Clustering of NSE indices using Dendrograms and KMeans Algorithm

* Dr. E.M. Naresh Babu
** Mrs. D. Hemalatha
*** Mr. Santosh Kumar G.

Abstract

Investors usually willing to park their funds in the stocks which gives them good returns with less risk. Whatever may be the financial position of the investor, he/she will be interested to invest in the stocks with less risk and higher return. With proper analysis, financial analysts will be identifying the stocks with the above qualities. In this process, analysts may use different charting or plotting techniques. After identifying that, they suggest investors to invest in single or multiple stocks depending upon their required Return and Risk pattern. This has lead to the concept of diversification. There are some investors who usually select some indices for mitigating the risk. Individual stocks may be at a higher risk but a bunch of stocks may be at a lesser risk is the concept behind selecting the index. Initially some mathematical formulae have been used to construct the portfolios to achieve diversification. Now the concept of Machine Learning has been included in construction of portfolios by many companies, since then lot of people attempted to categorize the assets, stocks, indices into different groups usually referred to as Clusters. There are many ways of clustering and KMeans clustering is one of the powerful ways of clustering the entities.

Keywords : Indices, Clusters, KMeans and Machine Learning

Introduction

Machine Learning has become the most fascinating word to many companies in the last couple of decades. But that is not a new term in the world of technology. We can trac back the inception of Machine Learning from early 1950s itself and we have seen a sea change in the last 7 decades, but with the advent of Industry 4.0 and industry revolution, we have been witnessing a sea change in the last two decades. Machine Learning, is the filed of computer science which is evolved by studying the pattern of a particular activity and computationally learning with the past experiences. (Annina Simon. et al, 2015)

History of Machine Learning

Table 1. History of Machine Learning

Year(s)	Historical improvement
1950s	Samuel's checker-playing program
1960s	Neural network: Rosenblatt's perceptron Minsky & Papert prove limitations of Perceptron
1970s	Symbolic concept induction Expert systems and knowledge acquisition bottleneck Quinlan's ID3 Natural language processing (symbolic)
1980s	Advanced decision tree and rule learning Learning and planning and problem solving Resurgence of neural network Valiant's PAC learning theory Focus on experimental methodology

* Associate Professor, ABBS School of Management, Bengaluru
** Assistant Professor, Oxbridge B School, Bengaluru
*** Associate Professor, ABBS School of Management, Bengaluru

Year(s)	Historical improvement
1990s (ML and Statistics)	Data Mining Adaptive agents and web applications Text learning Reinforcement learning Ensembles Bayes Net learning
1994	Self-driving car road test
1997	Deep Blue beats Gary Kasparov
2009	Google builds self driving car
2011	Watson wins Jeopardy
2014	Human vision surpassed by ML systems

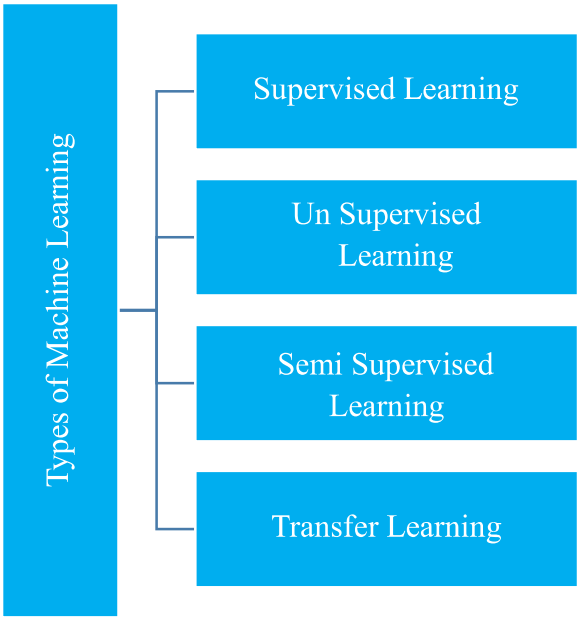


Figure 1. Types of Machine Learning

Supervised learning: Supervised learning is one in which Data (Features or Independent variables) will be available and also some labels (Target variables-Dependent variables) will be available for the model building and training purpose.

Semi supervised learning: Semi Supervised Learning is one in which some data (Features) will be available and very less some labels will be available for model building and training purpose.

