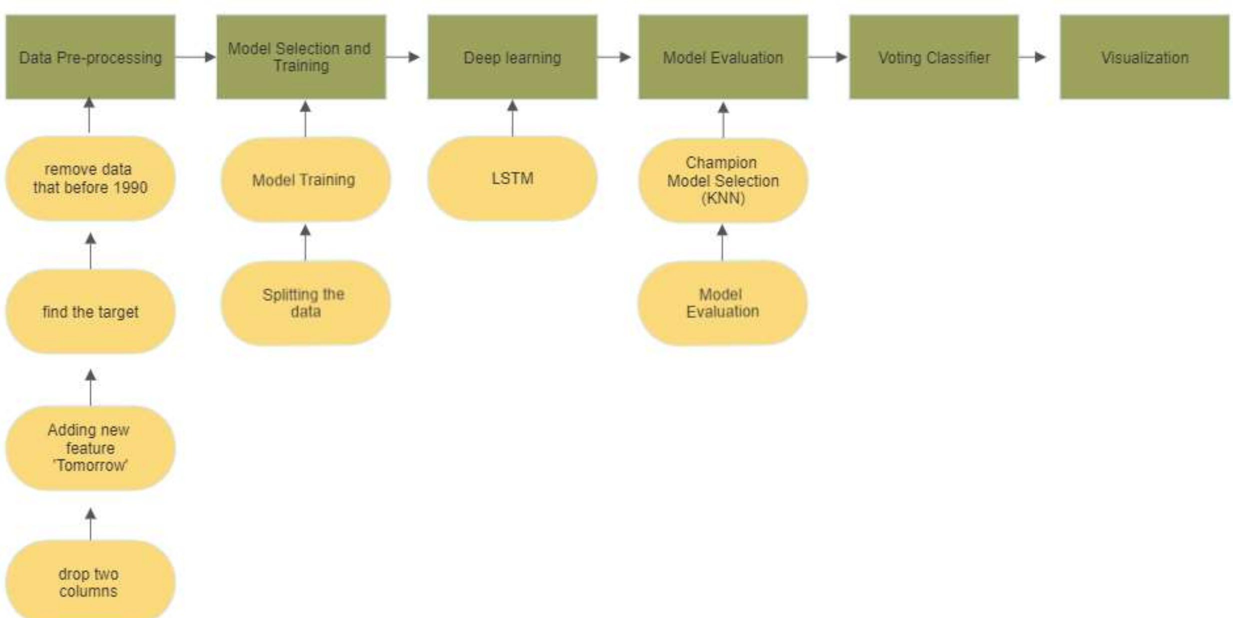
- After training different machine learning models (such as Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Tree, Ada boosting, Gradient Boosting, Naïve bayes, Logistic regression, and LSTM) in the earlier steps, the system utilizes a Voting Classifier to make a collective decision based on the predictions of these models.

- The Voting Classifier aggregates the predictions of each model and selects the final prediction based on a majority vote or weighted voting approach.

- By combining the outputs of multiple models, the Voting Classifier aims to improve prediction accuracy and reduce the impact of individual model biases or errors.

5. Visualization:

- Visualizations such as charts, graphs, and trend indicators are presented to help users understand the predicted trends and historical patterns.

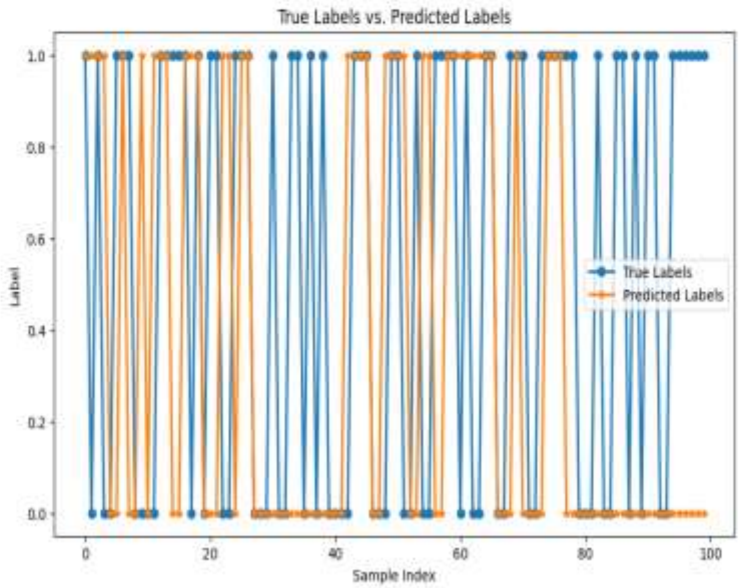


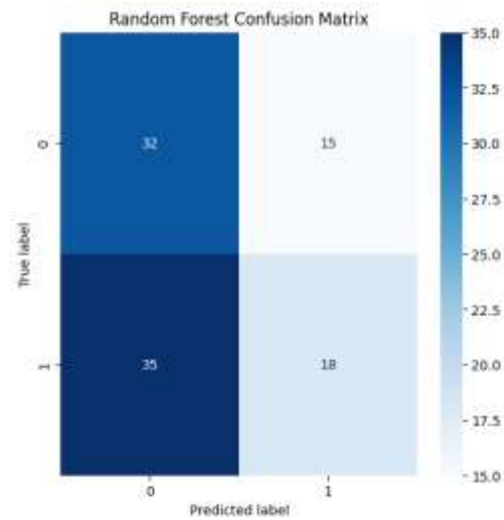
Models:

