ISSN: 2321-676X

STAR
Research Journal

*Available online at www.starresearchjournal.com (Star International Journal)*

**MATHEMATICS**

# USING MACHINE LEARNING TO MAKE PREDICTIONS REGARDING THE STOCK MARKET

**E. Jayanthi[1], S. Shunmugapriya[2],**

*1. Research Scholar, PG and Research Dept. of Mathematics, Rajah Serfoji Government College (Autonomous), Thanjavur.*

*2. Research Advisor, PG and Research Dept. of Mathematics, Rajah Serfoji Government College (Autonomous), Thanjavur.*

*(Affiliated to Bharathidasan University)*

**Abstract:**

The primary goal of this research is to examine and evaluate several stock market forecasting models. Despite the complexity of the problem space, we discovered that techniques such as random forest and support vector machine were underutilised. In this piece, we'll talk about a more realistic strategy for forecasting the direction of stock prices. Our first consideration is the stock market data set from the prior year. The data set was cleaned and fine-tuned before being used in the research. Therefore, we will also discuss preparing the raw data for our work. Second, when the data has been cleaned and prepared, we'll compare random forest and support vector machine, two popular machine learning methods. Random Forest Classifier mathematical modeling also looks at the accuracy of the overall values supplied and how they might be used in practice. In addition, this research presents a machine-learning strategy for forecasting stock prices in a volatile market. Financial institutions would benefit greatly from accurate stock forecasting, and investors would have real solutions to their problems.

**Keywords:** Stock Market, Data Mining, Data Pre-processing, Stock, Machine Learning, Dataset.

## I. INTRODUCTION

Investors and traders interact in the stock market to purchase and sell shares of stock. A stock, often known as shares, represents a part owner's claim to the assets of a corporation. Predictions about the stock market are attempts to estimate how much money will be made or lost in the future. The prediction is expected to be credible, accurate, and useful. The system has to be adaptable to many settings and work as intended in the actual world. Additionally, the system is anticipated to account for every factor that might influence the stock's performance and valuation.

Some techniques that might be utilised to build the prediction system include fundamental analysis, technical analysis, machine learning, market mimicry, and time series aspect structure. Mathematical modeling has become more sophisticated as the digital age has progressed. The most well-known and [3] promising technology now accessible is artificial neural networks that do machine learning, such as Recurrent Neural Networks. Machine learning employs AI to help computers learn from their errors and improve themselves without being explicitly programmed. One of the algorithms employed in conventional machine learning prediction methods is Backward Propagation, often known as Backpropagation Errors. Academic support for ensemble learning methodologies has grown in recent years. Another network would use low price and time [3] delays instead of delayed highs to predict future highs. Stock prices were derived using these forecasts. [1] It seems that predicting stock market prices over short time frames is a random process. Over time, a stock's price movement often follows a linear curve. People often invest in stocks whose values are anticipated to grow soon. People hesitate to invest in equities because of the stock market's instability. Therefore, making accurate stock market forecasts that may be used in the actual world is essential. Methods like time series forecasting, technical analysis, machine learning modelling, and fundamental analysis may be used to predict the volatile stock market. A stock market prediction model may be used to predict the object variable, the price on a particular day, using datasets that include information like the closing price, the starting price, data, and several other characteristics. The earlier model predicted future outcomes using well-established techniques, including multivariate analysis and a prediction time series model. Predicting the stock market is more successful when framed as a regression problem rather than a classification one. The objective is to use machine learning techniques applied to market data to generate a model capable of predicting changes in stock prices. The Support Vector Machine (SVM) may benefit both classification and regression. For classification problems like the one we have, SVMs are a common tool. The SVM technique converts each data point into a coordinate in an n-dimensional space (where n is the number of features in the dataset), with the value of each feature being the value of a specific coordinate.

Therefore, classification is finished by locating the hyperplane that separates the two groups. For this, analysts turn to predictive techniques like the Random Forest method. The random forest technique does classification and regression using an ensemble learning strategy. By averaging out the many subsamples of the dataset, the random forest improves prediction accuracy and reduces over-fitting.

## II. PROBLEM DEFINITION

Simply put, stock market prediction is making educated guesses about future stock prices and market conditions to help investors plan for the future. The financial ratio for the quarter included in the dataset might serve as an example. As a result, it's possible that making the forecast based on a single dataset might be insufficient and provide an incorrect result. We are thus considering studying machine learning and integrating different datasets to forecast market and stock movements. Until a more reliable stock market prediction algorithm is proposed, estimating stock prices won't be easy. It may be challenging to foresee the behaviour of the stock market. Thousands of investors' opinions may have a significant impact on the direction of the stock market. To accurately predict the stock market, one must consider the impact of unfolding events on the world's financiers. These may include political occurrences like a political leader's comment or news of a scam. A global occurrence, such as sudden changes in the value of currencies or commodities, might potentially be the cause. These many occurrences have an impact on business profits, which have an impact on market mood. The ability to consistently and properly estimate these hyperparameters is beyond the capabilities of practically all investors. Predicting stock prices is incredibly difficult due to all of these variables. After sufficient information has been amassed, it may teach a computer to make a prediction.

## III. LITERATURE SURVEY

After a literature review, we gathered some details on the present stock market prediction systems.

