Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7922992

ANALYSIS AND PREDICTION FOR STOCK MARKET WITH PHARAMA SECTOR USING DATA MINING AND MACHINE LEARNING APPROACHES

[1] M. Vijayakanth and [2] V. Veeramanikandan
[1] Research Scholar, Dept. of Computer and Information Science, Annamalai University,
Annamalainagar – 608 002, Tamil Nadu, India
Email: vijayakanth82@gmail.com
[2]Assistant Professor, Dept. of Computer Science, Thiru Kolanjiappar Govt. Arts College,
Vridhachalam-606 001, Tamil Nadu, India
Email: klmvmani@gmail.com

Abstract

The unpredictable nature of stock markets makes it hard to accurately predict the future. Nevertheless, there are countless individuals trying to enhance their chances of making a profit from their investments by developing a range of different models and methodologies. Despite being theoretically sound, most models and techniques don't work well in the real world due to their low hit rate. One of the primary reasons being the volatile nature of markets. Therefore, the focus of current research in the stock forecasting area is to improve the accuracy of stock trading forecasts. This paper introduces a system that addresses the particular need in the field of pharmaceutical stock market analysis using data mining and machine learning approaches. This system incorporates a range of data mining processes and helps to make informed decisions when it comes to pharmaceutical stock trades. This new system incorporates several machine learning algorithms, including the Gaussian Process, Linear Regression Model, Random Forest, Random Tree and Reduced Error Pruning Tree approaches to accurately predict the Pharmaceutical Stock Market Analysis in India. Its accuracy parameters include the Correlation Coefficient, MAE, RMSE, RAE and RRSE.

1. Introduction and Literature Review

Forecasting stock investment return is an important financial issue that has been given a lot of attention [1]. In the last decade, a number of intelligent systems and hybrid models have been proposed for making trading decisions in an attempt to outperform the main market and be profitable in stock investment [2]. The nature of stock market prediction requires the combining of several computing techniques synergistically rather than exclusively [3]. It is essential to clarify as predicting the "stock market trend." In reality, it is impossible to predict the future absolute value of the stocks on a daily basis. However, based on the assumption that is largely supported by real case studies that with appropriate training over any (uptrend, down-trend, and flat) horizon one could have enough indicators to forecast the trend with significant accuracy. Future trends may be predicted to some extent based on some key indicators and past behaviors.

Forecasting requires the knowledge of the dominant market variables that "explain" stock market behavior which is both dynamic and volatile. Due to system uncertainties and other unknown (random) factors, every stock market model is approximate. Thus, once model uncertainty is acknowledged, soft computing techniques emerge as the best candidates chosen over standard benchmark linear models to deal with such problems [4]. This paper presents a system that incorporates the top-down trading theory first introduced by [5] and various data mining techniques. Livermore believed that stock trends follow a trend line that can be used to forecast both in the long- and short-term. He published this particular idea in "How to Trade in Stock" in 1940.

Financial markets are one of the most fascinating inventions of our time. They have had a significant impact on many areas like business, education, jobs, technology and thus on the economy [6]. Over the years, investors and researchers have been interested in developing and testing models of stock price behaviour [7]. However, analyzing stock market movements and price behaviours is extremely challenging because of the markets dynamic, nonlinear, nonstationary, nonparametric, noisy, and chaotic nature [8]. According to [9], stock markets are affected by many highly interrelated factors that include economic, political, psychological, and company-specific variables. Technical and fundamental analysis are the two main approaches to analyse the financial markets [10] and [11]. To invest in stocks and achieve high profits with low risks, investors have used these two major approaches to make decisions in financial markets [12].

Data mining as ‘‘data mining is a process of discovering or extracting interesting patterns, associations, changes, anomalies and significant structures from large amounts of data which is stored in multiple data sources such as file systems, databases, data warehouses or other information repositories [13]. First, we have employed all widely used data mining techniques in this area [14], [15], namely, Decision Tree (DT), Artificial Neural Network (ANN), Naïve Bayes (NB), Support Vector Machine (SVM), and Bayesian Belief Network (BNN) Classifier while developing the predictive models. This allows us to explore the most suitable data mining technique - leading to the best possible predictive model in a comprehensive dataset.

2. Backgrounds and Methodology

2.1 Gaussian process

In probability theory and statistics, a Gaussian process is a stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution, i.e. every finite linear combination of them is normally distributed. The distribution of a Gaussian process is the joint distribution of all those (infinitely many) random variables, and as such, it is a distribution over functions with a continuous domain, e.g. time or space.The concept of Gaussian processes is named after Carl Friedrich Gauss because it is based on the notion of the Gaussian distribution (normal distribution). Gaussian processes can be seen as an infinite-dimensional generalization of multivariate normal distributions. Gaussian processes are useful in statistical modelling, benefiting from properties inherited from the normal distribution. While exact models often scale poorly as the amount of data increases, multiple approximation methods have been developed which often retain good accuracy while drastically reducing computation time.

To make choosing an appropriate noise level easier, this implementation applies normalization/standardization to the target attribute as well as the other attributes (if normalization/standardization is turned on). Missing values are replaced by the global mean/mode. Nominal attributes are converted to binary ones. Note that kernel caching is turned off if the kernel used implements cached kernel.

