Adv. Math. Fin. App., 2022, **7(**3), P. 793-809

Advances in Mathematical Finance & Applications

www.amfa.iau-arak.ac.ir
Print ISSN: 2538-5569
Online ISSN: 2645-4610
Doi: 10.22034/AMFA.2021.1877733.1304

Applied-Research Paper

Developing a Prediction-Based Stock Returns and Portfolio Optimization Model

Farzad Eivani[a], Davood Jafari Seresht[*, b], Abbas Aflatooni[a]

[a] *Department of Accounting, Bu-Ali Sina University, Hamadan, Iran*
[b] *Department of Economics, Bu-Ali Sina University, Hamedan, Iran*

ARTICLE INFO

Article history:
Received 2019-10-02
Accepted 2021-11-16

Keywords:
Stock Return Prediction
Portfolio Selection
Decision tree

ABSTRACT

The purpose of this study is to develop a prediction-based stock returns and portfolio optimization model using a combined decision tree and regression model. The empirical evidence is based on the analysis on 112 unique firms listed on the Tehran Stock Exchange from 2009 to 2019. Regression analyses, as well as six decision tree techniques including CHAID, ID3, CRIUSE, M5, CART, and M5 are used to determine the most effective variables for predicting stock returns. The results show that the six decision tree methods perform better than the regression model in selecting the optimal portfolio. Further analysis reveals that the CART model outperforms the other five decision tree models when compared using Akaike and Schwartz Bayesian. This finding is confirmed by comparing the actual returns of the selected portfolio across all six models in 2019. The findings indicate that the predicted returns on portfolio based on the CART model are not significantly different than the actual returns for 2019, suggesting that the selected model appropriately predicts the returns on the portfolio.

1 Introduction

As one of the main pillars of economy in any country, capital market has a significant role in financing and allocating resources to enterprises. Investors provide resource to capital market in exchange for ownership of shares of stock. The main challenge that all investors face is to select an optimal portfolio that produces highest returns and lowest risk. This decision is subject to a large amount of uncertainty as capital market is filled with unobservable parameters and variables. Thus, a constant challenge for investors is to develop ways through which they can predict future returns and select securities that form an optimal portfolio [13]. Prior research has developed various prediction models for stock returns. In the recent years, researchers have been able to increase the power and accuracy of their predictions by incorporating data mining and analysis techniques into their models. Since there are numerous variables that could affect stock returns, selecting the ones with highest prediction power is the key to the best prediction model. Decision tree is a new technique that has become very common in selecting prediction variables and forming optimal portfolios [13].

* Corresponding author. Tel.: +98 912 813 1175
E-mail address: *D.Jafariseresht@basu.ac.ir*

This research seeks to provide a model for selecting effective and appropriate variables for predicting stock returns and building optimal portfolio using regression and data mining techniques. The findings in this research can help determine the influence of variables on stock returns and select the most effective ones among a pool of variables using parametric and nonparametric techniques. Then utilizing scientific methods tested in Iran's capital market, best stock and portfolios in terms of returns are selected. This helps investors and analysts to easily make the best investment decisions with available information. The rest of the paper is as follows: in the next section, prior literature and theoretical framework are discussed. Then, the research method and empirical models as well findings are presented. The final section concludes the paper and makes suggestions for future research.

2 Theoretical Framework and Literature Review

Investors are one of the major sources of financing for companies in the stock market. Their effect on the stock market and the economy is vital. This important role has drawn a significant amount of attention from both regulators and academics in the last two decades. One important challenge in capital markets that concerns both individual and institutional investors is the selection of optimal portfolios. Investors constantly seek ways to form portfolios from lowest risk and highest return securities. In other words, investors view investment returns desirable and consider risk undesirable [1, 3]. Predicted return is a deciding factor in selecting investments and thus having an efficient prediction model is a way to improve the investment process [15, 39]. Given the importance of stock return predictions, scholars have contributed considerable amount of research to developing models to explain and predict stock return [4]. The assumption of predictability of stock returns has been documented and accepted in finance literature [19, 20]. In recent years, research has identified many variables that could facilitate stock returns prediction. An investment strategy, such as portfolio selection, should not only consider the stock return history, but must also consider the future potentials of the stock, highlighting the importance of stock price prediction for investors [38, 18].

Researches so far have developed many models and introduced an extensive list of variables that affect prediction of stock return. The most common method to predict stock return is regression. Regression models are based on the premise that investors are rational. However, researchers in the late 20s have come to the conclusion that investor rationality is no longer a valid assumption. They concluded that capital market behavior is impacted by many unknown and complex factors that make it almost impossible to predict with traditional linear regression models. Researchers believe that nonlinear and dynamic approaches such as data mining can create models that invalidate the previous theories [24]. To overcome the limitations of linear models, experts have used intelligent techniques such as artificial neural networks and genetic algorithms, to improve prediction of stock prices over the past two decades [10, 25]. Given the uncertainty in the stock market and the inability of the mean-variance model in today's markets, it seems necessary to use intelligent techniques to design and develop a specialized system that increases the accuracy of prediction models and ultimately help investors to build an optimal portfolio. In this paper a more efficient model for portfolio selection is proposed. Since the most important part in portfolio management is the prediction of expected return on each stock, this model emphasizes on stock price forecasting using decision tree methods.

2.1 Definition of Returns

Risk and returns are the most important concepts in investment decision making. Each stock or portfolio can yield specific returns if traded within particular periods. These returns include capital gains and dividends. The term “rate of return” is used to explain increase or decrease in investment within holding period. To calculate the rate, the yield on the investment is divided by initial investment. Yield on investment consists of two parts: 1) Increase/decrease in the stock price and 2) various forms of distribution to owners such cash dividend, stock dividends, and stock splits. In other words, the difference between all cash inflows and outflows divided by the cash outflows determines the rate of return on investment [28, 30]. The expected rate return is an important variable in analyzing financial aspects of firms. This variable plays a key role in valuing the firm, portfolio allocation, performance evaluation, risk control, capital budgeting, and other related issues; therefore, accurate measurement of this variable and identifying its components is one of the important issues in financial research [17].

2.2 Portfolio Returns

The expected return on capital (also called cost of capital) is the return that shareholders expect to achieve in order to feel sufficiently compensated. The expected returns on capital depend on factors such as interest rates and company risk [29]; the return on portfolio is equal to the weighted average of the expected returns of all portfolios.

$$\mu = \mathrm{E(Portfolio)} = \sum_{i=1}^{n} x_i E(R_i) = \sum_{i=1}^{n} x_i \mu_i \tag{1}$$

Where: $\mathrm{E(Portfolio)}$ or μ is the return average

$E(R_i)$ or μ_i is the average of i^{th} stock return in a specific period

x_i is the proportional budget value of i^{th} stock in the investment portfolio.

