INTERNATIONAL CENTRE
FOR APPLIED SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)

Department of B.Sc. Computer Science and Engineering

Machine Learning Laboratory

A mini project report on

Stock Trend Prediction Using Candlestick Charting and Ensemble Machine Learning Technique

by

Aryaman Sakthivel - 211627028
Ayusha Priyadarshani - 211627005
Satvik Rohira - 211627032
Aniket Mishra - 211627004

April 2023

Certificate

This is certified that mini project entitled "**Stock Trend Prediction Using Candlestick Charting and Ensemble Machine Learning Techniques**" submitted in partial fulfilment of the IV Semester Mini Project Machine Learning Laboratory is a result of bonafide work carried out by:
Aryaman Sakthivel (211627028), Ayusha Priyadarshani (211627005), Satvik Rohira (211627032), Aniket Mishra (211627004) during the academic year 2022-2023.

Examiner 1 **Examiner 2**

Table of Contents

Stock Trend Prediction Using Candlestick Charting and Ensemble Machine Learning Techniques

Aryaman Sakthivel, Ayusha Priyadarshani, Satvik Rohira, Aniket Mishra
International Centre for Applied Sciences-Manipal

ABSTRACT

Stock market forecasting is a challenging task due to the highly noisy, complex and chaotic nature of the stock price time series. This paper constructs a novel ensemble machine learning framework for daily stock pattern prediction, combining traditional candlestick charting with the latest machine learning methods. This framework can give a suitable machine learning prediction method for each pattern based on the trained results and incorporating machine learning techniques such as Logistic Regression, K-Nearest Neighbour, Random Forest and a multitude of others. Empirical results from 2000 to 2017 of the stock markets confirm that feature engineering has effective predictive power, with a prediction accuracy of more than 60%. An investment strategy based on forecasting framework excels in both individual stock and portfolio performance theoretically, but transaction costs have a significant impact on investment. Additional technical indicators, especially momentum indicators, can improve forecasting accuracy in most cases.

INTRODUCTION

The stock market, also known as the equity market, is a complex and dynamic financial marketplace where investors buy and sell ownership in publicly-traded companies. It provides a platform for companies to raise capital by selling shares of ownership to investors, and for investors to buy and sell those shares for potential profit. If the company performs well, the value of those shares may increase, allowing the investor to sell them for a profit. Conversely, if the company performs poorly, the value of the shares may decrease, resulting in a loss for the investor.

One of the primary functions of the stock market is to provide a means for companies to raise capital. By selling shares of ownership to investors, companies can raise the money they need to invest in their business and finance new projects, which can help drive economic growth. It is an integral part of the global financial system, providing companies with the capital they need to grow and investors with the opportunity to invest in these companies and potentially earn returns on their investment. The stock market serves as a barometer for the economy, reflecting the overall health and performance of the companies and industries that make up the market. It also helps in determining the fair market value of publicly traded companies based on supply and demand, which in turn helps to allocate resources efficiently and ensure that capital is allocated in the most productive manner.

Predicting the movement of stock prices can provide valuable insights into which companies and stocks are likely to perform well in the future. It is desirable as it enables investors, traders, and financial institutions in making informed investment decisions and managing their portfolios. Accurate prediction of stock prices can help investors identify opportunities to buy low and sell high, maximize their returns, manage their risks and potentially earn higher returns on their investments. By understanding market trends and predicting future movements, businesses can adjust their operations to take advantage of opportunities or mitigate risks. Companies may also use stock market predictions to plan their business strategies. Stock market predictions can also help policymakers for the future. Economic forecasts can help governments and central banks make informed decisions about monetary policy and interest rates.

Stock market prediction is a challenging task due to the non-linear and volatile nature of the market, which makes it difficult to forecast future prices with high accuracy. It is a complex problem that involves analysing multiple variables, such as economic indicators, company financial statements, and news events. The movement of stock prices can be influenced by a wide range of economic, political, and social factors, making it difficult to predict with accuracy. However, with the advent of technology and advanced analytical techniques, stock market prediction has

become an increasingly popular topic of research in finance and computer science. Machine learning algorithms have shown promising results in predicting stock market movements by analysing vast amounts of historical data and identifying patterns and trends. These algorithms can handle the intricacy of stock market and can successfully identify correlations between variables that may not be apparent to humans. Moreover, machine learning techniques can continuously learn and adapt to changing market conditions. As new data becomes available, the algorithms can update their predictions and adjust their models to account for any changes in market trends or conditions.

Investments in the stock market are mostly guided by many different prediction methods that are a part of two groups of technical analysis and fundamental analysis. Chart analysis is one of such method that includes the ancient method of Candlestick Charting. Candlestick charts are used by traders to predict price movements based on past patterns, each candlestick shows four price points: the opening price, the closing price, the highest price, and the lowest price during that period. Candlestick charting can reflect not only the changing balance between supply and demand, but also the sentiment of the investors in the market. A study revealed that almost 66% of stock market predictions were based on technical analysis. Traditional technical analysis is mostly limited to analysing the candlestick charting, performing statistical analysis from past historical data to obtain the probability of prediction.

After examining the effectiveness of five different candlestick reversal patterns in the stock market. After statistically analysing the data it was discovered that bearish harami, and crosssignals perform adequately in predicting head reversals for stocks of low liquidity, while bullish harami, engulfing, and piercing patterns were profitable when they were applied to small companies' stocks that were highly liquid. Upon examination the predictive power of single-day candlestick charting by using daily data for the Taiwan stocks for the period from 4 January 1992 to 31 December 2009. A study provided evidence that technical analysis can be improved by using automated algorithms. AI (Artificial intelligence) has been implemented to address the chaotic time series data.

ML (Machine learning) uses historical data (past information) for parameter fitting in order to predict new data. Many machine techniques have already been implemented to predict the stock market. For example, logistic regression (LR), Deep neural networks (DNN), decision trees (DTs), support vector machines (SVM), K-nearest neighbours (KNN), random forests (RFs), have been used to predict the stock market.

