Stock market forecasting using financial graph network and machine learning techniques

Report submitted in fulfilment for the requirement of
B. Tech degree in Computer Engineering

Name of the student	Roll Number
Adarsh Kumar	2020UCO1663
Rohit Sharma	2020UCO1697
Shubham Sharma	2020UCO1705

Under the supervision

Of

Dr. Preeti Kaur

Department of Computer Science and Engineering

Netaji Subhas University of Technology (NSUT)

New Delhi, India – 110078

CANDIDATES DECLARATION

DEPARTMENT OF COMPUTER ENGINEERING

We, Adarsh Kumar (2020UCO1663), Rohit Sharma (2020UCO1697) and Shubham Sharma (2020UCO1705) of B. Tech, Computer Engineering, hereby declare that the Project-Report titled "**Stock market forecasting using financial graph network and machine learning techniques**" which is submitted by us to the Department of Computer Engineering, Netaji Subhas University of Technology (NSUT), Dwarka, New Delhi in partial fulfilment of the requirements for the award of the degree of Bachelor in Technology is original and not copied from a source without a proper citation. The manuscript has been subjected to plagiarism check by software. This work has not previously formed the basis for the award of any degree.

Place:

Date:

Adarsh Kumar	Rohit Sharma	Shubham Sharma
2020UCO1663	2020UCO1697	2020UCO1705

CERTIFICATE

DEPARTMENT OF COMPUTER ENGINEERING

This is to certify that the work embodied in the project report titled "**Stock market forecasting using financial graph network and machine learning techniques**" by Adarsh Kumar (2020UCO1663), Rohit Sharma (2020UCO1697) and Shubham Sharma (2020UCO1705) is the bona fide work of the group submitted to Netaji Subhas University of Technology for consideration in 8th semester B. Tech project evaluation. The original Research work was carried out by the team under my guidance and supervision in the academic year 2023-24. This work has not been submitted for any other diploma or degree of any university. On the basis of declaration made by the group, I recommend project report for evaluation.

Dr. Preeti Kaur

(Associate Professor)

Department of Computer Engineering

Netaji Subhas University of Technology

CERTIFICATE OF DECLARATION

DEPARTMENT OF COMPUTER ENGINEERING

This is to certify that Project Report titled "**Stock market forecasting using financial graph network and machine learning techniques**" which is being submitted by Adarsh Kumar (2020UCO1663), Rohit Sharma (2020UCO1697) and Shubham Sharma (2020UCO1705) to the Department of Computer Engineering, Netaji Subhas University of Technology (NSUT), Dwarka, Delhi in partial fulfilment of the requirement for the award of Bachelor in Technology is a record of the work carried out by the students under my provision and guidance. The contents of this report, in full or in parts, has not been submitted for any other degree or diploma.

Place:

Date:

Dr. Preeti Kaur

(Associate Professor)

Department of Computer Engineering

Netaji Subhas University of Technology

Acknowledgement

We would like to express our gratitude and appreciation to those who made the completion of this project possible. We would like to thank our supervisor **Dr. Preeti Kaur** whose assistance has been invaluable throughout this journey. Her suggestions and encouragement lead us to the completion of the project and this report.

We would also like to thank our college administration for their support as well as the support from our family and friends.

Adarsh Kumar	Rohit Sharma	Shubham Sharma
2020UCO1663	2020UCO1697	2020UCO1705

ORIGINALITY REPORT

9%	7%	4%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.hindawi.com Internet Source	1%
2	Submitted to University of Durham Student Paper	1%
3	slogix.in Internet Source	1%
4	ideje-abre.com Internet Source	1%
5	www.mdpi.com Internet Source	1%
6	Ying Li, Hongduo Caö, Yong Tan. "Novel method of identifying time series based on network graphs", Complexity, 2011 Publication	1%
7	Hongduo Cao, Tiantian Lin, Ying Li, Hanyu Zhang. "Stock Price Pattern Prediction Based on Complex Network and Machine Learning", Complexity, 2019 Publication	<1%

8	28b15.budzianowski.eu Internet Source	<1 %
9	dataaspirant.com Internet Source	<1 %
10	web.eecs.utk.edu Internet Source	<1 %
11	Xiongwen Pang, Yanqiang Zhou, Pan Wang, Weiwei Lin, Victor Chang. "An innovative neural network approach for stock market prediction", The Journal of Supercomputing, 2018 Publication	<1 %
12	huskiecommons.lib.niu.edu Internet Source	<1 %
13	medium.com Internet Source	<1 %
14	yonsei.pure.elsevier.com Internet Source	<1 %
15	Submitted to The University of Manchester Student Paper	<1 %
16	lib.buet.ac.bd:8080 Internet Source	<1 %
17	cppsecrets.com Internet Source	<1 %

18	ijrpr.com Internet Source	<1 %
19	eprints.soton.ac.uk Internet Source	<1 %
20	www.isa.ru Internet Source	<1 %
21	aisberg.unibg.it Internet Source	<1 %
22	dokumen.pub Internet Source	<1 %
23	link.springer.com Internet Source	<1 %
24	www.coursehero.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

LIST OF CONTENTS

LIST OF FIGURES

List of Tables

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 ABSTRACT

Stock prediction has been one of the major target problems of the artificial intelligence domain. Since its beginning, people have tried to predict the share price values of stocks based on various parameters. However, predicting the exact value of any share is a near impossible task as it governed by infinitely many variables. Historical data, government policies, budgets, natural disturbances, to name a few of them. Hence, instead of predicting the exact value of a given share, we propose to predict the direction and intensity of change in the value of stock indices of India based on historical data. We propose to build a network of patterns seen in the 3 major stock indices of our country namely – SENSEX, NIFTY50 and NIFTY Consumption. After using the centrality measures of this network, we propose to use them as input variables to various classification methods such as KNN, K-means, SVM and certain deep learning strategies such as use of neural networks as well in order to classify the patterns of our test data. We also propose to combine the results of these techniques with some common trading strategies in order to enhance the results.