N is the number of stocks under study

2.3 Concept of Portfolio

The term "portfolio" is simply defined as a combination of assets that are collected by an investor for investment purposes. The two fundamental components in portfolio selection are risk and return. In other words, portfolio means the allocation of cash between different securities in a way that the risk and return of the portfolio cross at an optimal point. Given this, portfolio optimization can be considered the process of analyzing the portfolio and managing assets to achieve maximum returns at a certain level of risk [5]. Kaczmarek and Perez [19] show that both mean-variance and HRP optimizers outperform the random forest analysis. This is in contrast with a common criticism of optimizers’ efficiency and presents a new light on their potential practical usage. Davoodi Kasbi et al. [11] find a significant relationship between stock prices and earning per share, e/p ratio, firm size, inventory turnover ratio, and stock return. They also show that Chaid Rule-Based algorithm is a powerful tool that can be used to predict stock prices. Ramesh et al. [31] document that the fine-tuning and high accuracy of market value can be achieved using random forest algorithm.

Oztekin et al. [27] examined the prediction of daily stock returns using three methods namely adaptive neuro-fuzzy inference system, neural networks, and support vector machines. They find that the support vector machines provide more accurate predictions than the other two methods. Delen et al. [12], utilize four decision tree algorithms to evaluate the predictive power of financial ratios with regard to performance evaluation criteria (i.e. return on equity and return on assets). The results show that four financial ratios including income before tax to equity, net profit margin, financial leverage, and sales growth are the most importat ones in predicting return of equity. They also find income before tax to equity, net profit margin, debt ratio, and asset turnover ratio to be the most effective ratios in predicting return on assets. In another study, Uddin et al. [36] use regression analysis to investigate factors affecting stock prices in the Dhaka Stock Exchange. The results show that dividend per share, net asset value, net income after tax, and PE ratio are among the most important variables that affect stock prices. Karami and Talaie [21] studied the relationship between PE ratio, book to market ratio, cash return and investment return among listed companies in Tehran stock market during 1998 to 2007. The results indicate that book to market ratio and investment return have the ability to predict stock returns. Tiwari et al. [35], examined the power of a combined model based on heterogeneous Decision Tree and Markov Model in predicting stock returns in the Mumbai Stock Exchange. The results show that the accuracy of the decision tree model alone is 88.18% and this number is augmented by 3.92% with the decision tree model is combined with Markov model. Soroush Yar and Akhlaghi [33] in a study entitled "Comparative evaluation of the effectiveness of data mining techniques in predicting risk and stock returns of companies listed on the Tehran Stock Exchange" used four data mining algorithms and 16 independent variables to predict stock returns and systematic risk. Using the four most effective variables in predicting risk and return, they find nonlinear separator to be the best model for predicting returns. Wang et al. [37] used a decision tree model to predict stock returns using fifty financial ratios. In their research, they compared the models derived from several decision tree methods. They find that the bagging-decision tree technique outperforms other methods.

Salehi and Farrokhi Pilehrood [32] in a study titled "Earnings management prediction Using Neural Network and Decision tree” examined nine independent variables and 36 companies. They find that both neural network and decision tree methods are more accurate than linear methods. Their results also show that discretionary accruals and non-discretionary accrual as well as risk have the most significant relationships. Hosseinpour et al. [15] in a research titled "Identifying the Financial and Non-financial Variables Affecting the Basis of Audit Report Adjustment, Based on Accounting Estimates: Data Mining Approach" concluded that among three techniques of artificial neural networks, C5 decision tree and support vector machine, the decision tree has the highest prediction capability with an average of 91% accuracy. AliZadeh [13], in a research titled "The effect of macroeconomic variables on Tehran Stock Exchange returns volatility: observations based on the GARCH-X model", showed that the growth rate of money supply and logarithmic changes in exchange rate have a positive and significant effect on stock return volatility but no relation between inflation and stock returns was found. The results also show a significantly negative effect of growth rate of industrial productions on the stock returns instability. Barzegari Khaneghah and Jamali [6] in a study titled "Predicting Stock Returns Using Financial Ratios: Explorations in Recent Research", found that the profitability ratios had higher shares in predicting stock returns than other financial ratios. Their results show that return on equity and return on assets explain the majority of changes in stock returns. Ali Mohammadi et al. [2] investigated four decision tree algorithms (CHAID, ECHAID, QUEST, and CART) to predict stock returns using financial ratios. Their results showed that CART and ECHAID algorithms are the best in predicting current returns and the CHAID algorithm performs better in explaining future returns. Also, the models were more powerful in explaining current returns than predicting future returns.

Hejazi et al. [14] and Delan et al. [12] document that the results of the decision tree models are more reliable for ranking variables than those from regression analysis. In a study using four models based on data mining techniques, Keyghobadi et al. [23] examined the items and ratios for selecting optimal portfolios. The results indicated that the main balance sheet items and profitability ratios are both important for providing optimal portfolios, but their significance is different in each model; however, items such as total assets, profit to income, operating profit to income and share price to dividend ratio are of a higher priority in all models, suggesting that these items could be important indicators for investors. Tavasoli et al. [34] proposed a stock returns prediction model using twelve financial ratios and J48 decision tree algorithm. They tested twelve ratios in form of 12 hypotheses as to whether they were effective in predicting stock returns. Out of the twelve hypotheses, seven were rejected and five were not rejected. Moghadam et al. [26] investigated the power of market ratios in predicting stock returns. The results showed that there is a significant relationship between market price to dividend, market price to book value, market price to selling price, earnings per share and stock returns.

3 Research Hypotheses

The research hypotheses that are designed based on the literature review are as follows:

Main hypothesis 1: There is a difference between the average portfolio returns from the decision tree and regression.

Sub-Hypotheses:

1. There is a difference between the average portfolio returns obtained from the CHAID algorithm and the regression.
2. There is a difference between the average portfolio returns obtained from the CART algorithm and the regression.
3. There is a difference between the average portfolio returns obtained from the ID3 algorithm and the regression.
4. There is a difference between the mean portfolio returns obtained from the CRUISE algorithm and regression.
5. There is a difference between the average portfolio returns obtained from the M5 algorithm and the regression.
6. There is a difference between the average portfolio returns obtained from the M5 algorithm and the regression.

Main hypothesis 2: There is a difference between the average returns of portfolios derived from decision tree algorithms.

