International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Predicting Stock Exchange using Supervised Learning Algorithms

Sikkisetti Jyothirmayee, V. Dilip Kumar, Ch. Someswara Rao, R.Shiva Shankar

Abstract: ***The stock market price trend is one of the brightest areas in the field of computer science, economics, finance, administration, etc. The stock market forecast is an attempt to determine the future value of the equity traded on a financial transaction with another financial system. The current work clearly describes the prediction of a stock using Machine Learning. The adoption of machine learning and artificial intelligence techniques to predict the prices of the stock is a growing trend. More and more researchers invest their time every day in coming up with ways to arrive at techniques that can further improve the accuracy of the stock prediction model. This paper is mainly concerned with the best model to predict the stock market value. During the mechanism of contemplating the various techniques and variables that can be taken into consideration, we discovered five models Which are based on supervised learning techniques i.e.., Support Vector Machine (SVM), Random Forest, K-Nearest Neighbor (KNN), Bernoulli Naïve Bayes.The empirical results show that SVC performs the best for large datasets and Random Forest, Naïve Bayes is the best for small datasets. The successful prediction for the stock will be a great asset for the stock market institutions and will provide real-life solutions to the problems that stock investors face.***

Keywords: Stock Market, Machine Learning, Dataset, Data pre-processing, Supervised Learning Algorithms,Predictions.

I. INTRODUCTION

The stock market is probably an accumulation of multiple individuals and businesses of stock. The stock (also widely known as shares) usually reflects a person or a specific group of people ownership claims on the company. The attempt to define the long term value of the stock market is addressed as a stock market prediction. The stock market plays a key role in the fast economic boom of developing countries like India. So our country's progress and other Enterprise's growth may depend upon attainment of stock market. If stock market increments, then countries' economic growth would be huge. If stock market decrements, then countries' economic growth would be narrow.

One of the finance potentials is buying specific stocks, bonds, or mutual funds from certain markets. But, how to be (almost) reasonably sure, that this investment will pay off? How to choose marketplaces and exact paper funds, and deal with a precise portfolio investment plan for the near or future? That's why both buyers and researchers are now focusing for some period on the analysis for stock exchange movements [1].

Stock price prediction is one of the most widely studied and challenging problems, attracting researchers from many fields including economics, history, finance, mathematics, and computer science. The volatile nature of the stock market makes it difficult to apply simple time-series or regression techniques. Financial institutions and traders have created various proprietary models to try and beat the market for themselves or their clients, but rarely has anyone achieved consistently higher-than-average returns on investment. Nevertheless, the challenge of stock forecasting is so appealing because an improvement of just a few percentage points can increase profit by millions of dollars for these institutions [8].

There are two common ways of predicting stock market behavior. The first one is based on the prediction of future price values of the stock. This approach usually requires treating the historical data as time-series data, feeding the distinct time frame signals to an algorithm and trying to model the future time points in the signal. The second is based on predicting the future price direction of a stock, i.e. guessing whether the price will rise or fall the next day, or in a couple of days (trend forecasting) [1].

Forecasting or predicting stock prices may be done in accordance with one or a combination of four approaches: fundamental analysis, technical analysis, time series analysis, and machine learning. Fundamental analysis: The fundamental analysts perform this method tends to focus more on the company than the actual stock. The executives decide on the basis of the company's previous results, income prediction, etc. Technical analysis: the Technical Analysts are responsible for calculating the investment cost on the basis of previous inventory patents (have been using time-series appraisal) [3]. Time-series forecasting has been used globally to determine future stock rates and financial time series analysis and modeling. this impersonates a fundamental role in guiding customers' choices and trade [3].

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Sikkisetti Jyothirmayee*, M. tech Student,Department of CSE, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram, AP, India. Email: sikkisetti.jyothirmayee@gmail.com.
V. Dilip Kumar, Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram,AP, India. Email: dilipv510 @gmail.com.
Ch. Someswara Rao, Assistant Professor of Computer Science and Engineering, SRKR Engineering Collegeaffiliated to JNTU Kakinada, Bhimavaram, AP,India. Email: chinta.someswararao@gmail.com.
R. Shiva Shankar , Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada, AP,Bhimavaram, India. Email: shiva.srkr@gmail.com.

Many models of predictions had already predominantly been focusing on linear time series statistical systems like ARIMA[7].

However, linear technology becomes sub-optimal in terms of the variance behind inventory and other property and non-linear designs such as ARCH prefer to have a smaller predictive mistake Researchers are now turning to Big Data in the areas of software research and machine learning for stock pricing. These are computed to develop mathematical and statistical theories [8].

The growing popularity of machine learning in several sectors has made many traders more aware of the adoption of machine learning techniques and many of these have managed to bring very promising results. Machine learning and artificial intelligence applications are progressively used to forecast stock exchanges. More and more scientists every day spend their money in developing methods to enhance the precision of the inventory forecast model more accurately. Due to the wide range of accessible alternatives, there can be nine ways to forecast the inventory cost, but not all techniques operate in the same manner. The output varies for each technique even if the same data set is being applied.

Machine Learning offers a broad variety of algorithms that are claimed to be predictive for future stock prices. Recently, quite enough appealing work was done when using machine learning computation to analyze price patterns and to predict stock prices and changes in indexes. Today, most stock traders rely on intelligent trading systems to assist them to predict rates depending on diverse demands and circumstances, assisting them to take immediate investment choices.

The stock markets are vivid worldwide, so the moment to purchase and transfer stocks is difficult to choose. Over the years, many techniques for predicting stock trends had been developed. Classical regression techniques were initially trained to accurately predict stock trends. Since stock data could be classified as non-stationary time series data, non-linear machine learning techniques were used too[2].

