Analysis and Prediction of Stock Market Mining using Machine Learning Clustering Technique

Zahraa Elsayed Mohamed
Department of Mathematics,
Faculty of Science,
Zagazig University
P.O. Box 44519, Zagazig, Egypt

El-Amin Kamal El-Din El-Mesalamy
Department of Mathematics,
Faculty of Science,
Zagazig University,
P.O. Box 44519, Zagazig, Egypt

ABSTRACT
Stock market plays a vital role in a country's economy and it is an important consideration in all the fields due to its potential financial gain. This paper shows that data mining and unsupervised machine learning technique could be used to guide an investor's decisions. A model has been built using data mining future stock price, whether stock price go high or low can be predicted. Moreover, the best clustering indicators in Egypt Stock Exchange for all the 30 companies (EGX30) during first half year of 2019 has been identified.

Keywords
Unsupervised Machine Learning, Data Mining, Clustering, Stock Market, EGX 30 Index

1. INTRODUCTION
Due to high development of data accumulations and business applications. It is need to examine and concentrate valuable information from such data; where it is gathered under the term of data mining which is also called as data discovery or knowledge discovery [1]. Since the dynamic stock market leaves a huge impact on data, securing and checking terabyte data is always a challenge for experts. This can be done via using unsupervised machine learning purports to uncover previously unknown patterns in data. Mostly, these patterns are poor approximations of what supervised machine learning can achieve. The best usage of unsupervised machine learning is when there is no data on desired outcomes. Additionally, unsupervised machine learning technique finds all kind of unknown patterns in data and introduces features that can be useful for categorization, so all the input data to be analyzed and labeled in the presence of learners. Clustering approach in data mining is the most application type of unsupervised learning and represents important concept when it comes to unsupervised learning [2]. It mainly deals with finding a structure or pattern in a collection of uncategorized data. Based on past historical data and current information investors buy or sell shares of listed companies according to studying and analyzing stock data. The investor needs to pass a vast amount of data to discover hidden patterns that are important and not too cumbersome through algorithms using clustering technique [3]. The basic analysis refers to a company's financial information to predict whether it will meet its future expectations of incomes, and helps the investor in the selection of stocks. These analyses are classified by profitability ratios which measure firm's use of its strength in addition to control of expenses in order to generate an acceptable rate of return. This is due to the importance of the profits of a company to investors because these earnings are either retained or paid out in return to shareholders. This approach could be achieved by giving a trustworthy idea of how measuring a company's share price. This measure is called Price/Earnings Ratio (P/E ratio) which is determined by comparing the P/E ratio of a company with the averages in the industries and the markets, where investors can get a sense of the relative value of the stock [4]. This information is relatively static for financial quarter and can easily found on companies' official websites or with stock exchange [5].

The objective of this study is using the data mining techniques to classify the stocks data in the stock market through stock exchange organization. The paper is organized as follows: section 2 provides the related work [6]. The proposed method and the results are introduced in section 3. Section 4 presents a demonstration for the risk assessment and evaluation. Finally, conclusion is derived in section 5.

2. RELATED WORK
Researchers in data mining always try to find techniques that improve the performance of the extraction methods used in data mining by using history of the different transactions done in finding the data as it will be useful for future use. The dataset can be used to predict customer behavior and interests, Basaltoa et al. (2005) applied clustering approach to analyze Dow Jones companies, in order to determine the similar time behavior of quoted stock prices [7]. L.-K. Soon and S. H. Lee (2007) compared the implementation performance of numerical and symbolic representation of data usage in the same search period in order to produce a list of influential stocks on the Kuala Lumpur Composite Index (KLCI) [8]. Tola et al. (2008) underlined the importance of clustering technique in the advancement of the reliability of the portfolio considering the ratio between predicted and realized risk [9]. Chen and Huang (2009) applied cluster analysis to group the huge amount of equity mutual funds based on four evaluation indices in order to help investment decisions. In addition, they offered a fuzzy model, which gives the optimal investment proportion of each cluster [10]. Narayan et al. (2011) examined share price clustering on twelve largest companies listed on Mexican stock exchange and pointed out that volume and risk impact price clustering negatively [11]. Babu et al. (2012) analyzed the main clustering techniques to compare the performances and apply to 35 randomly selected stocks from a number of different sectors in India in order to be able to propose an effective method to predict the stock price movements. They indicated that the hierarchical agglomerative outperforms in terms of accuracy [12]. D'Urso et al. (2013) handled the clustering of financial time series and proposed a new approach which combines fuzziness and GARCH models [13]. M.Wadghule, et al. (2017) studied different methodologies for stock market prediction which will help the investor to making the correct decision for buy or sell the stocks [14].