2.2 Linear Regression Model

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

$$y = \theta 1 + \theta 2_x \qquad \ldots(1)$$

While training the model we are given: x: input training data (univariate – one input variable(parameter)) y: labels to data (Supervised learning) When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values. θ1: intercept θ2: coefficient of x Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2.3 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

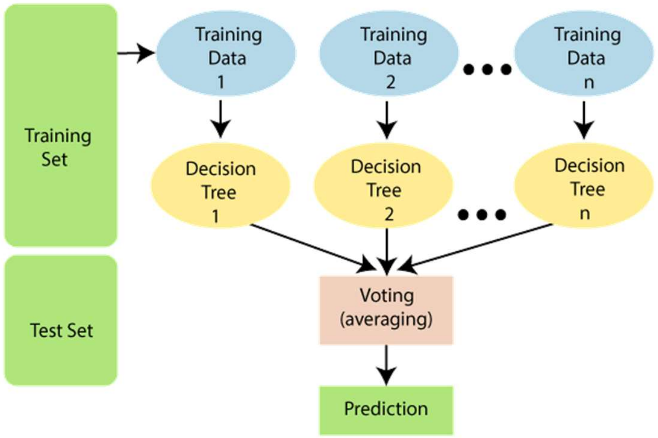

Fig. 1. Training and Testing approaches using Random Forest Tree

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase [18]. The Working process can be explained in the below steps and diagram:

Select random K data points from the training set.

i. Build the decision trees associated with the selected data points (Subsets).
ii. Choose the number N for decision trees that you want to build.
iii. Repeat Step 1 & 2.
iv. For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

2.4 Random Trees classifier

A single decision tree is easy to conceptualize but will typically suffer from high variance, which makes them not competitive in terms of accuracy.One way to overcome this limitation is to produce many variants of a single decision tree by selecting every time a different subset of the same training set in the context of randomization-based ensemble methods [19][24]. Random Forest Trees (RFT) is a machine learning algorithm based on decision trees. Random Trees (RT) belong to a class of machine learning algorithms which does ensemble classification. The term ensemble implies a method which makes predictions by averaging over the predictions of several independent base models[25][26].

The fundamental principle of ensemble methods based on randomization *"is to introduce random perturbations into the learning procedure in order to produce several different models from a single learning set L and then to combine the predictions of those models to form the prediction of the* ensemble*"* [20][23]. There is then a need to define stropping criteria to stop the growing of a tree before it reaches too many levels to prevent overfitting: "Stopping criteria are defined in terms of user defined hyper-parameters" [20]. Among those parameters, the most common are:[21][30]

- The minimum number of samples in a terminal node to allow it to split
- The minimum number of samples in a leaf node when the terminal node is split
- The maximum tree depth, that is, the maximum number of levels a tree can grow
- Once the Trees accuracy (defined by the Gini Impurity index) is less than a fixed threshold

2.5 Reduced Error Pruning Tree

RepTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree, the measure used is the mean square error on the predictions made by the tree. Basically, Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion and prunes it using reduced error pruning[31][32][33]. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances. The example of REP Tree algorithm is applied on UCI repository and the confusion matrix is generated for class gender having six possible values. [22][28][29]

2.6 Test Statistics

The authors discuss that the investigation based on recent hybridization of STDL and OKELM using short- and medium-term prediction for everyday sharing likes close price of the CRUDE OIL index. The parameter study of ELM done by utilizing the Gray Wolf Optimization Algorithm (GWO) to the predictive performance. The benefits in the intended work are done through the benefit of two related quantities called MASE and SMAPE. [23] and [24].

The coefficient of determination denoted R2 or r2 score which is used to moderation in the dependent variable means predicted from the independent variables. In this case, the r (CC) returns nearly 1.0 means strong positive correlation. If the value of r returns nearly -1 means strong negative correlation and return 0 means no correlation between all the variables.

$$\boldsymbol{r} = \frac{\boldsymbol{n(\sum xy)-(\sum x)(\sum y)}}{\sqrt{[\boldsymbol{n\sum x^2-(\sum x)^2}]-[\boldsymbol{n\sum y^2-(\sum y)^2}]}} \quad \ldots$$

(2)

In machine learning approach, MAE means the average of absolute error in future prediction which means error range between prediction and observations. In data mining and ML research, the MAE denoted as loss function. The given accuracy formula is:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad \ldots (3)$$

where n is called the no. of elements in the iterations, Σ which is used to add them all up and $|y_i - x_i|$ called absolute error between actual and predicted.

The RMSE is one of the familiar accuracy finding methods in data analysis, which is used to compute test the quality of prediction or forecasting. RMSE sometime name as root mean square deviation which is used to find the residuals between prediction and truth for all data points. The RMSE calculated using the following formula.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \|y(i) - \hat{y}(i)\|^2}{n}} \quad \ldots (4)$$

where *n* is called the number of elements in the iterations, *y(i)* means *i*th measurement, and *y ̂(i)* called the prediction.

The RAE is used to compute the accuracy for relatively comparison of each and every performance of a predictive model. The main reasons for calculating the RAE between actual and forecasted value. RAE is very useful to write the interpretation of the prediction which means if the RAE <1 means the model behavior is better. If RAE=0, the model behavior or accuracy is perfect.

$$RMSE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|} \qquad \ldots (5)$$

where *n* is called the number of elements in the observations, *y(i)* called the realized value and *y ̂(i)* called the prediction and $\bar{y}$ means the mean values of corresponding variables.