An intelligent trader would forecast stock returns and purchase stocks before prices rise or sell the shares before their value decreases. While it is difficult to supplement the knowledge acquired by an skilled trader, a precise predictor model can lead to a clear consequence that investors have great gain, suggesting that the forecast simulation is significantly related to the profit from its use.

When using Machine Learning on Stock Data, we would like to do a technical analysis in order to see if our algorithm can correctly understand the fundamental characteristics of the stock. Machine learning, however, can also serve an important part in assessing and forecasting the company's effectiveness and other comparable parameters that are useful in fundamental analysis. In fact, the most effective encrypted stock prediction and referral structures use some particular brand of a fusion analysis model involving both Fundamental and Technical Analysis.

Usually, the machine learning algorithms are of two classes , the first is supervised learning in which the training data is a sequence of labelled examples. Each sample is a collection of elements that are marked with the right value for the feature set. This implies that functions and inputs for a specific information collection will be provided to the machine. The results for certain other information sets which is the testing information are then predicted by what they learn from this specific dataset [8].

Unsupervised learning usually involves specific instances in which the feature set is unidentifiable or un labelled. The algorithms intend to group the data into communities. Supervised learning can be further divided into classification and regression problems. There are a number of results that can be labeled as a feature set in the classification, while in regression problems the output can take on constant values. [8].

The task focused in this work is to predict direction of movement for stocks and stock price indices of supervised algorithms. Prediction performance of four models namely, SVC, Random forest, KNN and Naive-Bayes is compared based on the dataset DJIA (Dow Jones Industrial Average). For analyzing the efficiency of the system we are used the Root Mean Square Error(RMSE) and rˆ2 score value. The historical data for dataset was downloaded from the Yahoo Finance. Then the results will be used to analyze the stock prices and their prediction in depth in future search efforts.

II. LITERATURE SURVEY

G. Zhang et al. [1] Portrayed a summary of the current status in implementations of neural networks for forecasting. Several researchers had already examined the artificial neural networks as models for predicting exchange rates and shown that neural networks could be one of the quite useful tools in foreign currency sectors data analysis. One significant use of ANNs is for time-series forecasting The prominence of ANNs is based upon the fact that they are simplified dynamical forecasting models. Linear statistical methods dominate forecasting since several centuries Although ANNs have the benefits of precise prediction, their results are incompatible in certain particular circumstances Several journals in the literature are about comparing ANNs with traditional exchange rate methods. Although several surveys show that ANNs are much faster than standard linear models and are far and continuously more precise in their prediction, some other surveys have shown incompatible outcomes. A number of potential applications suggest ANNs can be a promising alternative tool for researchers and professionals in forecasting.

K.kim et al. [2] represented the work as "Financial Time Series Forecasting Using Support Vector Machines". In this framework, SVM is used to forecast financial time series of equity markets. The analysis examines the effectiveness of the SVM application and compares it with background cellular systems and context-based reasoning to forecast the economic time series by using a risk function with uncertain mistakes, standardize it for the purpose of implementing the concept of organizational threat minimization. The test findings indicated that the SVM forecast assessment responded to this criterion's merit. In addition, this research found that SVM offers a successful option for prediction of economic time series.

T. Manojlović et al. [3]2015, employed the'Random-forest' computation to build the model used to forecast 5-days-ahead and10-days ahead bearings of the CROBEX record to select stocks. Using a random forest model, the suggested 5-day ahead and 10-day ahead designs are developed. The designs are based on the historic CROBEX inventory information and some businesses from multiple industries mentioned on the Zagreb Stock Exchange. Numerous technical indicators, prominent in quantitative research of stock markets, are chosen as model inputs. Their end results illustrate that random forests could be efficiently used for constructing computational models for predicting the course of equities trade patterns. Earlier work has shown successful outcomes on an opportunity to properly predict the inventory or market index value path.

Dai and Zhang et al. [4] (2013), initiated the work with 3 M Stock data as the training data used for their fieldwork. The data includes regular stock data from 1/9/2008to 11/8/2013 (1471 data points). The prediction system was trained using several algorithms. These algorithms include a logistic regression, a quadratic analysis of discrimination and an SVM. These algorithms were introduced to the next day model predicting the result of the stock cost on the preceding day large-term pattern predicting the inventory cost results on the next n days. The design of the subsequent day's forecast generated precision outcomes from 44.52 -58.2%. The findings were justified by Dai and Zhang (2013), which stated that the US equity is semi-strongly efficient, implying that no basic or technical assessment could be used to make inferior profits. The length-term forecast system generated stronger outcomes, however, which flourished when SVM showed a high precision of 79.3%.

Koosha Golmohammad**i**, Osmar R. Zaiane. et al. [5] University of Alberta .had extended the work of Diaz ***et. al***.Theinvestigators here adopted supervised machine learning algorithms in order to detect mistrustful bond market manipulation where machine learning algorithms can easily be used to detect market manipulation on the basis of a labeled, in-depth data connection to detect manipulated data. They submit a review and use a case study of fake stocks during 2003 of information mining methods to determining illicit operations and market manipulation, focused on supervised learning algorithms. They adopt CART, conditional inference trees, C5.0, Random Forest, Naïve Bayes, Neural Networks , SVM and KNN for classification of manipulated samples.