Random Forest

Accuracy = 53

Precision = 0.6053

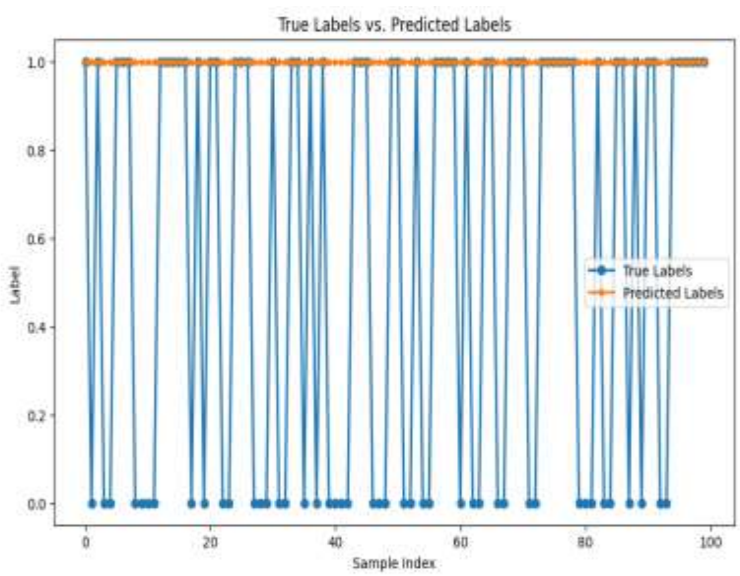


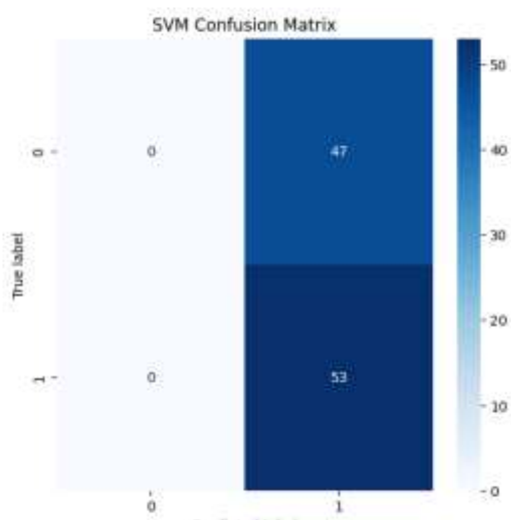


Support Vector Machines (SVM)

Accuracy = 55

Precision = 0.55

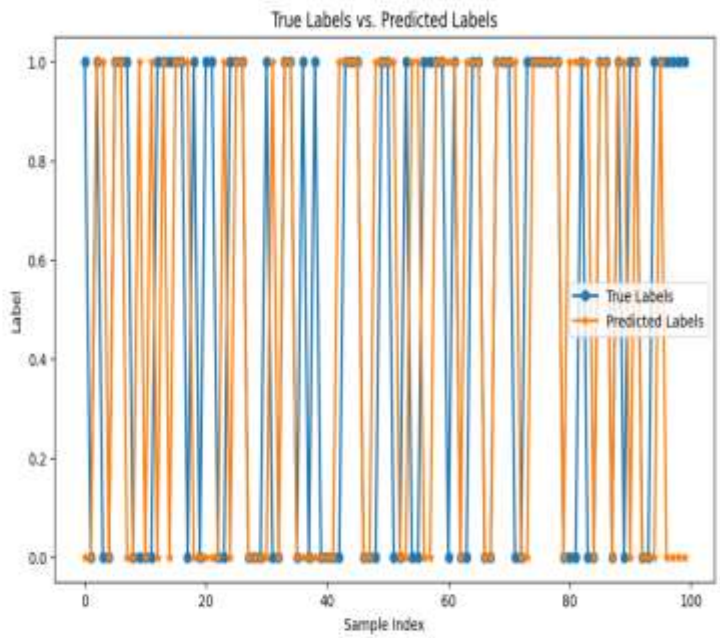


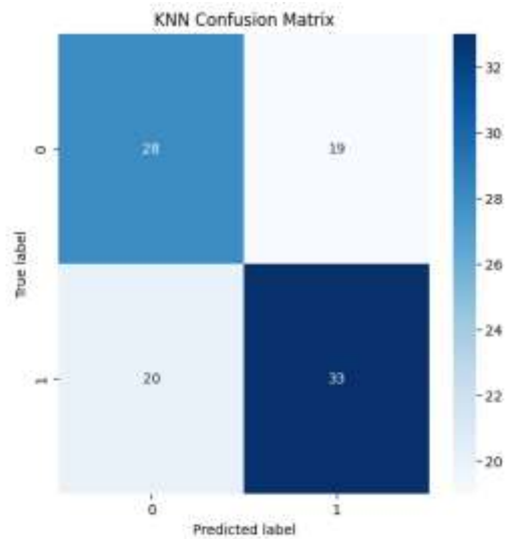


K-Nearest Neighbors (the winner)

