CHAPTER 2

LITERATURE REVIEW AND RESEARCH METHODOLOGY

2.1 INTRODUCTION

Prediction of the stock market has always been an interesting activity for many researchers all round the globe because of the lucrative gains involved in it. The ability of stocks to absorb and act on information that is immediately reflected in its prices makes them a very interesting investment option. Academicians and researchers have shown keen interest in studying the predictability of the stock prices, since it throws more light in understanding the behavior and dynamics of the stock market. Bachelier (1900) was the first to talk about the random characteristics of share price behavior followed by Working (1934) & Kendall (1953). It was Fama E (1965) who, in his notable study provided economic justification for the random walk model and coined the term, EMH. From then on, the number of studies trying to study the efficiency of stock markets was on the rise, and every study contributed in a significant way to understand the behavior and predictability of markets. This chapter tries to trace various research studies in the field of market efficiency, illustrating the empirical evidences for and against it, from developed and emerging markets. Research works with applications of data mining approaches to stock price behavior are included and inferences are drawn.
2.2 MARKET EFFICIENCY

Ever since the term EMH has come into the financial economics literature, the theory has stimulated a lot of interest and controversial views among both researchers and financial experts. There are a number of studies and research works undertaken from then on, till date, and this resulted in startling evidence both for and against "efficiency". The study on market efficiency purported that stock prices are unbiased in their reaction to information (Fama et al 1969). Fama (1970) in his influential work showed the growing body of evidence in favor of the EMH that has emerged as one of the most consistent and influential empirical areas in financial economics. Over time, there has been emergence of extensive "anomalous" evidence in return behavior. The size effect, day of the week, and dividend yield, seem to disprove the efficiency theory.

2.2.1 Meaning of Efficient Markets

Prior to Fama et al (1969), the literature on the time series of prices had increasingly but heuristically connected its random walk models with competitive markets. They defined efficient market as, "a market that adjusts rapidly to new information".

Fama (1970) defined an efficient market as one, "in which prices always fully reflect" available information and stated the sufficient conditions for efficiency being:

i. There are no transactions costs in trading securities.

ii. All available information is costlessly available to all market participants, and
iii. All agree to the implications of current information for the current price and distributions of future prices of each security.

Fama distinguished three nested information sets (past prices, publicly available information, and all information) and correspondingly distinguished the “weak, semi-strong, and strong” forms of efficiency.

Weak form efficiency : Prices fully reflect historical information of past prices and returns.

Semi-strong form efficiency : Prices fully reflect all information known to all market participants (public information).

Strong form efficiency : Prices fully reflect all information known to any market participant (public and private information).

Beaver (1981) states that, “A securities market is efficient with respect to a signal \( y_t \) if and only if the configuration of security prices \( (P_t) \) is the same as it would be in an otherwise identical economy (i.e, with an identical configuration of preferences and endowments), except that every individual receives \( y_t \) as well as the information of the individual. Beaver coined the term, “information system efficiency”.

Latham (1986) observed the logical feasibility of information that leads to offsetting revisions in individual investor’s portfolios, without any net effect on excess demand and therefore on prices. He defined efficiency relative to some information set, if revealing it to all agents would neither change equilibrium prices nor portfolios.
There have been many attempts to establish a suitable definition of market efficiency, and according to Ball (1989), each definition characterizes market equilibrium with respect to information. One source of difference lies in the type of equilibrium it characterizes. This definition specifies that either

1. Prices are in competitive equilibrium with respect to information; or

2. Prices and portfolios are in competitive equilibrium with respect to information; or

3. Trading rules defined over a set of information are not profitable, after deducting trading costs.

Most of the earlier works falls into the first category of definition, along with Fama (1970), whereas empirical researchers stick to the third set to define efficient markets.

2.2.2 Notable Empirical Studies in Market Efficiency

Samuelson (1965), in his seminal work, “Proof That Properly Anticipated Prices Fluctuate Randomly,” began with the observation that “in competitive markets there is a buyer for every seller. If one could be sure that a price would rise, it would have already been risen.” Samuelson asserted that “arguments like this are used to deduce that competitive prices must display price changes . . . that perform a random walk with no predictable bias.”

Building on the contribution from Harry Roberts (1967) & Samuelson (1965), Fama (1970) assembled a comprehensive review of the theory and evidence of market efficiency. The theory involves defining an efficient market as one in which trading on available information fails to provide an abnormal profit. A market can be deemed to be efficient, therefore,
only if we posit a model for returns. Fama suggested that the tests of market efficiency become joint tests of market behavior and models of asset pricing.

Fama (1970) summarizes the early random walk literature, with his own contributions and other studies of the information contained in the historical sequence of prices, and concludes that “the results are strongly in support” of the weak form of market efficiency. He also reviews a number of semi-strong and strong form tests, highlighting those that we cover in the next two sections, and concludes that “in short, the evidence in support of the efficient markets model is extensive, and (somewhat uniquely in economics) contradictory evidence is sparse.” He concedes, however, that “much remains to be done”, and indeed, Fama (1991) subsequently returned to the fray with a reinterpretation of the EMH in the light of subsequent research.

Fama et al (1969) made a study that particularly demonstrates that prices reflect not only direct estimates of prospective performance by the sample companies, but also information that requires more subtle interpretation.

Scholes (1972) made a study on the price effects of secondary offerings examines stock price movements when the seller may be in possession of non-public information. On average, share prices fall by an amount that reflects the value of this information. The impact of a secondary distribution on the stock price is largely unaffected by the size of the transaction that confirms the depth of the market and the substitutability of one security for another.

The EMH does not rule out small abnormal returns, before fees and expenses. Analysts could therefore still have an incentive to acquire and act on valuable information, though investors would expect to receive no more than an average net return. Grossman & Stiglitz (1980) formalize this idea,
showing that a sensible model of equilibrium must leave some incentive for security analysis.

2.2.3 Implication from Literature

The theory of EMH is simple in principle, but remains elusive. The EMH came to be supported by a growing body of empirical research demonstrating the difficulty of beating the market, whether by analyzing publicly available information or by employing professional investment advisors. The importance of EMH is demonstrated by the fact that apparently profitable investment opportunities are still referred to as “anomalies.” The EMH model continues to provide a framework that is widely used by financial economists.