3. THE PROPOSED METHOD AND THE ANALYSIS OF RESULTS

3.1 The Proposed Method

Stock market values continue to change day by day, hence an effective technique for predicting stock market values is thus implemented using the hybrid combinatory of clustering and

Where Earnings per share provides the profitability indication of a firm, and can be determined by dividing the firm's net income with its whole number of remaining stocks.
Stock Market data set is presented in Table1 and showed in figure 1.

Table 1. Price per earnings ratio of dataset

Company	Sector	P/ E
Amer Group Holding	Financial Services excluding Banks	126.00
Arab Cotton Ginning	Personal and Household Products	36.73
Arab Real Estate Investment CO.-ALICO	Real Estate	36.00
Arabia Investments Development Fin. Inv. Holding Comp.-Cash	Financial Services excluding Banks	15.17
Arabian Cement Company	Construction and Materials	N.A
Belton Financial Holding	Financial Services excluding Banks	10.52
Commercial International Bank (Egypt)	Banks	16.68
Egypt for Poultry	Food and Beverage	279.00
Egyptian Financial Group-Hermes Holding Company	Financial Services excluding Banks	N.A
Egyptian Kuwaiti Holding	Financial Services excluding Banks	4.80
El Ahli Investment and Development	Financial Services excluding Banks	28.85
El Shams Housing & Urbanization	Real Estate	16.18
Electro Cable Egypt	Industrial Goods and Services and Automobiles	62.50
Elsaeed Contracting& Real Estate Investment Company SCCD	Construction and Materials	11.86
ELSWEDY ELECTRIC	Industrial Goods and Services and Automobiles	84.57
Export Development Bank of Egypt (EDBE)	Banks	4.42
Ezz Steel	Basic Resources	15.53
Global Telecom Holding	Telecommunications	N.A
Heliopolis Housing	Real Estate	41.75
Juhayna Food Industries	Food and Beverage	30.11
Medinet Nasr Housing	Real Estate	37.37
Orascom Telecom Media And Technology Holding	Telecommunications	6.19
Oriental Weavers	Personal and Household Products	17.02
Palm Hills Development Company	Real Estate	24.50
Pioneers Holding	Financial Services excluding Banks	17.92
Pyramisa Hotels	Travel & Leisure	14.61
Six of October Development & Investment (SODIC)	Real Estate	N.A
South Valley Cement	Construction and Materials	25.39
T M G Holding	Real Estate	41.04
Telecom Egypt	Telecommunications	7.06

Classification method. The data collected were divided into two parts: qualitative and quantitative; the qualitative data were collected using the judgmental sampling. The quantitative data collection was carried out by means of secondary data, and this includes; list of EGX30 companies and (P/E ratio) of the stocks of the 30 EGX30 companies for the first half year of 2019 from stock exchanges, it is calculated based on following equation (1):

P/E ratio = Market price of the stock / Earnings per share (1)