Main hypothesis 3: The average portfolio returns predicted by the decision tree algorithm are equal to the actual average return of the portfolio.

4 Research Methodology

The sample used for testing the hypotheses consists of 122 companies listed in Tehran Stock Exchange during the period 2009 and 2019. Systematic method was used for sampling. For this purpose, all the companies in the population that have the following characteristics are selected as the sample and others are excluded.

A. The fiscal year of the companies end in March and the fiscal year has not changed during the research period.

B. The companies have been listed in the Stock Exchange before 2009 and were not removed from the stock exchange by the end of 2019.

C. Sample companies are not financial intermediaries (banks, investment companies, leasing companies).

D. Companies during the research period had been active continuously and had no interruption for more than six months.

E. Data on the research variables is available during the research period.

The purpose of this research is to select optimal portfolios using the most effective variables for predicting stock returns and the best model among decision tree and regression methods. To prepare the data for analysis, the initial data after extraction from the mentioned sources were entered into Excel to calculate the targeted variables. Then, the information calculated variables were analyzed in order to test the research hypotheses in several stages using decision tree and linear regression methods. The research processes are as follows:

Stage One: All variables affecting returns are evaluated using different decision tree methods, and the proposed variables for each method are determined to be used in testing the hypotheses. Also, variables are prioritized according to their importance using regression method.

Stage two: In this stage, the coefficient of significance and the relationship between selected and proposed variables from the previous stage (using decision tree methods) and stock returns will be examined using regression method.

At this stage, each decision tree method classifies and prioritizes the independent variables based on their significance. Also, important and fundamental variables are determined using the regression method. Then, using the significant variables and based on priority, portfolios are selected. Six portfolios are formed based on decision trees and one portfolio based on regression method. By comparing the average returns across all seven portfolios, it can be determined whether the portfolios based on the decision tree methods perform better in terms of predicting return and selecting optimal portfolios.

Stage three: In this step, to determine the best model for predicting the portfolio return, the models from the previous stage will be evaluated using Akaike and Schwartz Bayesian statistical methods. The best method and model will be selected among the six decision tree models.

Stage four: In this stage, predicted returns using the selected model and actual returns are compared to test the robustness of the selected model in predicting the portfolio return.

CART algorithm: classification and regression trees were discovered by Breiman et al. [8]. This model is a binary recursive partitioning procedure capable of processing continuous and nominal attributes as targets and predictors. The continuous processing means that the data is splitted into two subsets based on a variable to increase their homogeneity in each subset compared to the previous subset. These two subsets will then be splitted again, and this will continue until the homogeneity criterion and other stopping criteria meet the stopping rule. The ultimate goal of partitioning is to determine the proper variables with the desired threshold for maximizing the homogeneity of the sample subgroups. This algorithm creates a series of nested pruned trees, each of which is a candidate to be the optimal tree. The optimal tree is identified by evaluating the predictive performance of every tree in the pruning sequence on independent test data. This mechanism automatically and efficiently balances the classes. In other words, the CART method in the decision tree creates its branches in binary form and only based on one field, i.e. each non-leaf node is split into two other nodes. The first step is to determine which of the fields produces the best branch. The best way to create a branch occurs when each resulting branch has a variable which is dominant over other variables. The criterion used to evaluate the branches is diversity. There are many methods for calculating the diversity for a set of records, in all of which high diversity means collections that contain different variables, and low diversity means collections in which a variable dominates other variables. The best way to creating a branch is to minimize the diversity in collections. Next, there are two branches, each containing a set of records (each of the higher node records is located in one of the branches). Now for the branch, same as before, a field is selected again to create the best new branches with the lowest diversity. These steps will continue to the extent that produces nodes in each sub-branch, nodes are created in which the new branch does not reduce the diversity significantly. This final node is called leaf [9].

M5 algorithm: A new method for prediction (when output values are continuous), called Model Tree, was introduced by [28] in the learning algorithm called M5. The tree model combines the traditional decision trees with the probability of linear regression functions in the leaves. This method is relatively clear, since the decision structure is obvious, and regression functions normally do not include a large number of variables. The model was expanded by Quinlan [28] by combining the tree with the nearest neighboring models. Tree-based models are created using divide and conquer methods. The T set is also associated with a leaf, or some tests are selected to separate T into relevant subsets to test the outcomes, and the same process is performed recursively for subsets. This division often creates a lot of detailed structures that need to be pruned (for example, by replacing a tree with a leaf).

CHAID Algorithm: This method is a very effective statistical technique developed by Kass [22]. This method is a decision tree based on identifying the strongest relationships between independent and dependent variables, and for this purpose, the probability of the chi-square statistic is used to test the independence of contingency tables. In this method, among the existing variables, a variable with a smaller P-value is considered in the first step for divisions on a node [22]. The distinction of this algorithm with other decision tree methods is that CHAID can create more than two classes of trees per level. As a result, this algorithm is not a binary tree method. Therefore, the tree created in this method is broader than other methods. The output of this algorithm is very objective and its interpretation is simple, because in this method, by default, multiple branches are used. On the other hand, the weakness of this method is due to its inability to create optimal divisions based on existing variables [7].

ID3 Algorithm: One of the most commonly used symbolic learning methods is the decision tree induction, which was first developed by J. Ross Quinlan in 1986 as the ID3 algorithm, which is known as the Iterative Dichotomiser 3. ID3 is a commonly used method for categorizing symbolic data and is not suitable for numerical data. The ID3 algorithm has been proven to create a fuzzy decision tree as a general and effective algorithm for constructing decision trees from a set of data with discrete values [16]. In the decision tree, the ID3 uses a statistical value called the information gain to determine how much a feature is able to separate the learning examples based on their classification. The ID3 algorithm is very suitable for introducing and constructing a tree with multiple divisions in each node. This algorithm is designed for qualitative variables, but it can be used for a set of variables, both qualitative and quantitative. The decision criterion for this algorithm is based on the entropy index, which calculates the Information Gain and Gain Ratio indices. The quick, concise, useful and reliable results of this algorithm have made it an acceptable method for classifying observations used in medical science [27].

CRUISE Algorithm: This algorithm which was introduced by Kim and Loh [24] can develop a classification tree with multiple divisions. This algorithm works well with methods like CART, but it has a higher speed due to the use of multiple divisions, and a smaller tree is formed using this algorithm. The developed tree is unbiased in this way and is designed to work well despite missing values for data.

