Journal of Economics
and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

The Use of the Regression Tree and the Support Vector Machine in the Classification of the Iraqi Stock Exchange for the Period 2019-2020

Mohamed Hesham Ibrahim
Dept. of Statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq
mohammed.ibrahim1201@coadec.uobaghdad.edu.iq

Asmaa Ghalib Jaber
Dept. of Statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq
Drasmaa.ghalib@coadec.uobaghdad.edu.iq

Received: 30/1/2022 **Accepted: 2/3/2022** **Published: June / 2022**

Abstract:

The financial markets are one of the sectors whose data is characterized by continuous movement in most of the times and it is constantly changing, so it is difficult to predict its trends , and this leads to the need of methods , means and techniques for making decisions, and that pushes investors and analysts in the financial markets to use various and different methods in order to reach at predicting the movement of the direction of the financial markets. In order to reach the goal of making decisions in different investments, where the algorithm of the support vector machine and the CART regression tree algorithm are used to classify the stock data in order to determine the trend of the stock if it is a rising stock or a descending stock .The aim of the research is to classify the financial stock data using five variables where the data of the Iraqi Islamic Bank for investment and development was used where the results showed the accuracy of the algorithm, the support vector machine and the CART algorithm, and their performance was good. Also, the results showed that the Support Vector Machines algorithm is the best when compared with the CART algorithm, using the Classification Error and MSE criteria.

Keywords: Support Vector Machine, SVM, CART, Classification, Financial stock, Hyperplane

1. Introduction:

The financial markets is one of the important economic sectors in countries, the importance of which lies in their ability to revive .The economy of countries, because it is a source of money and investments, and this leads to the fact that it is characterized by its dense data. It is large and has constant movement most of the time[2], and that makes it more complex than other types of data. It contains the hidden relationships between its variables, and the classification is the direction of stock movement, which is characterized by being non-linear. Where the research aims to use the SVM algorithm and the CART regression tree algorithm in classifying Stocks and Support Vector Machine, which is one of the machine learning algorithms, which is characterized by its ability to control the decision function and its ability to use the kernel function, and The Support Vector Machine is considered to be very popular because of its ability to solve non-linear problems by transforming quadratic programming. And support vector machine, its solution is unique and the best in the world the cross-validation method has been used to avoid over-matching, as well as by specifying the maximization margin of Hyperplane and the CART regression tree algorithm, which is one of the tree algorithms classification regression is common in tree formation and is considered a special case of decision trees. Using a regression tree because of its potential and ability, the CART algorithm is characterized by its simplicity, flexibility and ease of understanding, as well as it is characterized by its ability to conduct an analysis without the need to understand each step of the program, and this is used in several types of the data including continuous, discontinuous, ordinal and nominal data. This method aims to improve the classification of the objective variable.

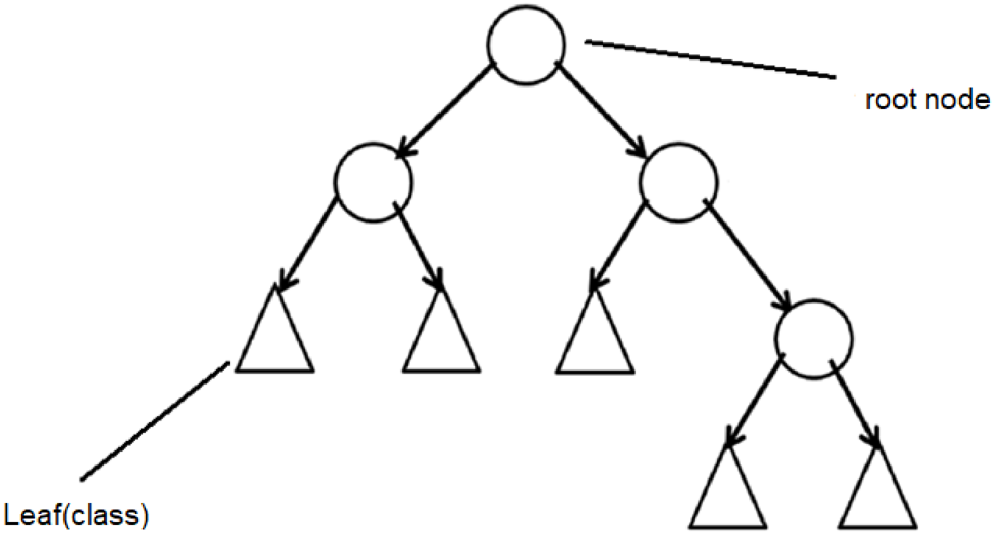
There are many studies that dealt with these two methods, and the following is a presentation of some of these studies [13]. The two researchers (Ghaida and Saja) in 2014 discovered the rate of desertification in the urban area based on the contents of the images. [12]The researchers (Rosllo & Fuente) presented a simulation of the stock market using the Support Vector Machine in 2014. [4]The researchers (Omar and Suhad) in 2016 presented the Bayesian method in the tree regression classifier in estimating the assembly model and comparing it with the logistic model with the application [3] In 2018, the two researchers (Rana and Ghaida) published a comparative study of algorithms for data mining and analysis of emotions and their applications [1] The two researchers (Asmaa and Baraa) in the year 2020 presented a comparison between tree regression and binomial regression methods using simulation.

