International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

# PREDICTION STOCK PRICE BASED ON DATA SCIENCE TECHNIQUE.

**Rushendran.S , Sai Prakash.S , Rajadurai.A , Dr.Saravanan**

[1]*Department of Computer Science and Engineering, SRM Institute of Science & Technology, Vadapalani, Chennai, Tamilnadu, India*
[2] *Department of Computer Science and Engineering, SRM Institute of Science & Technology, Vadapalani, Chennai, Tamilnadu, India*
[3] *Department of Computer Science and Engineering, SRM Institute of Science & Technology, Vadapalani, Chennai, Tamilnadu, India*

## Abstract

Predicting how the stock market will perform is one of the most difficult tasks. It can be described as one of the most important processes to predict. This is a difficult task with a lot of unknown things. To avoid this problem in one of the most interesting (or potentially profitable) time-series data sets, machine learning techniques have been used. As a result, stock price forecasts have emerged as an important research topic. The aim is to anticipate greater accuracy of stock price estimation systems based on machine learning. Suggest a method based on machine learning to accurately predict the value of a stock price index using the results of a stock price index or the highest accuracy scenario by comparing surveillance to differentiate machine learning algorithms. In addition, the effectiveness of a few machine learning methods from the departmental traffic database provided for testing will be compared and discussed. A set of data with a test breakdown report, identify the confusion matrix, and separate the data from the essentials, and the result shows that the efficiency of the proposed machine learning algorithm method can be compared with the best accuracy with precision, memory, and F1 Score.

Keywords: - Data Editing, Upcoming Stock Returns, Machine Learning Tips.

## I. INTRODUCTION

Data science is a multidisciplinary field that uses scientific techniques, processes, algorithms, and systems to extract information and data from structured and unstructured data, and then use that.information and possible details of various areas of the application. Data science can be identified as a combination of mathematical, business information, tools, algorithms, and machine learning methods, all of which help to discover hidden data or patterns in raw data that can be used to make important business decisions.

Artificial intelligence (AI) is the replication of human intelligence in robots designed to think and act like humans. The term can also refer to any machine that displays the same human characteristics as learning and problem solving.

Advanced Internet search engines and recommendation programs are examples of AI applications (used by YouTube, Amazon and Netflix), Personal speech comprehension (e.g., Siri or Alexa), self-driving vehicles (e.g., Tesla), and to play at the highest level all strategic gaming systems are possible (such as chess and go), As robots grow in skill, tasks that require "ingenuity" are often removed from the concept of AI, something known as the result of AI. For example, although it has become a common strategy, visual character recognition is often extracted from so-called AI.

AI is important because it can provide organizations with information about their operations that they did not know before, and because, in some cases, AI can perform tasks better than humans. AI tools, especially when it comes to duplicate operations, focus on details such as reviewing large amounts of legal paperwork to ensure that important forums are properly filled, often performing tasks quickly and with minimal errors.

Machine learning is used to predict the future based on past data. Computer literacy (ML) is a form of artificial intelligence (AI) that allows computers to learn without being explicitly programmed. Machine learning is about developing computer software that can change when exposed to new data, as well as the basics of machine learning, such as the use of a simple machine learning algorithm in Python. The training and forecasting process uses special algorithms.

Editing is a method of supervised learning in machine learning and mathematics where a computer program learns from a given input and then uses this learning to separate new observations. This data collection may be class (for example, classification if a person is male or female or whether email is spam or not) or multiple categories. Speech recognition, handwriting attention, biometric identification, document fragmentation, and other classification challenges are examples.


**Figure 1: - Process of machine leaning**

Supervised Machine Learning Techniquesuse techniques such as decompression, classification, Decision Trees, and vector support equipment, among others. The data needed to train the supervised learning algorithm should already be labeled with the appropriate answers. Separation problems are a subset of supervised learning problems. The purpose of this challenge is to create a short model that can predict the value of the attribute depending on the attribute variable.

## II. RELATED WORKS

According to Lobna Nassar et al., The ARIMA model has a large total error rate (MAPE) error and is therefore less effective compared to conventional ML models. In addition, among the standard methods, Gradient Boosting (GB) is the best as it has the lowest MAPE error. Finally, in both FPs, the performance of the basic LSTM DL model surpasses all traditional ML models tested (Watermelon and Bok Choy). This is because the markets for these two FP are small, which leaves us with very similar data over time series. According to the combined measure, the most effective model is the integrated DL model, ATTCNN-LSTM, which surpasses ML and basic DL models with accuracy.