RRSE is one of the accuracy metrics for predictive models called regression. It's an accuracy parameter which is used to compute the first result and behavior of model is performing. It is also an inheritance from RSE. The RRSE parameter for finding the process of square root for sum of squared errors for the corresponding predictive model with sum of squared errors.

$$RRSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \ldots (6)$$

where *n* is called the number of elements in the observations, *y(i)* called the realized value and *y ̂(i)* called the prediction and $\bar{y}$ means the mean values of corresponding variables.

3. Result and Discussions

The secondary dataset (table 1) is taken from https://in.investing.com/. The benchmark open-source dataset namely pharma, which includes five parameters namely date, open, high, low, and close. The dataset collection between 31.01.2011 to 12.08.2022 which include 2860 record. The dataset collection downloaded using [16].

Table 1: Pharama Stock Market Dataset

Date	Open	High	Low	Close
12-Aug-22	13016.5	13017.5	12830.45	12864
11-Aug-22	12998.4	13033.75	12961.4	13014.2
10-Aug-22	12986.25	13053.1	12915.5	12950.7
8-Aug-22	12952.05	12962.95	12878.1	12940.25
5-Aug-22	13003.85	13068.3	12920	12939.75
4-Aug-22	12720.05	12992.25	12681.3	12982.35
3-Aug-22	12789.05	12829.1	12618.2	12682
2-Aug-22	12765.5	12792.95	12699.95	12780.4
1-Aug-22	12835.4	12883.75	12707.7	12766.85
29-Jul-22	12677.75	12835.45	12526.3	12776.55
28-Jul-22	12665.6	12682.2	12500.2	12660.65
27-Jul-22	12341.65	12647.75	12301.85	12628.95
26-Jul-22	12495.65	12508.8	12325	12341.9
25-Jul-22	12612.35	12640.8	12454.25	12488.8
22-Jul-22	12687.8	12711.15	12573.9	12617.55
21-Jul-22	12710.1	12756.35	12599.5	12651.75
20-Jul-22	12786.4	12786.4	12685.6	12711.5
19-Jul-22	12680.4	12820.25	12663.15	12682.2
18-Jul-22	12827.35	12869.4	12668.45	12713.45

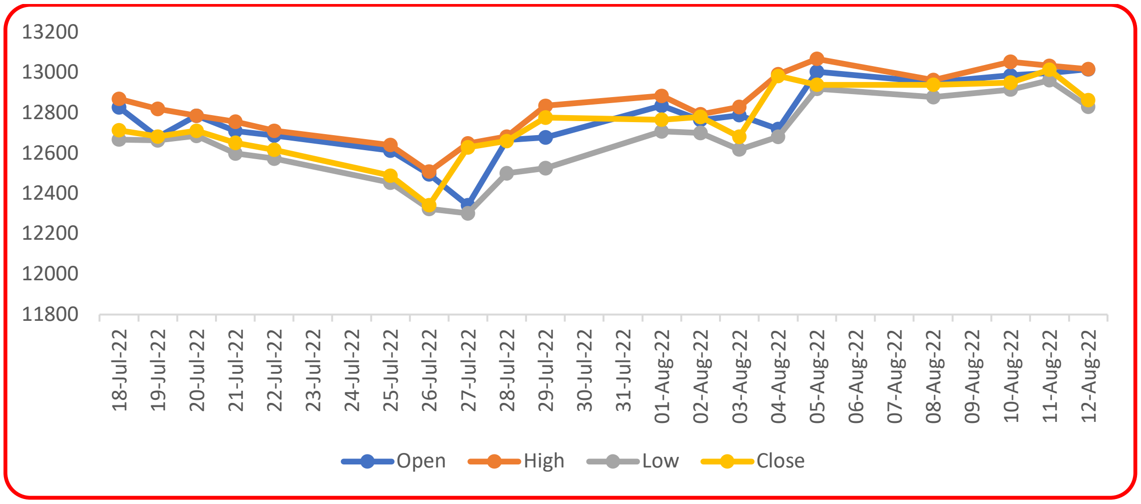

Fig. 2: Parameters Trends in Pharama Stock Market Dataset

Table 2: Data Analysis using Gaussian Process

Attributes	**Correlation Coefficient**	**MAE**	**RMSE**	**RAE (%)**	**RRSE (%)**	**Average Target Value**	**Time Taken (sec)**
Open	0.9988	2006.6901	2464.7687	84.4502	86.9484	0.4790	46.26
High	0.9970	2015.6814	2479.0366	84.3413	86.8978	0.4785	49.09
Low	0.9965	1938.1927	2435.3487	84.4111	86.8962	0.4792	47.03
Close	0.9976	2000.3868	2457.6856	84.4774	86.9661	0.4796	50.84