2.3 PREDICTABILITY OF STOCK MARKETS

There is a large body of research carried out suggesting the predictability of stock markets. Lo & MacKinlay (1988) in their research paper claim that stock prices do not follow random walks and suggested considerable evidence toward predictability of stock prices. Basu (1977), Fama & French (1992), Lakonishok et al (1997) in their various studies have carried many cross-sectional analyses across the globe and tried to establish the predictability of the stock prices. Studies have tried to establish that various factors like firm size, book to market equity, and macroeconomic variables like short-term interest rates, inflation, yield from short- and long-term bonds, and GNP help in the predictability of stock returns (Fama & French (1993), Campbell (1987), Chen, Roll & Ross (1986), Cochrane (1988)). Ferson & Harvey (1991) show that predictability in stock returns are not necessarily due to market inefficiency or over-reaction from irrational investors but rather due to predictability in some aggregate variables that are part of the information set. O’Connor et al (1997) demonstrated the
usefulness of forecasting the direction of change in the price level, that is, the importance of being able to classify the future return as a gain or a loss. Hence, at any given point of time, research on the predictability of stock indices is significant owing to the dynamic nature of the stock markets.

The stock market prediction task divides researchers and academicians into two groups, those who believe that they can devise mechanisms to predict the market and those who believe that the market is efficient and whenever new information comes up, the market absorbs it by correcting itself; thus there is no space for prediction (EMH). Furthermore, they believe that the stock market follows a random walk, which implies that the best prediction one can have about tomorrow's value is today's value.

In literature, a number of different methods have been applied in order to predict stock market returns. These methods can be grouped into four major categories: (i) Technical Analysis Methods, (ii) Fundamental Analysis Methods, (iii) Traditional Time Series Forecasting and (iv) Machine Learning Methods. Technical analysts, known as chartists, attempt to predict the market by tracing patterns that come from the study of charts that describe historic data of the market. Fundamental analysts study the intrinsic value of a stock and they invest on it if they estimate that its current value is lower than its intrinsic value. In Traditional Time Series forecasting, an attempt to create linear prediction models to trace patterns in historic data takes place. These linear models are divided into two categories: the univariate and the multivariate regression models, depending on whether they use one of more variables to approximate the stock market time series. Finally, a number of methods have been developed under the common label, Machine Learning. These methods use a set of samples and try to trace patterns in it (linear or non-linear) in order to approximate the underlying function that generated the data.
2.3.1 Data Mining Approaches to Stock Market Prediction

As the dynamic stock market leaves a trail of huge amount of data, storing and analyzing these terabytes of information has always been a challenging task for the researchers. After the advent of computers that have brute processing power, storage technologies such as databases, data warehouses, and the modern data mining algorithms, information systems play a pivotal role in the stock market analysis. Data mining provides nontrivial extraction of implicit, previously unknown, and potentially useful information from the data and thus it emerges as an invaluable knowledge discovery process. The four major approaches of data mining are classification, clustering, association rule mining, and estimation.

In the recent years, data mining has developed sophisticated applications that are capable of mining useful information for a large database. Soft computing techniques like ANN, Fuzzy Logic, Genetic Algorithm(GA), Machine Learning techniques, etc., have been generously employed to analyze time series data for various purposes like classification, prediction, association etc. Hence, lot of research work is attempted to apply comprehensive data mining tools to financial time series data like stock market/stock index prices, to identify if the behavior of prices can be better understood by the use of technology.

2.3.2 Application of ANN, k-NN, Decision Tree, and SVM Techniques in Stock Market Prediction and Classification

Brock et al (1992) found nonlinearities in market prices and showed that the use of technical analysis indicators, under certain assumptions, may generate efficient trading rules (Brock et al 1992). The prices of other financial commodities also have this nonlinear dynamic property. Savit (1989), for example, suggested a nonlinear dynamic model for option prices.

Kuan & White 1994; Bierens 1994; Lewbel 1994, The main focus on the ANN technology, in application to the financial and economic fields, has so far pertaining to non-linear relationship among variables. Many economists advocate the application of neural networks in different fields of economics (Kuan & White 1994; Bierens 1994; Lewbel 1994). According to Granger (1991) nonlinear relationships in financial and economic data are more likely to occur than linear relationships. New tests based on neural network systems therefore have gained popularity among economists. Several authors have examined the application of neural networks to financial markets, where the nonlinear properties of financial data pose many difficulties for traditional methods of analysis (Omerod et al 1991; Grudnitski & Osburn 1993; Altman et al 1994; Kaastra & Boyd 1995; Witkowska 1995). Yoon & Swales (1997) compared neural networks with discriminant analysis for the prediction of stock price performance and found that the neural network is superior to discriminant analysis in its predictions. Surkan & Singleton (1990) find that neural network models perform better than discriminant analysis in predicting future assignments of risk ratings to bonds. Trippi & DeSieno (1992) apply a neural network system to model the trading of standard and poor 500 index futures. They find that the neural network system outperforms passive investment in the index. Based on the empirical results, they favour the implementation of neural network systems into the mainstream of financial decision making.
Refenes et al (1995), for example, modeled stock returns with ANNs in the framework of Arbitrage Pricing Theory (APT) and compared their study with regression models. They showed that even the simple neural learning procedures such as back propagation (BP) algorithms easily outperform the best practice of regression models in a typical stock ranking application within the framework of APT. Tsibouris & Zeidenberg (1995) applied two models, a BP model and a temporal difference model, to six stock returns to test the weak form of EMH. They found some evidence to question the null hypothesis that the stock market is weakly efficient.

Stephen Lee et al (1996) proposed a trading system for the S&P 500 index, which consists of a feature selection component and a simple filter for data processing, two specialized neural networks for return prediction and a rule based algorithm for prediction integration. The objective was to explore if inclusion of additional knowledge for more sophisticated data filtering and return integration leads to further improvements in the system and presented evidences that by embedding some form of technical analysis knowledge into a neural network based trading system can improve its predictive capability.

Poddig & Rehkugler (1996) in their research work presented an economic approach to analyze highly integrated financial markets and the econometric methods, especially ANN, to realize it. The recurrent ANN networks are used to model integrated financial markets. This approach tries to develop a world model of integrated financial markets. The world consists of the stock, bond, and currency markets of the United States, Japan, and Germany. Traditional econometric methods and ANNs are used to develop various kinds of models that are rated by their ability to provide accurate forecasts for these financial markets in an out-of-sample test.
El Shazly, MR & El Shazly, HE 1997 (1997) employed neural network model to forecast currency prices and compared it with that of the forward rate of three currencies, the British pound, the German Mark, and the Japanese Yen. Tao criteria are applied to evaluate performance: accuracy and the ability to correctly predict the direction of the exchange rate movement. The results of the neural networks for the three currencies tested outperformed the forward rate both in terms of accuracy and correctness.