2. Materials and Methods

2.1 Regression Trees CART

It is one of the most popular classification regression tree algorithms for tree formation. It is considered a special case of decision trees, which is a tree diagram based on dividing the data set into more than one set to improve classification of the target variable. The way it works is by dividing the predictive data set into molecular groups subsets more homogeneous than the original group [1]and represents the decision rules in the so-called binary trees through the CART algorithm, where the resulting data set is a pyramidal shape and the top part is called the root node, while the base of the tree is represented by groups of small homogeneous observations that are represented by leaves, as the root node contains data.

The sample of the problem under study or part of it, that is, is a two-part tree, this algorithm continues work until a decision tree is formed[1] as shown in the figure below



[1]Figure 1: Represents the regression tree diagram for classification.

The interest in using the regression tree began because of its different capabilities and ability to deal with data from the rest of the traditional algorithms for classification and data analysis in many research studies about the rest of the traditional algorithms for classification and data analysis in many studies and research which needs big data to build a reliable decision base .The CART algorithm is characterized by its simplicity, flexibility and ease of understanding. If K is from the classes $(c_1 , c_2 , \dots , c_k)$ and a sample of the data is training data can be seen from the following:

1- If T contains one or more observations of the same class, $c_j$will be the tree has a leaf assigned to the class c_j

2- If T does not contain the views of those classes, then there is no tree for this data

3- If T contains a mixture of observations of those classes, there will be a testing data

based on the singular attributes of those observations that can give one or more results

The pairs are separated $(o_1 , o_2 , \dots , o_n)$ and the T group is divided into subgroups.

$(t_1 , t_2 , \dots , t_n)$ where $T_i$contains all the observations that have the results $o_i$from the test

that has been selected and the process is repeated on all subsets of the test data training data[1]. A tree classification can be represented by needing a large number of data

Suppose the data consists of the response variable (y) from a set of predictive variables
$x_i = (x_1, x_2, x_3, \dots\dots, x_m)$
It is in the form of a fixed matrix (m).
x_i**can be either quantitative variables (continuous or discontinuous) or descriptive variables (nominative or ordinal)**

<u>Algorithm steps of the operations that are performed at each node</u>
1- Choosing all allowed divisions of predictive variables usually binary divisions generate binary questions
2- Choosing the best division the word best in this step refers to the term selection of some criteria good division, as is the case with the concept of (good matching), and there are well-known methods of estimation:
(least squares) and (absolute least variance) both refer to comparison in terms of homogeneity
or reduce the application of the measurement at the node (father)
3- Splitting stops at the node that does not meet the conditions required for ordering variablesx_i**in the question in the first step**
4- Is $(x_i > c)$ **for all values of C that are within the range of** x_i **that is,**x_i**takes limited numbers**
$(b_0 , b_1 , b_2 ,\dots , b_i)$
The question here is whether ($\boldsymbol{x_m}$ **∈c) when C is within the range of the molecular groups**
$[b_0 , b_1 , b_2 ,\dots , b_i]$
These cases are in the tree T whose answer is either (yes) going to the left of the node is its answer
(No) it is going to the right node and the above methods stop at the third step when applying
It doesn't perform well the tree is too big at the node when there is little data in the
In each node there is an algorithm that searches the variables one by one
It starts from x_1**and continues until it reaches** x_m**. For all variables we find the best division and then compare with m**
The best division of a single variable and then selecting it is the best basic tree model
The first and second steps are repeated for the sons node until we reach the end of the tree

$$\boldsymbol{F^{\wedge} (x) = \sum_{m=1}^{n} cmI} \, [\, (\boldsymbol{x_1 , x_2}) \,] \, \in \boldsymbol{RM} \,] \qquad \dots\dots(1)$$