P. Hajek et al. [6] had proposed "Forecasting Stock Market Trend using Prototype Generation Classifiers "Where stock market planning is ongoing data mining research. In present work several prototype generation classifiers were used to identify the NASDAQ data trend and to review the NASDAQ Composite stock market index hit ratio for the selected methods by authenticating that the prototype generation classifiers outperform the vector and neural network endorse. Currently inventory cost trends are enforced by forecast techniques or pattern classifiers of time series. Having regard to the hit ratio of real expected pattern paths, Hajek thus validates that prototype classifiers outperform support vector machines and cellular networks. V.kranthi sai reddy et al. [7] refers to stock prediction as an attempt to identify the future value of an stock traded on the other financial instruments. In the approach, the stock forecast with machine learning is explained. Most stock brokers use the technical and essential analysis or time series analysis while making stock predictions. The programming language python is used to forecast Python's stock market. Proposed a machine learning approach (ML) to use data collected from various global financial markets to predict an stock index by means of machine-learning algorithms. The large dataset values gathered from various global financial markets are used for the SVM algorithm. Furthermore, SVM does not raise a suitable challenge. Different machine-based models for anticipating market stock trends are suggested. Numerical findings show elevated effectiveness. Our trained predictor was the basis of practical business models. In comparison to select criteria, the model produces greater profit.

Sahil madge et al. [8] Undertaken the major development on the SVM (Support Vector Machines) machine learning method. The model tries to predict that a stock price somewhere will be higher or lower in the future than it is at a given day and uses closing prices of 34 technology stocks for calculating price fluctuations and momentum of individual stocks and the sector to calculate these parameters. Into the short-run but drastic predictive ability in the long-run.

K. Hiba Sadia et al. [9] Developed a system to find the best classifier to predict the price of the stock market .During the process Of considering various techniques and variables that must be taken into account, Sadia et.al observed the models like Random Forest, Support Vector Machine were not employed fully. In, this manner they are going to present and review a more feasible method to forecast the stock movement with higher accuracy. The first step taken into account is the dataset of the stock market prices from previous year. The dataset was pre-processed and tuned up for real analysis. Hence, they focussed on data preprocessing of the raw dataset. Secondly, they have reviewed the use of the random forest, support vector machines as well as the results generated regarding pre-processing of the data. The suggested approach also examines the features of the forecast model in the practical world and problems related to the quality of the general statistics provided. A model of machine learning to require evidence stock persistence in a competitive market also introduces this current approach. Successful stock forecasting will be a major asset to stock market organizations and provide real-life alternatives to investor issues.

Shunrong Shen et al. [10] had proposed a new prediction algorithm that used the notion of temporal correlation among global markets and various important products to predict trend of next day stock. SVM was used as a classifier in this work. to maximize profit of stock purchased, minimizing risk & use of sentiment analysis for this purpose Shunrong Shen.et al discussed. Machine learning has been extensively used for prediction of financial markets.

Popular algorithms, such as support vector machine (SVM) and reinforcement learning have been quite effective in tracing the stock market and maximizing the profit of stock option purchase while keeping the risk low. It also includes sentiment analysis that take account of overall customers ' views as well as the worldwide inventory information to forecast the pattern for next day shares.

T.Shobana et al.[11] addressed the work as a completeness of methods and models used by many researchers in the application of data mining techniques for stock market prediction. By the use of this various available techniques, it is possible to create a new technique to predict the future trends in stock market. It is possible to utilize any of the discussed techniques and develop a hybrid system for the prediction of financial status of a company accurately. But it is important to design the system accordingly by which the accuracy and performance can be increased with less computational complexity. This survey will also help in choosing the best algorithm for time series and trend prediction method based on their prediction metrics.

J K. K. Patel et al. [12] mainly targets on task to predict the problem of direction of movement of stock and stock price index for Indian stock markets. Prediction performance of four models namely Artificial Neural Network (ANN), support vector machine (SVM), random forest and Naive-Bayes is compared based on ten years (2003–2012) of historical data of CNX Nifty, S&P BSE Sensex, Infosys Ltd. and Reliance Industries from Indian stock markets. Ten technical parameters reflecting the condition of stock and stock price index are used to learn each of these models. Trend Deterministic Data Preparation Layer is a distinct contribution to the research. This layer is employed to convert each of the technical indicator's continuous Process respectively. So this Layer is proposed in this work to exploit inherent opinion of each of the technical indicators about stock price movement prediction algorithms. the prediction algorithms are 'up' or 'down'. Multiple categories like 'highly possible to go up', 'highly possible to go down', 'less possible to go up', 'less possible to go down' and 'neutral signal' are worth exploring. This may give more accurate input to inference engine of an expert system.

Najeb Masoud et al. [13] Attempted to anticipate the direction in the emerging market of stock price movement such as the Libyan stock closing price levels by using daily data based ANN model to determine if LSM was predicted to be significant by Libyan learning techniques of the ANN model or not. The computational theory modeling instruments of an artificial neural network or neural network are lately made up of a variety of highly-connected computing systems known as artificial sensors or nodes. Each node conducts a straightforward procedure on an entry to produce an output forwarded to the next node. Firstly, it illustrates that by using the suggested model, how stock market price forecasts could be accomplished. In anticipating the path of market movement, this model of ANN indicated significant efficiency. This research has also specifies the importance of using twelve specific technical industry indices that have helpful outcomes for anticipating how inventory pricings work. ANN can thus be used as a stronger alternative technology to forecast regular stock market prices.