Accuracy = 61

Precision = 0.6538

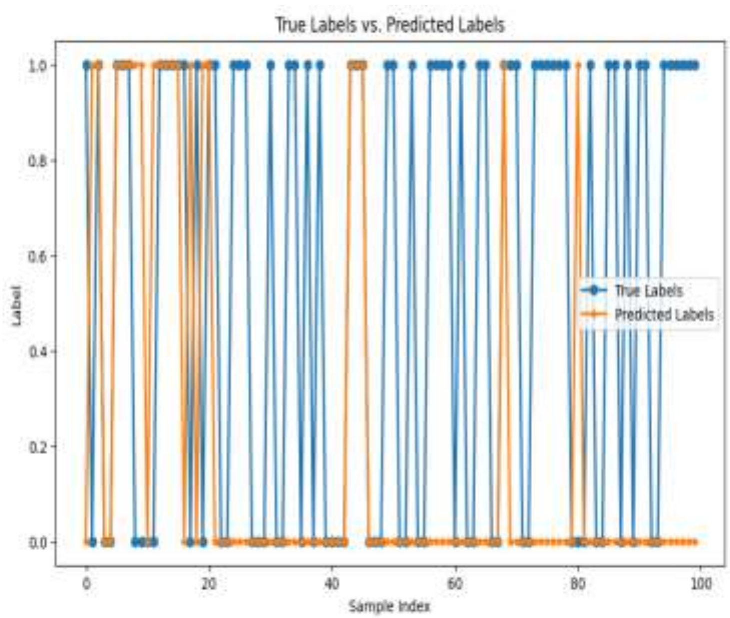


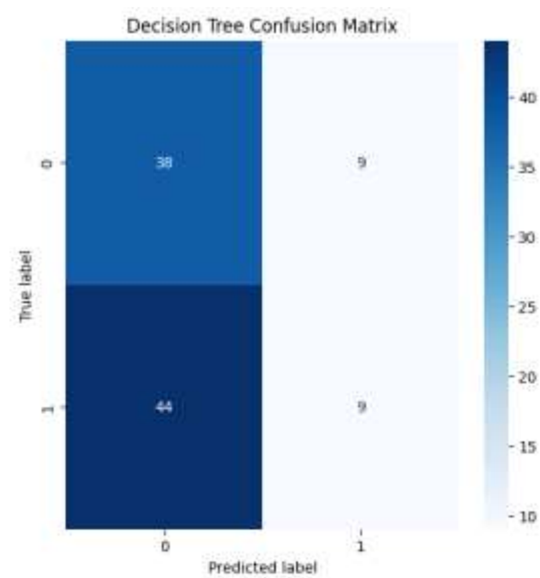


Decision Tree

Accuracy = 51

Precision = 0.65

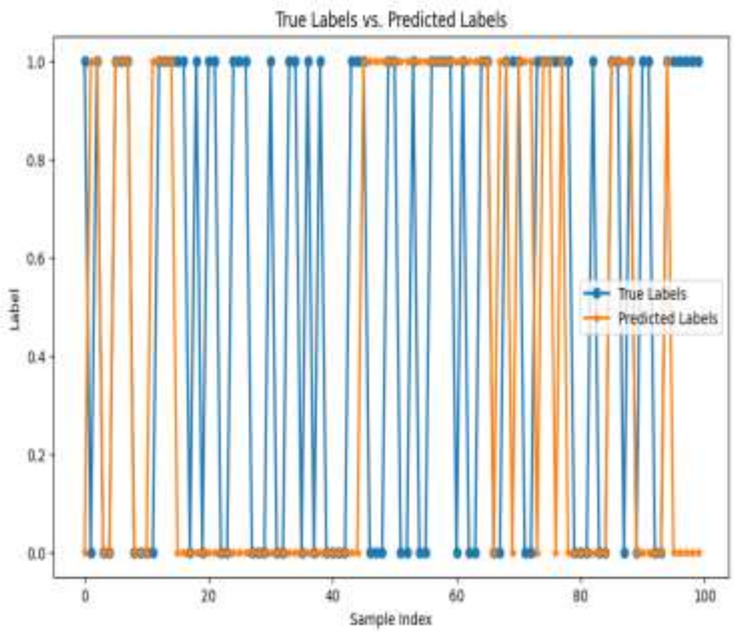


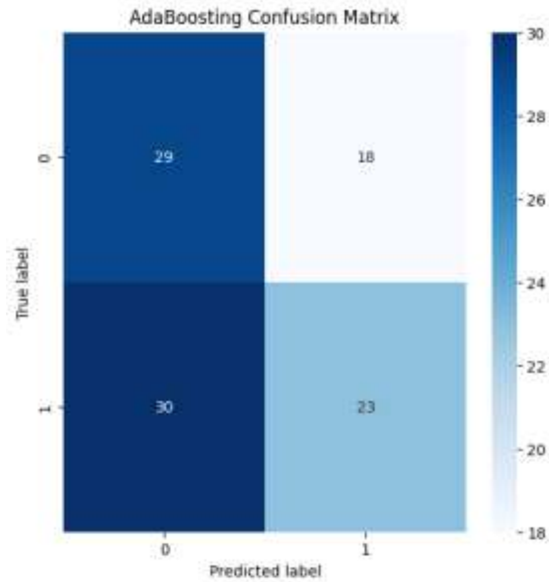


Ada boosting

Accuracy = 56

Precision = 0.6279

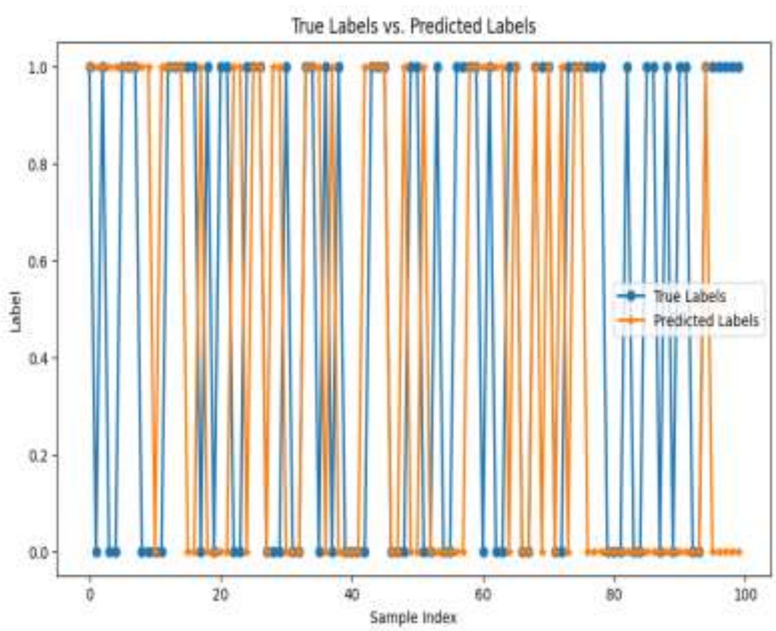