### 1. Analysis of the Machine Learning Method for Predicting the Stock Market

Today, stock market forecasting is a topic that is becoming more and more significant. Technical analysis is one of the tactics used. However, these approaches don't always provide correct outcomes. Therefore, strategies for a more precise forecast must be developed. Typically, projections based on stock price are used to make investments after considering all potential influences. In this case, regression analysis was the method used. Since the stock markets generate so much data at once, a thorough analysis of a vast amount of information is required before a prediction can be made. The many methods for analysing regression data each have their advantages and limitations. Linear regression was mentioned as a significant technique. While the least squares approach is often used to fit linear regression models, various methods exist. For instance, one may use a variant of the least squares loss function to minimise the "lack of fit" in a different norm. It is possible to apply the least squares method while fitting nonlinear models.

### 2. Stock Price Prediction Using Random Forests: The Role of Financial Ratios and Technical Analysis

Predictions of stock prices using AI and machine learning are becoming more common. More and more scientists are always working on new techniques to improve the stock prediction model's precision. There are many techniques to anticipate the stock price since so many possibilities are accessible, but each approach has a different mode of operation. Each approach produces a different result even when the same data set is used. Using financial statistics from the prior quarter and the random forest approach, the cited study was able to forecast the stock price. This is only one way to approach the problem using predictive modelling; another would be to anticipate the stock price's future performance without historical data. The price of a stock may also be affected by factors such as investor sentiment, public opinion of the company, external market factors, and other news and information. Combining the financial ratio with an emotional analysis model might improve the stock price prediction model.

**3. Predicting Stock Prices Using MULTIPLE INSTANCES OF LEARNING**

The internet has become an invaluable tool for simplifying this procedure in recent years. Predicting the stock market accurately is challenging. Some feelings are easy to extract because of the interconnected nature of the data; this makes it less difficult to establish associations between variables and, in effect, draw an investment pattern. The stock market may be accurately predicted by using the similarities between the data sets in the investment patterns of different companies. Additional techniques, such as sentiment analysis, may be used to correctly anticipate stock market information by employing more than just technical, historical data. Investors' feelings about certain firms and their subsequent performance are inextricably linked. Extrapolating significant occurrences from online news to see how they influenced stock prices was another crucial step in the forecast process.

**4. Forecasting the Stock Market using an Examination of Past Performance**

Since so many factors are at play, accurate stock market forecasting is challenging. As a result, the stock market has far-reaching consequences for business and finance. The emotional analysis procedure is used to do technical and fundamental analysis. Due to the rise in usage, social media data has a significant influence and may [6] be used to forecast stock market trends. Technical analysis [6] uses machine learning algorithms using historical stock price data. The process often entails collecting news and other social media data to extract the feelings expressed by people. In addition, other information is considered, such as stock price history. After considering the relationship between the various data pieces, a prediction is made using these points. The model has the potential to predict where stock prices are headed.

**5. Review of Support Vector Machines for Stock Market Prediction**

Most predictive regression models fail to perform well in tests of out-of-sample prediction, as shown by recent research. This inefficiency was brought on by parameter instability and model uncertainty. The investigations came to the same conclusions about the tried-and-true solutions to this issue. Support vector machines, or SVMs, provide the solution's

kernel, decision function, and sparsity. The multi-layer perceptron classifier and polynomial radial basis function are taught using it. It's a method for teaching classification and regression that uses a more extensive data set. Although several techniques exist, support vector machines (SVMs) provide superior accuracy and efficiency. The correlation study between SVM and the stock market shows strong ties between stock prices and the market index.

**6. A Mathematical Model Analysis for Estimating Stock Market Price Changes**

The different methods for estimation of parameters of Weibull distribution were examined using Mean Square Error (MSE) as a criterion for selecting the best model. The Method of Moments exceeded other methods. In the same circumstance, the estimated results were logically extended to form a matrix that would help in predicting different commodity price processes by exploring the properties of fundamental matrix solution where we obtained predicted stock prices and asset returns for 12 months. Finally, from the fundamental matrix system a theorem was developed and proved to show different levels of changes as it affects stock market in terms of short-run and long-run respectively.

**7. Using Support Vector Machines (SVMs) and Independent Component Analysis to Make Stock Market Predictions (ICA)**

The work centres at many financial institutions researched the time series prediction issue. The SVM-ICA prediction model, which combines SVM with independent analysis, is suggested to predict the stock market. Machine learning is the foundation of many time series analysis methods. The SVM was developed to resolve regression challenges in non-linear classification and time series analysis. A rough function based on the decreasing risk concept reduces the generalisation error. Consequently, the ICA approach retrieves several crucial properties from the dataset. Modelling time series using a support vector machine. No preprocessing was used to compare the SVM model and the ICA technique.

**8. Mathematical model of back propagation for stock price forecasting**

In order to establish a more accurate Stock Price Prediction Model, the Stock Price Prediction mathematical Model SPPM (Stock Price Prediction Model) based on BP neural network with high frequency data is proposed in this paper. The SPPM integrates several neural networks to predict the movement of stock prices over the next few days. The key problems in SPPM—such as data preprocessing, output fusion and the selection of nodes in the hidden layer of neural network—are discussed in detail. The experimental results show that the SPPM predicted the closing price of 2019-03-19 and 2019-03-20 as 207.16 and 207.22, respectively, which is very close to the actual observed value, and the back propagation mathematical model SPPM has a certain practical value. Our conclusion is that the back propagation model can predict the stock price with high accuracy.