Y. Zuo et.al et al. [14] introduces the Bayesian network or the up / down interpretation of the stock return. The regular stock exchange prices are used as the network points in three main industries, and then the K2 algorithm is used to define the networks with the K2 criterion. The network is used to forecast up-and down analyzes of the FTSE-100 stock index in 2007, followed by the results of traditional trade policies such as psychological line-up and trend assessment.This thesis aims at predicting weekly inventory rates through the use of the Bayesian network. In order to define the random variables for the Network, daily stock prices were computerized by the classification algorithm. Because the precision of the prediction depends on the clusters and the data size, huge data are required for good prediction precision.

D. Diaz et al. [15] Addresses demands Depicted with data mining to detect the use of stock prices and extends previous results by the intra-day price manipulation fusion analysis. Precisely, this work complements prior results by analyzing intra-day businesses ' imperial evidence and hourly data within hours. The analytical model outlined in the present statistics also confirms the outcomes of past market manipulation research, depending on traditional statistical and econometric methods that provide an option range of data mining and knowledge discovery methods and techniques. The represented work also proposes a number of policy recommendations towards increasing the effectiveness of the operational processes executed by stock exchange fraud departments.

III. METHODOLOGY

The purposed method for developing the system consists of mainly three steps

Firstly, data is collected and sorted for relevance from various sources. Secondly, the information gathered are analyzed by examining the present path of the industry. Tracking the industry sector and particular business after which the data is depicted and evaluated properly, an appropriate algorithm is finally modelled with supervised learning algorithms that provides the greatest strength to estimate the quality of the stock.

The program is designed to examine the number of distinct forecasting techniques to predict future stock returns based on historical returns and numerical news indicators in order to develop a portfolio of several stocks to consolidate the risk by using supervised learning methods for stock value forecasting by performing curiously with unpredictable market data.

Forecasting is the method of defining potential outcomes on the principle of historical information and evaluating present information trends. Computer processing capabilities have become sufficiently strong to handle big amounts of information. By attempting to run statistical analyses of future states based on present states, we can predict the trend of stock market.

The discrepancy of the stock market is destructive and there are many complex financial indicators, However, digitizing advancements in technology, providing an incentive to gain gradual prosperity from the stock market and assisting professionals to find a most succinct predictor direction concerning that predicting the market value is of paramount concern to help maximize profit.

Predictive Algorithm for Stock Market Prediction:

Input: COMP, D_RANGE, O_PRED, H_PRED, LOW_PRED, C_PRED, V_PRED N_PRED // [DJIA company, date range,(open,high,low,volume,close,n-day)-predictions]

Output : A vector of predicted prices and graph, RESULTS data ← fetch stock for COMP in date range D_RANGE, ,O_PRED,H_PRED,LOW_PRED, V_PRED ,C_PRED

1. plot(data)
2. data ← DF(Date,Close) //Creating data frame with date and target variable
3. train_data ← RMSE(data) //perform R MSE INDICATOR on data
4. RESULTS ← 0 //set accuracy scores(precison,recall,f1) to zeroes
5. for each day in N_PRED ,C_PRED:
6. model ← SVM(train_data) //pass the training data to SVM
7. model ← RF(train_data) //pass the training data to Random Forest
8. model ← NB(train_data) //pass the training data to Naïve Bayes
9. model ← B_NB(train_data) //pass the training data to Bernouli Naïve Bayes
10. model ← knn(train_data) //pass the training data to k-nearest neighbour
11. model←SVM_RF_NB_BNB_KNN(test_data) //pass the test data to all applied models
12. pred ← predict(model,close) //predict the price given model and target
13. train_data ← add pred to train_data
14. test_data ← add pred to test_data
15. RESULTS ← add pred to RESULTS
16. end for
17. print(RESULTS)
18. plot(RESULTS)

A. DATA PREPROCESSING

Real-world data is usually inaccurate and noisy and will probably comprise information or mistakes that are meaningless and useless. Data pre-processing, a significant stage in data mining and machine learning techniques, helps to convert genuine information into a clear and understandable structure.

Data is collected from YAHOO Finance. Further, The data is preprocessed for modeling. When the anomalies are removed, the required prediction algorithm will be modeled. The result will be compared to determine which algorithm is better for prediction.

Fig: pre-processing

➢ **Root Mean Squared Error (RMSE)**

1. Root *Mean Squared Error (RMSE)*

The square root of the mean/average of the square of all of the error. The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions. Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Fig. RMSE Value Calculation

2. R-Squared Value(r^2 value)

The value of R2 can range between 0 and 1, and the higher its value the more accurate the regression model is as the more variability is explained by the linear regression model. R2 value indicates the proportionate amount of variation in the response variable explained by the independent variables.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

IV. DATA MODEL

Machine Learning is a realm of Computer Science that affords Computers the ability to process information. There are two primary classes of algorithms for machine learning. They are Supervised Learning and Unsupervised Learning. A machine learning model methodology includes generating an algorithm and results to enable a machine to learn it's constraints from the training data supplied.

➢ **Supervised learning**

In supervised learning, they train the machine learning algorithm with a collection of input instances and their corresponding labels. The model attempts to replicate as closely as necessary the feature y= f(x), wherex is the instance of the input and y is its label. As we use a training dataset with proper labels to instruct the algorithm, this is called supervised learning. Supervised training models, depending on the input variable, are further divided into regression and classification. If the output constraint is a continuous variable, it is called a regression process. Most applications of machine learning use supervised learning algorithms in reality. Support vector machines, Naïve bayes, k-nearest neighbors and random forests are examples of supervised learning algorithms.

➢ **Classification**

Classification is an application of supervised learning where a group is evaluated and classified depending on a prevalent function. Classification derives some abstract from the predicted result from the characteristics of information provided. If more than one entry is provided then classification will attempt to forecast one or more results for the same.