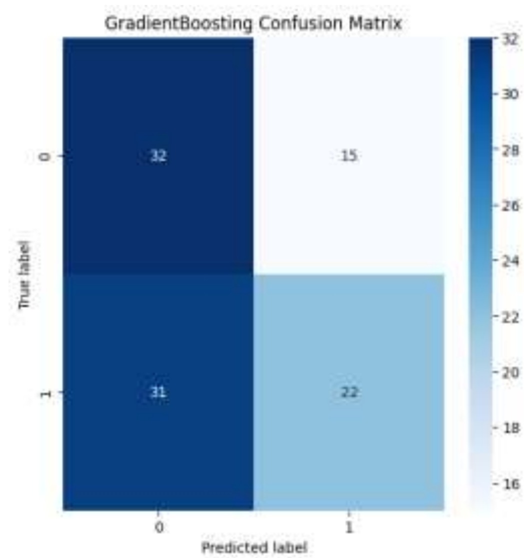


Gradient Boosting

Accuracy = 59

Precision = 0.5455

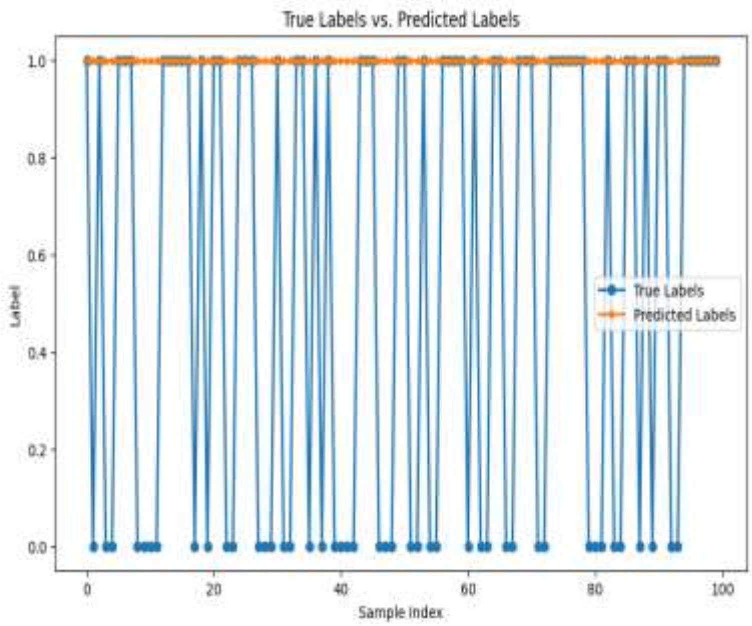


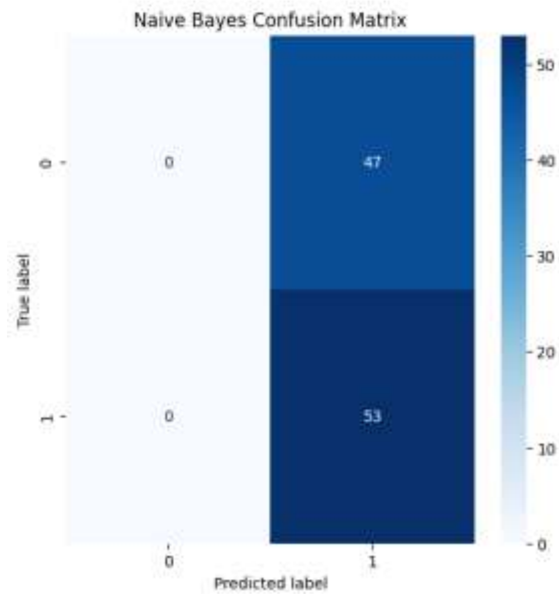


Naïve bayes

Accuracy = 55

Precision = 0.55

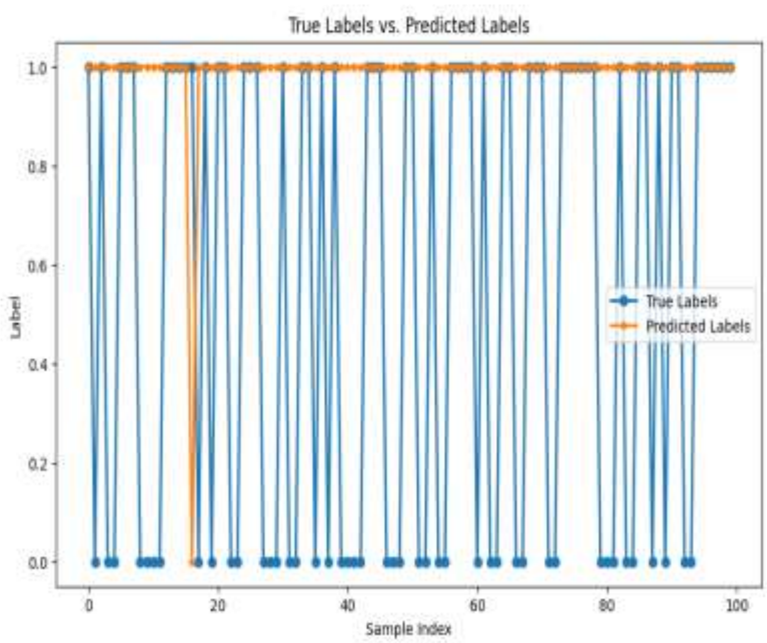


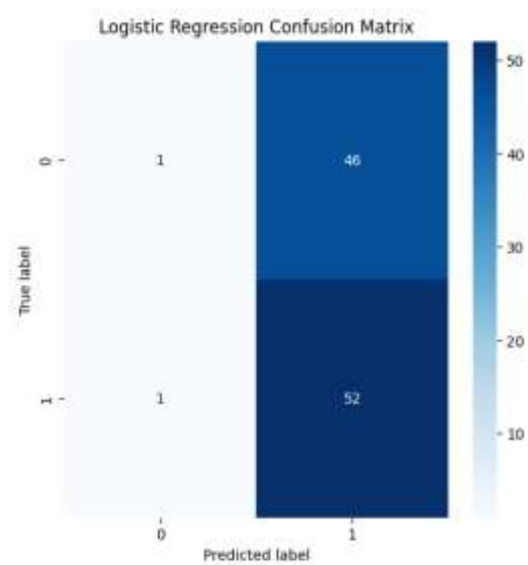