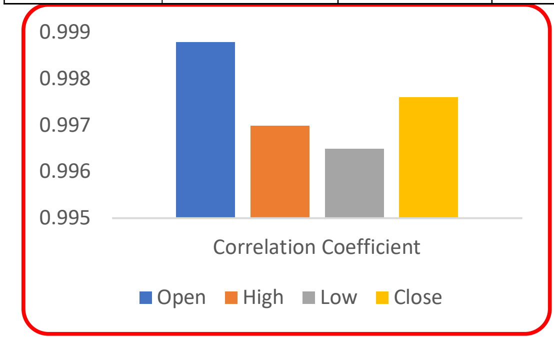

Fig. 3. Gaussian Process and their Correlations

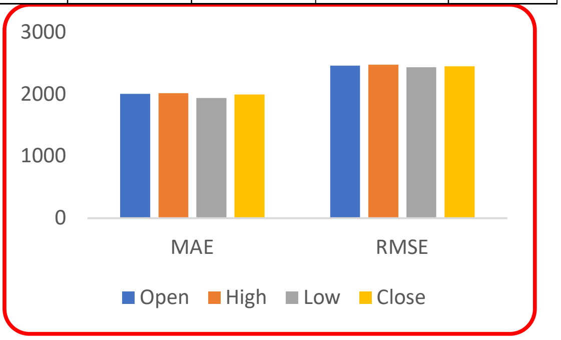

Fig. 4. Gaussian Process and their MAE and RMSE

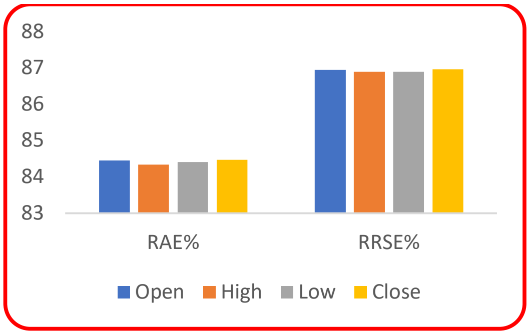

Fig. 5. Gaussian Process and their RAE and RRSE

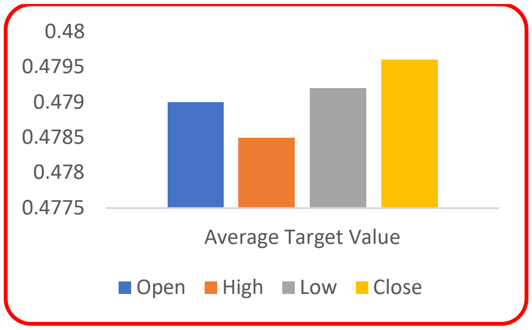

Fig. 6. Gaussian Process and their targets

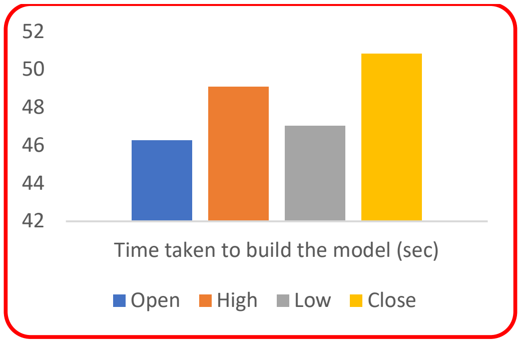

Fig. 6. Gaussian Process and time taken to build the model

The pharmaceutical stock trades are a major contribution of the share market. In this case the prediction about their finding also very difficult to predict the forecasting using open, high, low and close. The regression model is very useful to researchers for predict the future trends particularly in the field of share market prediction. The regression model is very useful and simple for predict the future using equation (1). The following prediction model in equations 6 to 9 which is used to find all the four parameters prediction using the regression model.

$$\text{Open} = 0.9437 * \text{High} + 7663 * \text{Low} - 0.7097 * \text{Close} - 12.6696 \quad \ldots (6)$$
$$\text{High} = 0.7062 * \text{Open} - 0.4507 * \text{Low} + 0.7481 * \text{Close} + 7.8652 \quad \ldots (7)$$
$$\text{Low} = 0.7163 * \text{Open} + 0.563 * \text{High} + 0.8415 * \text{Close} + 12.5609 \quad \ldots (8)$$
$$\text{Close} = -0.5963 * \text{Open} + 0.8399 * \text{High} + 0.7564 * \text{Low} - 7.0572 \quad \ldots (9)$$

Table 3: Data Analysis using Regression Model

Attributes	**Correlation Coefficient**	**MAE**	**RMSE**	**RAE (%)**	**RRSE (%)**	**Time taken (sec)**
Open	0.9998	34.6144	49.3777	1.4567	1.7419	0.05
High	0.9999	29.2756	42.6467	1.225	1.4949	0.03
Low	0.9999	31.0444	47.7248	1.3213	1.7029	0.02