**9. Stock Price Changes and the Company's Internal Communication Network: Insights from Data Mining**

The purpose of this paper is to argue that differences in communication styles may have serious consequences for an organisation's efficiency. Following the

study's recommendations, companies should be more forthcoming about their performance. The research strategy examines how often important employees contact one other and how that correlates with stock price performance metrics. An openly available dataset from Enron Corp. was analysed in this research to see whether or not there were any significant associations to be mined using data mining techniques. Public access is provided to stock data for The Enron Corporation, an energy, commodities, and services supplier headquartered in Houston, Texas.

**10. A mathematical model for stock price forecasting**

Many mathematical models of stochastic dynamical systems were based on the assumption that the drift and volatility coefficients were linear function of the solution. In this work, we arrive at the drift and the volatility by observing the dynamics of change in the selected stocks in a sufficiently small interval $\Delta t$ . We assumed that only one change occurs within $\Delta t = t_{i+1} - t_i$. During this time, a stock may gain one unit *[+1]*, remain stable *[0]*, or loss one unit *[-1]*. The likelihood of each change occurring were noted and the expectation (the drift) and the covariance (the volatility) of the change were computed leading to the formulation of the system of linear stochastic differential equations. To fit data to the model, changes in the prices of the stocks were studied for an average of 30 times. A simple checklist was used to determine the likelihood of each event of a loss, again or stable occurring. The drift and the volatility coefficients for the SDE were determined and the multi-dimensional Euler-Maruyama scheme for system of stochastic differential equations was used to simulated prices of the stocks for $1 < t < 30$. The simulated prices was compared to the observed price and we observed that the simulated prices is sufficiently close to the observed price and there for suitable for forecasting the price of the stocks for a short time interval.

## IV. CONSIDERATIONS OF THE CURRENT SYSTEM

- When dealing with out-of-the-ordinary results or predictors, the existing method fails due to the bootstrap sampling basis of the algorithm.
- According to previous research, using the standard classifier might lead to stock price uncertainty.
- The existence system gave highly predicted findings since the researchers could choose a suitable time frame to conduct their experiment and receive highly predictive assessments.
- When the current system is put into a new environment, it fails.
- It's not concerned with current events or issues like the news or social media.
- It places too much emphasis on a single source of information since that is all it uses.
- Since the existing system needs input interpretation, scaling is essential.
- It does not use pre-processing techniques to clean data of inconsistencies and gaps.

## V. PROPOSED SYSTEM

We focus on predicting stock prices in the proposed system using machine

learning techniques like Random Forest and Support Vector Machines. Using the random forest technique, we created a system called "Stock market price prediction" to anticipate the stock market's future value. Using the proposed approach, we used the many available historical data points to train the computer. We utilised last year's stock market data to teach the algorithm. Using two different machine-learning software, we successfully addressed the problem. Numpy was the first tool used to clean, alter, and organise the data in preparation for analysis. Scikit, the second tool, was useful for accurate analyses and forecasts. We utilised historical stock market data from a publicly available internet database, with 80% of the data used to train the computer and 20% used for testing. The fundamental concept underpinning the supervised learning model is the transfer of learned patterns and correlations from the training set to the test set. We utilised the Python pandas utility to merge many data sets into a single, manageable whole. The optimised data frame allowed us to prepare the data for feature extraction. Each day's closing price and date are shown in separate data frame columns. All of these attributes were used to train the computer's random forest model, which was used to predict the object variable, the price, on a certain day. We also calculated accuracy by comparing our models' predictions to the actual values in the test set. Data pre-processing, random forests, and other research areas are all brought together in the proposed method.

## VI. METHODOLOGIES

### 1. Classification

An instance of supervised learning is when data is analysed and categorised based on some common feature. Classification draws inferences about the observed value from the given values or data. Classification will try to predict multiple outputs based on multiple inputs. Classifiers like the SVM and Random Forest classifiers are used here to make stock market predictions.

**Random Forest Classifier**

One such supervised ensemble technique is the random forest classifier. In essence, it generates several decision trees, each of which can only end in one way. A random class classifier aims to provide a final classification or result based on the votes of a randomly selected subset of decision trees.

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

(1)

**Parameters**

Several decision trees (n estimators), random forest generalisation accuracy (oob-score), and maximum number of features for best-fit classification (max features) are the hyperparameters for the random forest classifier. The minimum weighted percentage of leaves specifies how many input samples must reside at leaf nodes. If there is no weighting scheme for the samples, they are all equal.

**SVM classifier**

An effective discriminative classifier is the support vector machine (SVM) classifier. Supervised learning, which is what the SVM does, requires labelled training data. The final product is a collection of classification hyperplanes applicable to the new data. They are supervised learning models

capable of combining regression and classification.

**Parameters**

The SVM classifier's tuning parameters are kernel, gamma, and regularisation parameters.