M5’ algorithm: Wang and Witten [38] introduced a new application of a model tree based on the Quinlan model [28] called the M5’ which works better than the previous model. The new model dramatically reduces the size of the tree, but slightly reduces the performance of the prediction. According to Wang and Witten, some of the details are not completely addressed and resolved in the M5. Additionally, handling the features (variables) with enumerated attribute and missing values must be specified. This model is described further as follows:

A. Variables with enumerated attribute: Before constructing a model tree, all enumerated attributes are transformed into binary variables. For each enumerated attribute, the average class value corresponding to each possible value in the enumeration is calculated from the training examples, and the values in the enumeration are sorted according to these averages. Then, if the enumerated attribute has k possible values, it is replaced by k-1 synthetic binary attributes. Thus, in M5' all splits are binary: they involve either a continuous-valued attribute or a synthetic binary[37].

B. Missing values: In order to take into account the missing values, the SD function in the M5 algorithm, is renamed as SDR and modified as follows [37]:

$$\mathrm{SDR} = \frac{m}{|T|} \times \beta(I) \times \left[\mathrm{SD}\ (T) - \sum\nolimits_{J \in \{L,R\}} \frac{|T_J|}{|T|} \times \mathrm{SD}(T_J)\right] \tag{2}$$

M denotes the number of examples without missing value for that attribute and T is the set of examples that reach this node. β_i is a moderating factor (which exponentially decreases with increasing number of values). T_L and T_R are collections that have been created as a result of decomposing this feature.

4.1 Research Variables

Based on previous research, the items and financial ratios that are effective in predicting stock return are as follows:

Table 1: Research Variables

No.	Variable name	Symbol	Formula
1	Earnings per share	EPS	Net profit on Weighted average of ordinary shares
2	Company Size	SIZE	Natural Log of the Commercial unit total assets
3	Book to market value	B / M	Book value of each share on the market price per share
4	Price to earnings per share	P / E	End of period Market price per share on earnings per share
5	Operating profit	EBIT	Enrings before interest and tax
6	Net profit after tax	NPAT	Net profit minus tax
7	Average dividend per share	DPS	Dividends approved by the General Assembly
8	Current ratio	CR	Current assets divided by current liabilities
9	Quick ratio	QR	(Current assets- inventories) divided by current liabilities
10	Debt ratio	DR	Total debt on the total assets
11	Return on sales or net profit ratio	ROS	Net profit divided by net sales
12	Gross profit ratio	GPR	Gross profit divided by net sales
13	Ratio of operating profit to sales	OPR	Operating profit divided by sales
14	Return on assets	ROA	Net profit on total assets
15	Return on equity	ROE	Net profit on equity
16	Volume of transactions	V	Volume of transactions over a period
17	Total assets turnover	AT	Net sales divided by total assets
18	Dividends per share to price	DP / P	Cash dividends divided by share price
19	Cash flow per share	CFPS	Operating cash flow on the number of shares
20	Cash return on Sales	CROS	Operating cash flow on the sales
21	Cash return on equity	CROE	Operating cash flow on equity
22	Cash return on assets	CROA	Operating cash flow on the total assets
23	The inflation rate	IN	According to Central Bank reports
24	Exchange rate in the free market	EX	According to Central Bank reports
25	Interest rates in the economy	IR	According to Central Bank reports
26	Rate of liquidity growth	RCASH	According to Central Bank reports
27	Earnings quality	EQ	Operating cash flow on earnings before interest and tax
28	Dividend yield trend	MD	Change in the dividend during a specified period of time
29	GDP	GDP	According to Central Bank reports
30	Oil prices	OP	According to Central Bank reports
31	Book value of shares	BV	Total equity divided by weighted average of ordinary shares
32	Price per share	P	Market value of each share
33	Operating profit to total Assets	OPA	Operating profit divided by total assets

Table 1: Research Variables

No.	Variable name	Symbol	Formula
34	Institutional owners rate	IOR	Number of Institutional owners shares on ordinary shares
35	Systemic risk	B	Portfolio yield with market return covariance divided by market return variance
36	Fixed asset turnover	FAT	Sales divided by fixed assets
37	Accounts receivable turnover	ACT	Net sales (credit) divided by average accounts receivable
38	Inventory turnover	IT	Cost of goods sold divided by the average inventories
39	Earnings per share growth rate	PEGR	The change in earnings per share during a specified time period
40	Sales growth rate	SGR	The change in the sales during a specified time period
41	Net profit growth rate	NPGR	The change in net profit during a specified time period
42	Dividend per share growth rate	DPR	Dividend per share divided by earnings per share
43	Tobin’s Q	TQR	Company's market value divided by total Assets
44	Liquidity	LI	The number of shares traded divided by the total issued stock
45	Earnings volatility	VE	Average absolute change in profits
46	Interest coverage rate	ICR	Profit before interest and taxes divided by interest cost
47	Net working capital	NWC	Current assets minus current liabilities
48	Cash ratio	CFR	Cash and equivalents divided by current liabilities
49	Independent Auditor opinion	AO	Unqualified, Qualified, Adverse and Disclaimer of Opinion
50	Assets growth rate	AGR	Change in the assets during a specified period

5 Results

5.1 Descriptive Statistics

Table 2 reports the descriptive statistics for the main variables. As can be seen in the table, the highest value of stock returns is 7.39 and its mean is 0.55, which shows that the companies in this study do not have high stock returns. The median of P/E ratio for each share is 19.03, which is an acceptable value. Standard deviation is 0.89 for the systematic risk, suggesting low distribution. The median of the variable Tobin's Q is 1.19, which indicates that the data is centred around this point.

Table 2: Descriptive Statistics for the Main Variables

Variable	Min	Max	Median	Mean	STD
RETURN	-0.63	7.32	0.56	0.56	1.13
B	-2.61	5.53	0.71	0.69	0.91
AT	0.03	3.37	0.85	0.78	0.45
ICR	-410	143245.1	240.32	2.22	5122.23
P/E	-161	1168.02	19.18	5.39	80.33
EBIT	-5055419	39651456	821421	795146	2935014
NPAT	-7204845	30884510	660407.8	0.11	2612156
TQR	0.05	6.79	1.23	0.94	0.95
AGR	-781.07	112045.4	140.38	0.13	4110.81
NPGR	483.13	1532.23	3.80	-0.23	71.22

5.2 Decision Tree Methods

5.2.1 CART Method

This algorithm includes two methods of classification and regression. For the classification method, the dependent variable should be a group and class variable, but for regression, there should be a numerical and continuous variable. In this research, for the classification method, the dependent variable (return) is divided into two groups of low and high, and then the algorithm is implemented. Companies with lower returns than the average, are in the low group and companies with higher returns than the average, are in the high group. After assigning the data to the training and test groups, the decision tree is drawn and the variables that are effective in the fitted model of the training group are specified. The initial 6 years data has been considered as training data.