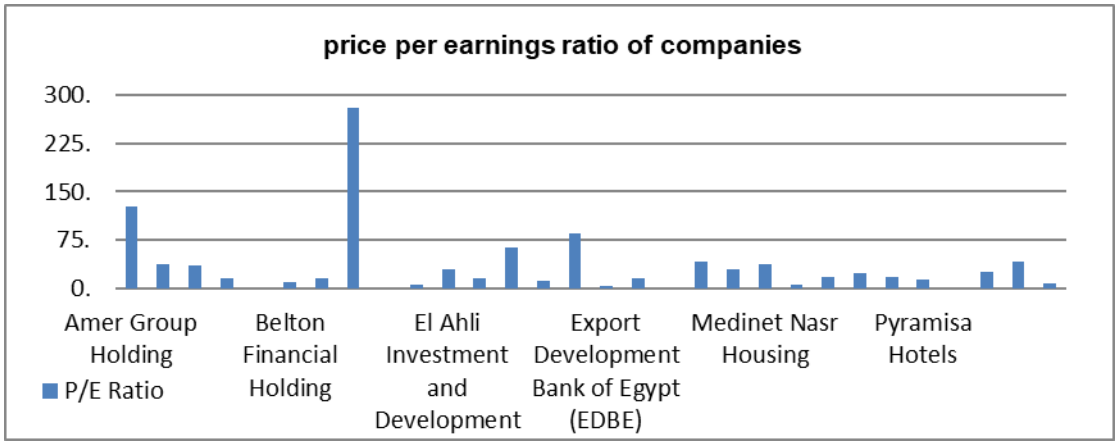

Figure1: Price per earnings ratio of companies

In this section we explain the proposed method (Sequential method) and this clustering, group of data elements may belong to one cluster or more, that is associated to each element as a group of membership levels. It indicates the strength of the relation between data element and a specific cluster. Clustering focus on assigning that membership levels and then use it to set data items to one or more clusters. Clustering algorithms makes the provision of means to make a rational division into small clusters by using a fraction of the work that is needed to calculate all possible combinations defined by sequential algorithms [15]. Sequential algorithms are direct and rapid methods of producing clustering. Typically, feature directives are introduced to the algorithm mentioned below one or several times.

The algorithm:

```
Input: cluster Ch    // cluster with high risk
       cluster Cm    // cluster with medium risk
       cluster Ch    // cluster with low risk
x=component of every cluster,
N =number of stocks in index,
E1, E2 thresholds = PE of stocks in index
Begin
    for each x from 1 to N
      if PE(x) =< E1 then
           Ch= PE(x) // add to cluster Ch
     else if PE(x) > E1 and PE(x) =< E2 then
           Cm= PE(x) // add to cluster Cm
    else if PE(x) > E2 then
           Cl= PE(x)  // add to cluster Cl
    End for
           collect all the clusters Ch, Cm and Cl
End algorithm
```

3.2 The Analysis of Results

Companies mentioned in Table 2 were clustered under sector wise based on (P/E), If P/E ratio < 10, the company does not grow and do not give expected profits / returns and investors lose their investment. Hence, such companies mentioned in the list are not recommended for investment. If P/E ratio 10 to 20, the company growth will take time and investor has to wait to get the benefits from his investment. Hence such company mentioned in the list involves risk for investment. If P/E ratio > 20, the company does well, growth is guaranteed, gives maximum profit and high returns. Hence, such companies mentioned in the list are recommended for investment [16].

The recommended EGX30 company for guaranteed return is reliance power for investment, the end result usually depends on the order of the presentation for (P/E) formative in Table 2 and in figure 2.

In this section we tested stated hypotheses on the data from the Egyptian stock exchange capital market, using the proposed data mining method. Empirical results verified the proposed method and they confirmed the hypotheses. When interpreting the obtained results, we identified the actual behavioral patterns of investors in the Egyptian stock exchange, which may be useful for future trading strategies.

The fundamental analyst uses the openly accessible facts about the stock to perform analysis of stock price movement in three dimensions, concerning the economy, its industry, and financial ratios of the firm which determined by (P/E)