where
$\boldsymbol{F^{\wedge} (x)}$**: represent the estimator model**

Cm =node means

$$\mathbf{c^{\wedge}\, m = \frac{1}{Nm} \sum x_i \in Rm\ Y_i}$$

2.2 Support Vector Machine

It is one of the classification techniques in data mining, and it is one of the machine learning algorithms that has been reached. It was mentioned by the scientist Vapnik in 1992[15] that the support vector machine includes line construction. The hyperplane between two groups in order to classify them is the decision surface. The separator between the positive group and the negative group is that its function is to construct the best separator level Hyperplane between the positive and negative classes and the distance between the boundary Hyperplane and the closest element of any of the two classes is called margin, where Hyperplane classifies data in a linear way by creating a classifier that classifies the data support vector machine SVM is easy to handle with little data but not with big data also, the support vector machine converts a non-linear data field into a linear data field. It is done using the kernel function, which converts the data field from non-linear to a field. Linear data and the support vector machine SVM does not depend on data dimensions but on properties engineering data support vector machine SVM consists of several stages is the representation of data in the drawing graph and finding the hyperplane and finding the biases that the concept of the support vector machine is separation between two classes where the first class (+1) represents the positive class and the second class represents (-1) negative class.

2.2.1 Methodology

Support vector machine SVM depends on several factors that directly and indirectly affect the creation of the final solution also affects the classification accuracy, including the Hyperplane and Lagrange multipliers and support vector

Support Vector:- They are points that lie on the boundary line between the two classes

The boundary can be explained by the following two equations[15]

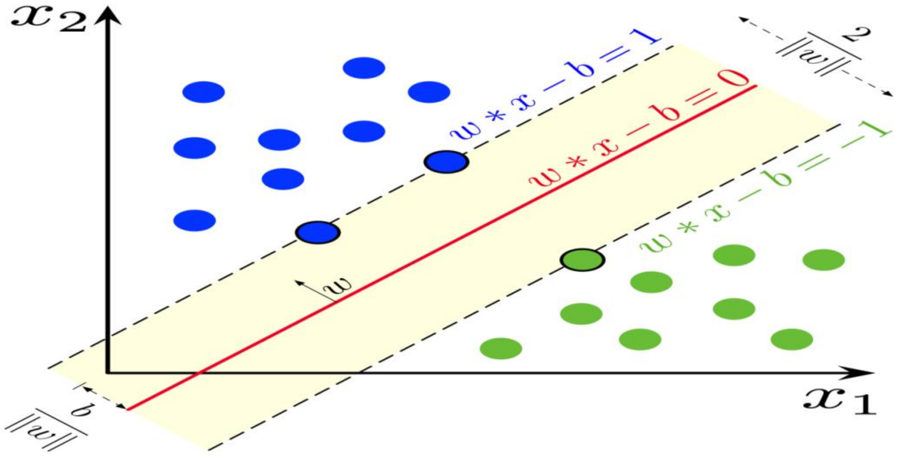
$$\boldsymbol{w^t x_i + b = +1 \qquad for\ \ y_i = +1 \qquad i = 1, 2 \ldots . N} \tag{2}$$

$$\boldsymbol{w^t x_i + b = -1 \qquad for\ \ y_i = -1 \qquad i = 1, 2 \ldots . N} \tag{3}$$

The separating surface level equation that lies at the extreme of each class can be represented as follows

$$\boldsymbol{w^t x_i + b = 0} \tag{4}$$

As shown in the following figure



[9]Figure 2: Separator diagram in SVM .

The separation limit first secondary can be represented in the support vector machine as follows

$$\boldsymbol{w^t x_i + b = +1} \qquad \textbf{(5)}$$

It represents the separation limit secondary in the support vector machine as follows

$$\boldsymbol{w^t x_i + b = -1} \qquad \textbf{(6)}$$

a- Primal equation in the following form

$$\boldsymbol{L(w,b,\alpha) = \frac{1}{2}\,\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i (y_i (x^{i^T}\, w + b) - 1)} \qquad \textbf{(7)}$$

$\boldsymbol{x^{i^T}}$**:- Represents training data points.**
$\boldsymbol{y_i}$**:- Represents training points a sign.**

When opening the equation, write as follows

$$\boldsymbol{L(w,b,\alpha) = \frac{1}{2}\, w^T w - \sum_{i=1}^{n} \alpha_i\, y_i\, x^{i^T}\, w + b(\sum_{i=1}^{n} \alpha_i\, y_i\,) + \sum_{i=1}^{n} \alpha_i\;\;)} \;\textbf{.....(8)}$$

We apply the terms and conditions

1-

$$\boldsymbol{\frac{\partial L}{\partial \mathrm{w}} = 0 \;\rightarrow\; w^* = \sum_{i=1}^{n} \alpha_i\, y_i\, x^i} \qquad \textbf{(9)}$$

$\boldsymbol{\alpha_i}$**:- Lagrange Multipliers**

$$\boldsymbol{0 \;\leq \alpha_{\mathrm{i}} \leq \mathrm{c}}$$

Where

c:- It represents a criterion that balances the estimated error with the amount of divergence from the main axis using a value of 1