Unsupervised learning: In Unsupervised Learning, partial data (Features) will be available but no labels are available as there is no surety of patterns

Transfer learning: As the name itself suggests, some data (Features) will be available which is used for building model and the model will be applied on other data set, and this type of learning is considered as very tough

Hence, in Supervised Learning, we can predict the output with the help of Features, which is considered as labelled data. Semi supervised data contains some labels and with the help of those we will be able to predict the output. But in Unsupervised and Transfer Learning we will not have labelled data, so we will create some algorithm to train the model to make the data into certain groups, which are also referred to as Clusters. Model or Algorithm will be built with the available data and that will be applied to the new dataset.

A *Cluster Algorithm* or *Cluster Analysis* is one of the ways of unsupervised Machine Learnings. *Clustering* is one of the Unsupervised Algorithms. A *Cluster* is a group of data points or observations which are grouped together based on some similarities. If we take the case of stocks or companies, there are multiple ways of clustering, the clustering can be done based on nature of company, years of existence of the company, return of the company, risk features of the company, value of beta of the company, assets of the company or some other characteristics of the company.

We can use clustering if the data is unlabelled, i.e. if we cannot have any expected output and there are many attributes for the observations then we can group the observations with the help of Clustering Algorithms. Clustering discovers certain patterns for the observations and groups into different clusters. We have many ways of clustering, such as Density-based Clustering, Distribution-based Clustering, Centroid-based Clustering and Hierarchical Clustering. There are different types of algorithms with which we can perform Clustering such as KMeans Clustering, DBSCAN Clustering Algorithm, Gaussian Mixture Model Algorithm, BIRCH Algorithm, Affinity Propagation Algorithm, Mean-Shift Clustering Algorithm, OPTICS Algorithm and Agglomerative Hierarchical Algorithm etc.

Clustering types K means clustering

Among the above algorithms widely used algorithm is KMeans Algorithm, in which each of the observations will be assigned to a group of 'k' categories, where 'k<n' (here 'n' is total number of observations). KMeans Algorithm is simplest among all the clustering algorithms and the best suitable for smaller datasets as it iterates over all the data points or observations. The main problem associated with the KMeans algorithm is it takes lot of time while processing as it considers all the options.

NSE Sectoral indices

An Index is one which reflects the overall behaviour and performance of the companies which are present in that sector. If we take NSE Auto index, it is a composition of 15 top performing companies in Automobile segment in India. Indices are usually used for ascertaining the performance and movement of particular segments, indices will be used for creation of portfolios by individuals as well as firms. There are many categories in NSE with respect to categorization, among those one of the important categorizations for which many of the investors will be willing to know is NSE Sectoral indices. If the investors come to know which sector is performing well, they will invest in that sector. But while considering the performance, most of the people consider "Return" and "Risk" as parameters.

Review of Literature

As per Mansoor, Maryam and Mansour (2015), Clustering is one of the important tools of knowledge discovery in modern machine learning. The Clustering of companies with respect to their performance is not only important for the investors for taking investment decisions, but also for various external parties such as Government, Creditors, Suppliers, Shareholders and other parties who are directly and indirectly involved with those companies. Clustering of companies can be useful for more comprehensive decision making with respect to the companies.

Bini B.S, Tessy Mathew (2016) proposed that validation index can be considered as good technique for assessing the performance of clustering techniques such as partitioning, hierarchical models. Also for the purpose of prediction of stock prices, multiple regression is good as it is best tool for choosing the companies for investment. K Means clustering and EM Clustering have given more accuracy when compared to other clustering tools.

Nanda, Mahanty, Tiwari (2010), has attempted to cluster group of securities with the help of Data Mining. Using different stocks from Bombay Stock Exchange, they have calculated the returns of stocks and then tried to cluster them for the purpose of portfolio creation there by reducing the risk. They have concluded that K-Means clustering can build robust and compact clusters when compared with Fuzzy C-means and SOM. The clustering with KMeans has reduced the portfolio risk when compared to SENSEX. They have also considered Markowitz model for calculation of risk of portfolio in their analysis.