Cao & Tay (2001) studied the use of SVM in financial forecasting by comparing it with a multilayer perceptron trained by the BPNN algorithm. They found that SVMs outperform better than BP based on the criteria of normalized mean square error (NMSE), mean absolute error (MAE), directional symmetry (DS), correct up (CP) trend and correct down (CD) trend for the S&P 500 daily price index. They concluded that it is advantageous to apply SVMs to forecast the financial time series.

Cao & Tay (2002) applied the SVM for financial time series forecasting. The feasibility of applying SVM in financial forecasting is examined by comparing it with the multilayer BP neural network and the regularized RBF neural networks. The variability in performance of SVM with respect to the free parameters was investigated experimentally using five future contracts collated from the Chicago Mercantile market. The simulation results showed that among the three methods, SVM outperforms the BP neural network in financial forecasting.

Irwin King et al (2002) applied the SVR to financial prediction tasks. They varied the margins of the SVR that could reflect the change in volatility of the financial data. They also analyzed the effect of asymmetrical margins in order to reduce the downside risk. The experimental results showed that the use of standard deviation for calculating the variable margin
of SVR produced good predictive result in predicting the stock index movement.

Ragusa et al (2002) in their work have brought out the application of combination of methods to evaluate stock market purchasing opportunities using the technical analysis school of stock market prediction. The results supported the effectiveness of the technical analysis approach through the use of the bull-flag price and volume pattern heuristics. The romantic approach of high performance computing, machine learning and soft computing techniques like neural network, Genetic Algorithm, and approaches recently developed that combine diverse classification and forecasting systems.

Kim (2003) applied SVM to predict the stock index price. The study further examined the feasibility of applying SVM in financial forecasting by comparing it with BPNN and case-based reasoning. The experimental results showed that SVM provided promising alternative to stock market prediction.

Kyoung-jae Kim (2006), applies SVM for predicting the stock price index. In addition, his study examines the feasibility of applying SVM in financial forecasting by comparing it with back-propagation neural networks and case-based reasoning. The experimental results show that SVM provides a promising alternative to stock market prediction.
Wu et al (2004) introduced the SVM in an attempt to provide a model with better explanatory power. They used the BPNN as a benchmark and obtained prediction accuracy around 80% for both BPNN and SVM methods for the US and the Taiwan markets. They obtained a slight improvement in the results of the SVM model.

Andrada-Felix et al (2005) tried to study the possibility of exploiting the nonlinear behavior of daily returns on the Spanish Ibex-35 stock index returns to improve forecasts over short and long horizons. They examined the out-of-sample forecast performance of smooth transition autoregression (STAR) model and ANNs. The results were evaluated with statistical and economic criteria. The results of the study showed that ANNs consistently surpass the random walk model and provide better forecasts that the linear Auto regressive model and the STAR models for some forecast horizons and forecasting methods.

Athappilly & Razi (2005) performed a three-way comparison of prediction accuracy involving nonlinear regression, neural network, and Classification and Regression Tree (CART) models using a continuous dependent variable and a set of dichotomous and categorical predictor variables using a large dataset of smokers. Different prediction accuracy procedures were used to compare performances of these models.

Hyun-jung Kim et al (2005) investigated the efficacy of applying SVM to bankruptcy prediction problems. This method overcomes the limitations of BPNN. They demonstrated that the SVM model captures geometric characteristics of feature space without deriving weights of networks from the training data, and is capable of extracting the optimal solution with small training set size. The results show that the proposed classifier of SVM outperforms BPN in the problem of bankruptcy prediction.
Rangold et al (2005) examined the predictability of stock returns using macroeconomic variables in 12 industrialized countries. The authors employed both in-sample and out-of-sample tests (McCracken variant of the Diebold & Mariano and West test for equal predictability and Clark and McCracken variant of the Harvey, Leybourne and Newbold test for forecast encompassing). Among the macro variables considered, it was found that interest rates are the most consistent and reliable predictors of stock returns across countries.

Schniederjans et al (2005) in their study used the ANN to predict stock price movements for firms traded on the Shanghai stock exchange. They compared the predictive power using linear modes from financial forecasting literature to the predictive power of the univariate and multivariate neural network models. The results show that neural networks outperform the linear models compared.

Thawornwong et al (2005) introduced an information gain technique used in machine learning for data mining to evaluate the predictive relationships of numerous financial and economic variables, neural network models for level estimation and classification for examining their ability to provide effective forecast of future values. A cross validation technique was employed to improve the generalization ability of several models. The results showed that the trading strategies guided by the classification models generate higher risk adjusted profits than the buy-and-hold strategy and those guided by the level estimation based forecasts of the neural network and linear regression models.

Suraphan Thawornwong et al (2005) discussed the issue on nonlinearity in the financial markets and neural ability of neural networks to uncover the relationship. They discussed about the alternative forecasting techniques, relevance of input variables or the performance of the models.
when using different strategies. The authors introduce an information gain technique used in machine learning for data mining to evaluate the predictive relationships of numerous financial and economic variables. Neural network models for level estimation and classification are then examined for their ability to provide an effective forecast of future values. The results show that the trading strategies guided by the classification models generate higher risk-adjusted profits than the buy-and-hold strategy, as well as those guided by the level estimation based forecasts of the neural network and linear expression models. Memariani et al (2011), proposed a modern neural network model for stock market mining and for the technical analysis of Japanese candlestick. The approach suggest a kind of regression model whose independent variables are important clues and factors for the technical analysis patterns and its dependent variable is the market trend in the near future.

Madden et al (2006) aimed to evaluate the effectiveness of using external indicators, such as commodity prices and currency exchange rates in predicting movements in Dow Jones Industrial Average (DJIA) Index using a neural network approach. The results of the experiments based on trading decisions on neural network training on a range of external indicators resulted in a return on investment of 23.5% per annum when DJIA grew by 3.3%.