Researchers proposed a forecasting model based on chaotic mapping, firefly algorithm and support vector regression to predict stock market price. The model performed best on mean squared error and mean absolute percent error, and was combined with ensemble adaptive neuro-fuzzy inference system ENANFIS. However, forecast research based on Artificial Intelligence methods and candlestick charting is still low.

In these applications of AI methods for financial market forecasting, many studies have used technical indicators as input features. We use a financial expert system that incorporated the historical stock prices, eight kinds of technical indicators, 1. Bear high, 2. Bear low, 3. Bullish harami, 4. Bear harami, 5. Bullish high, 6. Bullish low, 7. Bear horn, 8. Bullish horn. When comparing different machine learning effects, we used three machine learning methods, including DNN, SVM and RF. Results showed that RF perform the best when predicting one-day ahead stock price. In a study a statistical arbitrage strategy based on DNN, GBDT and RF to S&P 500 constituents from December 1992 to October 2015, finding that the RF outperform GBDT and DNN in their study. when contrasting ANN, SVM, RF, and naive-Bayes as four different prediction models. When the evaluation was done using 10 years' worth of historical data for two stocks and two stock price indices, experimental results showed that the RF beat the other three prediction models.

We attempt to provide a framework for automatically choosing the best machine learning approach, as the prediction of different machine learning techniques vary depending on the situation. Our research's primary goals are as follows: (1) To improve the research quality of stock market forecasting, this paper combines conventional candlestick charting with machine learning techniques. (2) Using the eight-trigram scheme, we employ a straightforward categorization. We further explained the function of technical indicators in machine learning models and compared several types of technical indicators used in short-term stock predictions. (3) Then a framework for selecting an adaptive machine learning method. We hope that the appropriate forecasting machine learning method can be automatically selected for each different candlestick charting patterns. We are planning on creating an investing plan using our prediction framework. Results are hoped to demonstrate that one can benefit economically from the technique described in this study.

LITERATURE REVIEW

In recent years, the use of machine learning algorithms has gained popularity in the field of finance for predicting stock market movements. Several studies have used machine learning algorithms to predict stock movements. When comparing a research paper with previously published projects, multiple components needed to be kept in consideration such as: Methodology, Accuracy, Performance, and datasets used. A presented research paper by YAOHU LIN, SHANCUN LIU, HAIJUN YANG and HARRIS WU has taken a unique approach. In their paper they use a two-step process for stock prediction. First by taking two-day candlestick patterns and matching those patterns with a pre-defined 8 trigram scheme by using an algorithm to grid search through feature engineering data and save the best performance model for each pattern. Secondly using a group of technical indicators and running them through a set of prediction models like: Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Random Forest (RF), Gradient Boosting Decision Tree (GBDT) and Long Short-term Memory (LSTM). In comparison our model takes a different approach where multiple types of Moving Averages such as: MA, EMA, WMA and etc are corelated with multiple candlestick patterns to see whether it is bullish or bearish. These moving averages are run through multiple machine learning algorithms. When the results of the paper by YAOHU LIN were evaluated the algorithm which ended up with the highest accuracy varied depending on the candlestick pattern identified however, Random Forest was the most consistent out of all. Another paper presented by Osman Hegazy, Omar S. Soliman and Mustafa Abdul Salam took a completely different approach where they narrowed down the machine learning models and focused on one. ANN is used to analyse fundamental and technical data from which feature extraction and selection takes place. Their proposed model architecture contains six inputs vectors represent the historical data and derived technical indicators and one output represents next price. Next an LS-SVM with PSO algorithm was trained and tested over for many companies which cover all stock sectors in S&P 500 stock market. After analysing the results their PSO-LS-SVM algorithm was capable to overcome the over-fitting problem which found in ANN, especially in case of fluctuations in stock sector. Other than this there have been multiple studies using various unique methods however when it came to the results there was a pattern which started to show. A study conducted by Li and Li (2016) used a combination of machine learning algorithms, including the support vector machine (SVM), random forest (RF), and artificial neural networks (ANNs), to predict stock movements. The study found that the combined model outperformed each individual model, indicating that combining machine learning algorithms could improve the accuracy of stock movement predictions. A study by Karamouzis and Maglogiannis (2019) used a hybrid model that combined the features of neural networks and decision trees to predict the movement of the Athens Stock Exchange. The study found that the hybrid model was able to accurately predict the stock market movements with a high degree of accuracy. In conclusion, recent research has demonstrated the effectiveness of machine learning algorithms in predicting stock movements. Studies reviewed in this literature review show that combining different algorithms, such as the SVM, LSTM, RF, and GAN, could improve the accuracy of stock movement predictions.

PROPOSED MODEL / TOOL

This paper proposes an adaptive prediction framework using an ensemble of machine learning models to predict the direction of the closing price, which is shown in the Figure Below. First, the Moving average and Exponential Moving average is calculated in the Stock data set. Then, all featured data is passed as input to multiple machine learning models and the method with the highest prediction accuracy is recorded. Finally, the data is passed through an Ensemble model that aggregate the predictions of the individual models and verifies the prediction accuracy of each pattern by processing the data.