Nguyen and Huynh (2020), used Exploratory Data Analysis(EDA) and Principle Component Analysis(PCA) for clustering the 20 companies from U.S. Stock Markets for a period of one year i.e. from March 2019 to March 2020. They have considered only returns to cluster the 20 selected companies and clustered into 5 groups depending upon the normalization of the returns.

Suresh Babu, Geetanjali and Satyanaraya(2012), used Hierarchical Agglomerative Clustering and Recursive K Means clustering to predict the short term price movements of stocks after the release of financial results of those stocks. In their study they also found that this method has outperformed the Support Vector Machine in terms of accuracy and average profits in terms of prediction. They have used both qualitative factors and quantitative factors for the purpose of analysis.

Methodology

The present study focuses on clustering the sectoral indices of National Stock Exchange. National Stock Exchange has many indices with 6 categories, but we have considered the sectoral indices as many of the investors will be willing to know the performance of a specific sector when compared to different ways of categorization. NSE has 15 indices with respect to sectoral indices which are shown in Table 2. Table 2 contains the basic details of sectoral indices such as index name, methodology of index calculation, number of companies listed in the index, index launching date, base date of index calculation, base value of the index, calculation frequency and index rebalancing time period. The data has been collected from the official site of NSE i.e. www1.nseindia.com.

The primary objective of the study is to cluster the indices depending on the Risk and Return characteristics for a period of 1 year ad 5 years. By this investors, or Asset Management Companies(AMCs) and Mutual Fund companies can create their investment strategies. It also focuses on clustering the indices with respect to the Beta of the indices for the same time period i.e. 1year and 5 years. This work covers the NSE Sectoral indices for the last 5 years i.e. from 2015-2020. Limitations of the study is this covering the Return, Risk and Beta aspects of indices and not considers the other economic aspects. The analysis is carried out with KMeans algorithm and Dendrograms.

Index_Name	Methodology	Num_of_companies	Launch_date	Base_date	Base_value	Cal_frequency	Index_rebalancing
NSE_Auto	Periodic Capped Freefloat	15	12.07.2011	01.01.2004	1000	Online Daily	Semi-Annually
NSE_Bank	Periodic Capped Freefloat	12	15.09.2003	01.01.2000	1000	Online Daily	Semi-Annually
NSE_CD	Periodic Capped Freefloat	15	15.01.2020	01.04.2005	1000	EOD Daily	Semi-Annually
NSE_FinServ	Periodic Capped Freefloat	20	07.09.2011	01.01.2004	1000	Online Daily	Semi-Annually
NSE_Finservs	Periodic Capped Freefloat	20	20.05.2020	01.01.2004	1000	Online Daily	Semi-Annually
NSE_FMCG	Periodic Capped Freefloat	15	22.09.1999	01.01.1996	1000	Online Daily	Semi-Annually
NSE_Healthcare	Periodic Capped Freefloat	20	18.11.2020	01.04.2005	1000	EOD Daily	Semi-Annually
NSE_IT	Periodic Capped Freefloat	10	NA	01.01.1996	100	Online Daily	Semi-Annually
NSE_Media	Periodic Capped Freefloat	10	19.07.2011	30.12.2005	1000	Online Daily	Semi-Annually
NSE_Metal	Periodic Capped Freefloat	15	12.07.2011	01.01.2004	1000	Online Daily	Semi-Annually
NSE_Oil_gas	Periodic Capped Freefloat	15	15.01.2020	01.04.2005	1000	EOD Daily	Semi-Annually
NSE_Pharma	Periodic Capped Freefloat	10	01.07.2005	01.01.2001	1000	Online Daily	Semi-Annually
NSE_Privatebanks	Periodic Capped Freefloat	10	05.01.2016	01.04.2005	1000	Online Daily	Semi-Annually
NSE_PSUbanks	Periodic Capped Freefloat	13	30.08.2007	01.01.2004	1000	Online Daily	Semi-Annually
NSE_Realty	Periodic Capped Freefloat	10	30.08.2007	29.12.2006	1000	Online Daily	Semi-Annually

Table 2. Details of NSE Indices based on sectors

Source : https://www1.nseindia.com/products/content/equities/indices/sectoral_indices.htm