Manish Kumar & Thenmozhi (2006) attempted to predict the direction of S&P CNX Nifty Market index of the NSE. The tested classification models that predict direction included LDA, Logit, ANN, Random forest, and SVM. The results of the experiment suggests that the SVM outperforms the other classification methods in terms of predicting the direction of the stock market movement and the Random forest method outperforms neural network, Discriminant analysis, and Logit model used in the study.
Kirley et al (2007) in their work proposed and implemented a fusion model by combining the Hidden Markov Model (HMM), ANN and GA to forecast financial market behavior. They used ANN to transform the daily stock prices to independent set of values that become the input to HMM. The GA was used to optimize the initial parameters of HMM. The trained HMM is used to identify and locate similar patterns in the historical data. The price difference between the matched days and the respective day are calculated. Then the weighted average of price difference of similar patterns is obtained to prepare a forecast for the next day.

Moreno et al (2007) have aimed to study the predictability of stock markets and checked whether the predictability is exploitable. They also wanted to check whether the economic significance of predictability is higher or lower in the emerging stock markets than in the developed markets. They used linear ANN models and perform a computationally demanding forecasting experiment to assess the predictability of returns. They included the explicit and implicit trading costs for emerging and developed stock markets. The results showed that the trading costs of both the emerging as well as the developed stock market returns are clearly nonpredictable.

Wang et al (2005) investigated the predictability of financial movement direction with SVM by forecasting the weekly movement direction of NIKKEI 225 index. The forecasting ability of SVM was compared with LDA, Quadratic Discriminant Analysis and Elman Back propagation neural networks. The results showed that the SVM outperforms the other classification methods and combination models offer the best performance among all the forecasting methods.
Wei Huang et al (2007) reviewed various articles and have discussed about the input variables, the type of neural network models, performance comparisons for the prediction of foreign exchange rates, stock market index, and economic growth. The authors have suggested that prediction performance of neural network can be improved by integrating it with other technologies.

Daouk et al (2008) attempted to model and predict the direction of return on market index of the Taiwan exchange. The probabilistic neural network (PNN) is used to forecast the direction of index return after getting trained with the historical data. The authors have compared the statistical performance of the PNN forecasts with that of the generalized methods of moments (GMM) with Kalman filter. The empirical results show that the PNN-based investment strategies obtain higher returns than that of the other investment strategies examined.

Arroyo & Mate (2009) in their research work adapts k-NN algorithm to forecast the Histogram Time Series (HTS) to generally deal with histogram data. The proposed k-NN relies on the choice of a distance that is used to measure dissimilarities between sequences of histograms and to compute the forecasts, considering the Mallows distance and the Wasserstein distance. The forecasting ability of the k-NN adaptation is illustrated with meteorological and financial data and promising results were obtained.

Atsalakis & Vakvanis (2009) surveyed more than 100 related published articles that focus on neural and neuro fuzzy techniques derived and applied them to forecast stock markets. Classifications are made in terms of input data, forecasting methodology, performance evaluation, and performance used. Through the surveyed papers it is shown that soft computing techniques are widely accepted for studying and evaluating stock market behavior.
Majhi et al (2009) developed a trigonometric functional line ANN model for short- (one day) and long-term (one month, two months) prediction of stock price of leading stock market indices (DJIA & S&P 500). The historical index data transformed into various technical indicators as well as macro economic data as fundamental factors are considered as inputs to the proposed model. The mean absolute percentage error (MAPE) with respect to actual stock prices is selected as the performance index to gauge the quality of prediction of the models. The empirical results show that the proposed model outperformed the other models.

Wang and On (2009) in their research work employed ten different techniques of data mining to predict the stock price movement. The results show that the SVM and LS-SVM generate superior predictive performance over other models.

Galit Shonuseli et al (2010) proposed a novel functional K-Nearest neighbor (fK-NN) forecaster for real-time forecasting of online auctions. They proposed a novel Beta growth model, and measured the distances between two price paths via the Kullback-Leibler distance and used it on e-Bay auctions and showed improved predictive performance over several competing models.

Ghanbai et al (2010) presented an integrated approach based on genetic fuzzy systems (GFS) and ANN constructing a stock price forecasting expert system. Initially, they carried out a stepwise regression analysis to determine the most influencing factors on stock prices. Then they divided the raw data into k-clusters by means of self-organizing map (SOM) neural networks. All clusters are then fed into independent GFS modes that have the ability of rule-based extraction and database. Data from IT and Airlines sectors are used and MAPE is used to compare the results. The results showed that the proposed approach outperforms all previous methods.
Hsiao et al (2010) combined multiple feature selection methods to identify more representative variables for better prediction. Three popular feature selection methods, the Principal Component Analysis (PCA), the GA, and Decision trees (CART) were used. The combination methods to filter out unrepresentative variables are based on Union intersection and multi-intersection strategies. Back propagation neural network BPNN was used for model prediction. The results showed that the intersection between PCA, GA and CART perform with better accuracy of 79% and 78.89%. These models were able to filter out 80% of unrepresentative feature from 85 original variables resulting in 14 and 17 important features. These variables form the important factors for stock prediction that can be used for future investment decisions.

Nambi & Subha (2010) made an attempt to compare the predictive performance of k-NN forecasting model and traditional regression model. Few actively traded companies in the Indian stock markets and BSE Sensex and NSE Nifty values were predicted with regression and k-NN model. K-NN model is found exhibit better forecasting capability in terms of standard error of prediction.

Sheha Soni et al (2010) combined three supervised machine learning algorithms, the CART, the LDA, and the QDA for classification of Indian Stock market data in the form of binary tree, linear surface, and quadratic surface, respectively. The performance comparison revealed that the CART model showed better classification results when compared to that of the LDA and QDA.

Wang et al (2010) investigated the statistical properties of the fluctuations of the Chinese stock index, HSI DJIA Index, IXIX and S&P 500 by comparison. The stochastic time-effective neural network is used to uncover the predictive relationships of numerous financial and economic
variables. They also introduced the Brownian motion so that the model has the effect of random movement while maintaining the original trend. The forecasting performance of the model is tested using different volatility parameters and it shows some predictive results on the global stock indices using the stochastic time-effective neural network mode.

Chang et al (2011) apply fuzzy logic as a data mining process to generate trees from a stock database containing historical information. The work established a novel case-based fuzzy decision tree model to identify the most important predicting attributes and extract a set of fuzzy decision rules that can be used to predict the time series behavior in the future. The fuzzy decision tree generated from the stock database is converted to fuzzy rule that helps in making decision about the stock market movement. The approach was experimentally applied on S&P500 index and some stocks S&P 500.