Close	0.9999	31.3081	45.3347	1.3222	1.6042	0.01

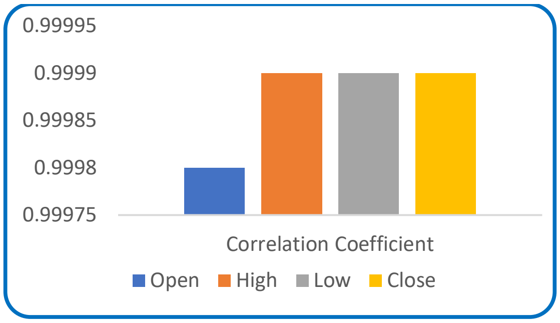

Fig. 7. Linear Regression and their Correlations

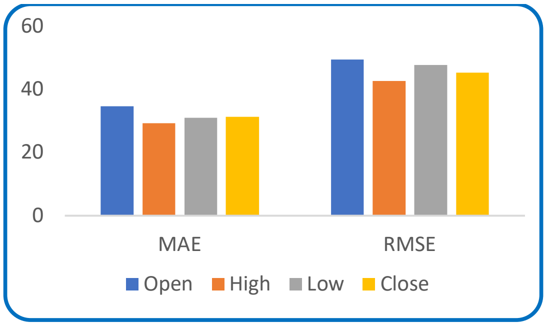

Fig. 8. Linear Regression and their MAE and RMSE

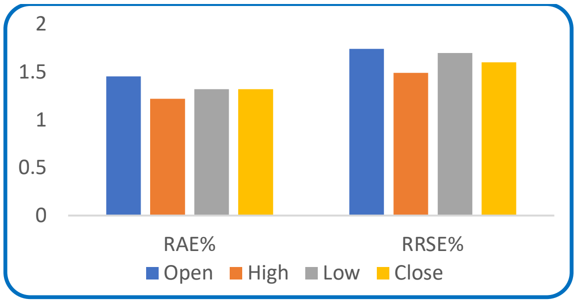

Fig. 10. Linear Regression and their RAE and RRSE

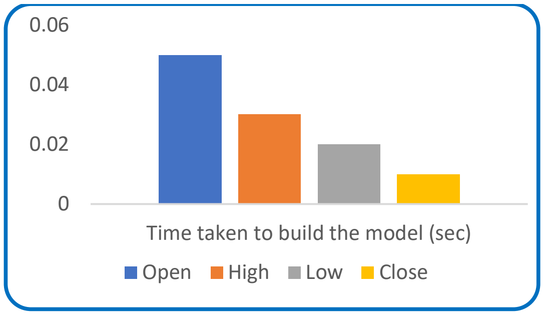

Fig. 11. Linear Regression and their time taken to build the model

Table 4: Data analysis using Random Forest Algorithm

Attributes	Correlation Coefficient	MAE	RMSE	RAE (%)	RRSE (%)	Time taken
Open	0.9997	54.3371	73.1883	2.2867	2.5818	0.34
High	0.9998	44.1529	62.0809	1.8475	2.1761	0.31
Low	0.9997	45.3806	66.2734	1.9315	2.3647	0.33
Close	0.9997	50.332	67.8736	2.1255	2.4017	0.31

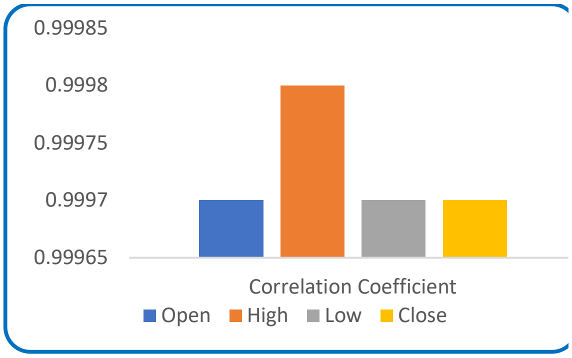

Fig. 12. Linear Regression with Correlations

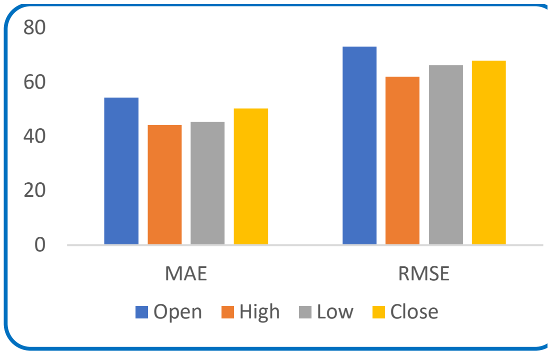

Fig. 13. Linear Regression with MAE and RMSE

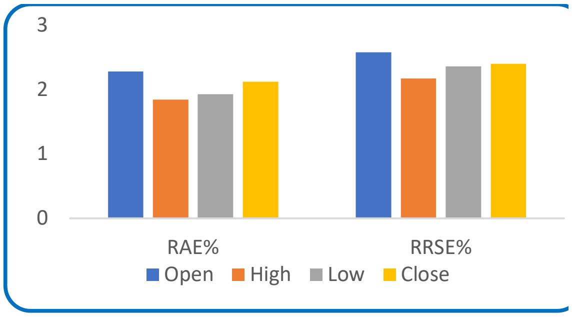

Fig. 14. Linear Regression with RAE and RRSE

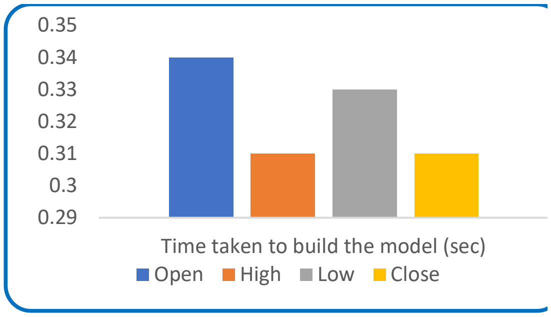

Fig. 15. Linear Regression with time taken to build the model

Table 5: Data analysis using Random Tree

Attributes	**Correlation Coefficient**	**MAE**	**RMSE**	**RAE (%)**	**RRSE (%)**	**Size of the Tree**	**Time taken**
Open	0.9994	73.3315	96.4396	3.0861	3.4021	181	0.1
High	0.9995	67.8683	89.0781	2.8398	3.1225	155	0.1
Low	0.9994	68.8307	95.2355	2.9297	3.3981	155	0.1
Close	0.9995	67.9881	90.8197	2.8712	3.2137	157	0.2