2-

$$\boldsymbol{\frac{\partial L}{\partial \mathrm{b}} = 0 \;\rightarrow\; \sum_{i=1}^{n} y_i\, \alpha_i = 0}$$

And the dual equation is in the following form[5]

$$\boldsymbol{\max_{\alpha_1.....,\alpha_n} \mathrm{L}(\mathrm{w}^*, \mathrm{b}^*, \alpha)}$$

Subject to

$$\boldsymbol{\alpha_{\mathrm{i}} \;\geq 0\,, \mathrm{i} \in 1\,, \ldots \mathrm{n} \;\; \text{and} \;\; \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\, \mathrm{y_i} \;= 0}$$

When substituting the equation, it will be in the following form[5]

$$\boldsymbol{max_{\alpha_1.....,\alpha_n} L(w^*, b^*, \alpha) = max_{\alpha_1.....,\alpha_n} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j\, x^{iT} x^j + \sum_{i=1}^{n} \alpha_i}$$

Subject to

$$\boldsymbol{\alpha_{\mathrm{i}} \;\geq 0\,, \mathrm{i} \in 1\,, \ldots \mathrm{n} \;\;\; \text{and} \;\; \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\, \mathrm{y_i} \;= 0}$$

$$\boldsymbol{\mathrm{w}^* = \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\, \mathrm{y_i}\, \mathrm{x^i}}$$

Hyperplane distance is as follows

$$\boldsymbol{\frac{2}{\|w^*\|}} \qquad \textbf{(10)}$$

And to find the limit of bias, through the following

The data point that satisfies equation (9), which takes the following form

$$\boldsymbol{y(w^t x + b) = 1} \qquad \textbf{(11)}$$

Then it is substituted into equation (14) to get the following equation

$$\boldsymbol{y(\sum_{i=1}^{n} \alpha_i\, y_i\, \mathrm{x}_i \mathrm{x} + \mathrm{b}) = 1} \qquad \textbf{(12)}$$

Then y is multiplied by the equation to get the following equation

$$\boldsymbol{\mathrm{y}^2(\sum_{i=1}^{n} \alpha_i\, y_i\, \mathrm{x}_i \mathrm{x} + \mathrm{b}) = y} \qquad \textbf{(13)}$$

This is because y^2**represents 1 as it is represented in the following two equations**

$\boldsymbol{w^t x + b \geq 1}$ **for** $\boldsymbol{y_i = +1}$
$\boldsymbol{w^t x + b \leq 1}$ **for** $\boldsymbol{y_i = -1}$
We get the following equation

$$\boldsymbol{b = y - \sum_{i}^{n} \alpha_i\, y_i\, x_i\, .\, x} \qquad \textbf{(14)}$$

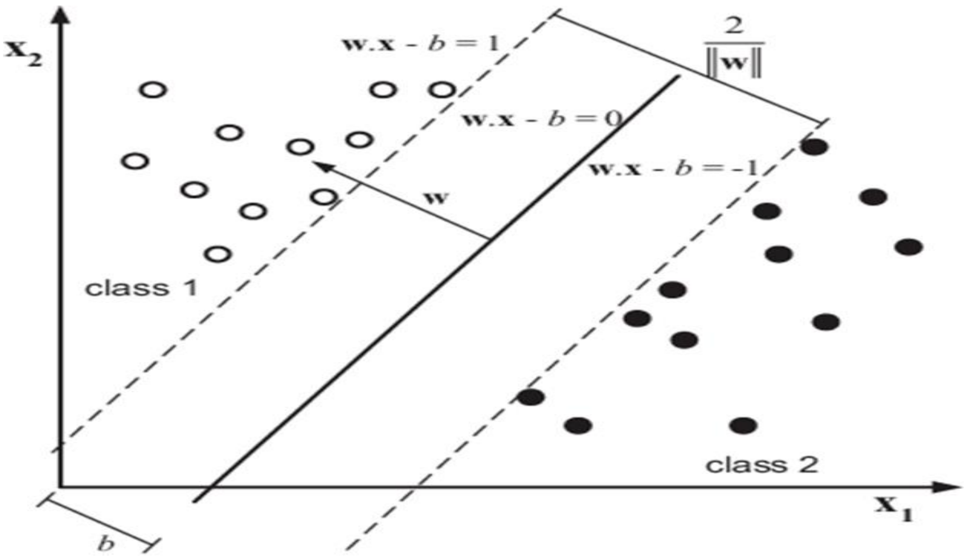
Then taking the average for each equation, we get the following

$$\boldsymbol{b = \frac{1}{N}\sum_{i=1}^{n}(y - \sum_{i}^{n} \alpha_i\, y_i\, x_i\, .\, x)} \qquad \textbf{(15)}$$