Desheng (2011) et al employed a radial basis function neural network to train data and forecast the stock indices of the Shanghai stock exchange. They also introduced the artificial fish swarm algorithm (AFSA) to optimize radial basis function (RBF). The forecasting efficiency is improved using K-means clustering algorithm. The results indicated that the RBF optimized by AFSA is easy to use and has considerable accuracy.

Salim Lahmiri (2011) applied the PNN and SVM to predict stock market daily trends. He examined the effect of macroeconomic information and technical analysis indicators on the accuracy of classifiers. He also aimed to study the joint effect of the two indicators on the classification performance when used together. He also employed the Granger tests to identify the causal relationships between the input variables and the predicted stock returns. The lagged returns that need to be considered for the input space were identified using autocorrelation function. The results showed that the macroeconomic variables were more suitable to predict stock market trends than the use of
technical indicators. The combination of both as inputs did not improve the prediction accuracy.

Subha & Nambi (2011) employs Classification and Regression Tree (CART) to predict the daily price direction of BSE Sensex and NSE Nifty. The results of the CART model are compared with that of the logistic regression model and it is seen that the decision tree classifier outperforms the logistic regression by more than 25%.

Ayodels et al (2012) presented a hybridized approach that combines the use of the variables of technical and fundamental analysis of stock market indicators for prediction of future price of stock in order to improvize the existing approaches. The hybridized approach was tested with stock data and the results showed remarkable improvement over the use of only technical variables. The prediction from hybridized approach was found to be a satisfactory guide for traders and investors in making qualitative decisions.

Sathyanarayana et al (2012) investigated the predictability of financial movement direction with SVM by forecasting the weekly direction of BSE SENSEX. They compared the performance of SVM with Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Elman Propagation Neural Networks. The experimental results show that the SVM outperforms the other classification methods.

Shom Prasad Das & Sudarsan Pathy (2012) employed two machine learning algorithms: the BPNN and the SVM to predict the futures prices traded in Indian stock market. They compared the performances of the two techniques and observed that SVM provides better performance results when compared with BPNN.
Subha & Nambi (2012) employed k-NN classifier for the task of classification of the stock index values of BSE-SENSEX and NSE-NIFTY. The results of k-NN classifier are compared with the Logistic regression model and it is observed that the k-NN classifier outperforms the traditional logistic regression method by more than 25% accuracy rate.

2.3.3 Inference from Literature

The review of various works brings out various interesting facts. There are many studies that applied data mining tools to investigate the predictability of various financial series. The reasons quoted for applying data mining tools are its ability to handle voluminous data and nontrivial extraction of implicit, previously unknown, and potentially useful information from financial data.

The literature review point out the fact that results of prediction varies among different financial markets. The other interesting fact is that data mining tools are found to consistently outperform other statistical approaches. The literature suggests that data mining tools are likely to predict stock market price movements better when compared to other methods. The most obvious advantage of the data mining techniques is that they can outperform the classical statistical methods with 5–20% higher accuracy rate (Ren et al 2006).

Majority of the studies incorporated technical factors, intraday movements and macro-economic indicators as input factors for predicting the stock price movements. However, not many studies are concerned with including the global cues and its influence on predicting the stock index movements.
The review suggests that the efficiency of predictability increased from k-NN and ANN with use of higher evolutionary computing tools like Random Forest and SVM. In the review by Atsalakis and Valvanis (2009), the results of the survey done by the authors find that till 2009, almost all soft computing studies focused on neural networks and neuro fuzzy approaches for prediction. About 60% of the articles used feed forward neural networks and recurrent networks.

The literature review highlights that both statistical and non-statistical measures were used to evaluate the efficiency of the data mining approaches adopted by researchers. The most commonly used statistical measures were, RMSE, MAE and Mean squared prediction error (MSPE). Non-statistical measures that are commonly used were the HIT rate and error rate that measures the percentage of correct/incorrect predictions of the model.

The review of previous works showed that the earlier works undertaken are highly empirical and it is inferred that new research works in new markets at different time periods are likely to show new results and insights. The review also points out that not many studies have been undertaken in Indian and other emerging market economies. Also a comprehensive comparative analysis of the above markets is not undertaken using a data mining approach.

The selection of tools initially started from ANN, and moved to Decision Tree, its ensemble variant random forest and Support Vector machines. The k-NN method was not used for prediction of stock market trend in many studies. Hence an attempt to use successful ANN and k-NN and other data mining tools to evaluate the predictability of emerging markets as against developed markets present an interesting topic of research.
2.3.4 Application of Data Mining Tools for Studying Association among Stock Indices

Liao et al (2001) investigated the co-movement between foreign exchange rates and Taiwan stock indices. They implemented the association ruling approach to explore the co-movement between foreign exchange rates and category stock indices in Taiwan. Data of Forex rates and stock indexes were collected and the Apriori algorithm was used to generate the association rules. The study proposed several possible portfolio alternatives in the Taiwan financial capital market including foreign exchange currencies and stock investment under different circumstances.

Lyhagen et al (2002) in their study compared forecasts from two different seasonal cointegration specifications in an empirical forecasting example and in a Monte Carlo study. The two seasonal cointegration specifications are specified by Lee with a parameter restriction included at the annual frequency, and the model proposed by Johansen and Schaumburg, with a general specification for the complex root frequency respectively. The results from the Monte Carlo favor the specification suggested by Johansen & Schuamberg did not receive much evidence in case of the empirical example chosen.

Pardatos et al (2006) considered a network representation of the stock market data referred to as the market graph that is constructed by calculating cross-correlations between pairs of stocks based on the opening prices of data over a certain period of time. Conclusions regarding the dynamics of the stock market developments were arrived at by studying the evolution of the structural properties of the market graph over time.

Toth & Kertezz (2006) analyzed the cross correlations of returns on the New York Stock Exchange. They showed that lead-lag relationships
between daily returns of stocks vanished in less than 20 years. Data constituted 190 most frequently traded stocks in this period from the NYSE stock exchange. The analysis revealed that even for high frequency data, the asymmetry of time dependent cross correlation functions has a decreasing tendency and the position of their peaks is shifted toward the origin while these peaks become sharper and higher resulting in the diminution of the Epps effect. The findings indicated that the market under study becomes increasingly efficient.