- The two types of kernels that compute the prediction line are linear and polynomial. In linear kernels, the dot product of the input and the support vector is used to determine the prediction for a new input.
- The regularisation or C parameter controls whether the model's accuracy grows or decreases. C is set to 10 by default. Misclassification results from lower regularisation values.
- The gamma parameter quantifies the effect of a single training session on the model. Values closer to one another suggest a closer closeness to the feasible boundary, while those further apart imply greater distance.

**2. Random Forest Algorithm**

Stock market predictions are made using the random forest technique. One of the most adaptable machine learning algorithms, it can be utilised by various individuals and deployed with relative ease, yielding accurate predictions. This method is often used for issues of classification. Since the stock market is so unpredictable, forecasting it is a challenging task. Since a random forest classifier's hyperparameters are similar to those of a decision tree for stock market forecasting, we use it instead. The decision-making interface is tree-like in design. Things like event outcomes, resource costs, and utility are all considered.

$$\hat{f}_{\mathrm{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b).$$

(2)

$$\hat{f}_{\mathrm{rf}}(x) = \mathrm{E}_{\Theta} T(x;\Theta) = \lim_{B\to\infty} \hat{f}(x)_{\mathrm{rf}}^{B}$$

(3)

To illustrate an algorithm that constructs several decision trees by randomly selecting different observations and features and then averaging the outputs of those separate decision trees, we have the random forest method. Parts of the data are separated based on questions about a label or a characteristic. We utilised a publicly available database that included stock market information for a whole year, with 80% of the information used to train the machine and 20% used for testing. Applying what has been learned in the training set to the test set is the core idea behind supervised learning models.

**3. Support Vector Machine(SVM) Algorithm**

To effectively classify the input points, the support machine algorithm must first locate an N-dimensional space that does so. In this context, N represents a set of attributes. Between any two sets of data, several possible hyperplanes exist. The algorithm's primary objective is to locate the plane with the largest margin. "Maximising margin" refers to the minimum distance between data points from both sets. Increasing the margin has the benefit of providing reinforcement, making it simpler to classify future data points. In the process of categorising data points, hyperplanes serve as decision boundaries. The data points are categorised differently depending on their distance from the hyperplane. If there are just two

characteristics, then the hyperplane is only a line; if there are three qualities, it is two-dimensional.

### Pseudocode for SVM Algorithm

```
def svm(data, labels):
  """ Trains a support vector machine model from the given data.
  Args:
    data: The training data.
    labels: The labels for the training data.
  Returns:
    The support vector machine model.
  """ mean = np.mean(data, axis=0)
  std = np.std(data, axis=0)
    data = (data - mean) / std
  w, b = find_separating_hyperplane(data, labels)
  support_vectors = find_support_vectors(data, labels, w, b)
  return SVMModel(w, b, support_vectors)
def find_separating_hyperplane(data, labels):
  """Finds the separating hyperplane for the given data.
  Args:
    data: The training data.
    labels: The labels for the training data.
  Returns:
    The weight vector and bias term of the separating hyperplane.
  """ w = np.zeros(data.shape[1])
  b = 0
  for i in range(len(data)):
    loss = labels[i] * (np.dot(w, data[i]) + b)
    if loss > 0:
      w += labels[i] * data[i]
      b += labels[i]
  return w, b
```