Keywords: stock, machine learning, graphs, centrality measures, financial graph.

1.2 INTRODUCTION

1.2.1 NEED FOR RESEARCH

Stock market forecasting or prediction refers to the prediction of the future prices of a stock or other parameter based on current information. Prediction in the correct range can lead to significant profits for investors.

The concept of Efficient Market Hypothesis (EMH) which suggests that the prices depend on not just mathematical factors but events in the future can also influence the prices heavily and thus cannot be predicted using available information. However, a lot of researchers do not believe in this hypothesis and have worked on a lot of technologies and models to correctly predict the movement of prices.

Stock market prediction is a separate domain of research now. In the past, any study used to be centric towards manual data analysis and analysing figures of patterns in price variations which were used to determine how the market "might" behave in the future. However, with the advent of artificial intelligence, using AI tools has now become a common idea. Although the idea of applying machine learning and deep learning techniques on stock prices data feels like an easy task, the way in which it is applied can significantly change the results.

1.2.2 AIM

The research area in which we are going to delve is that of graphs and networks. The research in this domain is interesting in the sense that people find general classification strategies to be much more effective in terms of stock market. The motivation behind our study is not to accurately predict the stock market for next few days but rather predict the behaviour which it "might" show in the near future based on historical data. This becomes a classification task as one needs to only consider whether the market will go up or down. The best parameter to consider in this regard is the volatility parameter which calculates the dispersion of share price values. It is going to be the centre of this study and will significantly affect the results.

We propose creating what is known as a financial graph which will envelope the patterns of our stock index price in the following way –

- Each node will represent the combination of patterns of the three stock indices which will be formed by using two of the major parameters of our research – 5-day returns and volatilities.
- A directed edge will be constructed between the nodes going from previous day pattern to the next day pattern with an edge weight of 1. If the edge repeats on any other day, the weight for that edge will increment.

1.2.3 CONTRIBUTIONS OF THE RESEARCH

The major contributions of this research include –

- To experiment with various classification techniques in machine learning to predict the pattern in which the market will behave.
- To have two separate classification methods, one to predict the pattern for a particular day and one to predict the behaviour of the entire graph.

1.3 LITERATURE SURVEY

Global stock market investment strategies based on financial network indicators using machine learning techniques [10]

This paper combines the varying markets around the world to form a network and use all of them to perform a time series forecasting on stock data using some simple machine learning algorithms such as regression, random forests and SVM. The paper uses the parameter of volatility for forecasting the Z-score of each stock indices and then applies two strategies to find out which one performs better with each algorithm.

Forecasting stock crash risk with machine learning [16]

This paper experiments with various features in order to find out which feature is responsible towards the financial distress of a stock. It also sheds light on the use of NLP techniques in order to extract data from news articles and find the features of stock market which has the highest variability in its SHAP score. It also uses distance-to-default parameter. It mainly focuses on news articles and business news in order to predict the directions of stock market and look for crashes.

Stock Price Pattern Prediction Based on Complex Network and Machine Learning [7]

This paper converts the problem of prediction into a classification one by not actually predicting the price but rather predicting the trend in the stock price. It considers 3 most popular stock indices of the US stock market. It finds the pattern of fluctuations in stock prices using returns and volatility and classifies them into 4 separate behaviours. It then constructs a graph for these parameters of 30 days for the entire training dataset. Centrality measures for these graphs are then calculated which act as input variables for KNN and SVM classification algorithms in order to perform prediction on testing data.

Novel Method of Identifying Time Series Based on Network Graphs [3]

This study experiments with various types of time series data and then converts them into graph. Each time series results in a separate kind of graph. The constant time series turns into a complete graph. The periodic

time series like a sine graph turns into a regular graph and so on. The properties of the graph such as their centrality measures, clustering coefficient etc. gives information about the time series.

A hybrid supervised semi-supervised graph-based model to predict one-day ahead movement of global stock markets and commodity prices [8]

This paper uses a semi-supervised approach by building a network of stock indices in the same time zone. The supervised portion of the model predicts the movement of stock market which then sends these results into the network. The research compares its results with the traditional classification methods such as KNN, SVM and Random Forests etc. with their model of HyS3 and Kruskal based graph construction.

Factors Affecting Stock Prices in the UAE Financial Markets [11]

This research focuses on the development of the stock market in the United Arab Emirates (UAE) and aims to identify the key factors influencing stock prices in this emerging market. Covering the period from 1990 to 2005 and based on data from 17 companies, the study employs regression analysis with five independent variables, excluding oil price and dividend per share due to multicollinearity issues. Notably, the findings align with previous research, revealing a strong and positive impact of earnings per share (EPS) on UAE stock prices. Money supply and GDP exhibit expected positive coefficients, albeit statistically insignificant, while the consumer price index demonstrates a significant negative relationship with stock prices, particularly at the 1% confidence level, unlike the interest rate, which remains statistically insignificant.

A method for automatic stock trading combining technical analysis and nearest neighbour classification [6]

This paper uses a nearest neighbour classifier and checks whether considering only historical data can be feasible in analysis of stock market or not. It uses technical indicators such as stop loss, stop gain, RSI filter as parameters for its own trading strategy. It compares the results of the traditional buy-and-hold strategy with its own. The variable to analyse here was profit which turned out to be better than the buy-and-hold strategy's profit.

Applications of deep learning in stock market prediction: Recent progress [14]

A review paper which summarizes the latest progress in deep learning approaches towards stock market prediction. The research pays special attention to implementation and reproducibility. It points out future directions of research in similar domains. The major models considered are some of the time series models like ARIMA, LSTM etc. and classification models like KNN, SVM, K-means etc. The review also combined all the repositories from GitHub into one single repository.