Pierdzioch et al (2008) compared forecasts of stock market volatility based on real time and revised macroeconomic data. They used statistical criteria, a utility-based criterion and an option-based criterion to evaluate volatility forecasts. The main results showed that the statistical and economic value of volatility forecasts based on real time macroeconomic data is comparable to the value of forecasts based on revised macroeconomic data.

Liao et al (2008) investigated stock market investment issues on Taiwan stock market using a two-stage data mining approach. The first stage used the Apriori algorithm, an association rule mining technique implemented to mine knowledge and illustrate knowledge patterns and rules in order to propose stock category association and possible stock category investment collections. The authors also applied the k-means algorithm cluster analysis to explore the stock clusters for investment information. They proposed several possible Taiwan stock market alternatives under different circumstances.

Liao et al (2011) attempted to understand the relationship among various mutual funds in Taiwan using association rule mining techniques. The mutual funds are classified based on the risk as high, medium, and low risk levels. Then the study evaluates the co-movement among funds within the same risk level and among funds across different risk levels. They concluded that within the same risk level, the performances of at least seven
funds exhibit strong co-movement. The study also shows the influence of
global economics on the correlations among different funds and investment
recommendations are made.

2.3.5 Inference from Review

The review of various articles that employed various statistical and
data mining association rule mining techniques showed that understanding the
association among various financial time series, especially the stock market
indices is a fascinating area of research. The review of past studies indicates
that researchers aim at studying the influence and relationship between the
stock prices and various macroeconomic variables and technical indicators. It
is notable that not many research attempts are made to study the influence of
global cues (gold prices, crude oil prices, LIBOR rate, leading stock indices,
etc) on the movement of a particular stock index price. Research studies
employ the cross correlation techniques and the Apriori algorithm popularly
to study the association among various stock price indices. The Indian stock
market being one of the favorite investment destinations for foreign
institutional investors, it would be very useful to study the relationship
between global cues and the Indian stock market. Also, the application of the
association rule mining technique employing Apriori algorithm, will be a
novel attempt that will help throw light on the global indicators that move
with the Indian stock market.
2.4 DATA AND SOURCES OF DATA

2.4.1 Stock Market Indices

Stock market indices are meant to capture the overall behavior of equity markets. A stock market index is created by selecting a group of stocks that are representative of the whole market or a specified sector of the market. A stock index is calculated with reference to a base period and a base index value.

Stock market indices help compare the returns on money invested in the stock market against other forms of investments such as gold or debt. They can be used as a benchmark for comparing the performance of an equity fund and others. As stock indices reflect highly up to date information, they are considered to be lead indicators of the performance of the overall economy or a sector of the economy. Thus stock market indices provide highly valuable inputs to financial analysts. In this study, we have used 13 major global stock indices for analyzing predictability of the markets.

The CNX NSE NIFTY is one of the most popular indices of Indian stock market. It is a well diversified 50 stock index accounting for 22 sectors of the Indian economy. The CNX NSE NIFTY is owned and managed by India Index Services and Products Ltd. (IISL). The IISL has marketing and licensing agreement with Standard & Poor's for co-branding equity indices. “CNX” stands for “CRISIL NSE Index”. The index was initially calculated on full market capitalization methodology but from June 26, 2009, the computation was changed to free-float methodology. The base period for the CNX NSE NIFTY index is November 3, 1995 and the base value of the index has been set at 1000.

The S&P BSE SENSEX also-called the BSE 30 or simply the SENSEX, is a free-float market capitalization-weighted stock market index of
30 well-established and financially sound companies listed on Bombay Stock Exchange (BSE) Limited. The 30 component companies represent various industrial sectors of the Indian economy. Published since January 1, 1986, the S&P BSE SENSEX is regarded as the pulse of the domestic stock markets in India. The base value of the S&P BSE SENSEX is taken as 100 on April 1, 1979, and its base year as 1978–79.

The NYSE Composite is a stock market index covering all common stock listed on the New York Stock Exchange (NYSE), the largest stock exchange in the world. Over 2,000 stocks are covered in the index, of which over 1,600 are from US corporations and over 360 are from foreign listings. It uses free-float market cap weighting for computing the index. It was originally given a value of 50 points, based on the market closing on December 31, 1965, and is weighted by the number of shares listed for each issue. It was re-introduced in January 2003 with a value of 5,000 points.

The NASDAQ Composite is an index of common stocks and similar securities listed in the National Association of Securities Dealers Automated Quotations exchange (NASDAQ) of America. It is the second largest stock market in terms of market capitalization, after the NYSE. It was the world's first electronic stock market that began its trading on February 8, 1971. This index has over 3,000 components and it is highly followed in the United States as an indicator of the performance of stocks of technology companies and growth companies.

The FTSE 100 index, informally, the “footsie”, is a share index of 100 companies listed in the London Stock Exchange with the highest market capitalization. The index began on January 3, 1984 at the base level of 1,000. The index is maintained by the FTSE Group, a wholly owned subsidiary of the London Stock Exchange that originated as a joint venture between the Financial Times and the London Stock Exchange.
The CAC 40 is a benchmark French stock market index and it takes its name from the Paris Bourse's early automation system Cotation Assistée en Continu (Continuous Assisted Quotation). The index represents a capitalization-weighted measure of the 40 most significant values among the 100 highest market caps on the Euronext Paris. Its base value of 1000 was set on 31 December 1987.

The NIKKEI 225, popularly known as the NIKKEI, is a stock market index of the Tokyo Stock Exchange (TSE). It is a price-weighted index and the components are reviewed once a year.

The DAX (Deutscher Aktien IndeX) is a blue chip stock market index consisting of 30 major German companies trading on the Frankfurt Stock Exchange. DAX measures the performance of the 30 largest German companies in terms of order book volume and market capitalization. DAX 30 was launched on December 30, 1987 and was started with a base value of 1,000.

The S&P/TSX Composite Index (^GSPTSE) is an index of the stock prices of the largest companies on the Toronto Stock Exchange (TSX) as measured by market capitalization. The TSX listed companies in this index comprise about 70% of market capitalization for all Canadian-based companies.

The IBOVESPA is a total return index comprising the most representative companies in the market, both by market cap and traded volume. It is the benchmark index of São Paulo Stock Exchange of Brazil.

The RTS Index (RTSI) is a free-float capitalization-weighted index of 50 Russian stocks traded on the Moscow Exchange of Russia. The RTS
Index value is calculated in a real-time mode. The index was introduced on September 1, 1995 with a base value of 100.