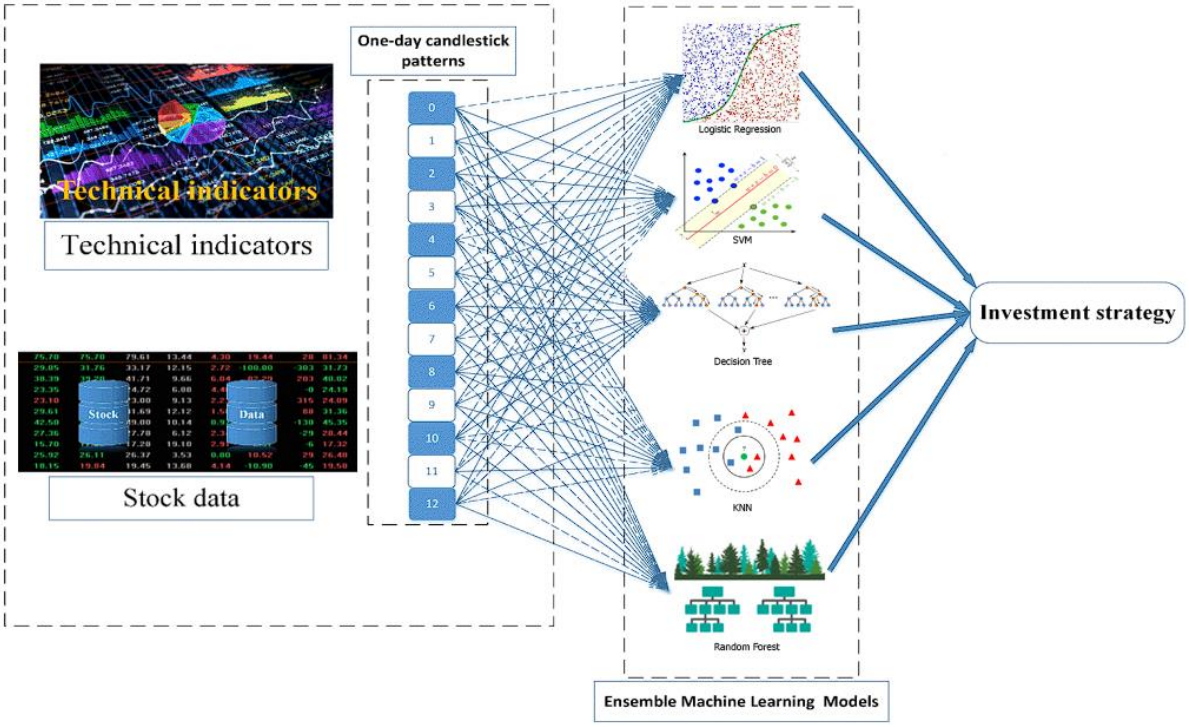
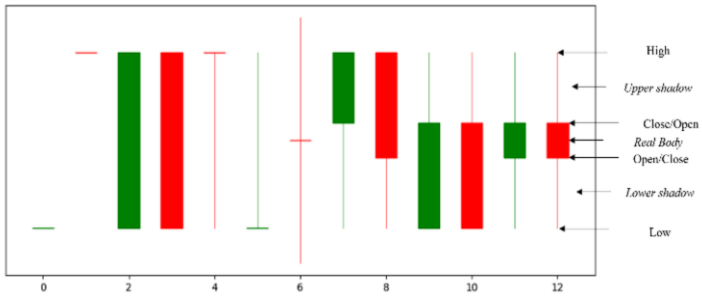


Figure: Overview of prediction framework

DATASETS AND PRE-PROCESSING

This research employs two distinct stock market datasets, namely the Axis Bank and Nifty50 datasets. We begin by pre-processing the datasets which involves handling Null and duplicate values. The research in this paper focuses on short-term forecasting, we choose 5 days or 10 days as the parameters of indicators. Each data set has the following dependent variables: Previous Closed, Open, High, Low, Lost, Last, Moving Average, Exponential Moving Average and Volume. These are technical indicators which are used to determine the Candlestick. We divide the candlestick patterns into 13 classes according to the candlestick basic elements.



GENERATION OF TRAINING AND TESTING SETS

We begin to extract the features after the data pre-processing process according to the feature engineering. In Machine Learning (ML), generating training and testing sets is a crucial step in developing predictive models. The purpose of splitting data into these two sets is to train the model on one set (training set) and test its performance on the other set (testing set). The process of creating these sets involves randomly splitting the original dataset into two non-overlapping sets, with the training set typically comprising 70-80% of the data and the remaining used for testing purposes. The goal of this process is to evaluate the model's generalization ability to new, unseen data. The training set is used to optimize the model's parameters by minimizing its prediction error, while the testing set assesses the model's performance by measuring its ability to accurately predict outcomes on data it has not seen before. This approach helps in identifying and mitigating overfitting, which occurs when the model performs exceptionally well on the training set but fails to generalize to new data.

EXAMPLE DATASETS

NIFTY50 Dataset:

	Date	Prev Close	Open	High	Low	Last	Moving Avg	Exponential Moving Avg	Volume	Close	Class
0	11/27/2007	440.00	770.00	1050.00	770.00	959.00	NaN	596.870	27294366	962.90	1
1	11/28/2007	962.90	984.00	990.00	874.00	885.00	NaN	942.200	4581338	893.90	0
2	11/29/2007	893.90	909.00	914.75	841.00	887.00	NaN	890.990	5124121	884.20	0
3	11/30/2007	884.20	890.00	958.00	890.00	929.00	NaN	895.405	4609762	921.55	1
4	12/3/2007	921.55	939.75	995.00	922.00	980.00	NaN	935.875	2977470	969.30	1
...	...	...	...	...	...	...	...	...	...	...	...
1993	12/17/2015	251.65	253.00	257.65	252.00	256.00	3600510.4	253.045	2430145	256.30	1
1994	12/18/2015	256.30	256.00	261.25	253.55	261.00	3547749.6	257.470	3384209	260.20	1
1995	12/21/2015	260.20	261.55	262.60	257.95	260.30	3402759.8	260.320	1249639	260.60	0
1996	12/22/2015	260.60	261.30	261.30	256.35	256.65	3327823.5	259.700	1992099	257.60	0
1997	12/23/2015	257.60	259.00	260.25	255.85	256.40	3097615.4	257.510	1924064	257.30	0

1999 rows × 11 columns

AXIS BANK Dataset:

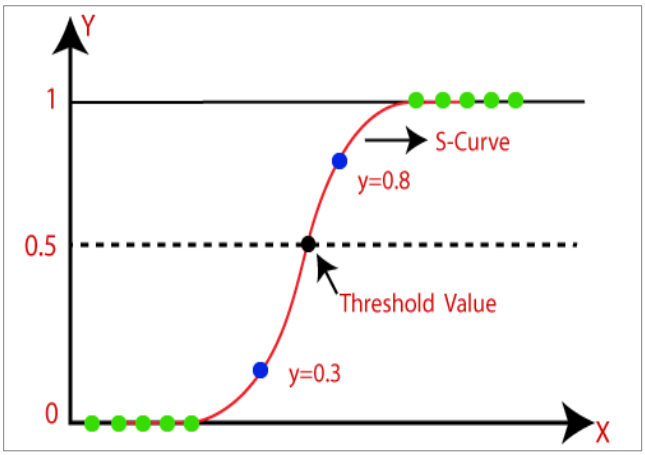
	Date	Prev Close	Open	High	Low	Last	Moving Avg	Exponential Moving Avg	Volume	Close	Class
0	3/28/2013	1294.20	1298.00	1311.20	1277.05	1305.00	NaN	1296.150	1365512	1300.70	1
1	4/1/2013	1300.70	1301.25	1318.50	1296.00	1312.00	NaN	1304.960	981343	1314.90	1
2	4/2/2013	1314.90	1319.90	1320.00	1293.10	1302.00	NaN	1310.835	1146904	1301.35	0
3	4/3/2013	1301.35	1301.90	1316.00	1271.20	1280.00	NaN	1294.390	1692725	1278.15	0
4	4/4/2013	1278.15	1266.00	1270.00	1240.10	1244.65	NaN	1267.860	1283541	1243.85	0
...	...	...	...	...	...	...	...	...	...	...	...
1994	4/26/2021	671.35	694.00	703.80	684.50	699.50	19117300.50	680.080	21646184	700.45	1
1995	4/27/2021	700.45	691.10	703.90	684.10	700.90	23058405.10	700.180	46559967	699.55	1
1996	4/28/2021	699.55	708.00	712.50	688.15	705.95	24378572.80	702.130	54060587	708.15	1
1997	4/29/2021	708.15	712.00	726.90	707.00	717.10	25160143.10	711.525	25939327	719.40	1
1998	4/30/2021	719.40	705.00	729.85	705.00	711.65	26595434.33	718.050	23011654	714.90	1

1999 rows × 11 columns

PREDICTION MODELS

Five machine learning models including Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), K-NearestNeighbor (KNN) and Random Forest (RF) are used to predict the direction of the closing price.

1) LOGISTIC REGRESSION (LR)



Logistic Regression is one of the simplest yet powerful algorithms used in machine learning. The algorithm involves constructing a model that can predict the relationship between independent variables and a binary dependent variable. The model returns an equation that can quantify the impact of each independent variable on the dependent variable. In Logistic Regression, the relationship between the input variables and the output is modelled using a logistic function, also known as the sigmoid function. This function maps any input value to a value between 0 and 1 (in this case, 0 indicates Fall while 1 indicates Rise in the direction of the price), which represent the probability of the input belonging to a particular class.

Logistic Regression has several advantages, such as being easy to implement, computationally efficient, and interpretable. It can also handle both linear and nonlinear relationships between the input variables and the output, making it a versatile algorithm for a wide range of problems. Logistic Regression can be regularized to prevent overfitting and improve the generalization performance of the model.

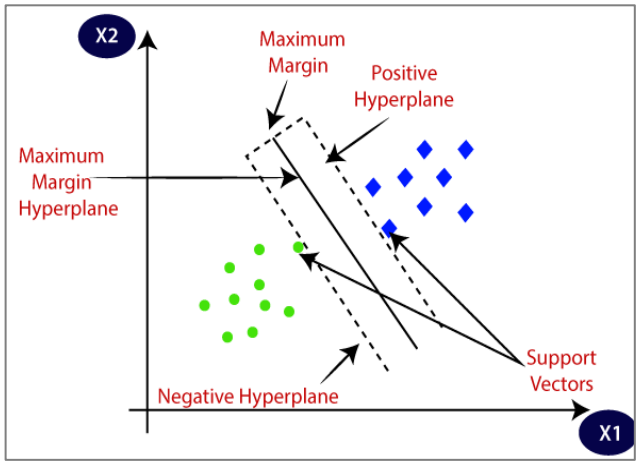
On the other hand, While Logistic Regression can handle both binary and multi-class classification problems, it may not be suitable for problems where the output variable has more than two classes and the classes are imbalanced. Logistic Regression cannot handle missing values in the input data and also requires large sample sizes to achieve good performance and avoid overfitting.

Despite these limitations, Logistic Regression is a powerful and flexible algorithm that can be used in a variety of domains and is often used as a baseline model for more complex machine learning models.

2) DECISION TREE (DT)

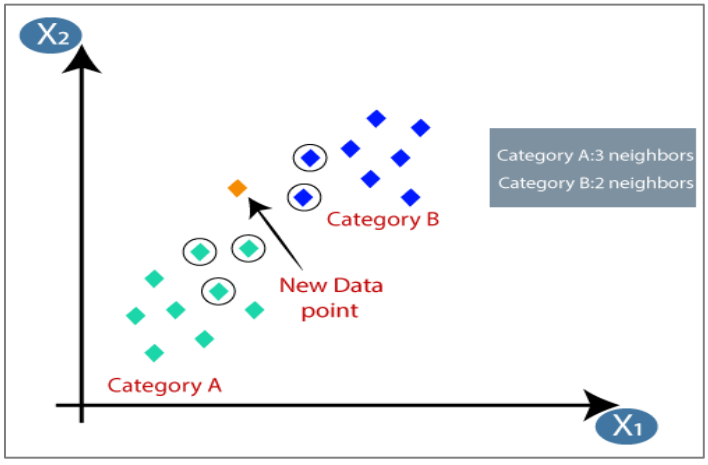
Decision Tree is a supervised machine learning algorithm used for classification as well as regression tasks. It works on the basic principle of splitting the main input dataset into smaller subsets based on the feature having the highest information gain. To compute the information gain of each feature, a measure of randomness in the dataset, known as entropy, is used. The algorithm selects a feature that results in highest information gain at each step and further partitions the dataset into two or more subsets. This process is repeated till the leaf nodes are reached. In classification problems, the algorithm predicts the class label of the input data based on the decision made at each node. For regression problems, the numerical value of output is predicted based on the decision made. Decision tree is a popular and powerful machine learning method that are easy to interpret and can handle non-linear relationships. It follows a non-parametric method and make no assumptions about the distribution of data. The algorithm is highly flexible and can work with various data types while performing feature selection and handling missing values. However, decision trees are prone to overfitting when the tree is too complex. In such cases, we must prune the tree or use ensemble methods to improve the overall performance.