And so, the classification equation is as follows[5]

$$\boldsymbol{y = (w * x_i + b)} \qquad \textbf{(16)}$$

The following is a diagram showing support vector machine



[11]Figure 3: Diagram showing Support Vector Machines Below are the types of support vector machine

2.2.2 Linear Support Vector Machine

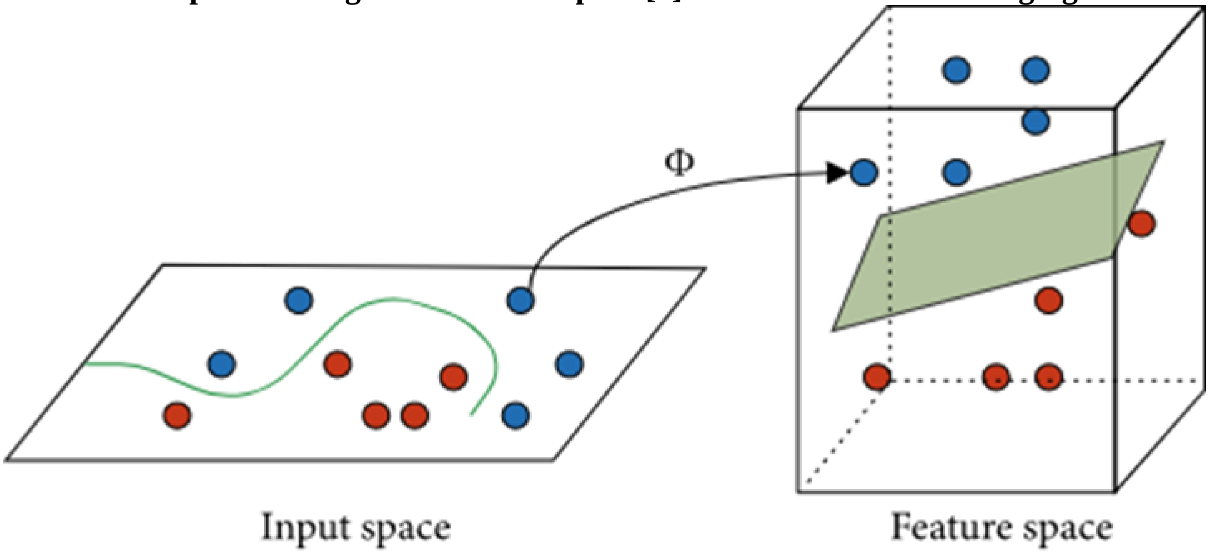
It is a linear classification problem that aims to find the best hyperplane between data with Maximization margin[5]. If the data is linearly separable where it is Hard Margin, and if the data is not separable, it is Soft Margin. If the data is two-dimensional, D = 2, then the data is divided into two parts

D:- Represents the number of dimensions

But if the data has dimensions greater than two D > 2, then the data is divided into three lines to configure the separator level

2.2.3 Nonlinear Support Vector Machine

In most cases, the classification process is non-linear, so the support vector machine is used which uses many procedures, and one of these procedures is the Kernel function. The kernel function is the function of converting data from a low-dimensional space to a high-dimensional space [6] as shown in the following figure



[7]Figure 4: Diagram Nonlinear classification in Support Vector Machines.

2.3. Comparison Criteria

i- Classification Error

$$Error\ of\ Classification = \frac{(N-sum(diag(cm)))}{N} \text{.................................}\textbf{(17)}$$

CM : It represents Confusion matrix

N : It represents the sum of the Confusion matrix

ii- Mean Square Error

$$MSE = \frac{1}{R}\sum_{i=1}^{R} MSE_i = \frac{1}{R}\sum_{i=1}^{R}\left[\frac{1}{N-P}\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2\right] \text{.................................}\textbf{(18)}$$

Y_i: It represents the real response variable.

$\hat{Y}_i$: It represents the estimated response variable.