An innovative neural network approach for stock market prediction [15]

The paper proposes two models, one time series model of LSTM with an embedded layer and the other one with an automatic encoder. The models are applied to the Shanghai A-share composite index and Sinopec. The LSTM model with the encoder gives better performance than a stochastic forecast. This research is also closely related to IMMT (Internet of Multimedia of Things) for financial analysis.

The research in this domain is definitely huge but each one of them lacks a concrete solution to the problem. The research below also does not provide a one-stop answer to the question of stock market prediction but we propose a new way of tackling the situation. Since most of the articles try to tackle the situation by actually predicting the stock price, they need to make a lot of tradeoffs which results in the problem becoming very centric. We have converted the problem of regression or time series into a classification problem which is simpler to solve and will required a lot less tradeoffs and as a result will be very generic in nature.

CHAPTER 2: THEORY AND METHODOLOGY

2.1 THEORY

In order to understand various aspects of the project, one needs to understand the following algorithms and concepts.

- **Floyd-Warshall algorithm [18]**

  Floyd-Warshall algorithm is used to find the shortest distances between all pairs of nodes in a graph. The way it works is that in order to find the shortest distance between the nodes *u* and *v,* it has to go through an intermediate node *k.* The idea is to consider each node from 1 to N as an intermediate node and try with every one of them.

- **K-nearest Neighbors or KNN [17]**

  K-nearest neighbors' or KNN classifier is a supervised machine learning classification algorithm which is used to locate clusters or groups of similar points. Its supervised nature is due to the fact that the labels for the training samples are already known and sorted. Whenever, a new sample point comes, the algorithm will classify it based on its nearest neighbors and will conclude that if *K* of its nearest neighbor have a particular label, then this new point must also have the same label.

  The parameters required to judge the "nearness" of the neighbors is based on various distances such as Euclidean distance, Manhattan distance etc.

- **Logistic regression** [17]

  Logistic regression is a supervised machine learning algorithm used for classification tasks. It particularly solves the problem of binary classification by considering a sigmoid function which results into a probability value of a data point having a particular label.

- **Decision Tree Classifier** [17]

  Decision Tree is a supervised learning classification algorithm which uses a tree like structure where each internal node is a feature classification which sets rules for the next classification to take place. At the final level, the label is decided based on all the internal nodes. At each level of nodes, information is gained about the final label.

- **Multi-layer Perceptron classifier** [17]

  Multi-layer Perceptron classifier is a supervised learning algorithm which converts a set of features into a classification. It is a very simple type of neural network which has an input layer, one or more non-linear layers called the hidden layers and the output layer. In the figure 2.1, we can see that the first layer (green) is the layer of features along with a bias, the next layer is a hidden layer and the final one is the output layer which in our case is a classification function.

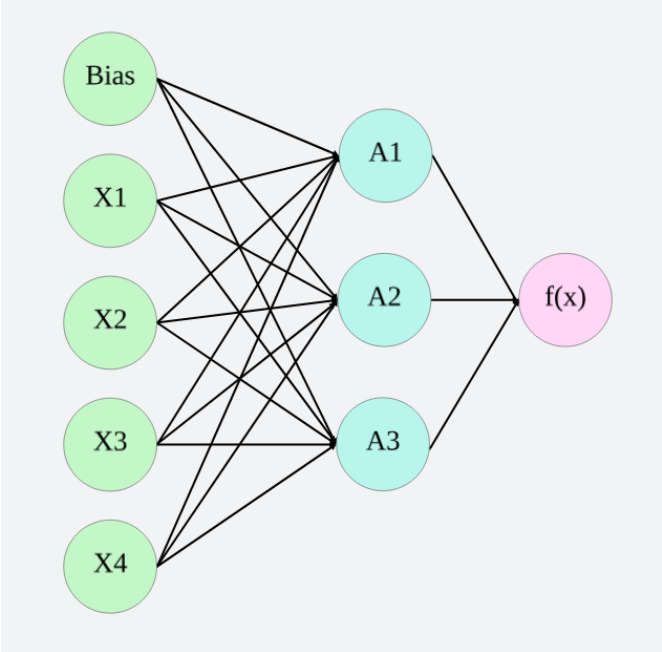


Figure 2.1: Multi-layer perceptron

- **Support Vector Machine or SVM** [17]

Support Vector Machine or SVM is a supervised machine learning regression and classification algorithm which works on the principle of dividing the space using a hyperplane. The type of hyperplane depends upon the number of features in consideration. If there are only 2 features then the hyperplane is a simple line (or curve). Similarly, if we have 3 features then the hyperplane is a plane. The objective of SVM is to keep the margin between the closest points of separate classes as large as possible. The hyperplane divides these classes into their respective labels. In figure 2.2, we can see the hyperplane dividing the two clusters of data points based on their labels.

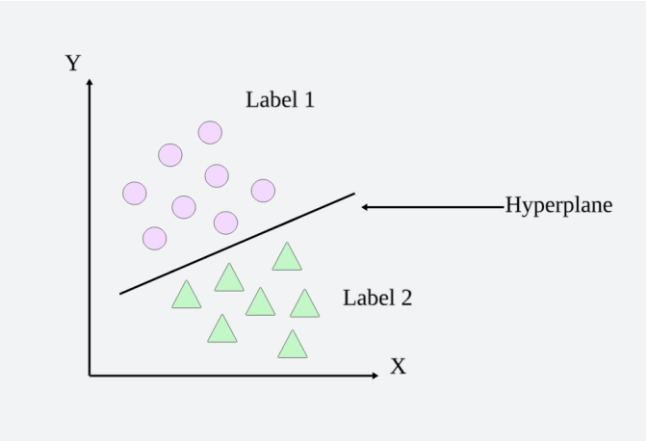


Figure 2.2: Support Vector Machine Classification

- **K-means Clustering** [17]

  K-means Clustering is an unsupervised machine learning algorithm which clusters similar unlabeled data points together using various distances. Initially, it randomly assigns the point a cluster. After that the centroids of the clusters are found and then the point is reassigned to a cluster due to change in distance. This process is repeated several times and stops when no change takes place in cluster assignment.