Index_Name	Return_1Year	Return_5Year	SD_1Year	SD_5Year	P/E	P/B	Div_Yield	Beta_1Year	Beta_5Year
NSE_Auto	69.8	4.9	23.54	24.34	211.6	4.77	0.53	1.02	1.09
NSE_Bank	84.1	15.62	29.18	25.26	24.41	3.23	0	1.44	1.24
NSE_CD	77.71	22.6	19.65	20.23	75.65	13.13	0.31	0.73	0.79
NSE_FinServ	75.64	19.14	26.38	24.15	24.51	4.3	0.4	1.34	1.21
NSE_Finservs	78.66	18.75	24.53	23.98	21.74	4.01	0.59	1.27	1.21
NSE_FMCG	22.6	12.93	14.22	17.75	41.6	8.77	2.89	0.48	0.68
NSE_Healthcare	49.08	7.78	20.22	19.92	38.48	5.66	0.53	0.49	0.6
NSE_IT	96.96	21.48	22.25	21.21	29.35	8.42	1.62	0.66	0.69
NSE_Media	49.88	-6.27	30.02	28.12	0	2.36	0.21	1.02	0.92
NSE_Metal	178.3	23.46	30.91	29.23	14.53	2.39	1.66	1.13	1.16
NSE_Oil_gas	61.44	21.83	21.5	22.52	12	2.14	2.37	0.86	0.95
NSE_Pharma	44.56	5.7	22.86	22.18	35.92	5.78	0.42	0.52	0.63
NSE_Privatebanks	76.65	13.87	29.34	25.81	29.04	3.31	0	1.44	1.26
NSE_PSUbanks	114.7	-0.24	38.16	36.48	22.48	0.9	0	1.35	1.2
NSE_Realty	87.14	13.39	31.2	29.99	0	2.79	0.14	1.17	1.11

Table 3. Risk, Return, Beta, P/E, P/B value and Dividend yield of NSE indices

Source : https://www1.nseindia.com/products/content/equities/indices/sectoral_indices.htm

Table 2. contains the details of different sectoral indices in NSE. The methodology of index calculation, the number of companies included in the index, launch date, base value date and index calculation frequency details etc. Table 3 contains the values of 1 year Return & Risk, 5 years Return & Risk and 1 year 5 year Beta values of NSE sectoral indices. It also contains NSE sectoral indices values for P/B Ratio, PE Ratio and Dividend Yield Ratios.

Figure 2. NSE indices 1 year return, 5 year return, 1 year risk, 5 year risk, 1 year Beta and 5 year Beta

Source : https://www1.nseindia.com/products/content/equities/indices/sectoral_indices.htm

Methodology of Clustering

In this analysis, we have opted for two methods of clustering one is Agglomerative (Dendrogram) to ascertain optimum number of clusters and the other is Elbow method, finally clustering has been done with the help of K-Means Algorithm.

We have considered 4 categories of Clustering in this analysis, they are

Case 1:1 year Risk and Return of all indices

Case 2: 5 years Risk and Return of all indices

Case 3: 1 year Beta of all indices

Case 4: 5 years Beta of all indices

For Case1, i.e. for 1 Risk and Return of indices, with the help of Dendrogram, we could get the optimum number of clusters as 3 and the same is the case with Elbow method. For Case 2, i.e. for 5 years Risk and Return got two clusters as optimum, for Case 3, 1 year Beta of indices, we can see only 2 clusters as optimum and for last case i.e. 5 years Beta the optimum number of clusters is 2

Figure 3. Dendrograms showing the optimum number of clusters with respect to 1 Year Risk- Return, years Risk-Return, 1 year Beta and 5 years Beta