3. Results and Discussion

On the practical side, Support Vector Machine svm and CART algorithm were used, as well as drawing to clarify the efficiency of Support Vector Machine svm. In classifying the trends of financial stocks, whether they were rising or descending, data was used. The financial stocks for the years 2019-2020 of the Iraqi Islamic Bank for Investment and development[16], which used variables related to stocks in order to perform the classification process, and these variables are as follows

1-X_1 : Represents the variable stock price
2-X_2: Represents the variable initial stock price
3-X_3 : Represents the variable highest stock price
4-X_4 : Represents the variable lowest stock price
5-X_5 : Represents the variable the real value of the stock

The Support Vector Machine Svm application, through the use of the data of the Iraq Stock Exchange, the Iraqi Islamic Bank for Investment and Development, and this data consists of five variables, which are the stock price and the initial stock price, the highest stock price, the lowest stock price, the real value of the stock, and 105 entries. The methods were applied by writing a program in Matlab 2018b, and the accuracy criteria were calculated depending on the criteria, which are as follows

3.1 Support Vector Machine Application

Here, the Support Vector Machine is applied, and the results are obtained as follows

Table 1: Results SVM for criteria Error classification and MSE

MSE	Error of Classification
0.000599	0.085714286

Note from the above table

1- The Support Vector Machine SVM method
The value of the Error Classification is equal to 0.0857
2- The Support Vector Machine SVM method
The value of the MSE is equal to 0.000599

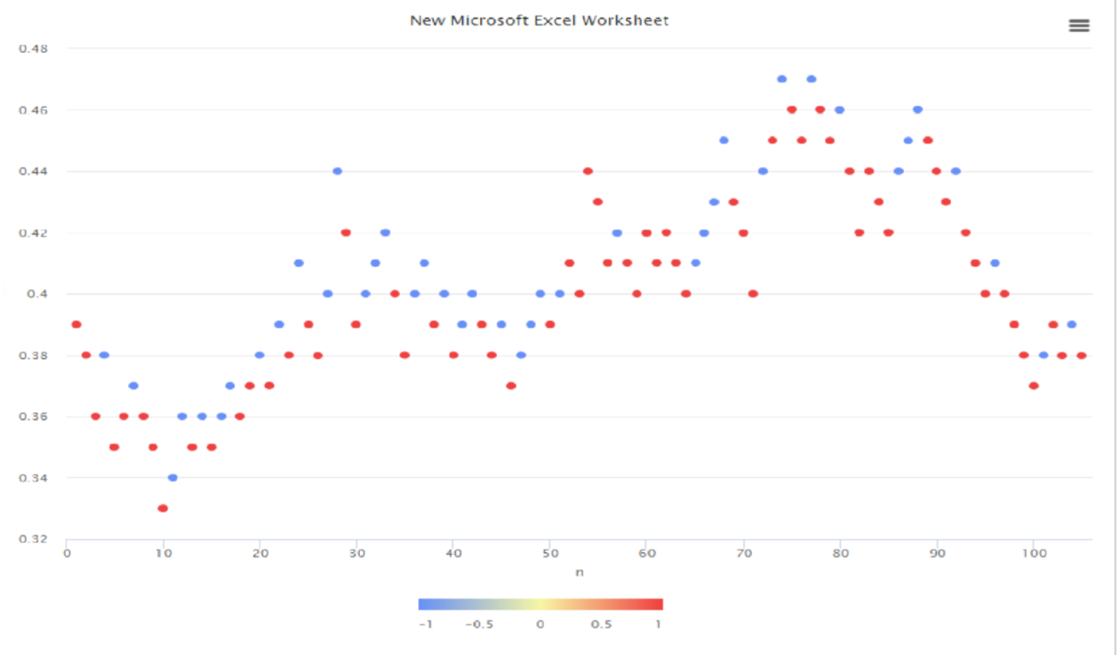


Figure 5: It shows the practical results of the SVM method

In the above figure, it was found that the SVM method classified the data into rising and declining stocks. Well, this indicates that the SVM method is highly efficient in classifying data

3.2 CART Application

Table 2: Results of Error classification and MSE Criteria for CART

MSE	Error of Classification
0.0054	0.5524

Note from the above table
1- The CART method
The value of the Error Classification is equal to 0.5524
2- The CART method
The value of the MSE is equal to 0.0054