Confusion matrix for test data: This matrix indicates that the data considered as experimental has been properly classified by a model that has fitted training data. Model accuracy and classification error: the confusion matrix criterion for test data is one of the main criteria to compare different models and algorithms, and the more accurate the prediction the better. In other words, it shows what percentage of data is correctly classified or predicted. Out of 112 companies, 51 companies are well classified, thus the accuracy of the model is 46%.

Table 3: The Significance of Variables Based on Classification Method

Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.
TQR	100	CROE	56.17	CFR	39.04	B.M	25.67	OP	19.62
DP.P	92.06	ROS	51.66	ICR	38.65	DR	25.11	SGR	18.71
P	83.14	CR	51.16	BV	37.77	AGR	24.25	AO	14.74
ROE	81.73	QR	50.41	ACT	36.65	ROA	24.16	Size	14.67
MD	74.62	OPA	47.24	VE	33.86	IT	23.78	EQ	13.1
EPS	63.51	NPGR	44.47	CFPS	31.21	FAT	22.64	GPR	10.5
P/E	60.36	PEGR	43.11	EBIT	29.81	LI	21.7	OPR	8.2
B	60.26	IOR	42.25	CROS	29.55	NWC	20.66	NPAT	7.82
V	59.66	DPR	39.42	AT	29.02	DPS	20.17	CROA	5.57
Classification error		53.28%			Model accuracy		46.30%		

CART Regression Method

MAE is the mean absolute error, MSE is the mean square error and RMSE is the Root mean square error. All three measures calculate the distance between the real value and the predicted value. A lower and near zero value indicates the suitability and accuracy of the model in prediction. Given the MSE value, this model is good and suitable.

Table 4: The Significance of Variables Based on Regression Method

Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.
B	100	NPAT	30.94	GPR	18.24	B.M	9.85	V	5.82
AT	58.02	DR	27.75	CFR	18.16	ACT	8.4	CROA	4.38
ICR	51.14	Size	24.77	OPA	17.05	DP.P	8.11	SGR	4.1
TQR	40.16	EPS	24.41	BV	15.44	MD	7.96	CR	3.48
AGR	39.65	ROE	24.34	VE	15.29	OPR	7.96	DPS	2.52
NPGR	35.39	P	23.9	PEGR	12.59	QR	7.96	FAT	1.51
ROA	34.36	DPR	22.97	EBIT	11.85	IT	6.68	CROE	0.75
P/E	31.23	ROS	22.91						
MAE	0.02147			**MSE**	0.00653			**RMSE**	0.080808

5.2.2 CHID Method

The CHAID algorithm is performed using Rapid Miner software. To perform this algorithm, all the variables are required to be grouped. Therefore, all independent variables are classified into three groups: Low, Medium, and High (Divided into three equal intervals). Also, the dependent variable (return) is splitted into Low and High groups and then the algorithm is performed. Companies with lower return than average are in the Low Group, and companies with higher returns than the average are high groups. After assigning the data to the training and test groups, the decision tree is constructed. The initial 6-year data has been considered as training set.

Table 5: The Significance of Variables Based on CHID

Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.
OPA	100	DPR	47.75	AT	32.78	CFPS	17.77	Size	5.08
B	90.68	TQR	46.23	BV	28.53	PEGR	14.86	OPR	3.64
LI	51.62	CROA	40.43	IOR	21.73	P	14.17	FAT	1.34
EPS	50.74	DP. P	34.47	ROS	21.67	NWC	12.87	ROA	-0.05
IT	47.92								
Classification error		32.84%					Model accuracy		65.13%

The confusion matrix shows that from 112 companies, 74 companies have been classified correctly and 38 companies are in the wrong group. The accuracy of this method is thus 66% and the classification error is 34%.

5.2.3 CRUISE Method

The CRUISE decision tree algorithm was performed using CRUISE v3.6.4 software. For this purpose, it is necessary that the dependent variable is a group and class variable; therefore, the dependent variable (return) is divided into two groups of low and high, and then the algorithm is performed.

Table 6: The Significance of Variables Based on CRUISE

Variable	**Sig.**	**Variable**	**Sig.**	**Variable**	**Sig.**	**Variable**	**Sig.**	**Variable**	**Sig.**
DP.P	100	B.M	57.38	AT	36.08	CROE	26.88	AGR	22.28
AO	95.38	GPR	48.38	OP	35.68	LI	25.68	QR	21.98
P	94.78	ROS	47.38	ACT	35.68	BV	25.58	CFR	20.18
TQR	92.08	CR	47.08	OPR	35.58	FAT	23.78	PEGR	19.48
P/E	89.88	IT	42.68	OPA	35.38	DPS	23.48	NWC	19.28
SGR	80.58	DR	39.78	VE	34.58	ICR	23.18	V	18.68
ROA	69.78	NPGR	39.58	B	30.48	CROS	22.98	CFPS	16.38
EPS	65.88	ROE	38.68	CROA	30.18	EQ	22.98	NPAT	14.18
DPR	62.48	MD	38.48	IOR	30.08	Size	22.98	EBIT	8.58
Classification error		45.11%					Model accuracy		57.45%

Companies with lower returns than the average are in the Low group, and companies with higher returns than average, are in the High group. After assigning the data to the training and test groups, the decision tree has been constructed. The initial 6-year data has been considered as training set. Based on the confusion matrix, 65 companies out of 112 companies are predicted accurately, and 47 companies are predicted in the wrong group. As a result, the accuracy of this model is 58 and its classification error is 42%.

5.2.4 ID3 Method

The ID3 algorithm was performed using Rapid Miner software. For this purpose, it is necessary that the dependent variable is a group and class variable; therefore, the dependent variable (return) is divided into two groups of low and high, and then the algorithm is performed. Companies with lower returns than the average are in the Low group, and companies with higher returns than average, are in the High group. After assigning the data to the training and test groups, the decision tree has been constructed. The initial 6-year data has been considered as training set. Based on the confusion matrix for training data, 60 companies out of 112 companies are in the correct group and 52 companies in the wrong group; therefore, the model accuracy is 52 and the model error is 45%.