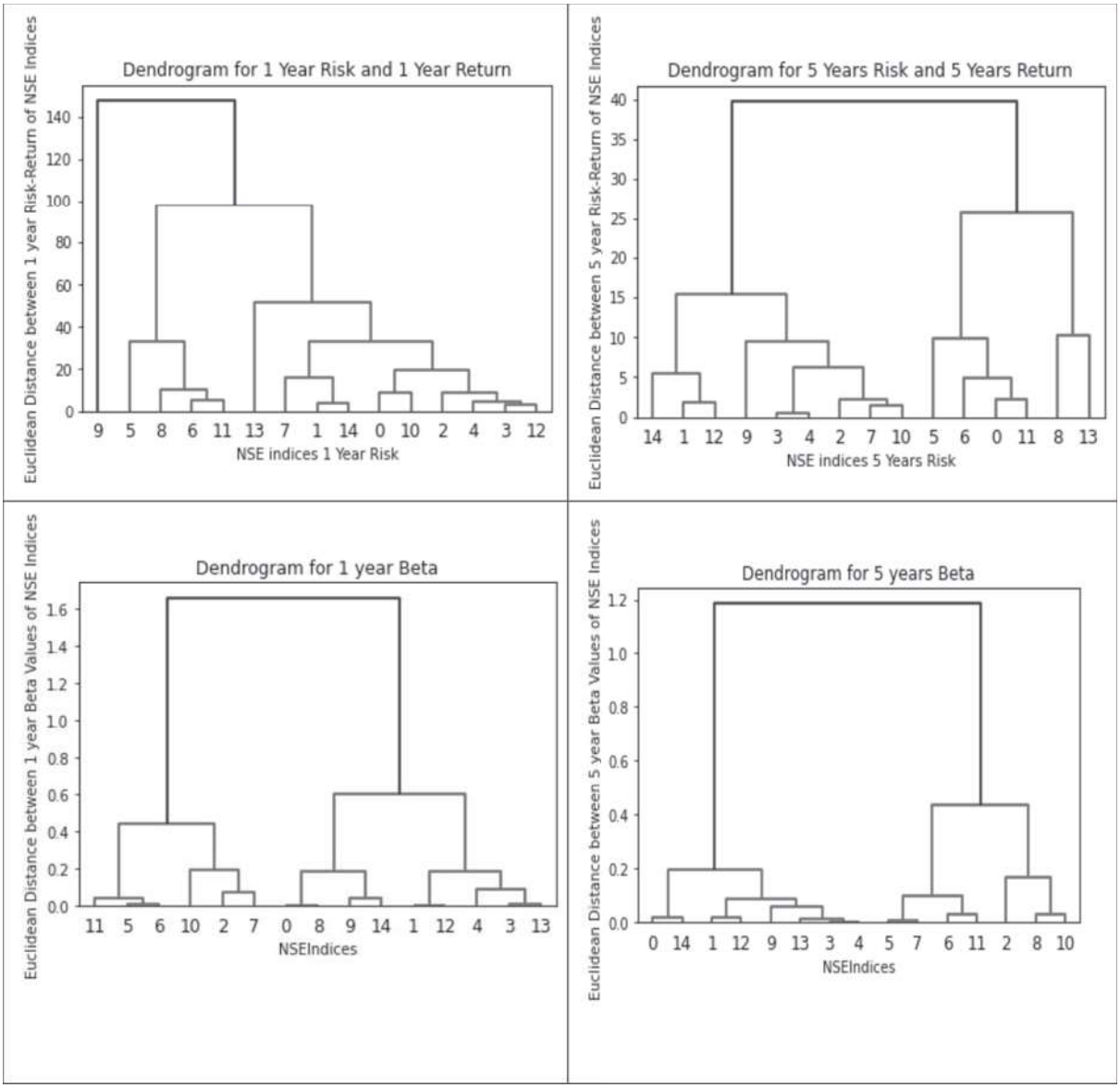