3) SUPPORT VECTOR MACHINE (SVM)



SVM (Support Vector Machine) is a popular machine learning algorithm used for classification and regression tasks. It is a powerful and flexible algorithm that can handle both linear and nonlinear relationships between the input and output variables. In SVM, the goal is to find a hyperplane in the input space that separates the data into different classes, such that the margin between the hyperplane and the nearest data points from each class is maximized. The hyperplane is chosen in such a way that it maximizes the margin while minimizing the classification error. The SVM model is to map the example as a point map in space, so that the examples of the separate categories are divided by a clear gap that is as wide as possible.
SVM has several advantages, such as being able to handle high-dimensional data and perform well on small to medium-sized datasets. It can also be customized using different kernel functions to handle nonlinear relationships between the input and output variables. SVM has a wide range of applications, including image classification, text classification, and bioinformatics. It is often used as a baseline algorithm for more complex machine learning models. However, SVM also has some disadvantages, such as being sensitive to the choice of kernel function and parameters, and being computationally expensive for large datasets. Nevertheless, SVM remains a popular and powerful algorithm in the field of machine learning.

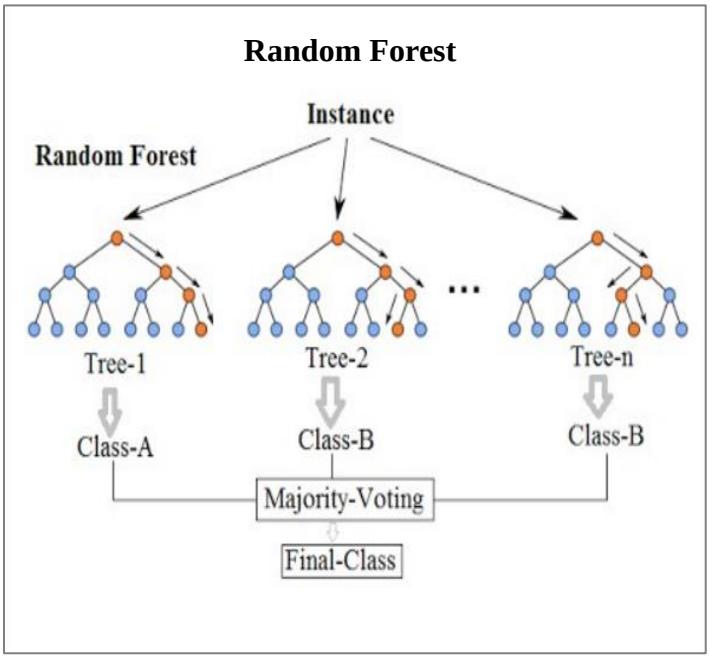
4) K-NEARESTNEIGHBOR (KNN)



KNN (K-Nearest Neighbors) is a simple yet effective algorithm that is based on the idea of similarity between data points. In KNN, the input data points are represented as vectors in a high-dimensional space, and the goal is to classify a new data point based on its similarity to the existing data points. The algorithm works by finding the K nearest neighbors of the new data point in the input space, and then assigning the class label of the majority of those neighbors to the new data point. With different parameter combinations we will get different clustering effects.
One of the main advantages of KNN is that it is a simple and intuitive algorithm that is easy to understand and implement. This makes it a good choice for beginners and for tasks that require a quick and simple solution. KNN can handle both binary and multi-class classification problems, as well as regression tasks. This flexibility makes it a versatile algorithm that can be applied to a wide range of problems. Additionally, KNN is a lazy learning algorithm, which means that it does not require any training or model building. Instead, it simply stores the training data in memory and uses it to classify new data points.
However, KNN also has some limitations, such as being sensitive to the choice of distance metric and the value of K, and being computationally expensive for large datasets.

5) RANDOM FOREST (RF)

RF (Random Forest) is a commonly-used machine learning algorithm which combines the output of multiple decision trees to reach a single result i.e., Random forests are a combination of tree predictors. The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. The algorithm works by creating a random sample of the input data and building multiple decision trees using different subsets of the data. Each decision tree is trained on a random subset of the features, which helps to reduce overfitting and improve the model's generalization performance. During the prediction phase, the algorithm combines the outputs of all the decision trees to make the final prediction. The majority vote of the decision trees is taken as the final prediction for classification problems, while the average of the outputs is taken for regression problems.



Random Forest is a robust and accurate algorithm that can handle a wide range of input data types and handle missing data. It is also resistant to overfitting, making it a good choice for high-dimensional datasets with many features. Additionally, Random Forest can provide feature importance scores, which can help in identifying the most important features in the input data. This makes it a useful algorithm for feature selection and feature engineering tasks. Although random forest can be used for both classification and regression tasks, it is not as suitable for Regression tasks.

ENSEMBLE MODEL

Ensemble models are a class of machine learning algorithms that combine the outputs of multiple models to make more accurate and robust predictions. These models have become increasingly popular in recent years due to their ability to improve the performance of individual models and provide better predictive power.
Ensemble models can be broadly categorized into two types: bagging and boosting.

Bagging, short for bootstrap aggregating, involves creating multiple independent models on different subsets of the training data and then combining their outputs to make the final prediction. This helps to reduce overfitting and improve the generalization performance of the model. Random Forest is a popular example of a bagging ensemble model.