Table 7: The Significance of Variables Based on ID3

Variable	**Sig.**	**Variable**	**Sig.**	**Variable**	**Sig.**	**Variable**	**Sig.**	**Variable**	**Sig.**
CROA	100	EBIT	46.59	V	23.87	NPGR	10.23	Size	10.23
DR	73.87	IOR	37.5	CFR	23.87	CR	10.23	BV	10.23
LI	60.23	NPAT	37.5	B	19.32	CROE	10.23	DPS	5.69
EPS	55.69	DPR	37.5	P	14.78	AO	10.23	PEGR	5.69
AT	51.14	CFPS	32.96	TQR	10.23	AGR	10.23	ACT	5.69
FAT	46.59	NWC	23.87						
Classification error		45.22%					Model accuracy		52.03%

5.2.5 M5 Method

The M5 algorithm is performed using the RWeka package in R software. For this purpose, it is necessary that the dependent variable is a group and class variable; therefore, the dependent variable (return) is divided into two groups of low and high, and then the algorithm is performed. Companies with lower returns than the average are in the Low group, and companies with higher returns than average, are in the High group. After assigning the data to the training and test groups, the decision tree has been constructed. The initial 6-year data has been considered as training set. Based on the confusion matrix 62 companies are predicted in the correct group and 50 companies in the wrong category. As a result, the model accuracy is 56 and the model error is 45%.

Table 8: The Significance of Variables Based on M5

Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.
P	100	NWC	53.71	SGR	38.79	B	28.91	ACT	14.81
EPS	81.13	NPGR	46.86	CROA	38.66	V	27.45	Size	14.46
OPA	78.65	ROA	45.91	CFR	37.87	VE	27.37	QR	13.17
ROS	75.29	MD	41.57	EBIT	37.7	GPR	27.34	LI	10.09
TQR	68.36	IT	41.26	CROE	37.2	CROS	26.2	DP.P	7.34
P/E	62.58	CR	40.65	ICR	36.53	EQ	26.2	IOR	6.84
NPAT	62.42	AO	40.34	B.M	36.39	DPR	24.75	OP	4.48
DPS	58.27	BV	39.17	FAT	36.3	OPR	22.25	AT	4.33
DR	57.67	CFPS	39.11	AGR	34.09	PEGR	16.86	ROE	1.94
Classification error		45.04%					Model accuracy		56.16%

5.2.6 M5' Method

The M5 decision tree algorithm is performed using Rapid Miner software. To perform this algorithm, the variables are required to be numeric data. After assigning the data to the training and test groups, the decision tree has been constructed. The initial 6-year data is considered as training data. MAE is the mean absolute error, MSE is the mean squared error and RMSE is the root mean square error.

Table 9: The Significance of Variables Based on M5'

Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.
NPGR	100	TQR	39.34	OPA	29.52	CFPS	21.49	P	11.38
PEGR	94.18	LI	39.3	OPR	28.51	B.M	20.49	CR	10.72
DR	87.38	DPS	38.52	EPS	24.06	GPR	19.25	ACT	10.38
CROE	51.79	Size	37.15	CROS	23.2	ROA	17.75	B	8.52
SGR	49.78	DP. P	36.1	EQ	23.2	QR	17.09	ROS	6.96
BV	46.91	VE	34.27	FAT	22.84	CROA	16.92	IOR	3.41
NPAT	44.69	EBIT	31.69	AT	22.53	NWC	15.11	AO	1.69
P/E	41.77	ROE	30.27	MD	21.59	CFR	12.09	DPR	1.27
MAE	0.03234			**MSE**	0.00721			**RMSE**	0.08491

All three measures calculate the distance between the real value and the predicted value. A lower and near zero value indicates the suitability and accuracy of the model in prediction. Given the value of MSE, this model is accurate and suitable.

Table 10: The Significance of Variables Based on Regression

Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.
B.M	100	IT	66.76	CFR	46.33	TQR	17.67	EBIT	7.8
P/E	92.58	Size	64.64	B	45.06	QR	17.5	ACT	7.74

Table 10: The Significance of Variables Based on Regression

Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.	Variable	Sig.
OPA	90.35	AT	63.95	EPS	32.1	V	17.23	CFPS	7.04
PEGR	90.07	P	57.89	GPR	30.93	IOR	16.49	ROS	6.98
OPR	88.38	ROA	52.57	DP.P	30.83	ROE	14.08	MD	5.51
OP	83.92	SGR	51.95	AGR	30.42	NWC	13.82	FAT	5.12
ICR	80.54	LI	50.47	VE	25.56	CR	11.01	DR	2.07
NPGR	73.5	CROA	49.59	CROE	20.06	DPS	9.83	NPAT	-0.86
BV	71.98	AO	46.73	DPR	19.37				

5.3 Data Analysis and Hypothesis Testing

After identifying the significant variables in each decision tree model and the regression method, according to existing data up to 2018, a portfolio is constructed. The output of the software arranged variables based on the Gini coefficient. For ease of use and comparison, using the relative significance of all variables in different methods, their significance has been identified out of 100. To select the number of significant variables in each method, the significance level of more than 30 was considered.

Table 11: Comparison of Different Decision Tree Models and Regression Method

No.	Data mining methods	No. of Sig. variables	No. of firms in portfolio	Data mining average return	Regression return (24 variables and 50 companies)
1	CART(reg)	9	78	25.667	14.531
2	CART(class)	9	47	16.768	14.531
3	CRUISE	26	50	22.892	14.531
4	M5	26	47	19.460	14.531
5	CHAID	11	44	22.809	14.531
6	ID3	11	58	19.473	14.531
7	M5prime	16	82	17.184	14.531

Table 12: Selecting the Best Model Based

No.	Decision tree method	AIC	BIC
1	CART(reg)	1041	1098
2	M5P	1054	1134
3	CHAID	1048	1105
4	CRUISE	1053	1189
5	ID3	1069	1131
6	M5P	1054	1191
7	CART(class)	1062	1198

To select the proposed portfolio in classification method, all high-return companies have been selected. For numerical methods, positive-return firms have been selected as portfolios. Using the actual returns in 2018, the average returns of each portfolio are calculated in different decision tree methods as well as in the regression model and used to test the first hypothesis. Independent t-test is performed to compare the averages in two methods of data mining and regression with each other. Given that the p-value is less than 0.05 (0.002602), the null hypothesis is rejected, suggesting a significant difference between the mean of the two methods, and therefore the hypothesis 1 cannot be rejected. Based on the results obtained, decision tree methods were better than the regression method, and the actual returns of their proposed portfolios were higher than that of the regression method. Now, in order to choose one of these methods, based on the number of proposed variables, we form models with return on one side the proposed significant variables of that method on the other side. In other words, according to the significant variables of each decision tree model, a linear regression model is formed and the best model is chosen using the Akaike and Schwartz Bayesian statistics.