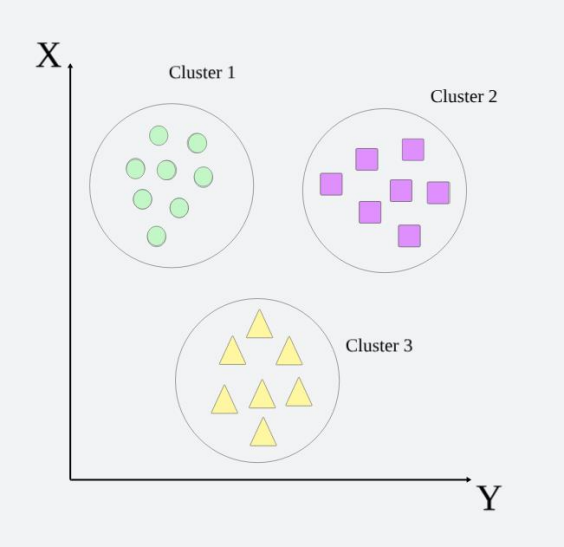


Figure 2.3: K-means Clustering

2.2 PROBLEM STATEMENT

"To classify stock price variation patterns using various machine learning techniques both for one-day classification and graph classification."

Our problem statement for this research involves a common process of creating financial graph using the three stock indices namely SENSEX, NIFTY50 and NIFTY Consumption using 5-day returns and volatilities and finding their centrality measures such as degree centrality, network strength and closeness centrality. Using these measures, we divide our work into two parts, one towards that of classifying for one particular day in consideration using supervised learning classification techniques such as KNN, SVM, Decision Trees etc. and the other one for classifying the behavior of the entire graph using clustering techniques.

2.3 METHODOLOGY AND WORK DONE

The workflow of the entire project is depicted in figure 2.4. It starts by collecting data of three major stock indices of the country namely SENSEX, NIFTY50 and NIFTY Consumption. The data is then used for graph construction through which the volatilities and returns are calculated. After this, the centrality measures of these graphs such as degree centrality, network strength and closeness centrality are used as input variables to our classification models.

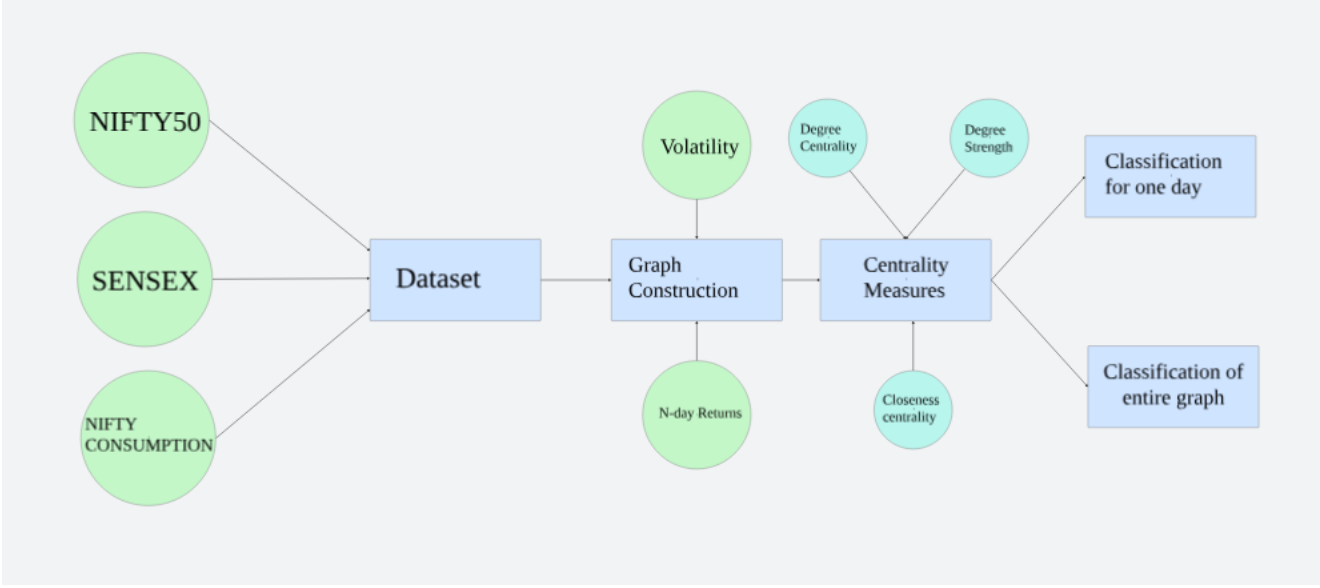


Figure 2.4: Workflow of the research

2.1 DATASET

For our purpose, we scraped the data from Marketwatch.com and Yahoo Finance. Since we wanted to create a synthetic dataset, we collected the data of 10 years from these websites. After collecting all the parameters including the opening price, closing price, maximum price, minimum price and volume we only worked

with the closing prices of the indices. We also combined all the closing prices of the three stock indices into one spreadsheet to make a hybrid dataset.

The dataset used in this research consists of the closing prices of 3 stock indices of India namely SENSEX, NIFTY50 and NIFTY Consumption every day from 01-01-2014 to 31-12-2023 (10 years). SENSEX is a free-float market capitalization consisting of 30 most traded and relatively liquid stocks which contribute towards the balance of the country's equity market. NIFTY50 on the other hand is a benchmark index of 50 companies. NIFTY Consumption reflects the performance of companies in the domestic consumption sector. The data for SENSEX and NIFTY50 is taken from MarketWatch and NIFTY Consumption is taken from Yahoo Finance.