Boosting, on the other hand, involves creating multiple models sequentially, with each subsequent model focused on correcting the errors of the previous model. This helps to improve the accuracy of the model and reduce bias. AdaBoost and Gradient Boosting are popular examples of boosting ensemble model.

Ensemble models offer several advantages over individual models. By combining the outputs of multiple models, ensemble models can reduce the variance and increase the stability of the predictions. They can also handle a wide range of input data types and perform well on high-dimensional datasets.
Additionally, ensemble models can provide valuable insights into the underlying relationships between the input and output variables. They can identify the most important features in the input data, which can help in feature selection

and feature engineering tasks. They can also provide uncertainty estimates, which can be useful in making decisions under uncertainty.
However, ensemble models also have some drawbacks. They can be computationally expensive and time-consuming to train, especially for large datasets. They can also be difficult to interpret and visualize, making it challenging to understand the underlying relationships between the input and output variables.
In summary, ensemble models are powerful machine learning algorithms that can significantly improve the accuracy and robustness of predictions. They offer several advantages over individual models, including improved stability, feature importance estimation, and uncertainty estimates. However, they also have some limitations, including computational complexity and interpretability.

MODEL EVALUATION

Since several machine learning algorithms are used to predict stock movement trends, to measure the performance of each prediction model we use Accuracy, Precision and Recall.

i. Accuracy:

It is used to evaluate the overall classification ability of the model and is defined as the number of correct predictions made by the model divided by the total predictions made.

Mathematically, accuracy is calculated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy is a commonly used performance metric but it may not be the most appropriate at times when the classes are imbalanced and one class has many more instances than the other(s).

ii. Precision:

It is used to estimate the accuracy of the positive samples in the input data and is defined as the number of correct positive predictions made by the model divided by the total positive predictions made.

The formula is as follows:

$$\text{Precision} = \frac{TP}{TP+FP}$$

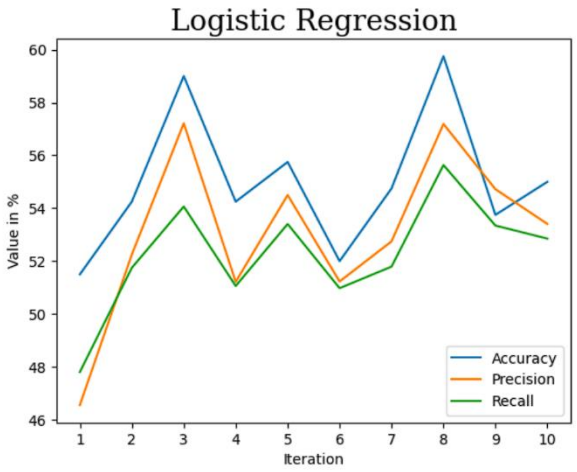
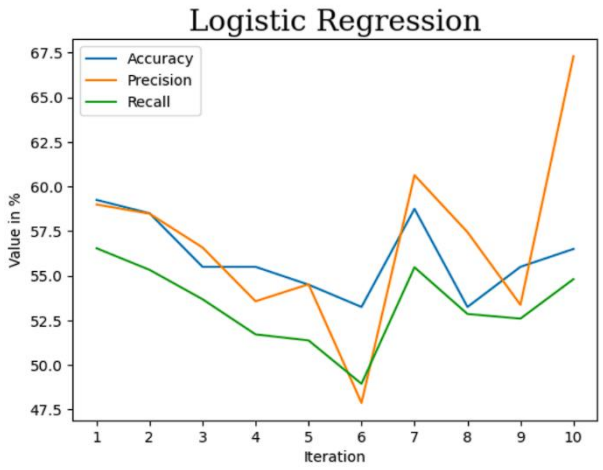
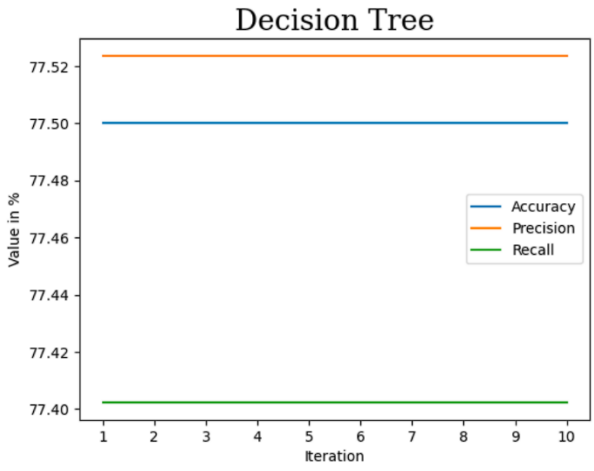
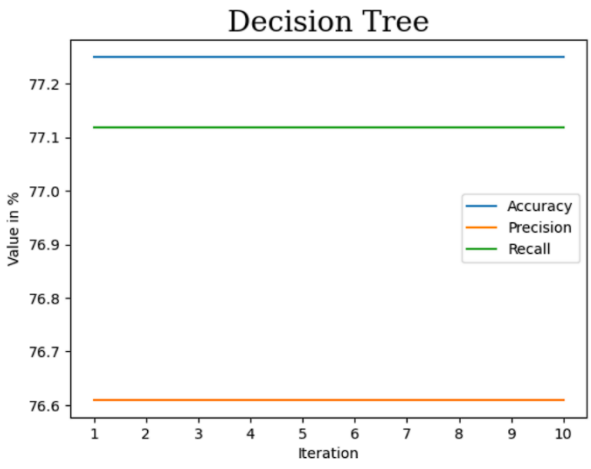
iii. Recall:

It is used to measure the proportion of actual positive instances that are actually identified against those that were not. Recall is defined as the total number of correctly identified positive samples divided by total positive samples in the prediction data.