Logistic regression

Accuracy = 54

Precision = 0.5455



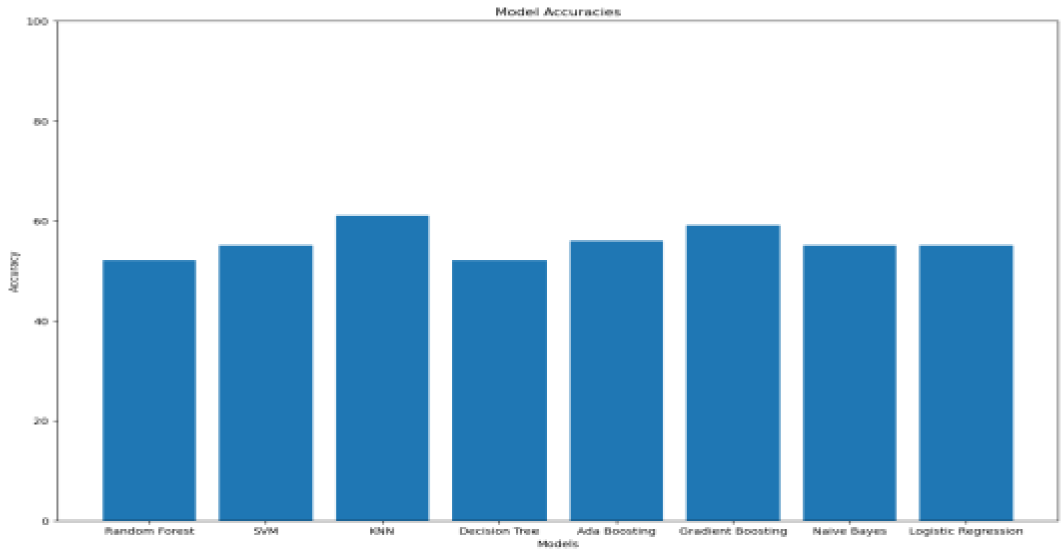


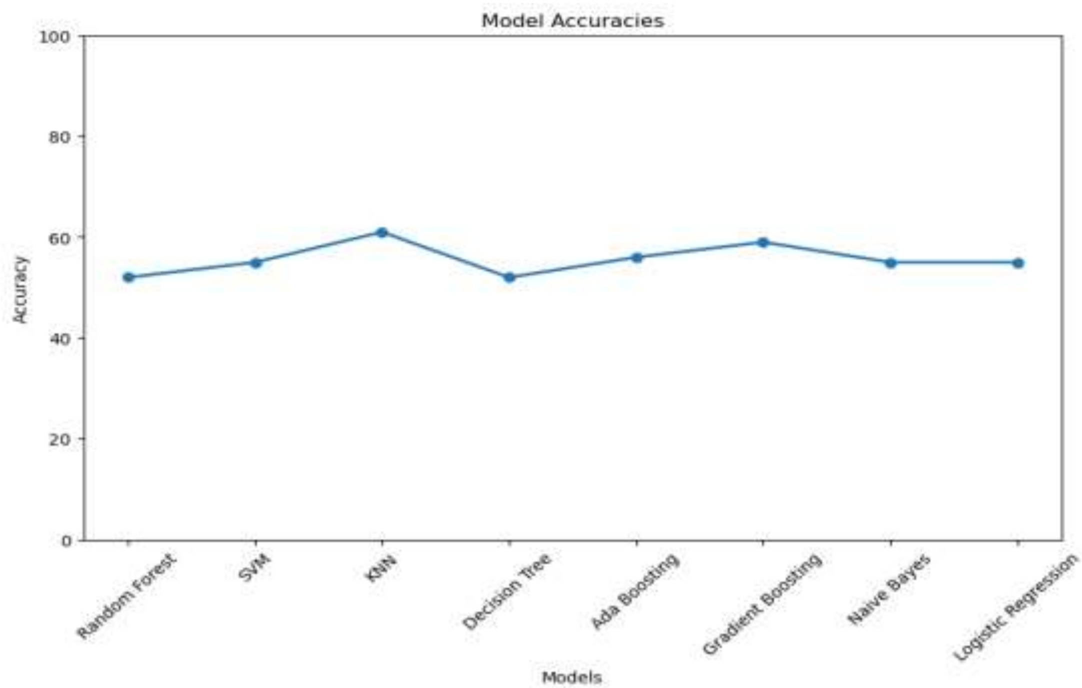
Performance evaluation:

Accuracy

The winner is KNN

Random Forest	SVM	KNN	Decision Tree	AdaBoost	Gradient Boosting	Naïve Bayes	Logistic Regression	LSTM
53	55	61	51	56	49	55	54	55