The random forest classifier, SVM classifier, k-nearest neighbor classifier, bernouli nb classifier are some of the classifiers used here for stock market prediction.

- **Random Forest Classifier**

Random forest classifier is a sort of ensemble classifier and also a supervised algorithm. It basically creates a set of decision trees, that yields some result. The basic approach of random class classifier is to take the decision aggregate of random subset decision tress and yield a final class or result based on the votes of the random subset of decision trees.

- **Random Forest Algorithm**

Random forest algorithm is often used for stock market prediction. Since it has been called one of the most user-friendly and versatile machine learning algorithm, it provides excellent forecast performance. Usually this is used in classification applications. The task of predicting is very challenging because of high volatility on the stock market. In Stock Market Prediction they were using a random forest classifier with the same multi parameters as a decision tree. The decision tool has a model similar to that of a tree. It takes the decision based on possible consequences, which includes variables like event outcome, resource cost, and utility. The random forest model depicts an algorithm from which it uniformly chooses distinct variables and features to construct several decision-trees and then gets results from the aggregate of all the other decision trees. The information is divided into partitions depending on label or domain considerations. The data frame we used was from past years stock markets gathered from the internet public database, 80 percent of information was used to train the device, and the remaining 20 percent were used to evaluate the inputs. This is the fundamental approach of the supervised learning.

- **SVM Classifier**

SVM was first created by (Vapnik). There are two primary classes for SVMs: Support Vector Classification (SVC) & Support Vector Regression (SVR) classifier. SVM is a sort of discriminative classifier. The SVM utilizes supervised learning, i.e. training data labeled. The result is hyperplanes that categorize the updated dataset. They are supervised learning techniques which are using complementary learning algorithms concerning Classification and Regression.

- **Support Vector Machine Algorithm**

In SVM, we plot each point in the dataset in a n-dimensional region where each aspect refers to a function. The data point value for that function will depict the positioning on the respective axis. SVM divides the collection of marked entry instances by an ideal hyperplane during training. During testing, the category of the unlabeled data point is determined by plotting it and testing which part the present point is on the hyperplane. A significant problem while using SVM is that, the input variables may lie in a very high dimensionality. Especially when feature dimensions range from hundreds to millions, training the model requires huge computation and memory.

- **k-nearest neighbor classifier**

The k-nearest neighbours classifier is indeed one of the brightest classifiers in machine learning. It is simply focused on the concept that "items that are' close' to each other will also have equivalent features. So if you understand the characteristics of one of the components, you can also forecast it for your closest neighbor. "k-NN is an progression over the closest neighboring approach. It is based on the assumption that every new instance could be classified by majority vote of its' k' nearest neighbours,here k is a +ve integer, probably a very small number.

- **K-NN Algorithm**

The current research tries to implement the k-NN algorithm to the objective of predicting and classifying stock index fluctuations. The output of k-NN is contrasted with the other models to estimate the effectiveness of the prediction and classification method. This model is introduced to the DJIA stock index to identify the stock index pattern that would demonstrate to shareholders whether stock indices are inclined to rise or fall in the up coming days.

The k-NN algorithm tests multiple' k' variables in the training data and discovers the optimum k value which generates the highest forecast outcome. Then this predictive model, with the optimum price of' k,' is introduced to the sample information frame to predict the ending price of the next day. The predictive model performance is summarized with the real sample dataset scores.

- **Naïve Bayes Classifier**

Naive Bayes Classifier is perhaps a simple straight forward statistic Bayesian classifier. It is called Naive as it implies that all factors lend to classification and are associated with each other. This premise is referred to as class conditional independence. It is sometimes termed as Idiot's Bayes, Bayes, and Independence Bayes. They can predict the high probability of class membership, such as the probability that somehow a given data item belongs to a particular class label. A Naive Bayes classifier believes that the existence (or lack) of a specific feature(attribute) of a category is equal to the existence (or lack) of any other function when the class variable is allocated.

- **Naïve Bayes Algorithm**

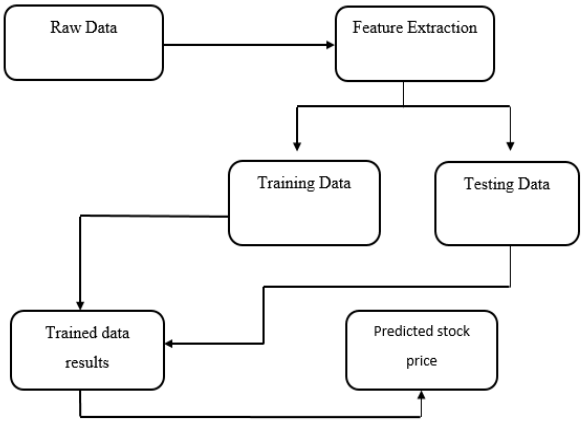
A Naïve Bayes model is a classification approach that produces Bayesian Networks for a known dataset relied on the Bayes theorem. It considers that the assigned dataset contains a specific function in a category that is trivial to any. For instance, because of some traits, the item is regarded to be A. These inclinations of characteristics may rely on each other or on other characteristics, but all the existence of the feature separately adds to the likelihood that this item is in. And that's why it's regarded as "Naïve." The advantages of the Naïve Bayes method are that it is simple to construct and helpful for every big dataset.

- **Bernouli-NB Classifier**

Bernouli-NB uses the simple Bayes training and classification algorithms for data processed as per the multivariate Bernoulli models; i.e. there may be numerous characteristics but each one is supposed to be a binary-valued (Bernoulli, Boolean) object. This category therefore needs samples to be displayed as binary-valued function vectors; if any other information is given over, a BernoulliNB instance may binarize its input (based on the binarize component).