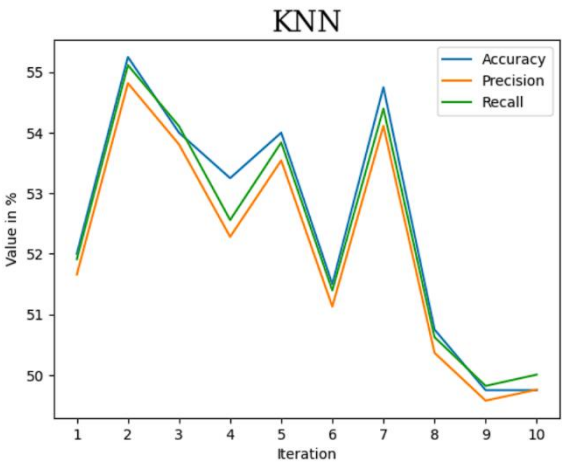
$$\text{Recall} = \frac{TP}{TP+FN}$$

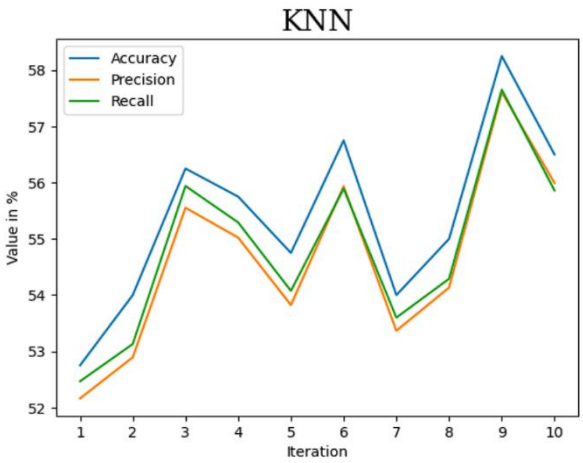
IMPLEMENTATION AND RESULTS

In this research paper, we have majorly worked on two datasets- Nifty50 and Axis Bank. Each dataset was cleaned and pre-processed to eliminate null and duplicate values. Further, we passed this prediction data through the above mentioned five machine learning algorithms for 10 iterations each and their respective Accuracy, Precision and Recall scores were documented and plotted in the form of a line graph.

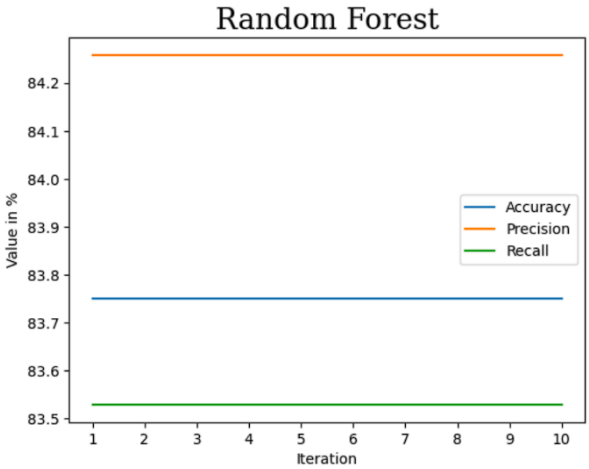
AXIS BANK	NIFTY50
I. Logistic Regression	
	