In this work, Xie Chen et al. discuss An in-depth hybrid fuzzy neural Hammerstein-Wiener (FNHW) model, proposed by Xie Chen et al. The fuzzy neuro-implication and speculation depend on a blurred legal framework established during training. Training data should be able to accurately reflect the behavior of the complete system. The test data, on the other hand, may change with the distribution of distribution across the time series domain. In addition, training data may be taken in a stable manner, but test data is constantly evolving and represents significant data changes under certain circumstances, such as a financial crisis.

Ruby Gupta and her colleagues described the analysis of Stock Twits data and the impact of emotions on stock price changes. They aim to improve their work in the following areas. To begin with, we use two types of emotions in this work: bullish (positive) and bearish (negative) (negative). Including a neutral attitude may reduce noise and may improve the accuracy of the work.

The stock market forecast has long been the focus of research in the field of big financial data, according to Jiannan Chen et al. Stock data is complex and indirect data, and stock prices fluctuate over time. This study provides a large financial data for the Stock Trend Prediction Algorithm based on a focus method based on stock data (STPA) features. To capture the long-term reliability of data over time, we use the Bidirectional Gated Recurrent Unit (BGRU) and monitoring method. The proposed reduction method is based on the attention-grabbing machine (STPA), and is divided into three levels. That is, there are three layers: a vector representation of the stock price conversion velocity, a BGRU feature extraction layer, and a stock price prediction monitoring layer.

According to Fitriyana et al., The price of a stock is an important factor in making a profit on a stock investment, and forecasts are often made by comparing the price of the stock with the volatility that affects it. The difficulty is that there are so many factors that can be used to anticipate stock prices that it is difficult for an investor to be able to decide which variables to use in predicting price movements. Key Component Analysis was employed as a way to reduce the size of this study to produce key components that affect stock prices without loss of information, and data from five firms was used.

## III. SYSTEM ANALYSIS

The stock market and its trends are volatile especially in the financial industry. According to a recent study, news media and social media analysis can have a significant impact on investors' perceptions of financial markets. As a result, the aim of this study was to investigate the link between news sentiment and stock market movements using data from various news agencies, business journals, and financial sites. Using previous knowledge about model design, this study provides the use of a more transparent Bayesian structural time (BST) model model that helps to better manage uncertainty than the autoregressive integrated moving average (ARIMA) and vector autoregression (VAR) model. ) method.

Using previous knowledge about model structure, this study provides the use of a more comprehensive Bayesian structural time (BST) model model that helps to better manage uncertainty than the autoregressive integrated moving average (ARIMA) and vector autoregression (VAR) method. This study is a special use of time series predictions in measuring the emotional effects of news on the stock market. We have included emotional estimates in the BST series model as the news captures ideas about the current market. The performance of the BST method is continuously improved using RNN (LSTM).

• They do not categorize stock prices using mechanical classification techniques and do not provide accurate results.

• It will not be able to better assess the familiarity of the stock price forecast data and get more accurate forecasting results as a result.

Multiple data sets from a variety of sources will be combined to form a standard data set, and then various machine learning methods will be used to reveal patterns and produce the most accurate findings.

The set of data collected to predict the data provided is divided into two parts: training and evaluation. Typically, 7: 3 scales are used to separate Training and Assessment sets. The Data Model, built using machine learning methods, is used in the Training set, and the test set prediction is based on the accuracy of the test result.

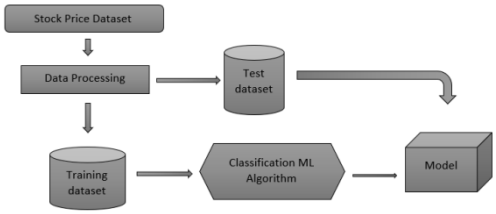



**Figure 2: Architecture of Proposed model**

These studies investigate the effectiveness of machine learning algorithms for predicting stock prices under operating conditions, and conclude with some assumptions about future research difficulties, challenges, and demands.