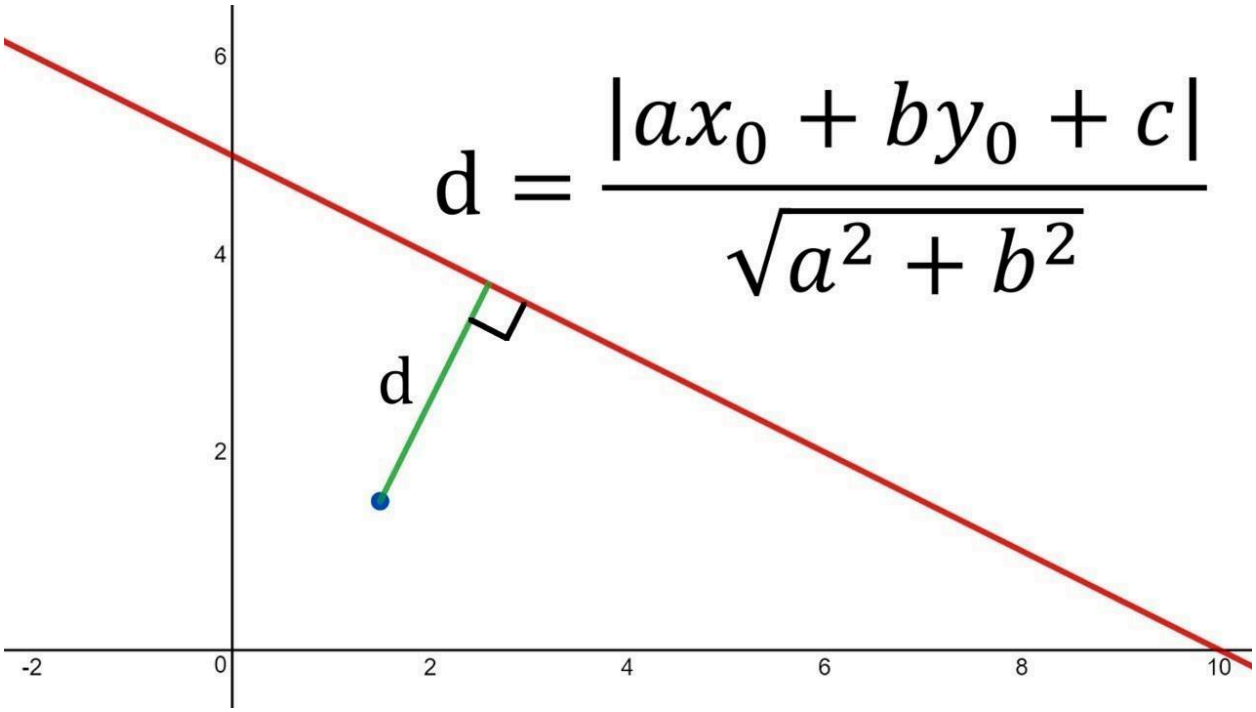
Equation of SVM

### Distance Measure

Now, we have seen how to represent data points and fit a separating line between the points. But, while fitting the separating line, we would want such a line that would be able to segregate the data points in the best

possible way, having the least mistakes/errors of miss-classification.

So, to have the least errors in the classification of the data points, that concept will require us first to know the distance between a data point and the separating line.




The distance of any line, ax + by + c = 0 from a given point, say, $(x_0 , y_0)$ is given by d.

Similarly, the distance of a hyperplane equation: $w^T\Phi(x) + b = 0$ from a given point vector $\Phi(x_0)$ can be easily written as :

$$d_H(\phi(x_0)) = \frac{|w^T(\phi(x_0)) + b|}{||w||_2}$$

**here ||w||2 is the Euclidean norm for the length of w given by :**

$$||w||_2 =: \sqrt{w_1^2 + w_2^2 + w_3^2 + \dots w_n^2}$$

The quality of the separating hyperplane's ability to distinguish between the two types of data is quantified by the hinge loss function, which serves as the objective function. Finding the proper weight vector and bias term is essential for minimising the hinge loss function.

The equation wx+b=0 characterises the separating hyperplane. The sign of the formula wx+b is the same for all data points on the same side of the separating hyperplane. The data points that lie along the segregating hyperplane are the support vectors.

In machine learning, the support vector machine method is useful for classification and regression. It's a flexible method that can solve linear and non-linear issues.

## VII. DESIGN OF COMPLEX SYSTEMS

Kaggle is an online competition for statistical modelling and data analysis. In addition, data miners from other fields have contributed datasets from various application areas. The best models for predicting and representing the data are developed by many data scientists in competition. It allows consumers to utilise their information to construct models and collaborate with other data scientists to address various actual data science concerns. The project's intended dataset has been obtained from Kaggle. The raw format of this data collection is available, however. A few firms' stock market statistics are included in the data set. This raw data must initially be transformed into processed data as the first stage. Since the received raw data has numerous characteristics, only a subset of those characteristics is useful for generating predictions, and feature extraction is used. The first step is feature extraction, in which critical properties from the raw dataset are chosen as candidates for inclusion in the final model. Using the raw state of the measured data as a starting point, feature extraction generates new values or features. These characteristics aim to teach and not repeat themselves, simplifying further stages of understanding and application. The feature extraction process is a kind of dimensionality reduction that takes a large number of raw variables and lowers them to a smaller set of characteristics that fully and correctly describe the original data set. Following the feature extraction step, a classification procedure divides the data gathered into two unique segments. Classification is difficult since it requires deciding which set of categories a new observation fits into. The model is taught using the training data, and its proficiency is evaluated using the test data. Compared to the training data, a smaller subset of the data is used for testing purposes. The random forest technique evaluates data using a collection of hypothetical decision trees. Put another way, a group of decision trees combs over the whole forest of decision trees, looking for a certain characteristic. This is known as "data splitting" in the industry. The ultimate goal of the proposed method is to predict the stock price by looking at historical data.

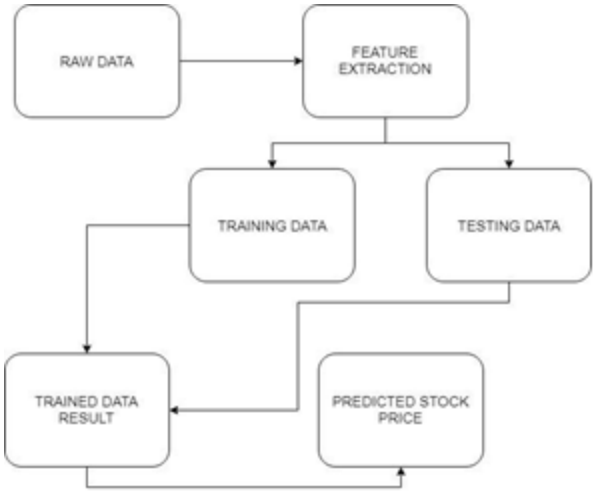



**Fig 1 DESIGN OF COMPLEX SYSTEMS**

## VIII. MODULE IDENTIFICATION

The project's numerous components would be split up into the pieces above.