2.2 CREATION OF GRAPH

The graph is constructed using the study of Cao, Lin et. al. who have used *N*-day volatility *V*, and *N*-day return *R* in order to divide the movement of stock index in 4 separate variations.

$$R = \ln\left(\frac{Close(t)}{Close(t-N)}\right) \quad \text{(Eq. 2.1)}$$

Here, *t* refers to the current day in consideration and *N* refers to the number of continuous trading days (generally a week if there is no national holiday) and *Close(t)* refers to the closing price of the stock index on t^{th} day. In order to find out *V*, we need to find one-day return, r which is given by

$$r = \ln\left(\frac{Close(t)}{Close(t-1)}\right) \quad \text{(Eq. 2.2)}$$

After calculating *r*, we can calculate *V* for *N* days by,

$$V = S.D.\left(r_1, r_{2,\dots} r_N\right) * \sqrt{N} \quad \text{(Eq. 2.3)}$$

Where $S.D.\left(r_1, r_{2,\dots} r_N\right)$ refers to the standard deviation of $r_1, r_{2,\dots} r_N$.

We can now calculate the average Volatility of entire stock index in question by simply averaging over the entire time series.

$$V' = \frac{1}{N}\sum V \quad \text{(Eq. 2.4)}$$

In addition to comparing with the entire time series, the volatilities will also be compared with an average of volatilities of a definite window size (30 days).

Now, in order to build the network using the two parameters mentioned earlier, we can devise 4 types of patterns. We can classify the changes in any stock index in the following way –

$$P = \begin{cases} P1, & when\ R \geq 0\ and\ V \geq V'\ (sharp\ rise) \\ P2, & when\ R \geq 0\ and\ V < V'\ (stable\ rise) \\ P3, & when\ R\ < 0\ and\ V \geq V'\ (sharp\ fall) \\ P4, & when\ R < 0\ and\ V < V'(stable\ fall) \end{cases} \quad \text{(Eq. 2.5)}$$

This classification is done for all the 3 indices and the combination of the patterns formed represents a node of a graph. Since the total number of combinations can be $4^3 = 64$, we used a 4-base number system as nodes *u* and *v* for the graph.

The graph is constructed for 30 days although experimenting with other window sizes is still a future prospect of this research.

Index	SENSEX	NIFTY50	NIFTY Consumption	Combined Pattern
1	P3	P2	P1	P3P2P1
2	P4	P1	P4	P4P1P4
3	P2	P1	P3	P2P1P3
4	P3	P2	P1	P3P2P1
5	P1	P2	P3	P1P2P3
6	P2	P1	P3	P2P1P3
7	P3	P2	P1	P3P2P1
…	…	…	…	…
30	P1	P2	P4	P1P2P4
31	P2	P3	P3	P2P3P3

Table 2.1: Patterns for the creation of graph

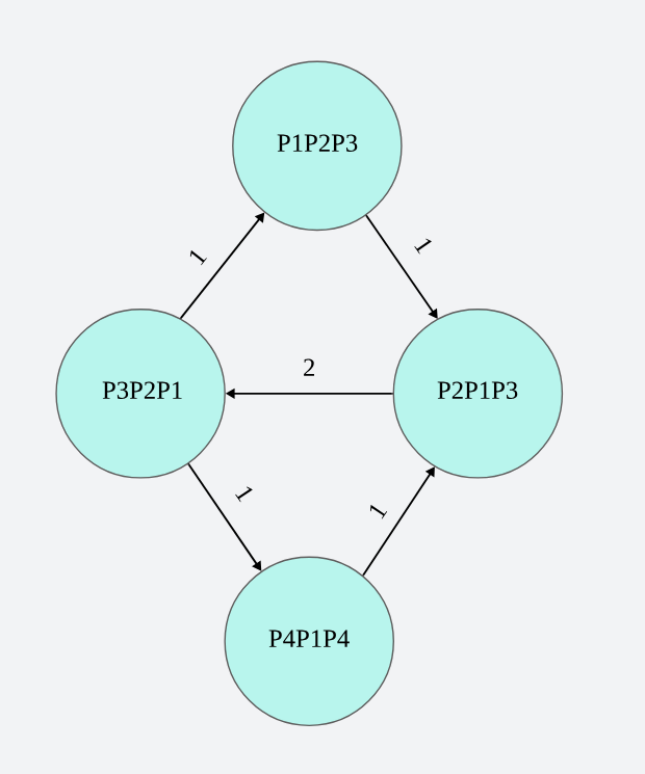


Figure 2.5: A depiction of how the graph will look

In figure 2.5, a small demonstration of how the graph will look like is given. The edge direction shows that towards which pattern the stock price index went. We see there is an edge from $P_4P_1P_4$ to $P_2P_1P_3$, which means that on one day the pattern was $P_4P_1P_4$ and then on the next day the pattern was $P_2P_1P_3$.

The base-4 indexing for the creation of graph is done using the following algorithm.

Algorithm 2.1: **Creation of graph using base-4 indexing**

---

1. Create a list to store all the **pattern graphs**
2. Find average of volatility of each window of days
3. Create an **adjacency matrix** of size 64 x 64 (this is the maximum number of nodes = $4^3 = 64$)
4. Two nodes *u* and *v* where *u* is the **previous node** and *v* is the **next node**.
5. Temporary variables are used for values ranging from 0 to 4 depending upon the comparison of returns and volatilities.
6. The nodes *u* and *v* are then found by multiplying these temporary variables by appropriate powers of 4.
7. An edge is constructed from the previous node to the next node.
8. Once this is done for the entire window, then the graph is appended in our list.