## IV. SYSTEM MODULES

- Data Processing
- Visual Data Analysis
- Comparing algorithm and predicting the outcome method for the best accuracy
- Distribution Using Flask

### Data processing

Machine learning verification methods are used to determine the error rate of the Machine Learning Model (ML), which may be considered as close to the actual data error rate as possible. If the data volume is large enough to represent the population, verification methods may not be required. However, in real-world situations, it is necessary to deal with data samples that may not be a real representation of the population of a particular database. To identify missing value, duplicate value, and definition of data type, whether a float variable or a total number. While modifying the hyper parameters of the model, a sample of data is used to provide an impartial test of model equity in the training database.

### Analysis of demonstration test data

In the calculations used and machine learning, data visibility is an important skill. Statistics, in fact, focus on the meanings of quantitative and quantitative data. Viewing data provides an important set of tools for gaining quality understanding. This can be useful when reading and knowing a set of data, as it can help identify trends, corrupt data, external factors, and other confusing issues. Data visualization can be used to transmit and display important links in structures and charts that are more visible and accurate than the relative dimensions or value of the sub-headline expertise. Data review and evaluation data analysis are self-contained topics, and will encourage the depth of some of the recommended literature in the conclusion.

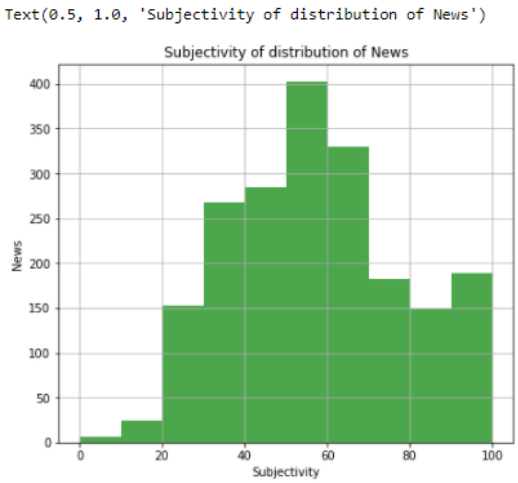



**Figure 3: - Data Visualization**

### Comparing algorithm and predicting the method of the result of the best accuracy

It is important to reliably evaluate the performance of different machine learning algorithms, and you will find out how to develop the test tool in Python using scikit-learn to compare learning algorithms for many different machines. This test tool can be used as a framework for your machine learning tasks, and you can add additional algorithms to others to compare them. Each model will have a set of performance indicators. You can check how accurate each model is on the invisible data using sampling methods such as reverse verification. It should be able to use these parameters to select

one or two of the best models from the list of models you have produced.

**In the example below compare 4 different algorithms:**

Logistic Regression

Random Forest

Decision Tree Classifier

Naive Bayes

**Shipping Using Flask**

Flask is an open source web framework. This means that Flash provides you with the tools, libraries, and technologies you need to create a web application. This online application can be as simple as a few web pages, a blog, or a wiki, or it can be as complex as a web-based calendar application or website.

• High compatibility with modern technology.

• Technical testing.

• Easy to use in most cases.

• Codebase is small in size.

• Excellent scalability for simple applications.

• Quick prototyping is easy.

• Moving URLs is easy.

• Applications are easy to create and manage.