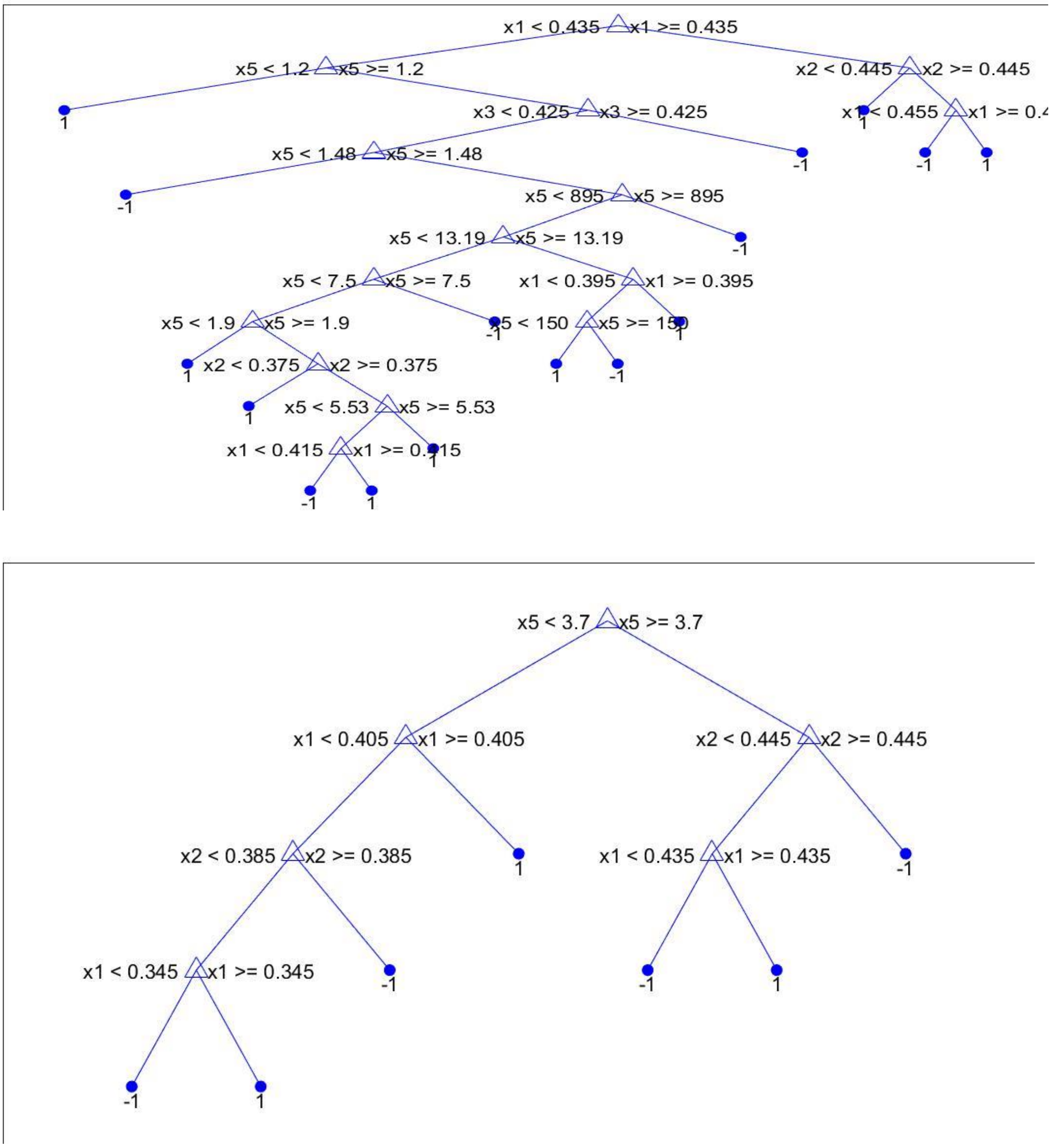


Figure 6: Shows the results of a CART regression tree.

We notice from the above figure that the CART method classified financial stocks into two categories, which are rising stocks and descending stocks where it can be observed at the ends of the tree, which indicates the completion of the classification process.

The results showed that the Support Vector Machine algorithm is the best when compared with the CART algorithm, using two criteria Classification Error and MSE.

4. Conclusions:

i- The results showed that the accuracy of the SVM algorithm was good, and the SVM algorithm was effective in classifying the categories of rising stocks and falling stocks.

ii- The SVM algorithm obtained the best results when compared with the CART algorithm by using the Classification Error criterion in classifying financial stocks.

iii- The SVM algorithm has recorded the best results when compared with the CART algorithm by using the MSE criterion in classifying financial stocks.

iv- It was found that the SVM algorithm obtained the best results when compared with the CART algorithm, using all criteria .

5. Recommendations:

i- Determination of large sample sizes for the methods used in order to obtain the best results.

ii- Using other artificial intelligence algorithms for comparison methods.

References

1-Abd-allah, A. N., & Abbas, B. K. (2020). Comparison Between Tree regression (TR), and Negative binomial regression (NBR) by Using Simulation. journal of Economics And Administrative Sciences, 26(119).

2-Al-Rubyee, H. M., Al-Mousawi, H. Y., & Hasan, M. F. (2018). Building and evaluating the performance of active momentum portfolios in the Iraq Stock Exchange. THE IRAQI MAGAZINJE FOR MANAGERIAL SCIENCES, 14(55).

3-Al-Obaidi, Rana Zuhair Abdel-Ghani and Al-Talib, Ghaida Abdel-Aziz. 2018 A comparative study of opinion mining algorithms in analyzing emotions and their applications. Al-Rafidain Journal of Computer Science and Mathematics, Vol. 12, p. 2, pp. p. 13-23.