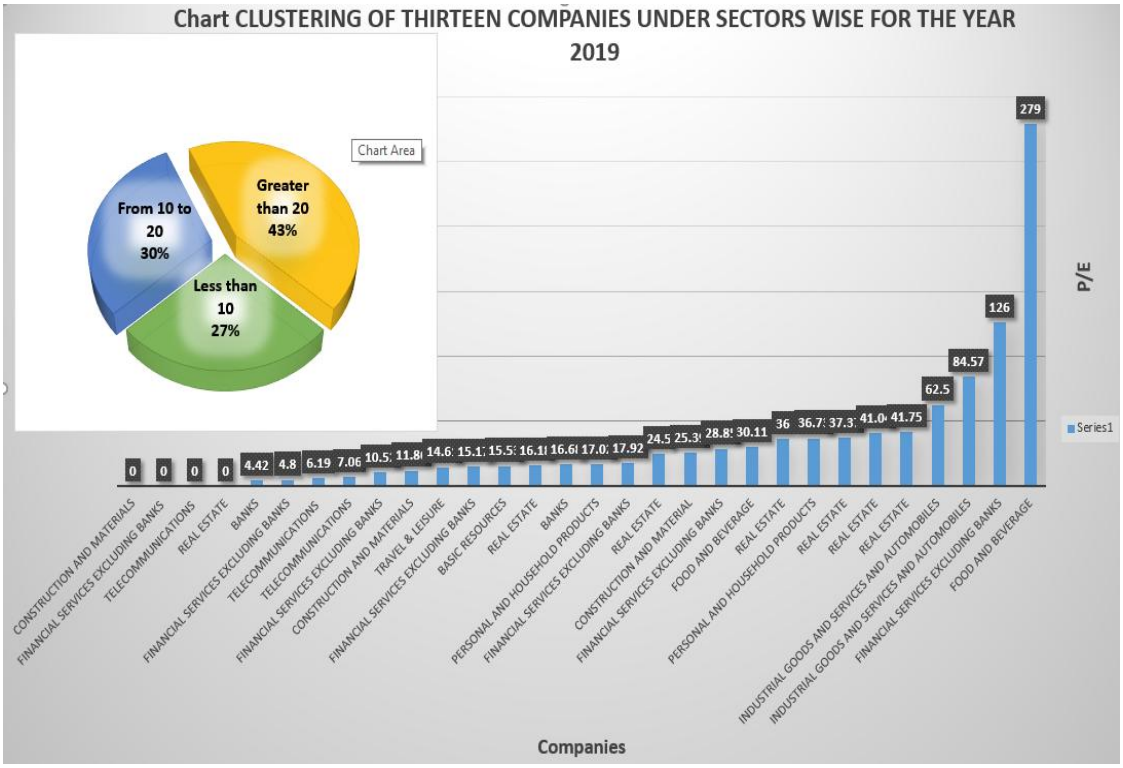

Figure 2: Clustering of thirteen companies

4. RISK ASSESSMENT AND EVALUATION FOR STOCK MARKET

Diversification is a well-known term amongst economists, and several investment companies put diversification theory into consideration when doing investment strategy. Diversification is a way investors can use to reduce overall portfolio uncertainty. If an investor invests in different types of stocks and the overall portfolio risk will be reduced. However, over time, the investor will have a higher backlog than he would get by investing in only one asset. This investment strategy is called diversification [17].

Here, the risks are separated into two categories: Level I (high risk) and Level II (low risk) risk types. They can be considered as (P/E Ratio) of the 30 companies listed in EGX30 as listed in table 2, and If P/E ratio is less than ten (<10), the company is not likely going to grow, and is not expected to have profits or returns, and most likely investors will lose their investments. So it belongs to Level I of risk's category. Hence, investors are not advised to invest in such companies. On the other hand, If P/E ratio is between ten and twenty (10 to 20), the company growth is considered slow, and investors will not benefit soon from this investment, that is why it belongs to the category of Level II of risk. In addition to the previous list of companies, companies with P/E ratio that is greater than 20 are also considered in the category of Level II of risk.

According to the Figure 2, 27% of the companies, are companies with high risk (Level I), while the remaining 73% are of low-risk companies (Level II), the percentage of risk for all the companies of EGX30, is calculated using the below equations (2) and (3):

$$\text{Percentage of risk} = (1 - \text{risk}) * 100 \quad (2)$$

where:

$$\text{Risk} = (0.27 * \text{Level I risk}) + (0.73 * \text{Level II risk}) \quad (3)$$

Clustering algorithm is used to classify the risk levels as low, medium and high, based on the percentage of risk values obtained. A threshold value is set, so that the P/E ratio below the threshold value is high risk and remaining percentage are low-risk.

Table 2. Clustering of thirteen companies under sectors wise for the year 2019

Sector	Price per Sector earnings ratio is		
	< 10	**10 –20**	**> 20**
Banks	Export Development Bank of Egypt (EDBE)	Commercial International Bank (Egypt)	
Basic Resources		Ezz Steel	
Construction and Materials	Arabian Cement Company	Elsaeed Contracting& Real Estate Investment Company SCCD	South Valley Cement
Financial Services excluding Banks	Egyptian Financial Group-Hermes Holding Company	Arabia Investments Development Fin. Inv. Holding Comp.-Cash	Amer Group Holding
	Egyptian Kuwaiti Holding	Belton Financial Holding	El Ahli Investment and Development
		Pioneers Holding	
Food and Beverage			Egypt for Poultry
			Juhayna Food Industries
Industrial Goods and Services and Automobiles			Electro Cable Egypt
			ELSWEDY ELECTRIC
Personal and Household Products		Oriental Weavers	Arab Cotton Ginning
Real Estate	Six of October Development & Investment (SODIC)	El Shams Housing & Urbanization	Arab Real Estate Investment CO.-ALICO
			Heliopolis Housing
			Medinet Nasr Housing
			Palm Hills Development Company
			T M G Holding
Telecommunications	Global Telecom Holding		
	Orascom Telecom Media And Technology Holding		
	Telecom Egypt		
Travel & Leisure		Pyramisa Hotels	