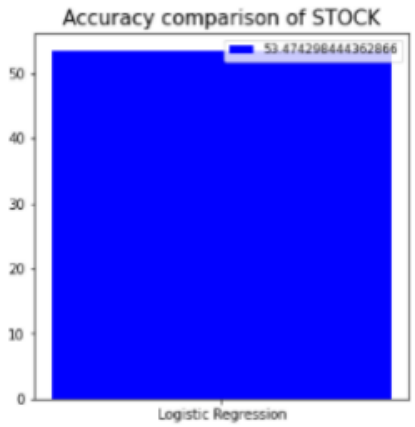
## V. ALGORITHM

Depression Logistic A statistical strategy for analyzing a set of data with one or more independent factors influencing an outcome. A dichotomous variable is used to test the outcome (where only two possible outcomes). The purpose of the retrospective is to obtain a model that better describes the relationship between the dichotomous element of interest (dependent variance = response or outcome variability) and a set of independent variables (predictor or descriptive). Systematic regression is a classification algorithm used in Machine Learning to measure the probability of phase-dependent variability. The variance that depends on a regression is a binary variable that contains coded data such as 1 (yes, success, etc.) or 0. (no, failure, etc.). Random Forest Classifier Random Forests, also known as deciduous forests, are an integrated learning method for dividing, descending, and other activities that work by creating a large number of deciduous trees during training and extracts classroom mode. (planning) or the average prediction (descent) of individual trees. Random tropical forests compensate for the sloping of the pruning trees to fit their training set. Random Forest is a machine learning program based on integrated learning. Shared learning is a form of learning where several types of algorithms or similar methods are combined several times to create a more efficient guessing model. Decision Tree Classifier is a very powerful and well-known algorithm. Decision tree algorithm is a type of supervised learning method. It works on both continuous output variables and phase. Decision tree speculation: Initially, we look at all the training set as root. For information, features are categorized; if not, the attributes are considered continuous. Records are still distributed over and over again based on the amount of responsibility. We use mathematical

methods to classify attributes as root or internal node. Naive Bayes Algorithm The Naive Bayes algorithm is a simple method that predicts the use of individual attributes that belong to each category. If you wanted to model the problem of predicting probability, you would use a supervised learning method. By assuming that the attributes of each attribute belonging to a given category value are independent of all other aspects, the absurd bayes make it easier to calculate the probability. This is a great idea, yet it presents a quick and effective way. Conditional opportunities are the category value options given the attribute value. Naive Bayes is a mathematical division based on the Bayes Theorem. It is one of the most sought-after

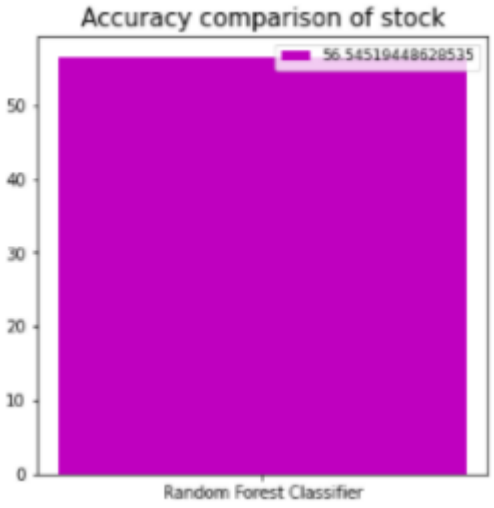
supervised learning methods

## VI. RESULT



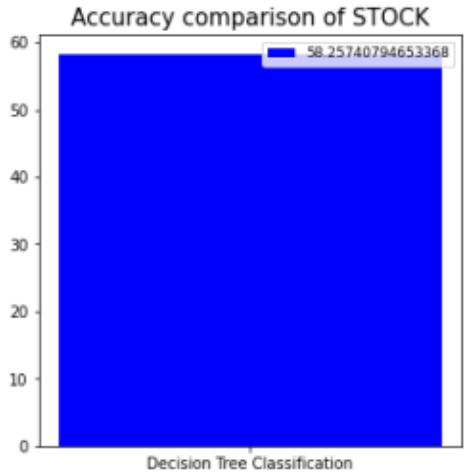

```
Accuracy result of Logistic Regression is: 54

Classification report of Logistic Regression Results:

              precision    recall  f1-score   support

           0       0.00      0.00      0.00       277
           1       0.54      1.00      0.70       319

    accuracy                           0.54       596
   macro avg       0.27      0.50      0.35       596
weighted avg       0.29      0.54      0.37       596


Confusion Matrix result of Logistic Regression is:
 [[  0 277]
 [  0 319]]

Sensitivity :  0.0

Specificity :  1.0

Cross validation test results of accuracy:
[0.53517588 0.53652393 0.53400504 0.53400504 0.53400504]

Accuracy result of Logistic Regression is: 53.474298444362866
Cross validation of Logistic Regression is: 0.535234899328859
```

Figure 4:- **Logistic Regression**




```
Accuracy result of Random Forest Classifier is: 67

Classification report of Random Forest Results:

              precision    recall  f1-score   support

           0       0.66      0.61      0.63       277
           1       0.68      0.73      0.70       319

    accuracy                           0.67       596
   macro avg       0.67      0.67      0.67       596
weighted avg       0.67      0.67      0.67       596


Confusion Matrix result of Random Forest Classifier is:
 [[168 109]
 [ 87 232]]

Sensitivity :  0.6864081048458483

Specificity :  0.7272727272727273

Cross validation test results of accuracy:
[0.57708945 0.52896725 0.5768262  0.58940622 0.5440806 ]

Cross validation of Random Forest Classifier is: 56.54519448628535
```