4-Ali, O. O. A. M., & Ahmed, M. S. A. (2016). Albezi method in regression to estimate the tree synthesis model classified and compared to the model Logistics with the application. Journal of Administration and Economics, (109).

5-Deng, N., Tian, Y., & Zhang, C. Support vector machines: optimization based theory, algorithms, and extensions. CRC press.2012)).

6-Fahad, H. E., & Asmaa, G. J. (2021, May). Using Support Vector Machine To Determine the Limits of Multivariate Control Charts. In Journal of Physics: Conference Series (Vol. 1897, No. 1, p. 012009). IOP Publishing.

7- Hussain, L., Awan, I. A., Aziz, W., Saeed, S., Ali, A., Zeeshan, F., & Kwak, K. S. (2020). Detecting congestive heart failure by extracting multimodal features and employing machine learning techniques. BioMed research international, 2020.

8-Kecman, V. Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. MIT press.2001)).

9- Noyum, V. D., Mofenjou, Y. P., Feudjio, C., Göktug, A., & Fokoué, E. (2021). Boosting the Predictive Accurary of Singer Identification Using Discrete Wavelet Transform For Feature Extraction. arXiv preprint arXiv:2102.00550.

10-Pham, H. (Ed.). Springer handbook of engineering statistics. Springer Science & Business Media.2006)).

11-Papadonikolakis, M., & Bouganis, C. S. (2012). Novel cascade FPGA accelerator for support vector machines classification. IEEE transactions on neural networks and learning systems, 23(7), 1040-1052.

12-Rosillo, R., Giner, J., & De la Fuente, D. (2014). Stock market simulation using support vector machines. Journal of Forecasting, 33(6), 488-500.

13- Student, Ghaida Abdel Aziz and Hamed, Saja Younes. 2014. Discovering the percentage of desertification in the urban area based on the contents of the image. Al-Rafidain Journal of Computer Science and Mathematics, Vol. 11, p. 2, pp. p. 37-51.

14-Steinwart, I., & Christmann, A. Support vector machines. Springer Science & Business Media.2008)).

15-Sonmez, R., & Sözgen, B. (2017). "A support vector machine method for bid/no bid decision making". Journal of Civil Engineering and Management, 23(5), 641-649.

16- Iraq securities commission, Reports for the year(2019-2020).

استعمال شجرة القرار والة الموجه الداعم في تصنيف سوق العراق للأوراق المالية للفترة 2020_2019

الباحث/ محمد هشام ابراهيم
كلية الادارة والاقتصاد / جامعة بغداد / قسم الاحصاء
بغداد ، العراق
mohammed.ibrahim1201@coadec.uobaghdad.edu.iq

ا.م.د. اسماء غالب جابر
كلية الادارة والاقتصاد / جامعة بغداد / قسم الاحصاء
بغداد ، العراق
Drasmaa.ghalib@coadec.uobaghdad.edu.iq

Received: 30/1/2022 Accepted: 2/3/2022 Published: June / 2022

مستخلص البحث:

ان الاسواق المالية تعد من القطاعات التي تمتاز بياناتها بانها ذات حركه مستمرة في اغلب الاوقات وانها تكون متغيره بصوره مستمرة لذلك يكون من الصعب التنبؤ بأتجهاتها وذلك ادى الى وجود الحاجه الى طرائق ووسائل وتقنيات لاتخاذ القرارت مما يؤدي الى دفع المستثمرين والمحللين في الاسواق المالية لاستخدام الطرائق والأساليب المتنوعة والمختلفة وذلك للوصل الى التنبؤ بحركة اتجاه الاسواق المالية واوصل لغايه اتخاذ القرارات في الاستثمارات المختلفة حيث تم استخدام خوارزميه اله المتجة الداعم وخوارزميه شجره الانحدار وذلك لتصنيف بيانات الاسهم المالية لتحديد اتجاه الاسهم فيما اذا كانت اسهم صاعده او اسهم هابطه ان الهدف من البحث هو تصنيف بيانات الاسهم المالية وذلك باستخدام خمسه متغيرات اذ تم استخدام بيانات مصرف العراقي الاسلامي للاستثمار والتنمية حيث اظهرت النتائج دقه خورازميه اله المتجه الداعم وخورازميه الانحدار الشجري وكان ادائهم جيد وكذلك اظهرت النتائج ان الخوارزميه الة المتجة الداعم كانت الافضل عند مقارنتها مع خوارزمية الانحدار الشجري وذلك باستخدام المعايير MSE و Classification Error

*البحث مستل من رسالة ماجستير