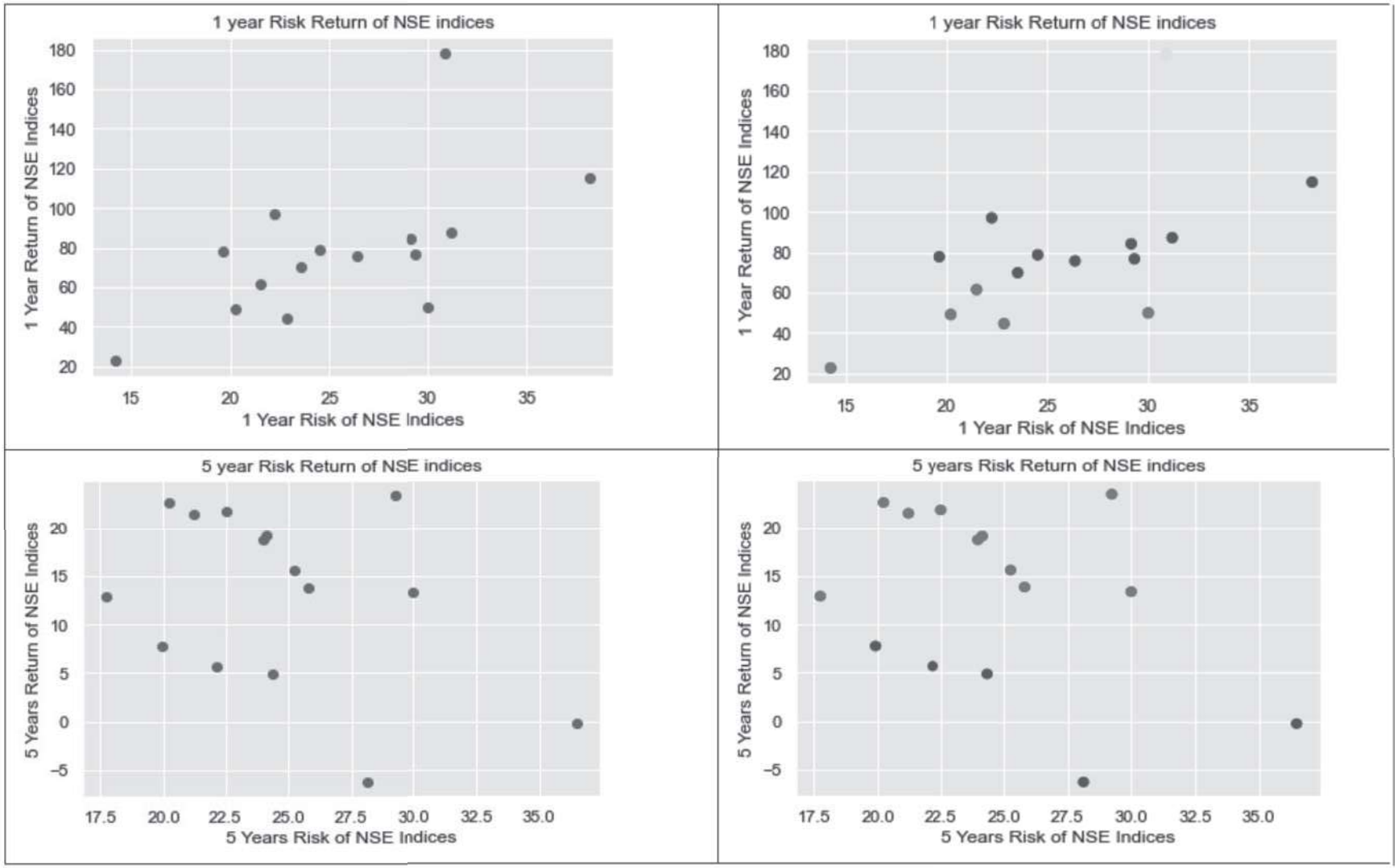