V. SYSTEM ARCHITECTURE

System Architecture defines "the conceptual framework of the model and how the framework delivers conceptual integrity." Architecture is the hierarchical framework of program parts (parts), how these parts communicate, and the structure of information used by some parts.



VI. DATASET

The dataset used in the project namely DJIA STOCK NEWS which is downloaded from the kaggle . Kaggle is an online association for data analysis and forecasting patterns. It also contains dataset of different fields, which is assigned by data prospector. Several data scientists oppose to create the better models for predicting and detailing the information. It allows the users to use their datasets so that they can build models and work with various data science engineers to solve various real-life data science demands.

The initial step is to resolve this raw information into fixed information. This is achieved by using removal of features, as there are various characteristics in the raw information gathered, but only a few of those characteristics are helpful for forecast purposes. So the first stage is removal of the trait, where the main characteristics are obtained from the whole collection of characteristics accessible in the stored dataset. The extraction of features starts from the starting state of the empirical data and builds derived values or features. These characteristics are designed to be important and reiterating, supporting the following measures of training and generalization.

Here we have collected the historical stock market data for DJIA STOCK NEWS Historical data from the Reddit World News report websites and the dataset information is focused on the Yahoo Finance company, which is exclusively obtained from the internet.

➢ Attribute Description:

Dataset consists of 6 columns out of which 5 are features and one is target variable.

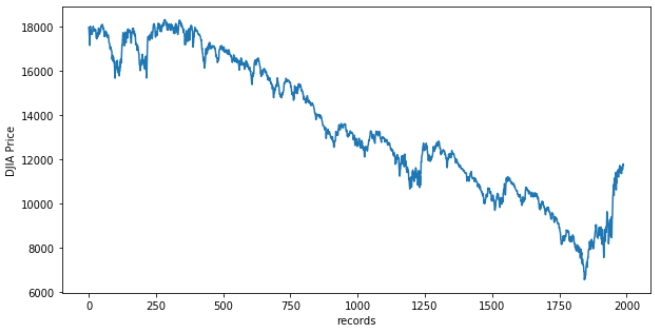
➢ Features

- Open – It is the starting cost of the stocks for that specific day.
- High – It is the amount in which the stock has passed throughout the day.
- Low – It is the smallest value in which the stocks have dropped throughout the day.
- Close – It is the finishing cost of the stocks of that dy
- Date – Date of request, mostly the sum of our information.
- Volume – Number of stocks purchased throughout the day.

➢ Target

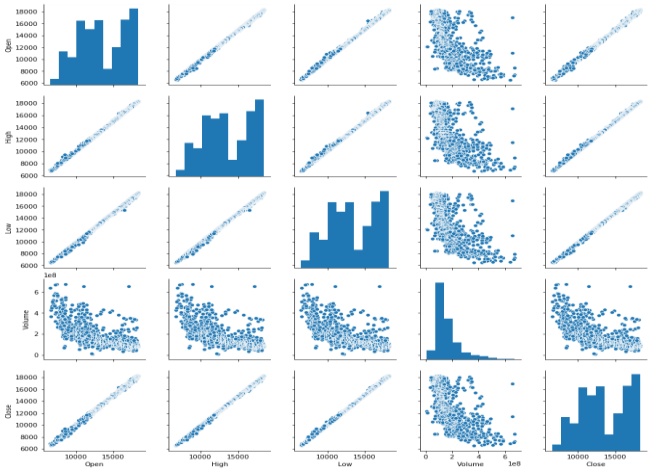
Stock Trading-The price(close) of complete business stocks for that day. Since our information has a' time' element, it is highly probable that our information is a time-series data in our data set, we have basically stock trading information. That indicates, our prediction is relying on the business practices of every day, stock situations. We can therefore strongly believe that our data will not show any pattern for more than a day. And it's also obvious from the data set that our observer frequency is one day. Therefore, we have chosen the frequency.

➢ Graph



From the above graph, we can see the individual behaviour of the feature with respect to time we can also have the correlation graph between every feature with each other and understand the inter dependent relation between every feature.

➢ Histogram



In the above diagram, the diagonal matrix used each column's histogram values which are not useful to us as we are interested in distributing features or targets (as no variables are high and are also co-dependent).

A graph is said to be correlated if it sounds like a curve (whether linear or non-linear).

➢ Result &analysis

The csv file contains the raw data based on which we are going to publish our findings. There are five columns or five attributes that describe the rise and fall in stock prices. Some of these attributes are (1) OPEN is the value of the stock at the very beginning of the trading day (2) HIGH, which describes the highest value the stock had in previous year. (2) LOW, is quite the contrary to HIGH and resembles the lowest value the stock had in previous year (3) (4) CLOSE stands for the price at which the stock is valued before the trading day closes. There are other attributes such as, VOLUME and Date, but the above mentioned four play a very crucial role in our findings.