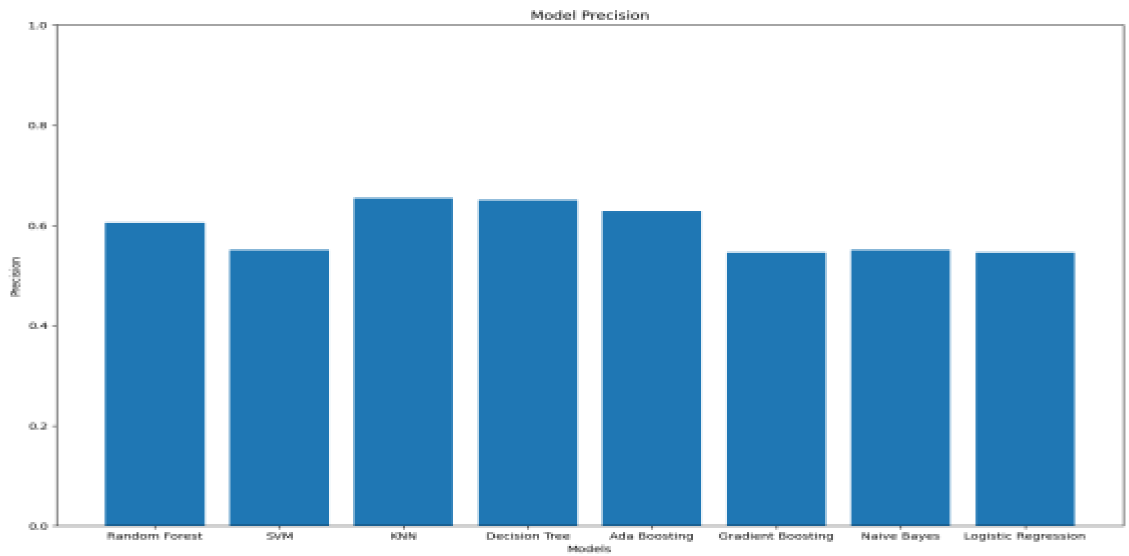


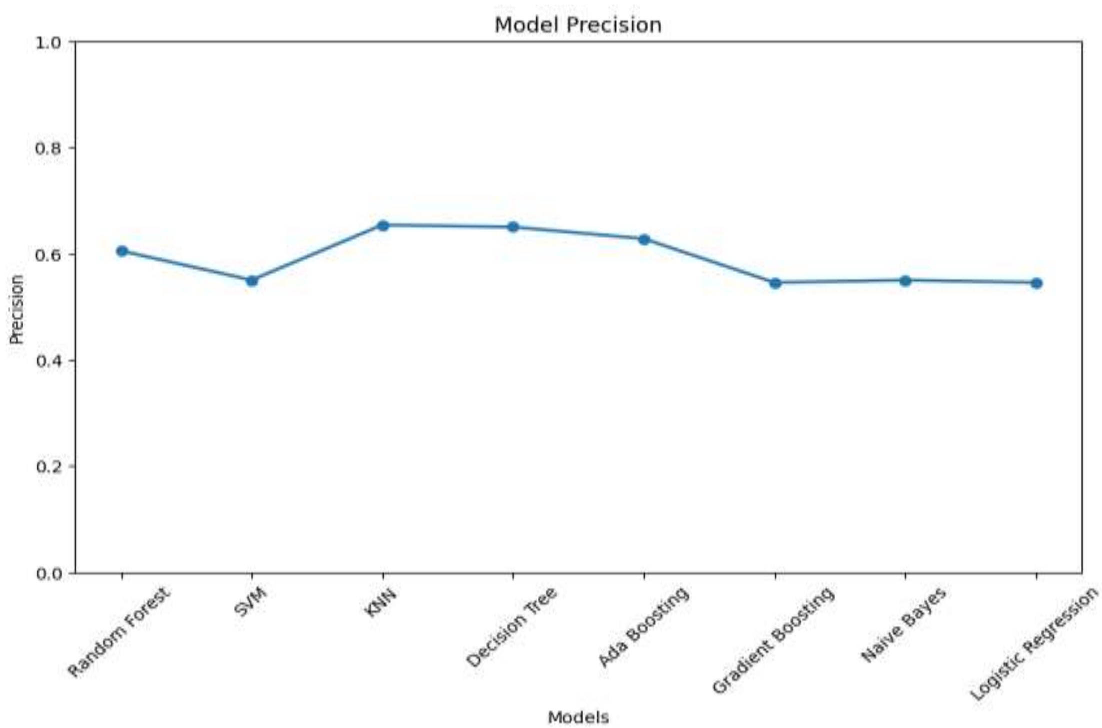


Precision

The winner is KNN

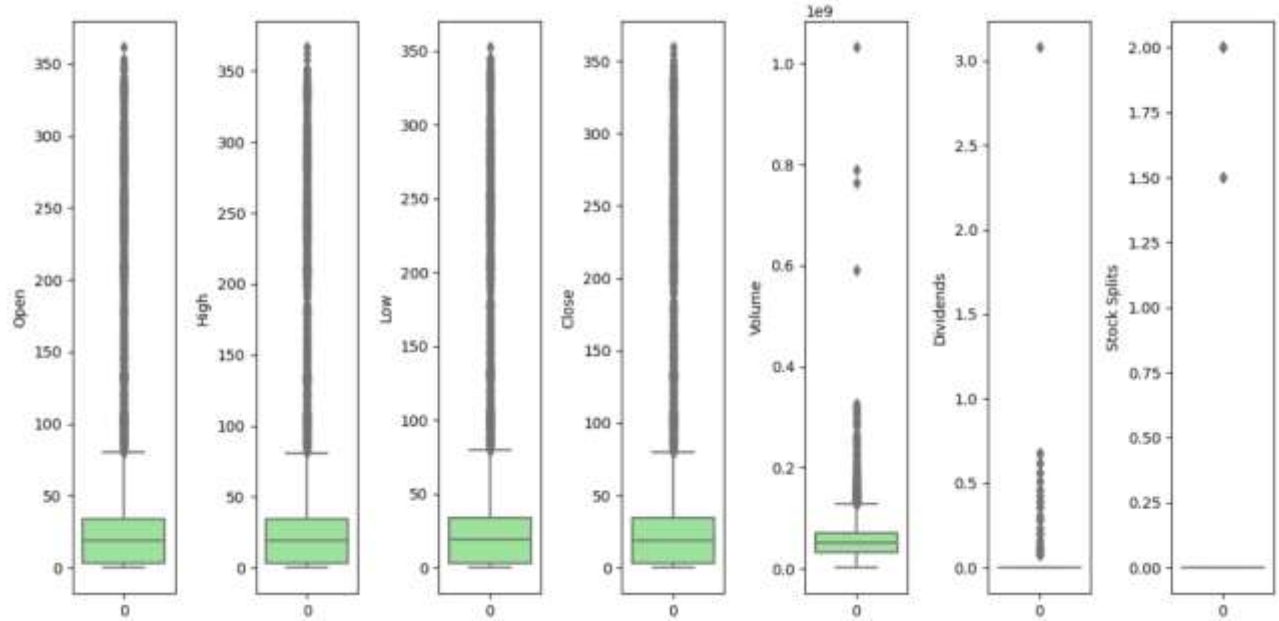
Random Forest	SVM	KNN	Decision Tree	AdaBoost	Gradient Boosting	Naïve Bayes	Logistic Regression
0.6053	0.55	0.6538	0.65	0.6279	0.5455	0.55	0.5455