Figure 4. Clustering of NSE Indices with 1 Year Risk and Return & Clusters with 5 Year Risk and Return



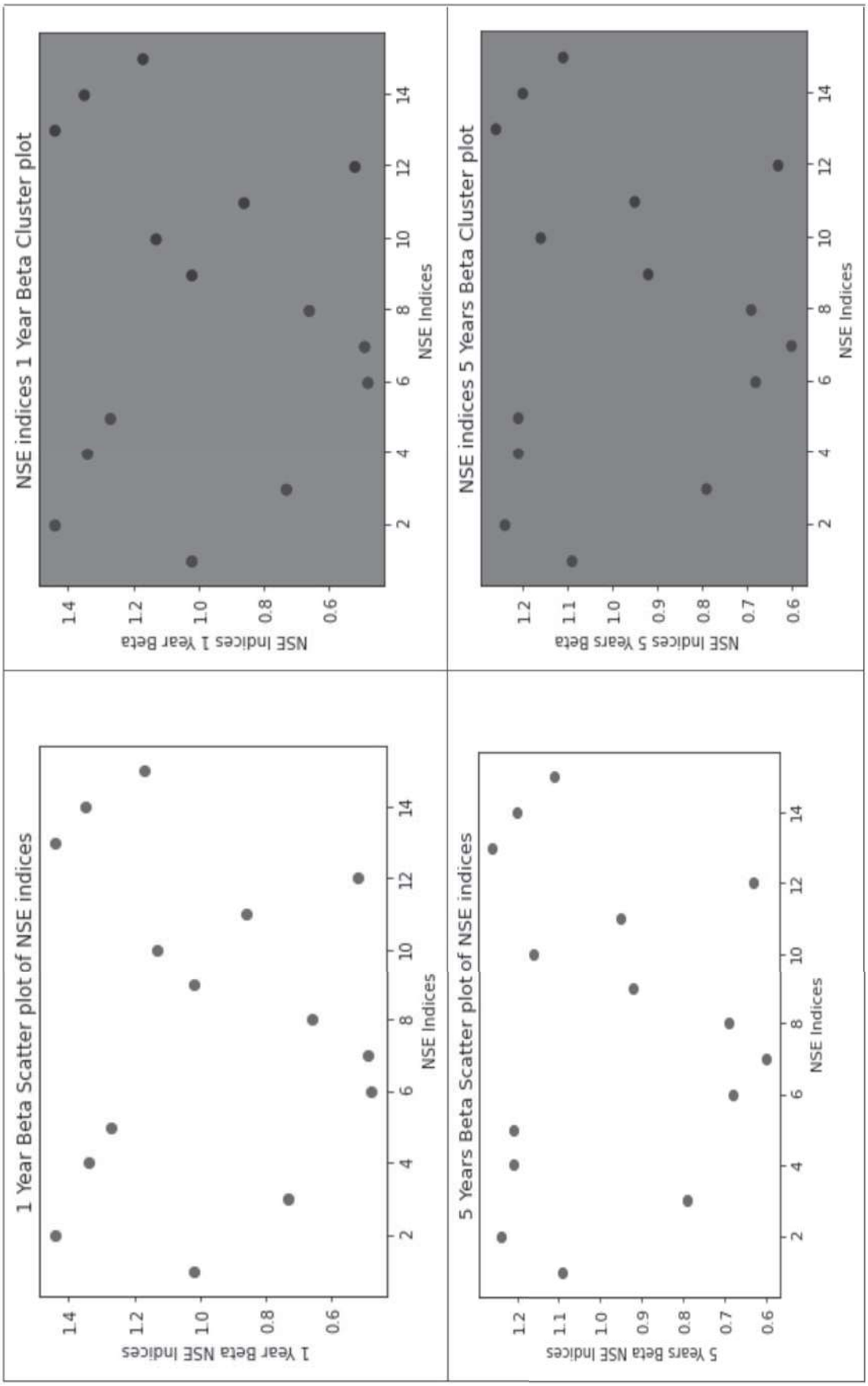


Figure 5. Clustering of NSE Indices with 1 Year Beta & Clusters with 5 Years Beta

Findings and Observations

With the above analysis, we can observe that with respect to the 1 Year Risk and Return, 3 clusters could be made, and with respect to 5 years Risk and Return, 2 clusters could be made which is shown in the Table 4 and Table 5. Figure 3 represents Dendrogram which advises the optimum number of clusters to be made out of NSE indices with respect to the 1 year Return & Risk, 5 years Return & Risk and also the Beta Values for 1 year and 5 years.

Clusters with respect to 1 Year Risk and Return – 3 Clusters

Cluster Number	Indices	SD_1 Year	Return_1 Year
1	NSE_FMCG	14.22	22.60
1	NSE_Healthcare	20.22	49.08
1	NSE_Media	30.02	49.88
1	NSE_Oil_gas	21.50	61.44
1	NSE_Pharma	22.86	44.56
2	NSE_Auto	23.54	69.80
2	NSE_Bank	29.18	84.10
2	NSE_CD	19.65	77.71
2	NSE_FinServ	26.38	75.64
2	NSE_Finservs	24.53	78.66
2	NSE_IT	22.25	96.96
2	NSE_Privatebanks	29.34	76.65
2	NSE_PSUbanks	38.16	114.70
2	NSE_Realty	31.20	87.14
3	NSE_Metal	30.91	178.30