	A	B	C	D	E
1	Open	High	Low	Volume	Close
2	17924.24023	18002.38086	17916.91016	82160000	17949.36914
3	17712.75977	17930.60938	17711.80078	133030000	17929.99023
4	17456.01953	17704.50977	17456.01953	106380000	17694.67969
5	17190.50977	17409.7207	17190.50977	112190000	17409.7207
6	17355.21094	17355.21094	17063.08008	138740000	17140.24023
7	17946.63086	17946.63086	17356.33984	239000000	17400.75
8	17844.10938	18011.07031	17844.10938	98070000	18011.07031
9	17832.66992	17920.16016	17770.35938	89440000	17780.83008
10	17827.33008	17877.83984	17799.80078	85130000	17829.73047

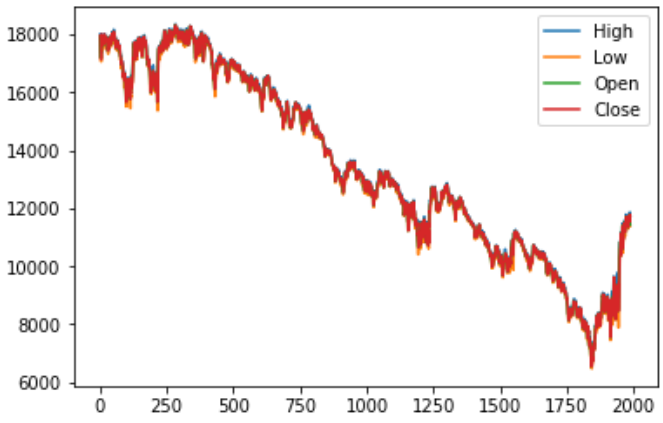
Fig:Raw data

This is a pictorial representation of the data present in our CSV(DJ1A_stock_news) file. This particular file contains 1990records. There are more than ten different trading codes available in the dataset and some of the records do not have relevant information that can help us train the machine, so the logical step is to process the raw data. Thus we obtain a more refined dataset which can now be used to train the machine.

```
print(df.head())
```

```
         Open         High          Low      Volume         Close
0  17924.24023  18002.38086  17916.91016  82160000  17949.36914
1  17712.75977  17930.60938  17711.80078  133030000  17929.99023
2  17456.01953  17704.50977  17456.01953  106380000  17694.67969
3  17190.50977  17409.72070  17190.50977  112190000  17409.72070
4  17355.21094  17355.21094  17063.08008  138740000  17140.24023
```

This is the result of using the head(). Since we are using the pandas library to analyse the data, it returns the first five rows. Here five is the default value of the number of rows it returns unless stated otherwise. The trading code in the processed data set is not relevant so we use the strip () to remove it and replace all of the trading codes with a value „GP



This is a time series plot generated from using the "matplotlib.pyplot" library. The plot is of the attributes "HIGH'',''LOW'',''OPEN'', "CLOSE".

➢ Performance Evaluation

For comparative study of the supervised learning algorithms for stock market prediction, accuracy , precision ,recall and F-measure are used.

➢ Accuracy

Accuracy is the most intuitive metric of results and is merely a proportion of accurately predicted assessment to complete results. One can believe that our model is the best if we are highly accurate. Yes, accuracy is a major step, but only if you have matched data sets that almost have the same values of true positive and true negative. So you need to examine other parameters in order to assess your model's performance. We have 0.803 for our model, implying that our model is approximately.80%accurate.

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

➢ Precision

Precision is the ratio of properly forecast positive findings to the total predicted positive findings. High precision refers to the low false positive rate.

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

➢ Recall (Sensitivity)

Recall is the ratio of correctly predicted positive observations to the all observations in actual class -

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

➢ F1 Score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

➢ Supervised Classification (Training Dataset)

The data has been divided into two parts i.e., training and testing data in the 80:20 ratios. Learning algorithms have been applied on the training data and based on the learning, predictions are made on the test data set.

➢ Supervised Classification (Test Dataset)

The test dataset is 20%of the total data. Supervised learning algorithms have been applied on the test data and the output obtained is compared with the actual output.

a) Comparison of individual performances of different classifiers with Training and Testing Data Set.

Fig. Train Data

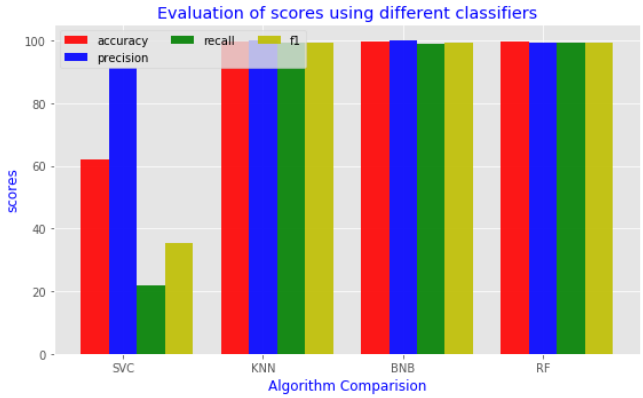
CLASSIFIER	ACCURACY	PRECISION	RECALL	F1
SVC	62.06%	58.36%	98.10%	73.18 %
KNN	98.49%	100.00%	94.29%	97.06 %
BERNOULI NB	99.50%	100.00%	98.94%	99.47 %
RANDOM FOREST	99.50%	99.47%	99.47%	99.47 %

Fig. Test Data

CLASSIFIER	ACCURACY	PRECISION	RECALL	F1
SVC	99.87%	99.88%	99.88%	99.88 %
KNN	98.55%	99.88%	99.88%	99.88 %
BERNOULI NB	98.11%	98.48%	97.47%	97.94 %
RANDOM FOREST	99.81%	99.86%	99.73%	99.80 %

Comparison of different algorithms for Train Dataset and Test Dataset has been shown in Figure 1 & Figure2. It verifies our result that Accuracy measure for SVC performs best for Train dataset. For Test dataset is shown in Figure 2. It verifies our result that Accuracy Measure for Random Forest ,Bernouli NB performs best.