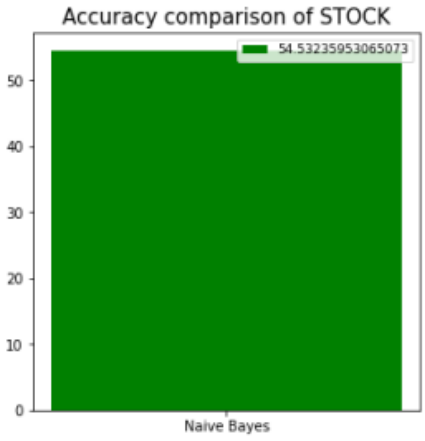
Figure 5: - **Random Forest Classifier**




```
Accuracy of  Decision Tree Classifier 63

Classification report of Decision Tree Results:

              precision    recall  f1-score   support

           0       0.60      0.63      0.61       277
           1       0.66      0.64      0.65       319

    accuracy                           0.63       596
   macro avg       0.63      0.63      0.63       596
weighted avg       0.63      0.63      0.63       596


Confusion Matrix result of Decision Tree Classifier is:
 [[174 103]
 [116 203]]

Sensitivity :  0.628158844765343

Specificity :  0.6363636363636364

Cross validation test results of accuracy:
[0.59045226 0.59697733 0.55667506 0.63224181 0.53652393]

Cross validation of Decision Tree Classifier is: 58.25740794653368
```

Figure 6: - **Decision Tree Classifier**

Figure 6: - **Decision Tree Classifier**




```
Accuracy result of Naive Bayes Algorithm is 56

Classification report of Naive Bayes Results:

              precision    recall  f1-score   support

           0       0.58      0.15      0.24       277
           1       0.55      0.91      0.69       319

    accuracy                           0.56       596
   macro avg       0.57      0.53      0.46       596
weighted avg       0.57      0.56      0.48       596


Confusion Matrix result of Naive Bayes is:
 [[ 42 235]
 [ 30 289]]

Sensitivity :  0.15162454873646208

Specificity :  0.9059561128526645

Cross validation test results of accuracy:
[0.53266332 0.53904282 0.5465995  0.55415617 0.55415617]

Cross validation of Naive Bayes Algorithm is: 54.53235953065073
```

Figure 7: - **Naive Bayes algorithm**

## VII. CONCLUSION

The analysis process began with the cleaning and processing of data, followed by deficit analysis, test analysis, and finally model creation and testing. The best accuracy in a public test with high accuracy scores will be obtained. This program can assist in determining stock prices.  Stock price forecast for cloud connection.  Expand your work in the field of Artificial Intelligence.

## VIII. REFERENCES

1] Wang, Y.F., (20 03) "Price of mining stock using incorrect set system", Expert Systems with Applications, 24, pp. 13-23. [2] Wu, M.C., Lin, S.Y., and Lin, C.H., (2006) “Effective use of the decision tree in stock trading”, Expert Systems with Applications, 31, pp. 270-274. [3] Al-Debie, M., Walker, M. (1999). "Analysis of Basic Information: Extensions and Evidence in the UK", Journal of Accounting Research, 31 (3), pages 261-280. [4] Lev, B., Thiagarajan, R. (1993). “Basic Information Analysis”, Statistical Research Journal, 31 (2), 190–215. [5] Tsang,

P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong, K., and Wong, T.L. (2007) “Designing and implementing NN5 Hong Kong Stock Predictability”, Artificial Intelligence Engineering Applications, 20, pp. 453-461. [6] Ritchie, J.C., (1996) Key Analysis: A Guide to Investing Back to Basics in Choosing Quality Stocks. Irwin Professional Publishing. [7] Murphy, J.J., (1999) Financial Markets Technical Analysis: A Comprehensive Guide to Trading Methods and Applications. New York Financial Institution. [8] Wang, Y.F., (2002) “Predicting stock prices using a dull gray forecasting system”, Expert Systems with Applications, 22, pp. 33-39. [9] Han, J., Kamber, M., Jian P. (2011). "Ideas and Strategies for Data Mining". San Francisco, CA: Morgan Kaufmann Publishers. [10] Enke, D., Thawornwong, S. (2005) "Use of mining data and neural networks for forecasting stock market returns", Expert Systems with Applications, 29, pp. 927-940