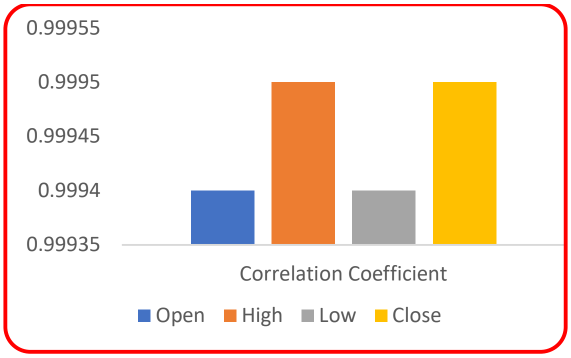

Fig. 16. Random Tree with Correlations

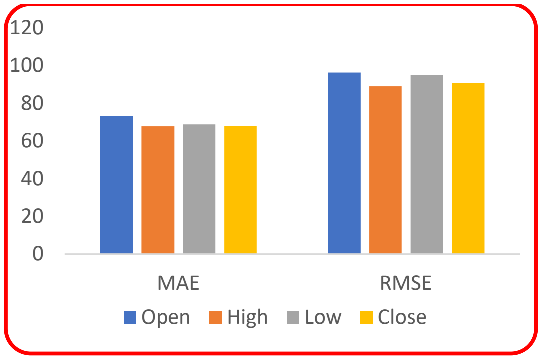

Fig. 17. Random Tree with MAE and RMSE

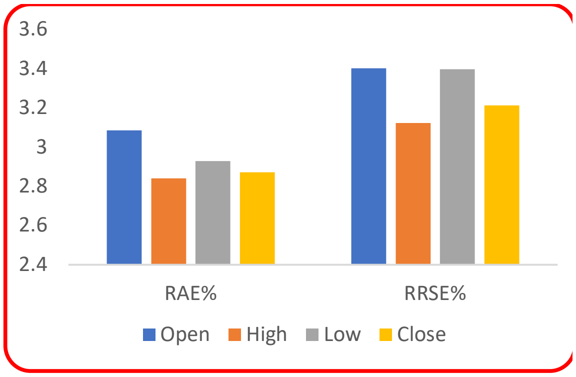

Fig. 18. Random Tree with RAE and RRSE

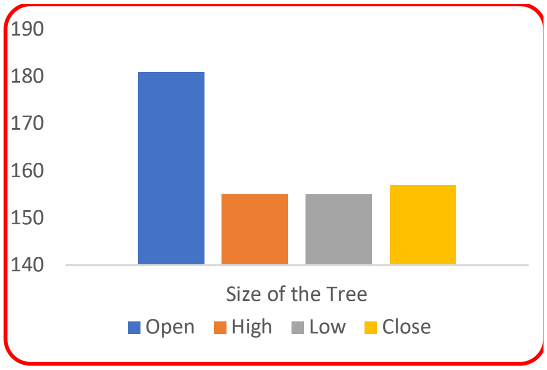

Fig. 19. Random Tree with Size of the Tree

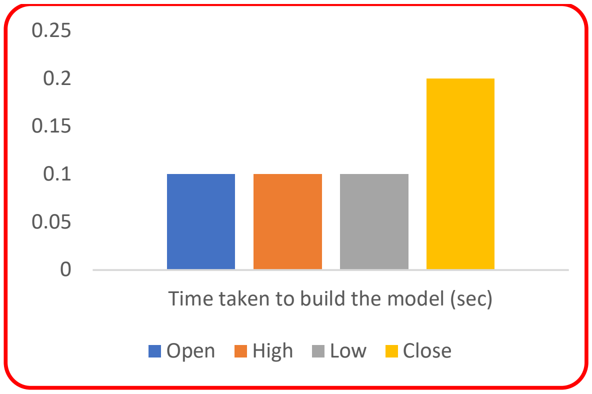

Fig. 20. Random Tree with Size of the Tree

Table 6: Data analysis using REP tree

Attributes	Correlation Coefficient	MAE	RMSE	RAE (%)	RRSE (%)	Size of the Tree	Time taken

Open	0.9994	73.3867	97.8187	3.0884	3.4507	123	0.08
High	0.9995	71.8535	94.3616	3.0065	3.3077	117	0.02
Low	0.9994	70.2800	94.2763	2.9913	3.3639	121	0.02
Close	0.9994	71.8381	94.6849	3.0338	3.3505	111	0.01