➢ *Performance Results using different classifiers:*



From the above results this study has achieved a high accuracy equals 99.87% for svc These results will enable the decision makers and investors in the domain of stock market exchange to make a safe decision with low risk because these results depend on facts regarding the domain of stock market. Facts like the necessary to spatial and temporal features besides the role of the stock market experts in achieving real stock result. High accuracy with real stock prediction will lead to generating accurate and reliable reports and indicators on the company stock. Moreover, the simulation of assumptions in the domain of stock market decision making like the importance of timestamp and location features lead to add more reliability for the stock prediction result.

CONCLUSION

The stock market prediction process is filled with uncertainty and can be influenced by multiple factors. Therefore, the stock market plays an important role in business and finance. This paper proposed a technique to reveal the performance of a company (DJIA STOCK NEWS) by measuring the accuracy measures of the different algorithms, we found that the most suitable algorithm for predicting the market price of a stock based on various data points from the historical data is the SVM algorithm. Results have been evaluated where accuracies and F-measure values for each learning algorithm have been calculated and the results reveal that for large dataset SVC outperforms all the other algorithms in terms of accuracy and for small data set Bernouli, Naïve Bayes and random forest performs best.

REFERENCES

1. G. Zhang, B.E. Patuwo, M.Y. Hu*, Forecasting with Artificial Neural Networks: the state of the art, Int. J. Forecasting 14 (1998) 35-62.
2. K. Kim, Financial Time Series Forecasting Using Support Vcctor Machines. Neurocomputing, 55, pp. 307 319, 2003.
3. T.manoj lovic* and I.stajdvhar*, Predicting Stock Market Trends Using Random Forest: A Sample of the Zagreb Stock Exchange, IEEE International Convention, pp. 1189- 1193, 2015
4. Yuqing Dai, Yuning Zhang, Machine Learning in Stock PriceTrend Forecasting, 2013.
5. KooshaGolmohammadi, Osmar R. Zaiane and David Díaz, Detecting Stock Market Manipulation using Supervised Learning Algorithms, IEEE International Conference on Data Science and Advanced Analytics, pp. 435-441, 2014.
6. P. Hajek, Forecasting Stock Market Trend using Prototype Generation Classifiers, WSEAS Transactions on Systems, Vol.11, No. 12, pp. 671-80, 2012.
7. V Kranthi Sai Reddy, "Stock Market Prediction using Machine Learning", International Research Journal of Engineering and Technology (IRJET), vol. 08, no. 4, pp. 1033-1035, Oct 2018, ISSN 2395-0056.
8. Saahil Madge, Predicting Stock Price Direction using Support Vector Machines Independent Work Report Spring 2015, Princeton University
9. K. Hiba sabia, Aditya Sharma, Adarrsh Paul, Sarmistha Pardhi, Saurav Sanyal, "Stock Market Prediction using Machine Learning", International Research Journal of Engineering and Technology (IREAT), vol. 05, no. 4, pp. 1033-1035, April 2019, ISSN 2249-8958.
10. Shen, Shunrong, Haomiao Jiang, and Tongda Zhang. "Stock market forecasting using machine learning algorithms." (2012)
11. https://www.researchgate.net/publication/313408030_A_Review_on_Prediction_of_Stock_Market_using_Various_Methods_in_the_Field_of_Data_Mining.
12. J. Patel, S. Shah, P. Thakkar, and K. Kotecha, Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques, Expert Systems with Applications, Volume 42, Issue 1, Pages 259-268, ISSN 0957- 4174, 2015.
13. N. Masoud, Predicting direction of stock prices index movement using artificial neural networks: the case of Libyan financial market, British Journal of Economics, Management & Trade, 4(4): 597-619, 2014.
14. Y. Zuo, E. Kita, Up/Down Analysis of Stock Index by Using Bayesian Network, Engineering Management Research, Issue 2, Vol.1, 2012, pp.46–52.
15. D. Diaz, B. Theodoulidis, and P. Sampaio, "Analysis of stock market manipulations using knowledge discovery techniques applied intraday trade prices," Expert Syst. Appl., vol. 38, no. 10, pp. 12757–12771, Sep. 2011.

16. K. Golmohammadi and O. R. Zaiane, “Data Mining Applications for Fraud Detection in Securities Market,” in 2012 European Intelligence and Security Informatics Conference, 2012, pp. 107–114.
17. F. Neri, Agent based Modeling under Partial and Full Knowledge Learning Settings to Simulate Financial Markets, AI Communications, Issue 4, Vol.25, 2012,pp.295-304.
18. P. Hajek, Municipal Credit Rating Modelling by Neural Networks, Decision Support Systems,Issue 1, Vol. 51, 2011, pp.108–118.
19. P. Hajek, V. Olej, Credit Rating Modelling by Kernel-based Approaches with Supervised and Semi-Supervised Learning, Neural Computing & Applications, Issue 6, Vol.20, 2011, pp.761–773.
20. F.E.H. Tay, L. Cao, Application of support vector machines in -nancial time series forecasting, Omega 29 (2001) 309–317.

AUTHORS PROFILE

Sikkisetti Jyothirmayee, M.tech Student, Department of CSE, SRKR Engineering College affiliated to JNTU Kakinada, AP,Bhimavaram, India. Email: sikkisetti.jyothirmayee@gmail.com

V. Dilip Kumar, Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada,AP,Bhimavaram, India. Email: dilipv510 @gmail.com

Ch.Someswara Rao, Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada,AP, Bhimavaram, India. Email: chinta.someswararao@gmail.com.

R. Shiva Shankar Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram,AP, India. Email: shiva.srkr@gmail.com