2.3 CENTRALITY MEASURES AS INPUT VARIABLES

The significance of this graph is that the denser this graph is, more is the dispersion and hence more care needs to be taken by investors while investing. Hence, we consider certain centrality measures in order to feed them as characteristics of our graph for classification on unseen data.

The measure we are considering for this research are degree centrality, strength (as described by Cao, Lin et. al.), closeness centrality and betweenness centrality. So far, we have applied the KNN algorithm for degree centrality and strength of the network.

The centrality measures considered for this research include –

- *Degree centrality*

  Degree centrality is defined for a particular node of a graph which is equal to the total number of edges connected to it.

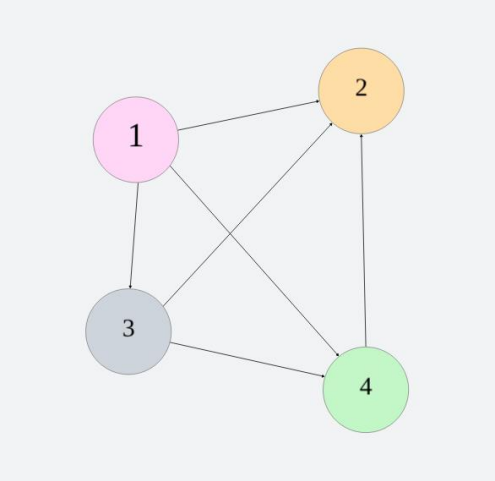


Figure 2.6: A small graph for demonstration

$$\rho = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} a_{ij} \qquad \text{(Eq. 2.6)}$$

Where *N* represents the total number of nodes and "edges" represents the total number of edges in the graph.

In the figure, the degree centrality of node labelled *2* is 3 as it has 3 edges.

- *Closeness centrality*

Closeness centrality denotes how "close" a particular node is to other nodes. Mathematically, it is calculated as the inverse of the summation of all closest distances of nodes from the node in consideration.

$$avgcls(v) = \frac{N-1}{\sum d_{ij}} \quad \text{(Eq. 2.7)}$$

Here, N represents the total number of nodes in the graph and d_{ij} represents the shortest distance between nodes i and j.

In order to calculate the shortest distance d_{ij}, we need to create a matrix using the famous Floyd Warshall algorithm in order to construct a matrix of shortest distances.

- *Degree strength*
  Although, not a very prominent centrality measure, it is very similar to the degree centrality but the difference is that it considers weighted graph where instead of the number of edges connected to the node, the total weight connected to the edges is considered.

These three centrality measures will act as input variables to our classification algorithms with the closing price of the 30th day being our target variable.

Since this is a classification problem, we are going to focus on whether the closing price went up or down after a particular window period. If the price went up, the target variable is going to be 1 and if the price went down then the target variable is going to be 0.

After this the research is mainly divided into two parts –

- *Pattern Recognition for One-day*
  In one phase of the project, we treat the window of N days as a training portion which we use to classify N^{th} day closing price. This generally corresponds to a "buy-and-hold" strategy in which the time period for which the share is being held corresponds to the window size of N days.
- *Pattern Recognition for the entire graph*
  Second phase of the project will cater to the short-term investment where we classify the graphs constructed using these windows of N days and investors will look at the behavior of the market and will act according to the corresponding graph constructed.

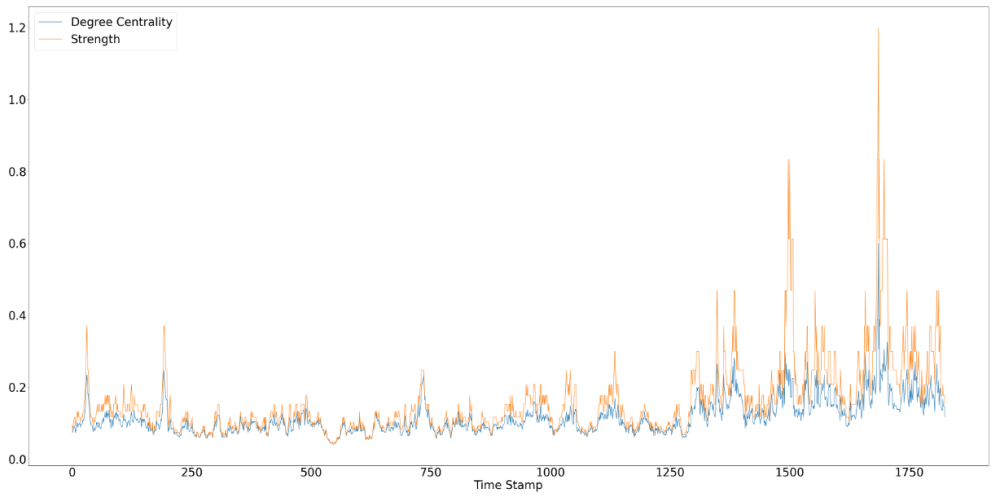
2.4 Pattern recognition for One-day

In order to recognize the pattern for a single day, we considered the window as a training sample and the label of the day in consideration (N^{th} day in this case for a window of N days). The centrality measures of the graph constructed for 30 days acts as input variables to supervised classification algorithms where we