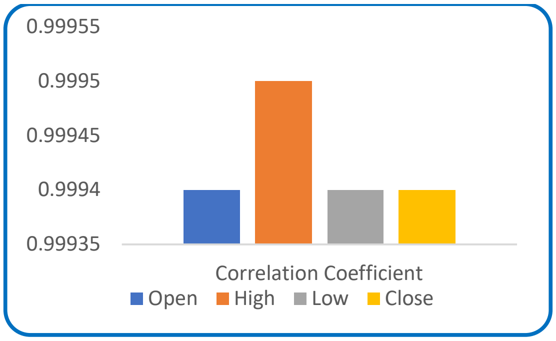

Fig. 21. REP with Correlations

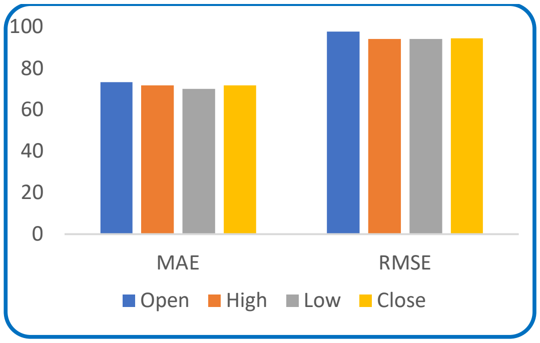

Fig. 22. REP Tree with MAE and RMSE

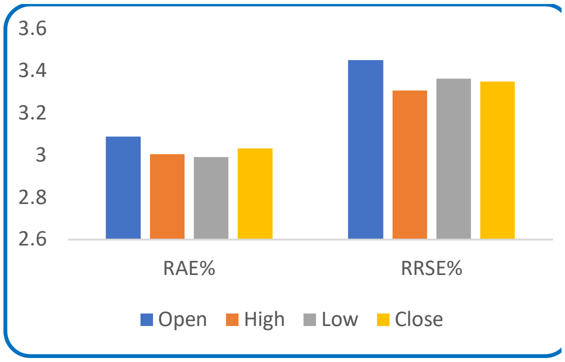

Fig. 23. REP Tree with RAE and RRSE

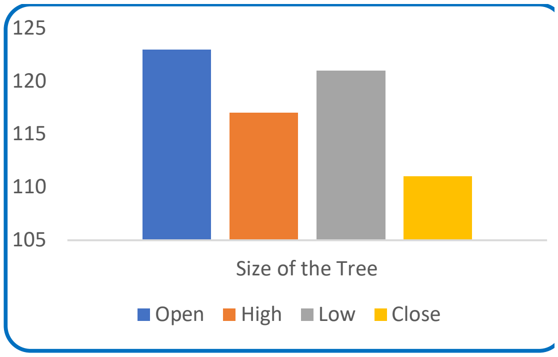

Fig. 24. REP Tree with Size of the Tree

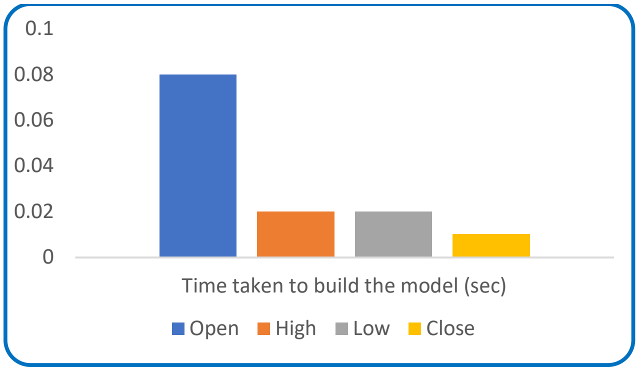

Fig. 25. REP Tree with Size of the Tree

Figure 1 graphically shows the training and testing approaches for decision tree approaches. Table 1 and Figure 2 called the primary dataset which is downloaded using the open-source machine learning data warehouse namely Kaggle. In this table indicates the fluctuation in pharama stock market with five different parameters namely date, open, high, low and close share values. According to table 2, figure 2 to figure 6, explain various data analysis using gaussian process. In this case, finding the correlations between four different numerical parameters like open, high, low, and close using gaussian process, the given algorithm returns the open parameters is 0.9988, which is technically called strong positive correlation and also time taken to build the model also low (46.26 seconds).

The regression model is very useful to researchers for predict the future trends particularly in the field of share market prediction. In this data analysis, the all the four-prediction parameter returns strong positive correlation nearly 0.9999. Time taken to build the model between 0.01 to 0.05 seconds. In this case, the four different test statistics namely MAE, RMSE, RAE, and RRSE returns less error compared to other algorithms. The related results and discussions shown in table 3, figure 7 to figure 11.

Results analysis of all the two decision tree methods namely random forest and random tree classifiers returns the strong positive correlations between four different parameters. In this case the random forest and their CC for all the four parameters returns nearly 0.9999 and time taken to build the model (0.31 to 0.34 seconds). The random forest result and discussion shown in table 4, figure 12 to figure 15. Decision tree approaches namely random tree shown in table 5, the correlations for four parameters return strong positive correlation nearly 0.9995. The time taken to build the model also very low compared to other decision tree approaches. In this algorithm return four different errors namely MAE, RMSE, RAE and RRSE the related results and discussion based on random tree shown in table 5, figure 16 to figure 20. Decision tree approaches namely REP tree shown in table 6, the correlations for four parameters return strong positive correlation nearly 0.9995. The time taken to build the model also very low using close parameter compared to other three parameters. In this algorithm return four different errors namely MAE, RMSE, RAE and RRSE, the related results and discussion based on REP tree shown in table 6, figure 21 to figure 25.

4. Conclusion and future work

This research presents a framework for stock market future trends prediction using Pharma sector. We examined the effect of machine learning approaches on stock prediction in future. By including four different attributes, we found that compare to banking sector the Pharma sector having strong positive correlation for all the parameters combinations and all decision tree approaches. We also concluded that by combining various machine learning approaches and its test statistics also prove the results and discussions. For future study, the use of a more machine learning and stochastic model for determining other related area like oil and gas, software predictions with its accuracy compare their effects on the stock market prediction.

Reference

1. Matias, J.M. and Reboredo, J.C., 2012. Forecasting performance of nonlinear models for intraday stock returns. *Journal of Forecasting*, *31*(2), pp.172-188.
2. Keller, W.J. and Keuning, J.W., 2016. Protective asset allocation (PAA): a simple momentum-based alternative for term deposits. *Available at SSRN 2759734*.