5. CONCLUSIONS AND FUTURE SCOPE

Data mining has been extensively used to extract vital information from history stock data to analyze and predict its future trends. This paper presents stock market related information, process, technical indicators and tools to analyze stock exchange data. Based on this information, confirmed firstly that machine learning technique can be used to predict which stock are likely to go up and proposed system model can be built to formulate various stock trading strategies with suitable data mining techniques to investors' requirements and The results obtained are very encouraging, proving the practical applicability of the stock market. Not only the investors who do financial investments but the organizations which want to predict the sale value of its products based on certain factors can also use this model. There is a scope to create new analysis that can be used for the companies listed under EGX100 and EGX70 indexes of Egypt Stock Exchange.

6. REFERENCES

[1] Ramez, E.; Shamkant, B. N., 2011. "Fundamentals of Database Systems," sixth edition, Addison-Wesley,

Publishing Company, USA ISBN:0136086209 9780136086208.

[2] Xiao, C.; Chaovalitwongse, W. A., 2016. "Optimization Models for Feature Selection of Decomposed Nearest Neighbor," IEEE Trans. Syst. Man Cybern. Syst. 46, 2168–2216.

[3] R. Peachavanish, 2016, "Stock Selection and Trading Based on Cluster Analysis of Trend and Momentum Indicators," Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong, vol. I, IMECS, Mar.

[4] Archana, G.;Pranay, B.;Kashyap, D.;Pritesh, J., 2019, "Stock Market Prediction Using Data Mining Techniques", 2nd International Conference on Advances in Science & Technology (ICAST).

[5] https://www.egx.com.eg .

[6] Shweta, N.;Ritesh, M.;Mirza, B.; Shila J., 2020."Prediction on Stocks Using Data Mining" Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST).

[7] Basaltoa, N.; Bellottib, R; De Carlob, F.;Facchib, P.;Pascazio, S. , 2005. "Clustering stock market companies via chaotic map synchronization," Physica A 345, pp. 196–206.

[8] L.-K. Soon and S. H. Lee, "Explorative Data Mining on Stock Data - Experimental Results and Findings," Lecture Notes in Computer Science, pp. 562-569, 2007.

[9] Tola,V.;Lillo, F.; Gallegati, M.;Montegna, R. N. 2008. "Cluster analysis for portfolio optimization," Journal of Economic Dynamics & Control.

[10] Chen, L. H., & Huang, L. 2009. "Expert Systems with Applications," 37203727.

[11] Nrayan, P. K.,Narayan, S.,Popp,S., & D'Rosario, M. 2011."Share Price Clustering in Mexico, International Review of Financial Analysis," pp. 113-119.

[12] Wang, B.; Wang, X., 2012. "Deceptive Financial Reporting Detection: A Hierarchical Clustering Approach Based on Linguistic Features, Procedia, Engineering" 29, pp. 3392–3396.

[13] D'Urso, P.; Cappelli, C.; Di Lallo, D.; Massari, R., 2013. "Clustering of Financial Time Series," Phsica A 392: pp. 2114–2129.

[14] Y. M. Wadghule and V. R. Sonawane, 2017, "Stock Market Prediction and Forecasting Techniques: A Survey," International Journal of Engineering Sciences & Research Technology (IJESS7).

[15] Jukka, K., 2002. "Clustering Algorithms: Basics and Visualization" HELSINKI UNIVERSITY OF TECHNOLOGY, Laboratory of Computer and Information Science, T-61.195 Special Assignment 1,47942F, 1.8.

[16] D. Venugopal Setty, T. M. Rangaswamy and Dr. A. V. Suresh 2010, "Analysis and clustering of nifty companies of share market using data mining tools" jers/vol. i/issue.

[17] K. Kala and E. Ramaraj, 2013, "ERPCA: A Novel Approach for Risk Evaluation of Multidimensional Risk Prediction Clustering Algorithm," International Journal on Computer Science and Engineering (IJCSE), vol. 5, no. 10.