The Hang Seng Index (HSI) is a free float-adjusted market capitalization-weighted stock market index in Hong Kong. It is used to record and monitor daily changes of the largest companies of the Hong Kong stock market and is the main indicator of the overall market performance in Hong Kong. HSI was started on November 24, 1969, and is currently compiled and maintained by Hang Seng Indexes Company Limited, which is a wholly owned subsidiary of Hang Seng Bank.

The Shanghai Stock Exchange (SSE) Composite Index is a capitalization-weighted index. The index tracks the daily price performance of all A-shares and B-shares listed on the Shanghai Stock Exchange. The index was developed on December 19, 1990 with a base value of 100.

The daily close price of all the above mentioned indices along with their intraday movements such as open price, high price, low price, and close price for the period of seven-and-half-a-year from January 1, 2006 to May 31, 2013 are downloaded from their respective websites of the stock exchanges.

2.4.2 Other Global Cues

It is a known fact that daily exchange rates of foreign currencies also influence the behavior of stock markets. Many studies confirm that stock markets are found to be cointegrated with exchange rates and there exists causal relationship among them. Exchange rates are found to Granger cause the stock indices in many emerging markets and in the case of developed markets, the causal relationships are found to be bidirectional. In this study, exchange rates of four strong currencies of US Dollar, Euro, British Pound, and Japanese Yen are considered as major global cues as they are found to be
the highly correlated American, European, and Asian currencies with Indian stock markets.

The World Gold Council is the market development organization for the gold industry and the global voice of authority for gold. Based in UK, with operations in India, the Far East, Europe and the United States, the World Gold Council is an association whose 23 members comprise the world’s leading gold mining companies. The daily gold price for the study period is downloaded from the following official site of World Gold Council: (http://www.gold.org/download/value/stats/statistics/xls/gold_prices.xls).

Brent is the leading global price benchmark for Atlantic basin crude oils. It is used to price two-thirds of the world’s internationally traded crude oil supplies. The daily price of Brent oil for the study period is downloaded from the following official site of British Petroleum: (http://ir2.flife.de/data/bp/hpl_brentoil.ph)

The London Interbank Offered Rate (LIBOR) is the world’s most widely used benchmark for short-term interest rates. It is the interest rate at which banks can borrow funds, in marketable size, from other banks in the London interbank market. The LIBOR is fixed on a daily basis by the British Bankers’ Association.

The daily exchange rates of the four currencies and LIBOR for the study period are downloaded from the following official data store of Reserve Bank of India: (http://dbie.rbi.org.in/DBIE/dbie.rbi?site=statistics).
2.5 RESEARCH TOOLS

2.5.1 k-Nearest Neighbor Algorithm

Nearest Neighbors is a data mining prediction technique that looks for records with similar predictor values in the historical database and uses their prediction values for making prediction for an unclassified record. As the name implies, this algorithm looks for ‘k’ nearest cases in sample dataset and uses the outputs of those nearest cases for forecasting the output for a new instance of data. All neighbors output can be given equal weights or the closest neighbors can be given more weights which is inversely proportional to its distance from the new data record. One of the most widely used metrics for identifying the nearest neighbor is Euclidean distance.

2.5.2 Artificial Neural Network

An artificial neural network is an efficient information processing system that resembles the function of human brain. An ANN contains several nodes with an input layer and an output layer. Each node emulates a neuron as in the brain and is connected to the other by a connection link. Each connected link is associated with weights that contain information about the input signal. First inputs from the training datasets are propagated through the network, getting affected by the arbitrary initial weights assigned to links. The output of the model is compared with the actual outputs in the training data. This error is then fed back into the network and the connection weights are adjusted so that errors are minimized. A neural net is trained by repeated passes of the data until it reaches the desired level of accuracy. Neural network can be used for both forecasting and classification.
2.5.3 Support Vector Machines

The SVM is a promising new machine learning algorithm that uses nonlinear mapping to transform original training data into higher dimensional plane so that a demarcating plane can be found between class labels. It then employs optimization techniques to optimize the width of this hyperplane. By identifying support vectors, which are instances touching the boundary of the marginal hyper plane, a new instance can be easily classified. By modeling the input-output relationship using sequential marginal optimization regression technique, the SVM can be adapted also for forecasting.

2.5.4 Decision Trees

Decision trees are represented by a set of questions that splits the entire dataset into smaller and smaller subgroups. It searches for one predictor variable and its particular value that splits the entire dataset into two parts with maximum homogeneity in terms of decision variable. The choice of split variable is based on impurity function. This splitting process is then repeated with each of the resulting data fragments until “leaves” or decision nodes are reached. The resulting tree can be used for making decision rules for prediction. A powerful ensemble variant of decision tree called random forest is used as a classifier in this study.

2.5.5 Apriori Algorithm

Apriori is a classic algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent item sets in transactional databases. The name of the algorithm is based on the fact that it uses prior knowledge of the frequent item set properties. It proceeds by identifying the frequent individual items in the database that satisfy minimum support level. It then employs an iterative approach by extending them to larger and larger item sets until no
more frequent item sets are found in the database. The frequent item sets determined by Apriori lead to the discovery of association rules and correlation among the items in large transactional or relational data sets.

The association rules are often expressed in the form \( A \Rightarrow B \), where “A,” is the antecedent and “B” is the consequent. These rules have minimum support and confidence levels. Support is the percentage of transactions that contain both antecedent and consequent (“A” and “B”) and thus it is a measure of frequency of association. Mathematically, support can be expressed as support \((A \Rightarrow B) = P(A \cap B)\). Confidence is the percentage of times “B” that is present in “A”. Mathematically, it is the conditional probability i.e., confidence \((A \Rightarrow B) = P(B | A)\). Confidence is the measure of the strength of the association. Generally, association rule mining involves the following two-step process:

- Finding all frequent item sets with predetermined minimum support.
- Generating strong association rules from the frequent item sets that satisfy minimum support and confidence level.

Detailed explanations on the concept, theory, and application of these tools are given in the subsequent sections.