### I. Data Collection

The project's first stage is data collecting, a highly fundamental module. The primary goal is to compile a useful data collection. Several criteria must be used to filter the dataset that will be used for market prediction. The data collected may be further enhanced by including new external information. Most of what we have in stock values from the previous calendar year. We will first analyse the Kaggle dataset to assess the forecasts thoroughly. We will employ the model with the data if it has a high enough degree of accuracy.

### II. Pre-Processing

As a part of data mining, data pre-processing involves transforming raw data into a more manageable format. Raw data always has many discrepancies, inconsistencies, and missing information. Data pre-processing includes steps like identifying category values, segmenting the dataset into training and test sets, and feature scaling to normalise the range of variables.

### III. Training the Machine

Training a machine entails altering the test data, like training an algorithm entails providing the algorithm with test data. The models are fine-tuned and fitted using the data from the training sets. Since a model shouldn't be judged on unobserved data, the test sets haven't been changed. Cross-validation is used throughout the training process to provide a reliable estimate of the model's performance on the training set. Tuning models may be used to fine-tune hyperparameters like the tree size in a random forest. We do a full cross-validation cycle on each unique combination of hyperparameter values. The next step is calculating a cross-validated score for each combination of hyperparameters. The optimal hyperparameters are then selected. Starting with values extracted from the dataset, we optimise the model's target parameters during training. This process is repeated

until the desired parameters are met. So, by combining the test dataset's inputs with the training model's predictions. This causes data partitioning, with 80% going to the training set and 20% to the testing set.

## IV. Data Scoring

Data is "scored" when a prediction model is applied to it. The Random Forest Algorithm was used to analyse the information gathered here. Random forest is a popular method for classification and regression because it uses an ensemble approach. Based on the learning models, we get some fascinating results. In the last unit, you'll learn how to utilise the model's predictions to assess a stock's upside and downside potential. The vulnerabilities of a company or stock are also exposed. Only authorised parties can see the results after implementing the user authentication system control.

## IX. FINDINGS FROM EXPERIMENTS

Our research relies on the raw data included in the xlxs file. Eleven columns, or characteristics represent the rise and fall of stock prices. One of these features is a stock price (1) HIGH, or the highest price it reached in the preceding calendar year. (2) LOW is similar to the stock's yearly low, while HIGH is similar to its annual high. The stock's value at the opening of trading is denoted by (3) OPENP, while (4) CLOSEP denotes its value at the close of trading. We also consider YCP, LTP, TRADE, VOLUME, and VALUE; nonetheless, those above four are crucial to our findings.

| DATE | TRADING CODE | LTP | HIGH | LOW | OPENP | CLOSEP | YCP | TRADE | VALUE (mi | VOLUME |
|---|---|---|---|---|---|---|---|---|---|---|
| 28-12-2017 | 1JANATAMF | 6.4 | 6.5 | 6.4 | 6.4 | 6.4 | 6.5 | 79 | 1.888 | 2,94,720 |
| 27-12-2017 | 1JANATAMF | 6.5 | 6.5 | 6.4 | 6.5 | 6.5 | 6.5 | 73 | 1.295 | 2,00,062 |
| 26-12-2017 | 1JANATAMF | 6.5 | 6.6 | 6.4 | 6.5 | 6.5 | 6.5 | 103 | 4.119 | 6,30,548 |
| 24-12-2017 | 1JANATAMF | 6.6 | 6.6 | 6.4 | 6.5 | 6.5 | 6.5 | 46 | 0.654 | 1,01,104 |
| 21-12-2017 | 1JANATAMF | 6.6 | 6.6 | 6.4 | 6.4 | 6.5 | 6.4 | 24 | 0.241 | 37,098 |
| 20-12-2017 | 1JANATAMF | 6.4 | 6.5 | 6.4 | 6.4 | 6.4 | 6.4 | 37 | 0.296 | 45,885 |
| 19-12-2017 | 1JANATAMF | 6.4 | 6.6 | 6.4 | 6.5 | 6.4 | 6.5 | 55 | 1.387 | 2,16,529 |
| 18-12-2017 | 1JANATAMF | 6.4 | 6.5 | 6.4 | 6.4 | 6.5 | 6.4 | 36 | 0.141 | 21,817 |
| 17-12-2017 | 1JANATAMF | 6.5 | 6.5 | 6.4 | 6.5 | 6.4 | 6.6 | 118 | 2.904 | 4,52,125 |
| 14-12-2017 | 1JANATAMF | 6.5 | 6.6 | 6.5 | 6.6 | 6.6 | 6.6 | 36 | 0.596 | 90,597 |

**Fig 2 Raw Data**

The information included in our xlxs file is graphically represented here. There are 121608 duplicates in this file. More than ten different trade codes may be found in the dataset, and some entries lack data that might be utilised to train the machine. Processing the raw data makes sense as a consequence. Consequently, we can produce an accurate dataset and use it to train the computer.