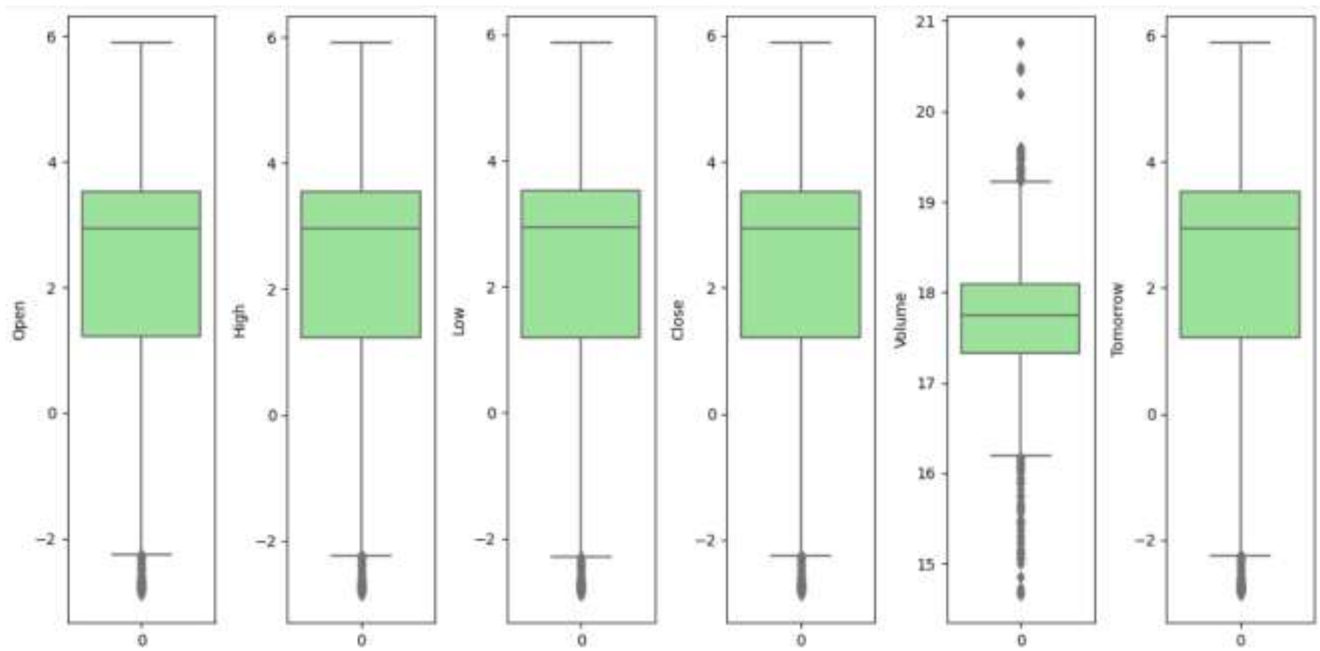


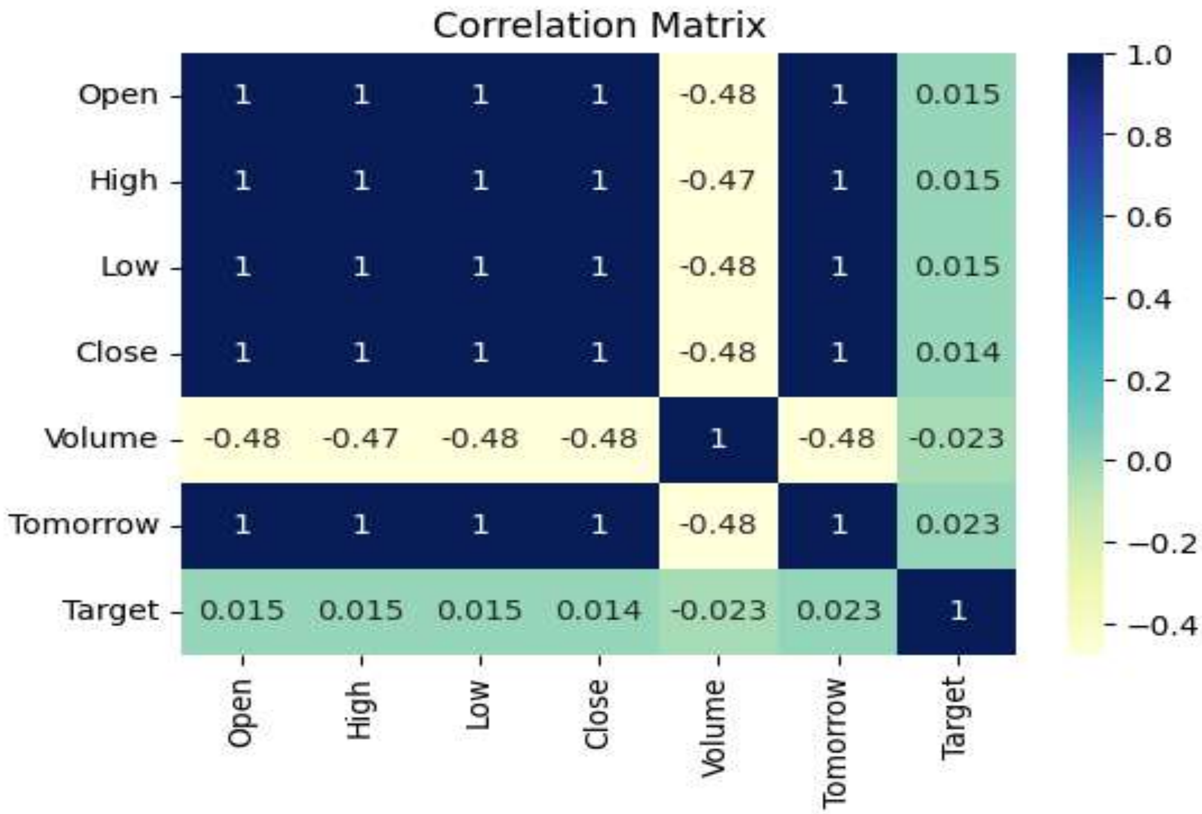
Visualisation:

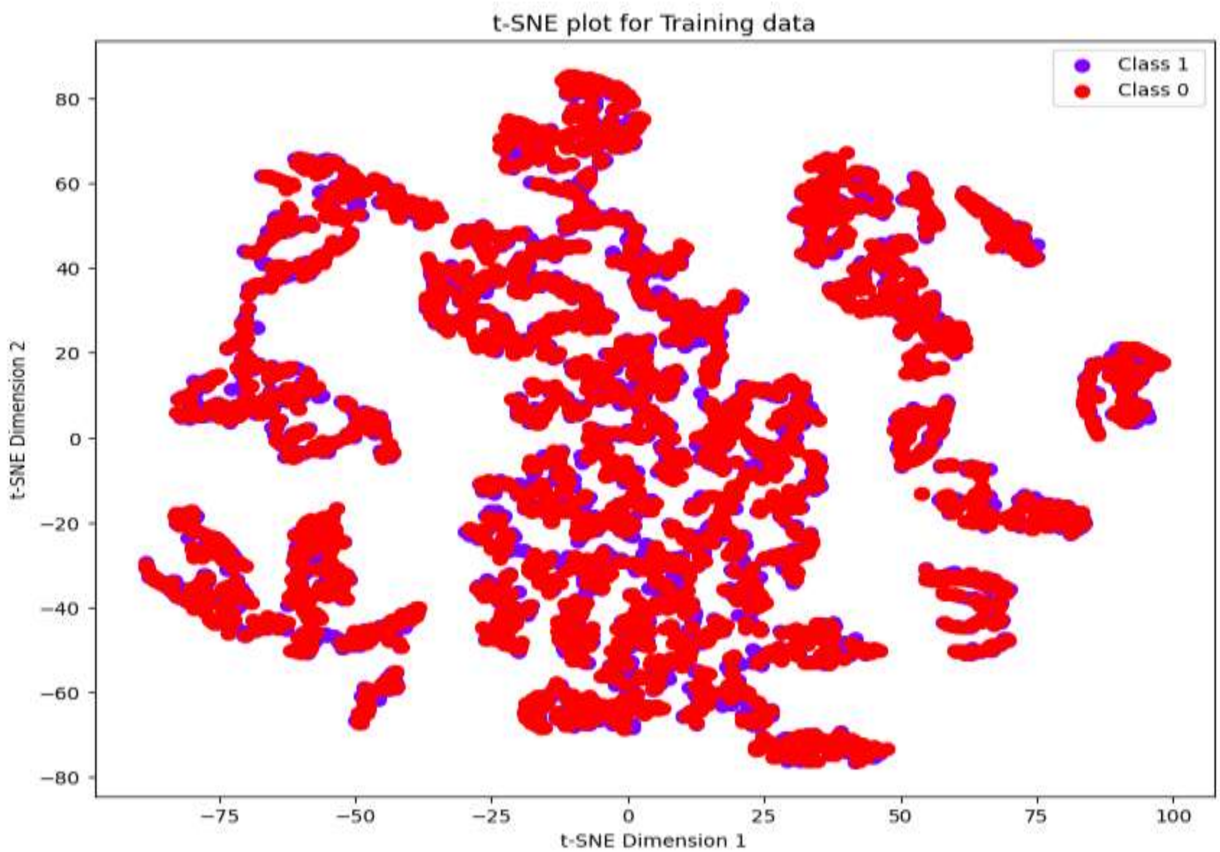
features before data preprocessing.



features after data preprocessing.




Correlation Matrix
Open
High
Low
Close
Volume
Tomorrow
Target
1 1 1 1 -0.48 1 0.015
1 1 1 1 -0.47 1 0.015
1 1 1 1 -0.48 1 0.015
1 1 1 1 -0.48 1 0.014
-0.48 -0.47 -0.48 -0.48 1 -0.48 -0.023
1 1 1 1 -0.48 1 0.023
0.015 0.015 0.015 0.014 -0.023 0.023 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


t-SNE plot for Training data
Class 1
Class 0
t-SNE Dimension 2
t-SNE Dimension 1
80
60
40
20
0
−20
−40
−60
−80
−75
−50
−25
0
25
50
75
100

Summary and Conclusion:

In the challenging task of predicting the stock market, our project utilized various machine learning models to identify the most accurate predictor. After thorough evaluation, the K-Nearest Neighbors (KNN) model emerged as the winner, achieving an accuracy of 61%. Despite the model's modest accuracy, it is essential to recognize the inherent difficulty of predicting stock market movements, as they are influenced by numerous dynamic and unpredictable factors.

Our project aimed to address the significant need for accurate stock market predictions by exploring the potential of machine learning models. The KNN model demonstrated the highest accuracy among the models tested, achieving 61%. However, it is crucial to acknowledge that predicting the stock market is an intricate task, as it is influenced by a myriad of ever-changing and unpredictable factors, including economic conditions, geopolitical events, market sentiment, and unexpected news.

While the accuracy of 61% may seem relatively modest, it reflects the complex nature of the stock market and highlights the challenges in accurately forecasting its movements. Even with sophisticated machine learning algorithms and extensive historical data, achieving high accuracy in stock market prediction remains elusive.

Therefore, as investors and financial analysts, it is essential to exercise caution and maintain a prudent approach when making investment decisions based on stock market predictions. Rather than relying solely on predictive models, incorporating diverse sources of information, fundamental analysis, and risk management strategies can enhance the decision-making process in the volatile and uncertain world of the stock market.

Our project's findings serve as a stepping stone for further research and development in the field of stock market prediction. While the KNN model demonstrated promising results, future endeavors may explore ensemble methods, deep learning architectures, or the integration of alternative data sources to potentially improve prediction accuracy.

In conclusion, predicting the stock market will always present a formidable challenge due to its complex and dynamic nature. Our project's focus on machine learning models, with the KNN model as the winner, underscores the importance of ongoing research and development in the pursuit of more accurate predictions. Ultimately, making well-informed decisions in the stock market requires a comprehensive understanding of the inherent uncertainties and a balanced blend of data-driven insights and informed judgment.

Bibliography:

- Subasi, Abdulhamit & Amir, Faria & Bagedo, Kholoud & Shams, Asmaa & SARIRETE, Akila. (2021). Stock Market Prediction Using Machine Learning. Procedia Computer Science. 194. 173-179. 10.1016/j.procs.2021.10.071.
- Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H. C. (2021, November 8). Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. MDPI.
- DAORI, HIND & ALHARTHI, MANAR & ALANAZI, ALANOUD & ALZAHRANI, GHAIDA & ABOROKBAH, MAJED & Aljehane, Nojood. (2022). Predicting Stock Prices Using the Random Forest Classifier.
- Ghazanfar, Mustansar ali & Alahmari, Saad & Aldhafiri, asmeen & Mustaqeem, Anam & Maqsood, Muazzam & Azam, Muhammad Awais. (2017). Using Machine Learning Classifiers to Predict Stock Exchange Index. International Journal of Machine Learning and Computing. 7. 24-29. 10.18178/ijmlc.2017.7.2.614.
- Ravikumar, S., & Saraf, P. (2020). Prediction of Stock Prices using Machine Learning (Regression, Classification) Algorithms. 2020 International Conference for Emerging Technology (INCET), 1-5.
- Yasmin, S. Kamalakkannan and P. Kavitha, "Stock Market Prediction using Machine Learning Models," 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 2022, pp. 1399-1405, doi: 10.1109/ICECAA55415.2022.9936188.
- Vanukuru, Kranthi. (2018). Stock Market Prediction Using Machine Learning. 10.13140/RG.2.2.12300.77448.