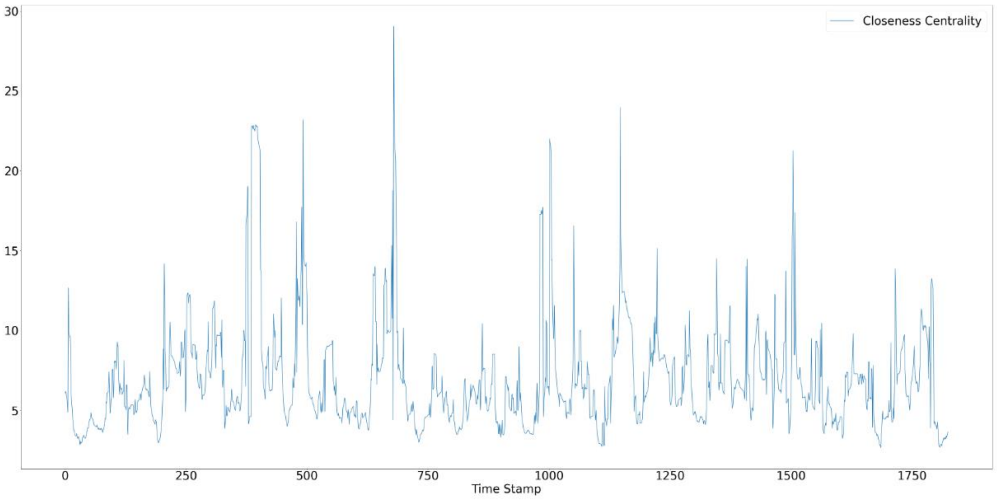
already have the direction of closing price on the N^{th} day. The algorithms considered here include KNN, SVM, Logistic Regression, etc.

The way in which we achieved this was by creating a *pandas dataframe* with the features as degree centrality, degree strength and closeness centrality.

The target variables are considered one-by-one for all three indices namely SENSEX, NIFTY50 and NIFTY Consumption.



(i)



(ii)

(iii)

Figure 2.7: (i) shows the line graph for degree centrality and degree strength (ii) shows the curve for closeness centrality and (iii) shows the trend of the three stock indices

	Degree Centrality	Degree Strength	Closeness Centrality	SENSEX Target pattern	NIFTY50 Target pattern	NIFTY Consumption Target pattern
0	0.077160	0.092593	6.076122	1	1	1
1	0.074792	0.083102	6.171934	0	0	0
2	0.080247	0.092593	6.156658	1	0	1
3	0.086505	0.103806	5.860013	1	1	1
4	0.093750	0.117188	5.684800	1	1	1

Figure 2.8: Data frame constructed for classification

The SENSEX Target pattern, NIFTY50 Target pattern and NIFTY Consumption Target pattern denotes whether the price of the went down after the window of days was over or did it go up. If it went up, the value is 1 and if it went down or remained constant the value is 0.

2.5 Pattern recognition for the entire graph

Pattern recognition for the entire graph is done by clustering similar graphs together. In order to do this, only two centrality measures were considered namely – degree centrality and degree strength as both of them show a strong correlation with each other.

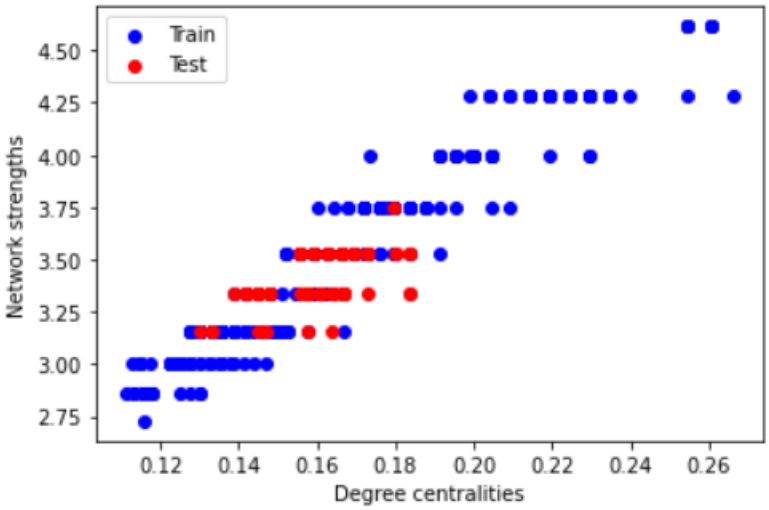


Figure 2.9: Graph showing the correlation between network strengths and degree centralities

Figure 2.9 shows the similarity between the network strength and the degree centralities pertaining to which only these two metrics are considered for graph classification.

In order to predict for the entire graph, all the graphs with the same strength were considered together and a hybrid graph (positive edge weight which was present for the maximum number of times for that edge) was constructed of these graphs. After this the edges in the testing graph of the same strength and the hybrid graph are compared for their weights.

CHAPTER 3: RESULTS AND DISCUSSION

3.1 RESULTS

a. Pattern recognition for one day

The results for the classification for a window size of 30 days without any fine tuning of the algorithms is given below.

Algorithms	**Parameters**	**SENSEX**	**NIFTY50**	**NIFTY Consumption**
KNN	Accuracy	0.70	0.72	0.70
	Precision	0.63 & 0.77	0.63 & 0.79	0.61 & 0.76
	Recall	0.70 & 0.71	0.70 & 0.72	0.68 & 0.71
SVM	Accuracy	0.54	0.54	0.57
	Precision	0.42 & 0.67	0.46 & 0.59	0.51 & 0.60
	Recall	0.45 & 0.55	0.45 & 0.55	0.49 & 0.55

Logistic Regression	Accuracy	0.54	0.54	0.57
	Precision	0.42 & 0.67	0.47 & 0.60	0.51 & 0.60
	Recall	0.45 & 0.54	0.46 & 0.43	0.52 & 0.45
Multi-layer perceptron	Accuracy	0.54	0.54	0.57
	Precision	0.42 & 0.67	0.51 & 0.60	0.51 & 0.60
	Recall	0.45 & 0.54	0.52 & 0.45	0.49 & 0.55
Decision Tree	Accuracy	0.51	0.51	0.51
	Precision	0.45 & 0.56	0.43 & 0.58	0.44 & 0.57
	Recall	0.47 & 0.54	0.46 & 0.54	0.44 & 0.57

Table 3.1: Results for supervised classification of one day based on the window size of 30 days

We can see that the best models for this type of classification is coming out to be KNN with an accuracy of 70%, 72% and 70% for the three stock indices.

b. Pattern recognition for the entire graph

Since the problem of clustering the entire graph was not a problem of supervised classification, its evaluation can simply be done by comparing how many edges in our hybrid graph have the same weight as our testing graph.