| | DATE | TRADING CODE | LTP | HIGH | LOW | OPENP | CLOSEP | YCP | TRADE | VALUE (mn) | VOLUME |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2018-08-16 | 1JANATAMF | 6.2 | 6.3 | 6.1 | 6.2 | 6.2 | 6.2 | 56 | 0.757 | 122741 |
| 1 | 2018-08-16 | 1STPRIMFMF | 11.2 | 11.2 | 10.9 | 11.0 | 11.1 | 10.9 | 145 | 2.640 | 238810 |
| 2 | 2018-08-16 | AAMRANET | 80.1 | 80.4 | 78.5 | 78.5 | 79.7 | 78.3 | 545 | 15.488 | 195035 |
| 3 | 2018-08-16 | AAMRATECH | 30.8 | 31.6 | 30.7 | 31.0 | 30.9 | 31.0 | 195 | 5.100 | 164899 |
| 4 | 2018-08-16 | ABB1STMF | 6.1 | 6.1 | 5.9 | 6.0 | 6.1 | 6.0 | 109 | 11.214 | 1857588 |

**Fig 3 head()**

The outcome of utilising the head is this (). Due to our use of the pandas package, only the top five rows will be returned. Unless otherwise provided, the default number of rows returned is five. The strip() method is used to remove the trade code from the processed data set, and the value "GP" is substituted for any trading codes that are no longer relevant.

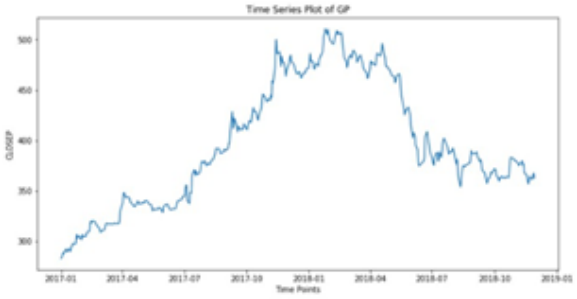



**Fig 4 Time series plot of GP**

The "matplotlib.pyplot" package was used to create the time series plot you see here. The narrative pits "CLOSEP" vs. "DATE" as the two qualities. This illustrates the trajectory

of the stock's closing price over a two-year period. The candle stick plot, shown in the illustration below, was produced using the "mpl finance" package.

**Fig 5 Chandelier diagram**

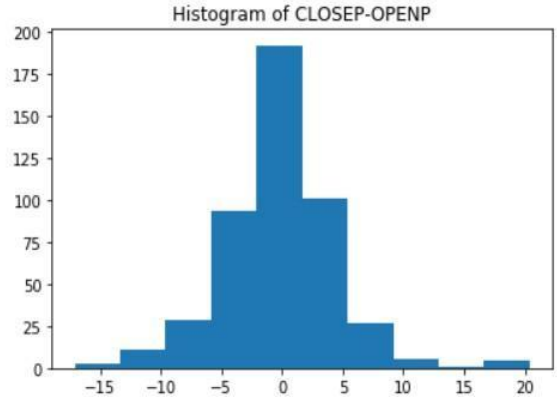



**Fig 6 Histogram of CLOSEP-OPENP**

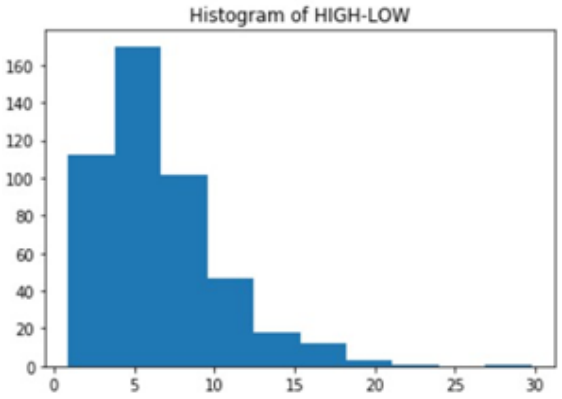



**Fig 7 High-low histogram**

Histograms of the relationship between "CLOSEP" and "OPENP" and "HIGH" and "LOW" attributes are shown in the two figures at the top. We're doing this because we expect today's high and low stock prices to play a role in determining the company's future value, along with the stock's opening and closing prices. Based on this reasoning, we determined that DEX should be set to 1 if the CLOSEP from today is greater than the CLOSEP from yesterday, and -1 otherwise. On this basis, the whole dataset is processed, and a preview of the collected data may be obtained through head().The following stage included establishing the feature and target variables as well as the train size. SVC classifier is imported and fitted with training data using the sklearn packages. The confusion matrix that was created is displayed below after the model had been trained using the data and tested using test data.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1.0 | 0.76 | 0.93 | 0.84 | 28 |
| 1.0 | 0.85 | 0.58 | 0.69 | 19 |
| micro avg | 0.79 | 0.79 | 0.79 | 47 |
| macro avg | 0.81 | 0.75 | 0.76 | 47 |
| weighted avg | 0.80 | 0.79 | 0.78 | 47 |

**Fig 9 Confusion Matrix**

Also, we train a different model using the same dataset. The ensemble technique's Random Forest Classifier is used in this model. Given that this is version 0.20, the default parameters for the decision trees leave the "n estimator" value at 10. However, in version 0.22, "n estimator" will have a value of 100. We discovered that this has an accuracy score of 0.808 after running the model on expected data and fitting the model with the data. Finally, we find that the random forest classifier has an accuracy score of 0.808, whereas the SVC Model

only achieves an accuracy score of 0.787 on the test set.