As can be seen, based on both indicators, the best method in the tree decision models is the CART method with the regression branch. The number of significant variables of this method is 9. As a result, the best model for choosing the optimal portfolios is the CART model with regression branch with 9 significant variables namely B, AT, ICR, TQR, AGR, NPGR, ROA, PE and NPAT, respectively. When analyzing the first hypothesis, the portfolios of this method have the highest returns among all methods. To test the selected and proposed model as stated in hypothesis 2, which selects the CART model with regression branch as a better model, the expected return on each share and the average predicted returns for the year 2019 are calculated. Then, based on the actual return in 2019, the average actual return of the portfolio was calculated and using t-test the third hypothesis was tested.

Table 13: Comparison of Actual and Predicted Returns of Companies in 2019

Companies average predicted returns in 2019	**Companies average actual returns in 2019**
20.27%	24.18%

The t-test is used to compare the mean of the predicted and actual data groups of 2019; considering p-value = 0.3248, the null hypothesis is accepted implying that there is no significant difference between the two groups.

6 Discussion and Conclusions

This study examined one of the most important and fundamental challenges investors face in their investment decisions. Prediction of stock returns has always been one of the most attractive issues in financial research. Two factors affecting the decision making of investors on the stock exchange are risk and return. To reduce the risk, a portfolio is formed and at the equal risk conditions a portfolio that yields more return is selected. This research emphasized the returns on portfolio and seeked to provide an appropriate model for predicting stock returns and creating optimal portfolios given the most significant variables. By reviewing the domestic and international research, a list of 44 accounting variables and 6 economic variables shown to affect returns was created. Due to the weakness of linear and parametric methods, nonlinear and non-parametric data mining method have been used, which nowadays is one of the most important aspect of empirical research in different fields. The decision tree method is one of the newest data mining methods and has been relatively overlooked in Iran.

The results of the research indicate that the six methods of decision tree models are better than regression method. The returns on portfolio that was formed based on all decision tree methods is greater than the returns on the portfolio formed on the basis of the regression method. As expected, decision tree methods that are non-parametric and nonlinear performed better than regression methods, which is a classical and linear method. Therefore, investors and analysts should take advantage of data mining methods instead of regression methods when predicting stock returns. This result is consistent with the results of research by Oztekin et al. [27]. The CART method with regression branch is chosen as the best model for predicting stock returns. The significant variables of this method are as follows according to their importance, 1. Systemic risk, 2. Asset turnover, 3. Interest coverage rate, 4. Tobin's Q, 5. Assets growth rate, 6. Net profit growth rate, 7. Return on assets, 8. Price to earnings per share, and 9. Net profit after tax. Out of six methods of the decision tree, one that was superior to the other methods should be selected. Based on the returns on selected portfolio, the CART method with regression branch method was selected. One of the advantages of these models is that they require only a small number variables and thus are very cost effective. The results of this study are consistent with the results of the research by Ali mohammadi et al. [2].

The findings in this study suggest that data mining methods be used to predict stock returns and optimum portfolio should be selected using lower amount of data (9 variables) to save costs. It is also recommended that the selected model in this research be compared to other data mining methods such as genetic algorithm, neural networks and expert systems. Also, future research should consider risk factors alongside return. Investors and analysts are recommended to pay more attention to systematic risk, asset turnover, interest rate, Tobin's Q ratio, asset growth rate, net profit growth rate, PE ratio, and net profit after tax. The results also suggest that financial institutions can benefit from decision tree and CART-regression methods when selecting stocks or forming their portfolios.

References

[1] Abzari, M., Ketabi, S., & Abbasi, A. *Optimization of investment portfolio using linear programming methods and presentation of an applied model*. Journal of Social Sciences and Humanities, 2005, **22**(2), P. 1-17. https://dx.doi.org/10.22099/jaa.2005.3467

[2] Ali Mohammadi, A.M., AbBasimehr, M.H., & Javaheri, A. *Prediction of stock return using financial ratios: A decision tree approach*. Financial Management Strategy, 2015, **3**(4), P. 151-129. https://dx.doi.org/10.22051/jfm.2016.2349

[3] Ali Zadeh, Z. *Particle swarm optimization algorithm and optimal portfolio selection*. Iranian capital market, 2016, **107**, P. 80-82.

[4] Aouni, B. *Multi-attribute portfolio selection: New perspectives*. Information Systems and Operational Research, 2009, **47**(1), P. 1-4. https://doi.org/10.3138/infor.47.1.1

[5] Barkhordari, M.H., Rezaei, M. *Optimal portfolio determination of stuck efficient industry using cover analysis of data from the perspective of institutional investors (Case Study: Ansar Bank).* Journal of Development In Monetary and Banking Management, 2015, **2**(5), P. 53-72.

[6] Barzegari Khaneghah, J., Jamali, Z. *Predicting stock returns with financial ratios; An exploration in recent researches*. Journal of Accounting, Accountability and Society Interests, 2016, **6**(2), P. 71-92. https://dx.doi.org/10.22051/ijar.2016.2432

[7] Behnampour N, Hajizadeh E, Semnani S, Zayeri F. *The introduction and application of classification tree model for determination of risk factor for esophageal cancer in golestan province*. Jorjani Biomed Journal, 2013; **1**(2), P. 38-46. http://goums.ac.ir/jorjanijournal/article-1-183-en.html

[8] Breiman L., Friedman, J., Stone, C.J., & Olshen, R.A. *Classification and Regression Trees*. 1ST edition. New York, N.Y.: Chapman and Hall/CRC; 1984. https://doi.org/10.1201/9781315139470

[9] Chalaki, P., Uoosefi, M. *Earnings management prediction by decision trees*. Accounting and Auditing Studies, 2012, **1**(1), P. 110-123. https://dx.doi.org/10.22034/IAAS.2012.105369

[10] Chang, T.S. *A comparative study of artificial neural networks, and decision trees for digital game content stocks price prediction*. Expert systems with applications, 2011, **38**(12), P. 14846-14851. https://doi.org/10.1016/j.eswa.2011.05.063

[11] Davoodi Kasbi, A., Dadashi, I. *Stock price prediction using the Chaid rule-based algorithm and particle swarm optimization (pso)*. Advances in Mathematical Finance and Applications, 2020, **5**(2), P. 197-213. https://doi.org/10.22034/amfa.2019.585043.1184

[12] Delen, D., Kuzey, C., & Uyar, A. *Measuring firm performance using financial ratios: A decision tree approach*. Expert systems with applications, 2013, **40**(10), P. 3970-3983. https://doi.org/10.1016/j.eswa.2013.01.012

[13] Fallahpour, S., Pirayesh Shirazinejad, H. *Portfolio formation using diagonal quadratic discriminant analysis and weighting based on posterior probability*. Financial Engineering And Securities Management (Portfolio Management), 2018, **9**(34), P. 85-103.