3. Jang, J.S.R., Sun, C.T. and Mizutani, E., 1997. Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]. *IEEE Transactions on automatic control*, *42*(10), pp.1482-1484.
4. Atsalakis, G.S., Dimitrakakis, E.M. and Zopounidis, C.D., 2011. Elliott Wave Theory and neuro-fuzzy systems, in stock market prediction: The WASP system. *Expert Systems with Applications*, *38*(8), pp.9196-9206.
5. Leung, M.T., Daouk, H. and Chen, A.S., 2000. Forecasting stock indices: a comparison of classification and level estimation models. *International Journal of forecasting*, *16*(2), pp.173-190.
6. Hiransha, M., Gopalakrishnan, E.A., Menon, V.K. and Soman, K.P., 2018. NSE stock market prediction using deep-learning models. *Procedia computer science*, *132*, pp.1351-1362.
7. Fama, E.F., 1995. Random walks in stock market prices. *Financial analysts journal*, *51*(1), pp.75-80.
8. Abu-Mostafa, Y.S. and Atiya, A.F., 1996. Introduction to financial forecasting. *Applied intelligence*, *6*(3), pp.205-213.
9. Zhong, X. and Enke, D., 2017. Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, *67*, pp.126-139.
10. Park, C.H. and Irwin, S.H., 2007. What do we know about the profitability of technical analysis?. *Journal of Economic surveys*, *21*(4), pp.786-826.
11. Nguyen, T.H., Shirai, K. and Velcin, J., 2015. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), pp.9603-9611.
12. Arévalo, R., García, J., Guijarro, F. and Peris, A., 2017. A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting. *Expert Systems with Applications*, *81*, pp.177-192.
13. Han, J. and Kamber, M., 2006. Data mining: concepts and techniques, 2nd. *University of Illinois at Urbana Champaign: Morgan Kaufmann*.
14. Ngai, E.W., Hu, Y., Wong, Y.H., Chen, Y. and Sun, X., 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, *50*(3), pp.559-569.
15. Albashrawi, M., 2016. Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015. *Journal of Data Science*, *14*(3), pp.553-569.
16. https://in.investing.com/
17. Deringer, V.L., Bartók, A.P., Bernstein, N., Wilkins, D.M., Ceriotti, M. and Csányi, G., 2021. Gaussian process regression for materials and molecules. *Chemical Reviews*, *121*(16), pp.10073-10141.
18. Nimje, S., Mayya, R., Baig, M.N.A. and Jawale, S., 2020, April. Prediction on stocks using data mining. In Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST).
19. Breiman, Leo. 2001. *Random Forests*. Machine Learning. Vol-45, p.5-32.
20. Louppe, Gilles. 2014. *Understanding Random Forests, From Theory to Practice*. University of Liège. Faculty of Applied Sciences. Department of Electrical Engineering and Computer Science. 223 pages.

21. Natarajan, V. Anantha, et al. "Prediction Of Soil Ph From Remote Sensing Data Using Gradient Boosted Regression Analysis." Journal of Pharmaceutical Negative Results (2022): 29-36.
22. Kumar, M. Sunil, et al. "Deep Convolution Neural Network Based solution for Detecting Plant Diseases." Journal of Pharmaceutical Negative Results (2022): 464-471.
23. Ganesh, D., et al. "Implementation of AI Pop Bots and its allied Applications for Designing Efficient Curriculum in Early Childhood Education." International Journal of Early Childhood 14.03: 2022.
24. Kumar, M. Sunil, et al. "APPLYING THE MODULAR ENCRYPTION STANDARD TO MOBILE CLOUD COMPUTING TO IMPROVE THE SAFETY OF HEALTH DATA." Journal of Pharmaceutical Negative Results (2022): 1911-1917.
25. Prasad, Tvs Gowtham, et al. "Cnn Based Pathway Control To Prevent Covid Spread Using Face Mask And Body Temperature Detection." Journal of Pharmaceutical Negative Results (2022): 1374-1381.1911-1917.
26. P. Sai Kiran. "Power aware virtual machine placement in IaaS cloud using discrete firefly algorithm." Applied Nanoscience (2022): 1-9.
27. Malchi, Sunil Kumar, et al. "A trust-based fuzzy neural network for smart data fusion in internet of things." Computers & Electrical Engineering 89 (2021): 106901.
28. Sangamithra, B., P. Neelima, and M. Sunil Kumar. "A memetic algorithm for multi objective vehicle routing problem with time windows." 2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE). IEEE, 2017.
29. Sunil Kumar, M., and A. Rama Mohan Reddy. "An Efficient Approach for Evolution of Functional Requirements to Improve the Quality of Software Architecture." Artificial Intelligence and Evolutionary Computations in Engineering Systems. Springer, New Delhi, 2016. 775-792.
30. https://catalyst.earth/catalyst-system-files/help/concepts/focus_c/oa_classif_intro_rt.html
31. Kalmegh, S., 2015. Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. *International Journal of Innovative Science, Engineering & Technology*, *2*(2), pp.438-446.
32. Veeramanikandan, V.; Jeyakarthic. M. (2020). Hybridization Of Stdl with Optimal Kernel Extreme Learning Machine (Okelm) Based Short Term Crude Oil Price Forecasting In Commodity Futures Market. International Journal of Scientific & Technology Research, 9(2), pp. 4029- 4036.
33. Jeyakarthic. M.; Veeramanikandan, V. (2020). Forecasting of commodity future index a hybrid regression model based on support vector machine and grey wolf optimization algorithm. International Journal of Innovative Technology and Exploring Engineering, 9(2), pp. 2856-2862.