II. Decision Tree	
	

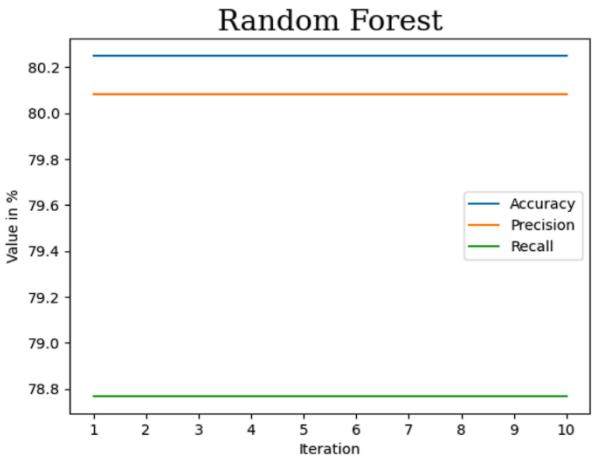
III. K-Nearest Neighbors



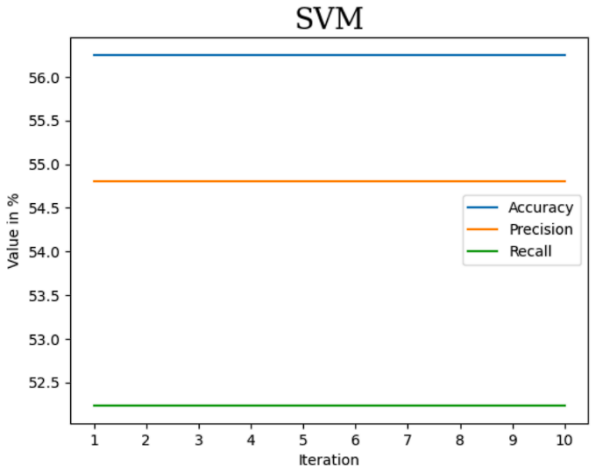


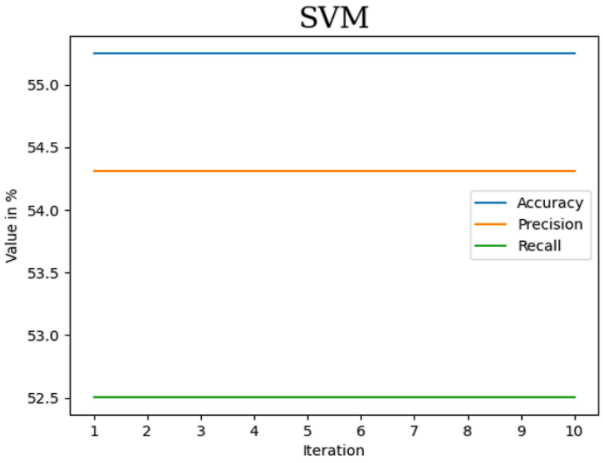
IV. Random Forest





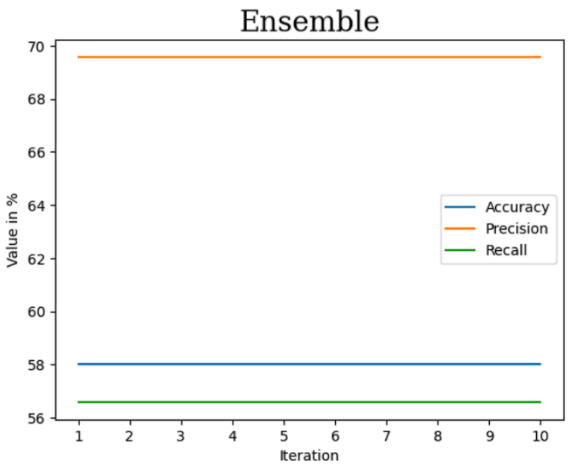
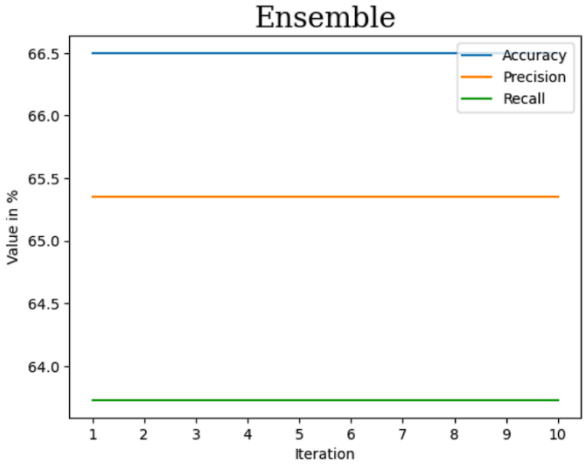
V. SVM





USE OF ENSEMBLE MODEL

Ensemble models are widely used in machine learning to improve the stability and accuracy of predictions, especially in cases where individual models such as KNN and logistic regression are prone to high variance. Ensemble models work by combining the predictions of multiple models, which can help to reduce variance and improve generalization performance. KNN and logistic regression are both prone to overfitting, which can lead to high variance in their predictions. Overfitting occurs when a model is too complex and fits the training data too closely, resulting in poor generalization to new data. This can lead to unstable and unpredictable results. For example, if we have a dataset where KNN and logistic regression are both producing unstable results due to high variance, we can use an ensemble model to combine the outputs of multiple models and produce a more stable prediction. The ensemble model can take into account the strengths and weaknesses of each individual model and produce a more robust and accurate prediction.

AXIS BANK	NIFTY50
	

In summary, ensemble models can be a powerful tool to reduce variance and increase the stability of predictions, especially when working with complex datasets or models prone to overfitting. By combining the outputs of multiple models, ensemble models can provide more reliable and accurate predictions, even in cases where individual models are producing unstable or inconsistent results.

CONCLUSION

This paper constructs a novel machine-learning framework that predicts daily stock patterns by combining the latest machine-learning technique with traditional candlestick charting. This framework can analyse each pattern and make predictions based on trained results, using algorithms such as Logistic Regression, Decision Tree, K-Nearest Neighbours, SVM as well as Random Forest.

Implementing an ensemble model that combines these algorithms can lead to improved stock market prediction accuracy. For example, an ensemble model can use the predictions from each algorithm as input, and then use a weighted average or a voting mechanism to make the final prediction. This can help to mitigate the limitations of individual algorithms and improve overall prediction performance. Additionally, ensemble models can benefit from the diverse perspectives of multiple algorithms, leading to more robust and accurate predictions in the unpredictable stock market environment. By leveraging the strengths of different algorithms in an ensemble model, it is possible to create a more reliable and accurate tool for predicting stock market trends and making informed investment decisions.

This study makes contributions in the following aspects:

Firstly, this article combines the method of candlestick charting with a plethora of machine-learning algorithms to bolster the forecasting research of the stock market.

Secondly, by using the candlestick patterns alongside different machine learning methods, we combine traditional techniques of technical analysis methods with modern datasets.

Thirdly, we also concluded that certain algorithms, for example, Random Forest and Decision Tree, have apparent predictive effects in the stock market.

Additionally, to improve the forecast level, we introduced the following criteria: Date, Previous Close, Open, High Low, Moving Average, Exponential Moving Average, Volume, Close and finally Class.

We find that the momentum indicators are significantly better when we used empirical testing and no other indicators while short-term forecasting.

The Candlestick Patterns were made on both the datasets of Axis Bank and Nifty50. The variables considered were, Moving Average and Days (calculated based on Date)

The forecasting framework of this paper has very useful predictions on the stock market, although slightly disputable in specific periods. As the datasets used contain data from the economic crisis in India during the 2000s.

For example, the Support Vector Machine (SVM) is not suitable for predicting large-scale stock data.

In the future of this thesis, we may implement more appropriate machine learning methods for prediction, namely Reinforcement Learning into our Ensemble Model.

Furthermore, we plan to incorporate various other factors which may or may not affect the stock market, such as predictive factors, such as major news events, and market sentiment, to improve forecasting results.

REFERENCES

[1] YAOHU LIN , SHANCUN LIU, HAIJUN YANG, AND HARRIS WU "Stock Trend Prediction Using Candlestick Charting and Ensemble Machine Learning Techniques With a Novelty Feature Engineering Scheme," IEEE, 2021
[2] L. Breiman, "Random forests," 2001
[3] M. V. Subha and S. T. Nambi, "Classification of stock index movement using k-nearest neighbours (k-NN) algorithm," 2012
[4] D. Brownstone, "Using percentage accuracy to measure neural network predictions in stock market movements," 1996
[5] M.-C. Lee, "Using support vector machine with a hybrid feature selection method to the stock trend prediction," 2009
[6] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," 2015
[7] R.S. Latha, G.R. Sreekanth, R.C. Suganthe, M. Geetha, R. Esakki Selvaraj, S. Balaji, K.R. Harini and P. Priya Ponnusamy, "Stock Movement Prediction using KNN Machine Learning Algorithm," IEEE, 2022