## IX. CONCLUSION

We found that the random forest algorithm is the most effective in predicting a stock's market price from a variety of historical data points, after comparing the accuracy of the other methods. The algorithm will be a great help to brokers and investors who are seeking to make stock market investments since it was selected after being reviewed on a sample of historical data and trained on a big collection of historical data. The results of the experiment demonstrate that a machine learning model can provide more accurate stock price predictions than previous machine learning models.

## X. FUTURE ENHANCEMENT

Financial ratios, a large number of instances, and other factors will be added to the project's scope in the near future. The more parameters you use, the more precise your results will be. The algorithms may also be used to analyse public remarks in search of patterns or connections between consumers and firm workers. Company-wide performance might potentially be predicted using standard algorithmic approaches and data mining techniques.

## Reference:

1. Agrawal, J. G., V. Chourasia, and A. Mittra. "State-of-the-art in stock prediction techniques." International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 2.4 (2013): 1360-1366.
2. Ou, Jane A., and Stephen H. Penman. "Financial statement analysis and the prediction of stock returns." Journal of accounting and economics 11.4 (1989): 295-329.
3. Rajkumar, V., and V. Maniraj. "HYBRID TRAFFIC ALLOCATION USING APPLICATION-AWARE ALLOCATION OF RESOURCES IN CELLULAR NETWORKS." Shodhsamhita (ISSN: 2277-7067) 12.8 (2021).
4. Nayak, Aparna, MM Manohara Pai, and Radhika M. Pai. "Prediction models for Indian stock market." Procedia Computer Science 89 (2016): 441-449.
5. Siew, Han Lock, and Md Jan Nordin. "Regression techniques for the prediction of stock price trend." 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE). IEEE, 2012.
6. Rajkumar, V., and V. Maniraj. "RL-ROUTING: A DEEP REINFORCEMENT LEARNING SDN ROUTING ALGORITHM." JOURNAL OF EDUCATION: RABINDRABHARATI UNIVERSITY (ISSN: 0972-7175) 24.12 (2021).
7. Kaboudan, Mark A. "Genetic programming prediction of stock prices." Computational Economics 16.3 (2000): 207-236.
8. Rajkumar, V., and V. Maniraj. "PRIVACY-PRESERVING COMPUTATION WITH AN EXTENDED FRAMEWORK AND FLEXIBLE ACCESS CONTROL." 湖南大学学报（自然科学版) 48.10 (2021).
9. de Oliveira, Fagner Andrade, et al. "The use of artificial neural networks in the analysis and prediction of

stock prices." 2011 IEEE international conference on systems, man, and cybernetics. IEEE, 2011.
10. Boyacioglu, Melek Acar, and Derya Avci. "An adaptive network-based fuzzy inference system (ANFIS) for the prediction of stock market return: the case of the Istanbul stock exchange." Expert Systems with Applications 37.12 (2010): 7908-7912.
11. Rajkumar, V., and V. Maniraj. "Software-Defined Networking's Study with Impact on Network Security." Design Engineering (ISSN: 0011-9342) 8 (2021).
12. Gandhmal, Dattatray P., and K. Kumar. "Systematic analysis and review of stock market prediction techniques." Computer Science Review 34 (2019): 100190.
13. Ambika, G., and P. Srivaramangai. "REVIEW ON SECURITY IN THE INTERNET OF THINGS." International Journal of Advanced Research in Computer Science 9.1 (2018).
14. Adriaans, Pieter. Data mining. Pearson Education India, 1996.
15. Rajkumar, V., and V. Maniraj. "HCCLBA: Hop-By-Hop Consumption Conscious Load Balancing Architecture Using Programmable Data Planes." Webology (ISSN: 1735-188X) 18.2 (2021).
16. Ambika, G., and P. Srivaramangai. "A study on data security in Internet of Things." Int. J. Comput. Trends Technol. 5.2 (2017): 464-469.
17. Chen, Ming-Syan, Jiawei Han, and Philip S. Yu. "Data mining: an overview from a database perspective." IEEE Transactions on Knowledge and data Engineering 8.6 (1996): 866-883.
18. Seifert, Jeffrey W. "Data mining: An overview." National security issues (2004): 201-217.
19. Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349.6245 (2015): 255-260.
20. Rajkumar, V., and V. Maniraj. "Dependency Aware Caching (Dac) For Software Defined Networks." Webology (ISSN: 1735-188X) 18.5 (2021).
21. Carleo, Giuseppe, et al. "Machine learning and the physical sciences." Reviews of Modern Physics 91.4 (2019): 045002.
22. Ambika, G., and D. P. Srivaramangai. "A study on security in the Internet of Things." Int. J. Sci. Res. Comput. Sci. Eng. Inform. Technol 5.2 (2017): 12-21.
23. Liakos, Konstantinos G., et al. "Machine learning in agriculture: A review." Sensors 18.8 (2018): 2674.
24. Ambika, G., and P. Srivaramangai. "Encrypted Query Data Processing in Internet Of Things (IoTs): CryptDB and Trusted DB." (2018).
25. Ferryman, James, and Ali Shahrokni. "Pets2009: Dataset and challenge." 2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance. IEEE, 2009.
26. Kille, Benjamin, et al. "The plista dataset." Proceedings of the 2013 international news recommender systems workshop and challenge. 2013.