[14] Hejazi, R., Mohamadi, S., Aslani, Z., Aghajani, M. *Earnings management prediction using neural networks and decision tree in TSE.* Accounting and Auditing Review, 2012, **19**(2), P. 31-46. https://doi.org/10.22059/acctgrev.2012.29198

[15] Hosseinpour, R., Bagherpour, M.A., & Salehi, M. *Identification of financial and non-financial variables affecting the bases for adjusting audit reports related to accounting estimates: data mining approach.* Audit knowledge, 2017, **17**(1), P. 107-130. http://danesh.dmk.ir/article-1-1494-fa.html

[16] Izadinia, N., Ramesheh, M., & Yadegari, S. *Forecast for stock returns based on trading volume*. Journal of Financial Accounting, 2012, **4**(16), P. 174-160. http://qfaj.ir/article-1-263-fa.html

[17] Jafari, B., Azar, A. *Fuzzy decision tree; the new approach in strategy formulation*. Management Researches, 2013, **6**(19), P. 25-39. https://doi.org/10.22111/jmr.2013.1257

[18] Jahanshad, A., Parsaei, M. *Analysis of factors affecting expected stock returns based on the implied cost of capital.* Journal of Investment Knowledge, 2015, **4**(14), P. 125-144.

[19] Kaczmarek, T., & Perez, K. *Building portfolios based on machine learning predictions*. Economic Research-Ekonomska Istraživanja, 2021, **1**(1), P. 1-19. https://doi.org/10.1080/1331677X.2021.1875865

[20] Karami, G.R., Moradi, M.T., Moradi, F., & Mosalanezhad, A. *Study of linear and nonlinear relationships between financial ratios and stock returns in tehran stock exchange*. Accounting and Auditing Reviews, 2007, **13**(4), P. 19-46.

[21] Karami, G.R., Talaeei, L. (2013). *Predictability of stock returns using financial ratios in the companies listed in Tehran Stock Exchange.* International Research Journal of Applied and Basic Sciences, 2013, **5**(3), 360-372.

[22] Kass, G.V. *An exploratory technique for investigating large quantities of categorical data*. Journal of the Royal Statistical Society: (Applied Statistics), 1980, **29**(2), 119-127. https://doi.org/10.2307/2986296

[23] Keyghobadi, A.R., Fathi, S., & Seif, S. *The impact ranking of key balance sheet items and ratios profitability on the optimal portfolio selection (with data mining technique)*, Financial Accounting and Audit Research, 2015, **7**(28), P. 86-75.

[24] Kim, H., Loh, W.Y. *Classification trees with unbiased multiway splits*. Journal of the American Statistical Association, 2001, **96**(454), P. 589-604. https://doi.org/10.1198/016214501753168271

[25] Mahmoudiazar, M., Raei, R. *Prediction of stock market returns with out of sample data: Evaluating out of sample methods (regression method and wavelet neural network).* Journal of Asset Management and Financing, 2014, **2**(2), P. 1-16.

[26] Moghadam, A., Ghadrdan, E., & Rashedi, M. *Predicting stock return by using the market ratios in tehran stock exchange*. Accounting and Auditing Research, 2014, **6**(24), P. 104-117. https://doi.org/10.22034/iaar.2014.104331

[27] Oztekin, A., Kizilaslan, R., Freund, S., & Iseri, A. *A data analytic approach to forecasting daily stock returns in an emerging market*. European Journal of Operational Research, 2016, **253**(3), P. 697-710. https://doi.org/10.1016/j.ejor.2016.02.056

[28] Quinlan. J.R. *Induction of decision trees*. Machin Learning, 1986, **1**(1), P. 81–106. https://doi.org/10.1007/BF00116251

[29] Rahnemaei Roudeposhti, F., Chavoshi, K., Ibrahim, S. *Optimization of portfolios consisting of shares of Tehran Stock Exchange mutual funds with the genetic algorithm approach*. Journal Of Investment Knowledge, 2014, **3**(4), P. 231-218.

[30] Rai, R., Pouyanfar, A. *Advanced investment management*. SAMT Press, 2018, Tehran, Iran.

[31] Ramesh, K., Vinitha, A., Dhamodharan, M., & Shanmuga vadive, M. *An improved random forest algorithm for effective stock market prediction trending towards machine learning*. International Journal of Grid and Distributed Computing, 2020, **13**(1), P. 873-881.

[32] Salehi, M., Farrokhi Pilehrood, L. *Prediction of earnings management using the neural network and the decision tree.* Financial Accounting and Auditing Research, 2018, **10**(37), P. 1-24.

[33] Soroushyar, A., Akhlaghi, M. *The comparative assessment of data mining methods effectiveness to forecasting return and risk of stock in companies listed in tehran stock exchange*. Journal of Financial Accounting Research, 2017, **9**(1), P. 57-76. https://doi.org/10.22108/far.2017.21746

[34] Tavasoli, N., Javid, D. *Stock return prediction by decision tree*. The 5th National Management and Accounting Conference, 2015. Tehran, Iran.

[35] Tiwari, S., Pandit, R., Richhariya, V. *Predicting future trends in stock market by decision tree rough-set based hybrid system with HHMM*. International Journal of Electronics and Computer Science Engineering, 2010, **1**(3), P. 1578-1587. https://doi.org/10/1/1/261/3105

[36] Uddin, M. R., Rahman, Z., & Hossain, R. *Determinants of stock prices in financial sector companies in Bangladesh-A study on Dhaka Stock Exchange (DSE).* Interdisciplinary Journal of Contemporary Research in Business, 2013, **5**(3), P. 471-480.

[37] Wang, H., Jiang, Y., & Wang, H. *Stock return prediction based on Bagging-decision tree*. International Conference on Grey Systems and Intelligent Services, 2009, P. 1575-1580. https://doi.org//10.1109/gsis.2009.5408165

[38] Wang, Y., & Witten, I. H. *Inducing model trees for continuous classes*. In Proceedings of the 9TH European conference on machine learning, 1997, 9, London: Springer-Verlag.

[39] Zamani, M., Afsar, A., Saghafnejad, S.V., & Bayat, E. *Expert system Stock price prediction and portfolio optimization using fuzzy neural networks, fuzzy modeling and genetic algorithm*. Financial Engineering And Securities Management (Portfolio Management), 2014, **5**(21), P. 107-130.