In our case out of the total 2057 positively weighted edges, the weights of 1490 of them were equal to each other.

CHAPTER 4: CONCLUSION AND SCOPE FOR FUTURE WORK

4.1 CONCLUSION

The final take of the research is that predicting the stock price is not a simple machine learning task. One has to innovate with other ideas in order to build a foolproof system to correctly predict the stock market. The combination of graphs and their parameters with machine learning techniques introduced a lot of new aspects. Higher values of centrality measures in the research indicate the pattern in the market is repeating and hence a clear prediction can be made for such a time series. Lower values of centrality measures indicate that the market is not that predictable and we can skip such periods from our data frames. We see that deep learning technique of multi-layer perceptron is the best in actually classifying the pattern. Since predicting patterns is a relatively simpler task than actually predicting the price, the research gives a new direction for future. For all different values of centrality measures we find that KNN is producing the best results with the highest accuracies as well as precision and recall.

4.2 FUTURE WORK

As with every research, there are future prospects associated with this one as well. The way we leveraged the idea of graphs in our research still requires more refining. The advent of graph neural networks is more prominent for stock price prediction which can be integrated with this idea. Better implementation of models by fine tuning them or combining them with other models in order to improve results is also a future prospect of this research. Integrating these machine learning and deep learning models with some real-life investment strategies such as buy-and-hold, value investing or quality investing can also be done in order to validate the results even further. The models can be integrated with these strategies in order to build an application which can act as a "Investment companion" for investors.

4.3 REFERENCES

[1] Lucas Lacasa, Bartolo Luque, Fernando Ballesteros, Jordi Luque and Juan Carlos Nun. From time series to complex networks: The visibility graph. Albert-Laszlo Barabasi, Northeastern University, Boston, MA. January, 2008. DOI: 10.1073/pnas.0709247105

[2] Yung-Keun Kwon, Sung-Soon Choi and Byung-Ro Moon. Stock Prediction based on Financial Correlation. GECCO'05, Washinton DC, USA. June 2005. DOI:

[3] Ying Li, Hongduo Cao and Yong Tan. Novel Method of Identifying Time Series Based on Network Graphs. Department of Management Science, Business School, Sun Yat-Sen University, Guangzhou 510275, China. DOI: 10.1002/cplx.20384

[4] Minggang Wang, Ying Chen, Lixin Tian, Shumin Jiang, Zihao Tian, Ruijin Du. Fluctuation behaviour analysis of international crude oil and gasoline price based on complex network perspective. Accepted at Elsevier, May 2016. DOI: 10.1016/j.apenergy.2016.05.013

[5] Michel Ballings, Dirk Van den Poel, Nathalie Hespeels, Ruben Gryp. Evaluating multiple classifiers for stock price direction prediction. Elsevier, May 2015. DOI: 10.1016/j.eswa.2015.05.013

[6] Lamartine Almeida Teixeira, Adriano Lorena Inácio de Oliveira. A method for automatic stock trading combining technical analysis and nearest neighbour classification. Elsevier, 2010. DOI: 10.1016/j.eswa.2010.03.033

[7] Hongduo Cao, Tiantian Lin, Ying Li, and Hanyu Zhang. Stock Price Pattern Prediction Based on Complex Network and Machine Learning. Accepted at Wiley, May 2019. DOI: 10.1155/2019/4132485

[8] Arash Negahdari Kiaa, Saman Haratizadeha, Saeed Bagheri Shourakib. A hybrid supervised semi-supervised graph-based model to predict one-day ahead movement of global stock markets and commodity prices. Accepted at Expert Systems with Applications, March 2018. DOI: 10.1016/j.eswa.2018.03.037

[9] Jan Grudniewicz, Robert Slepaczuk. Application of machine learning in algorithmic investment strategies on global stock markets. Elsevier, July 2023. DOI: 10.1016/j.ribaf.2023.102052

[10] Tae Kyun Leea, Joon Hyung Chob, Deuk Sin Kwonb, So Young Sohnb. Global stock market investment strategies based on financial network indicators using machine learning techniques. Accepted at Expert Systems with Applications, September 2018. DOI: 10.1016/j.eswa.2018.09.005

[11] Hussein A. Hassan Al-Tamimi, Ali Abdulla Alwan & A. A. Abdel Rahman. Factors Affecting Stock Prices in the UAE Financial Markets. Accepted at Transnational Management, March, 2011. DOI: 10.1080/15475778.2011.549441

[12] Vangipuram Radhakrishna,C.Srinivas, Dr.C.V.Guru Rao. Document Clustering using Hybrid XOR Similarity function for Efficient Software Component Reuse. Procedia Computer Science, December 2013. DOI:10.1016/j.procs.2013.05.017

[13] Weiwei Jiang. Applications of Deep learning in Stock Market prediction: Recent progress. Expert Systems with Applications, June 2021. DOI: 10.1016/j.eswa.2021.115537

[14] Xiongwen Pang, Yanqiang Zhou et. al. An innovative neural network approach for stock market prediction. Springer, January, 2018. DOI: 10.1007/s11227-017-2228-y

[15] Laurens Swinkels. Forecasting stock crash risk with machine learning

[16] Andreas C. Müller, Sarah Guido (2016). *Introduction to Machine Learning with Python.* O'Reilly Media, Inc.

[17] Kathleen McKendrick. The Application of Artificial Intelligence in Operations Planning

[18] Mark Needham, Amy E. Hodler. (2019). *Graph Algorithms.* O'Reilly Media, Inc