2.6 RESEARCH SCHEME

The daily closing price of 13 major indices, including their intraday movements for the period of seven and half years from January 1, 2006 to May 31, 2013, comprising of more than 1,800 trading days, are considered for this study. This data set split in the ratio of 80:20 is to be used as training dataset and test dataset, respectively. The training dataset covers the time period from January 2006 to December 2012, with roughly about 1,500
trading days and the remaining data of the period from December 2012 to May 2013, with more than 300 trading days’ data, are used as test dataset. The major stock indices of India (BSE SENSEX, NSE NIFTY), stock indices of developed countries including America (NASDAQ & NYSE), Canada (GSPTSE), England (FTSE), France (CAC40), Germany (DAX), Japan (NIKKEI) and stock indices of emerging economies of BRIC countries (IBOVESPA, RTSI, SSE, HSI) are chosen for the study. The stock markets are chosen on the basis of the total market capitalization and the most widely followed indices in the respective markets are chosen for the study. (Vide. Appendices 2 and 3). The table below lists out the 13 popular global indices considered for the study. Seven indices representing developed markets and six representing emerging markets from BRIC countries are chosen for the study.

Table 2.1 List of global stock indices

<table>
<thead>
<tr>
<th>NYSE</th>
<th>NSE NIFTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASDAQ</td>
<td>BSE SENSEX</td>
</tr>
<tr>
<td>NIKKEI</td>
<td>IBOVESPA</td>
</tr>
<tr>
<td>FTSE</td>
<td>RTSI</td>
</tr>
<tr>
<td>CAC 40</td>
<td>HSI</td>
</tr>
<tr>
<td>DAX</td>
<td>SSE</td>
</tr>
<tr>
<td>GPSTSE</td>
<td></td>
</tr>
</tbody>
</table>

This data set split in the ratio of 80:20 is to be used as training dataset and test dataset respectively. The training dataset covers the time period from January 2006 to December 2012 with roughly about 1,500 trading days and the remaining data of the period from December 2012 to May 31, 2013, with more than 300 trading day’s data, are used as test dataset.
The training dataset that is meant for the machine learning purpose as the predictive models based on the four data mining algorithms, use the data in the dataset for learning purpose to improve the accuracy of the model. The learning process is usually iterative, where all the data are repeatedly scanned by the model to fine-tune their model attributes to get the best possible output. Once a satisfactory level of accuracy is reached, forecasting models are put to actual use. The test dataset is earmarked for the purpose of validating the functioning of the trained model. The data in the test dataset are predicted with the trained forecasting models and their output is compared with the actual values in the test dataset. Using various error measures, the performance of the forecasting models can be validated.

Data preprocessing is the first step, where data collected are made ready for further processing. This involves computation of ten widely used technical indicators such as stochastic %K, stochastic %D, Momentum indicator, Price rate of change, Williams’ %R, Moving average convergence and divergence (MACD), Price minus Moving Average, A/D oscillator, Price oscillator, and Relative strength index. All these technical indicators are calculated from the intraday movements of the index such as opening price, high price, low price, and closing price and together they constitute the predictor set for the model.

Data reduction aims at achieving parsimony and eliminating multicollinearity problem. Principal component analysis is a popular technique of dimensionality reduction. It groups attributes that contain similar information into a single set of attribute and thus reduces the number of input attributes to the predictive model. The grouping is done on the basis of correlation between the two attributes. If the correlation between the two attributes is high, it implies that the two attributes convey the same
information. In such cases, we can use one attribute instead of using both as more attributes increase the complexity and noise level in the predictive algorithm. The number of distinct groups in the attributes can be identified from the Eigen values of correlation matrix of all the attributes. If all the Eigen values have similar order of magnitude, it is an indication that each attribute carries a significant amount of information individually and none of them could be ignored from the modeling process. On the other hand, if some Eigen values are high and some are low, it implies that some attributes carry similar information. In that case, we can discard the attributes that do not contribute significant additional information by specifying the level of variance covered. A sample output of the principal component analysis is given in the appendix-1.

Normalization is a process in which attribute data are scaled so as to fall within a small specified range such as -1.0 to 1.0 or 0.0 to 1.0. Normalizing the input values for each attribute in the training dataset help in speeding up the learning phase. It also helps prevent attributes with large range values from outweighing attributes with smaller range values such as binary values. The most commonly used normalization process is Min-Max normalization. It performs a linear transformation of the original data. Suppose, \( \max_A \) and \( \min_A \) are the maximum and minimum values of an attribute \( A \), Min-Max maps a value \( v \) of \( A \) to \( v' \) in the range \((\text{new}_\min_A, \text{new}_\max_A)\) by using the following formula:

\[
v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new}_\max_A - \text{new}_\min_A) + \text{new}_\min_A
\]  

(2.1)

In the training phase, predictive models are built and trained to predict both numeric values and class label attributes of the next day’s closing
value of the index using the training dataset. The four different predictive models used in this study are k-NN algorithm, Feed forward BPNN, SVM, and decision tree. In this learning phase, these models scan each and every record in the dataset and study the input-output relationships so that they can predict the future values or states. Predicting future values that are numeric and continuous in nature is referred to as forecasting and predicting future trends or states that are categorical or discrete in nature is known as classification. Usually, this phase involves large number of iterations that enable models to learn and alter the model configuration such as weights for each predictor, to improve prediction accuracy. This learning process continues until the desired level of prediction accuracy is reached.

These models are then put to actual use to predict the movement/value of the next day's index price of the test dataset. The actual values in the test dataset are compared with the predicted outputs from the models to evaluate the performance of the models. Forecasting errors are measured in terms of correlation coefficient between prediction and actual value, mean absolute error, and root mean squared error, whereas classifier errors are measured in terms of hit ratio, error rate, and kappa statistics.
Figure 2.1 Schematic diagram of the research methodology
2.7 CONCLUSION

The review of various literatures in the field of financial forecasting and prediction revealed that numerous interesting studies are being carried out to understand the phenomena of predictability of financial markets. The notable work of Eugene Fama paved as a stepping stone that raised interesting queries questioning the predictability to stock market movements. From then on, there have been many research works that talk in favor and also question the theory of market efficiency. Researchers have employed statistical and econometric methods with the advent of hardcore computing data mining tools to predict or forecast stock prices. Though large number of studies employs data mining tools, a comprehensive analysis to study the predictability of Indian stock markets in comparison with other developed and emerging markets have not yet been undertaken. Also the influence of global cues on the stock prices and the association of global cues on Indian stock prices have not yet been undertaken. Hence, a data mining approach to study the predictability of Indian stock markets with other global indices is undertaken to understand the nature and predictability of Indian stock markets during the time period of the study.