Table 4. Clusters with respect to 1 Year Risk and Return

Clusters with respect to 5 years Risk and Return -2 Clusters

Cluster Number	Indices	SD_1 Year	Return_1 Year
1	NSE_Auto	24.34	4.90
1	NSE_Healthcare	19.92	7.78
1	NSE_Media	28.12	-6.27
1	NSE_Pharma	22.18	5.70
1	NSE_PSUbanks	36.48	-0.24
2	NSE_Bank	25.26	15.62
2	NSE_CD	20.23	22.60
2	NSE_FinServ	24.15	19.14
2	NSE_Finservs	23.98	18.75
2	NSE_FMCG	17.75	12.93
2	NSE_IT	21.21	21.48
2	NSE_Metal	29.23	23.46
2	NSE_Oil_gas	22.52	21.83
2	NSE_Privatebanks	25.81	13.87
2	NSE_Realty	29.99	13.39

Table 5. Clusters with respect to 5 Year Risk and Return

In terms of Beta(Systematic Risk), it is observed that the NSE Indices can be clustered in to two clusters with respect to 1 Year Beta values and two clusters with respect to 5 year Beta values also, which are shown in the Table 6 and Table 7.

Clusters with respect to 1 year Beta – 2 Clusters

Cluster Number	Indices	Beta_1 Year
1	NSE_Media	1.02
1	NSE_Metal	1.13
1	NSE_Oil_gas	0.86
1	NSE_Pharma	0.52
1	NSE_Privatebanks	1.44
1	NSE_PSUbanks	1.35
1	NSE_Realty	1.17
2	NSE_Auto	1.02
2	NSE_Bank	1.44
2	NSE_CD	0.73
2	NSE_FinServ	1.34
2	NSE_Finservs	1.27
2	NSE_FMCG	0.48
2	NSE_Healthcare	0.49
2	NSE_IT	0.66

Table 6. Clusters with respect to 1 Year Beta

Clusters with respect to 5 years Beta – 2 Clusters

Cluster Number	Indices	Beta_5 Years
1	NSE_Media	0.92
1	NSE_Metal	1.16
1	NSE_Oil_gas	0.95
1	NSE_Pharma	0.63
1	NSE_Privatebanks	1.26
1	NSE_PSUbanks	1.20
1	NSE_Realty	1.11
2	NSE_Auto	1.09
2	NSE_Bank	1.24
2	NSE_CD	0.79
2	NSE_FinServ	1.21
2	NSE_Finservs	1.21
2	NSE_FMCG	0.68
2	NSE_Healthcare	0.60
2	NSE_IT	0.69

Table 7. Clusters with respect to 5 Year

Conclusion

Clustering algorithms are new ways to learn the old things and new ways to segregate the data based on which the investors can take timely investment decision. With the help of KMeans clustering the indices can be clustered and the AMCs as well as investors can take investment decisions by identifying the right clusters. As the primary objective of the study is to cluster the NSE indices into the relevant groups so that the investment decision can be taken with relevant

Return and Risk factors. Portfolios can be created by AMC and Mutual Fund companies by observing the clusters and hence the Risk can be mitigated. If the index movement can be predicted for the future, the risk can be still decreased which can be considered as the future scope of this research..

References

Babu, Suresh.M., Geetanjalai.N., & Satyanarayana, B. (2012). Clustering approach to stock market prediction. International Journal of Networking and Applications,3(4),1281-1291

Basalto, N., Bellotti, R., De Carlo, F., Facchi, P., & Pascazio, S. (2005). Clustering stock market companies via chaotic map synchronization. Physica A: Statistical Mechanics and its Applications, 196-206

Bini, B. S., & Mathew, T. (2016). Clustering and regression techniques for stock prediction. *Procedia Technology*, *24*, 1248-1255

Momeni, M., Mohseni, M., & Soofi, M. (2015). Clustering stock market companies via K-means algorithm. *Kuwait Chapter of the Arabian Journal of Business and Management Review*, *4*(5), 1

Nanda, S. R., Mahanty, B., & Tiwari, M. K. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications*, *37*(12), 8793-8798

Nguyen, T., & Huynh, N. (2020). Clustering stock market data using K-means clustering algorithm

Simon, Annina., Singh, Deo Mahima., Venkatesan, S., & Babu, Ramesh D.R. (2015). An overview of machine learning and its applications. *International Journal of Electrical Sciences & Engineering(IJESE),* 1(1), 22-24

Weblinks

1. https://www.niftyindices.com/indices/equity/sectoral-indices
2. www.niftyindices.com