Stock market prediction by applying big data mining

Bedour M. Alshammari
Arabian Gulf University, Manama, Bahrain
Fairouz Aldhmour
Department of Innovation and Technology Management, Arabian Gulf University, Manama, Bahrain and
MIS, Mutah University, Karak, Jordan
Zainab M. AlQenaei
Information Systems and Operations Management Department, College of Business Administration, Kuwait University, Sabah Al-Salem University City, Kuwait, and
Haidar Almohri
Data Science Department, Gulf Bank, Kuwait City, Kuwait

Received 11 May 2022
Revised 27 July 2022
Accepted 28 July 2022

Abstract

Purpose – There is a gap in knowledge about the Gulf Cooperation Council (GCC) because most studies are undertaken in countries outside the Gulf region – such as China, India, the US and Taiwan. The stock market contains rich, valuable and considerable data, and these data need careful analysis for good decisions to be made that can lead to increases in the efficiency of a business. Data mining techniques offer data processing tools and applications used to enhance decision-maker decisions. This study aims to predict the Kuwait stock market by applying big data mining.

Design/methodology/approach – The methodology used is quantitative techniques, which are mathematical and statistical models that describe a various array of the relationships of variables. Quantitative methods used to predict the direction of the stock market returns by using four techniques were implemented: logistic regression, decision trees, support vector machine and random forest.

Findings – The results are all variables statistically significant at the 5% level except gold price and oil price. Also, the variables that do not have an influence on the direction of the rate of return of Boursa Kuwait are money supply and gold price, unlike the Kuwait index, which has the highest coefficient. Furthermore, the height score of the variable that affects the direction of the rate of return is the firms, and the accuracy of the overall performance of the four models is nearly 50%.

Research limitations/implications – Some of the limitations identified for this study are as follows: (1) location limitation: Kuwait Stock Exchange; (2) time limitation: the amount of time available to accomplish the study, where the period was completed within the academic year 2019-2020 and the academic year 2020-2021. During 2020, the coronavirus pandemic (COVID-19), which was a major obstacle, occurred during data collection and analysis; (3) data limitation: The Kuwait Stock Exchange data were collected from May 2019 to March 2020, while the factors affecting the stock exchange data were collected in July 2020 due to the corona pandemic.

Originality/value – The study used new titles, variables and techniques such as using data mining to predict the Kuwait stock market. There are no adequate studies that predict the stock market by data mining in the GCC, especially in Kuwait. There is a gap in knowledge in the GCC as most studies are in foreign countries, such as China, India, the US and Taiwan.

Keywords Logistic regression, Decision tree, Support vector machine, Random forest, Classification

Paper type Research paper

Arab Gulf Journal of Scientific Research
Vol. 40 No. 2, 2022
pp. 139-152
Emerald Publishing Limited
e-ISSN: 2536-0051
p-ISSN: 1985-9899
DOI 10.1108/AGJSR-05-2022-0053

Introduction

Computers are used in all aspects of daily transactions, and, as a result, a large amount of data is generated. The volume of the data is expected to continue to grow in the future, which leads to the need to be able to analyze big amounts of data. Big data analysis is a data mining technique that can be used in many sectors – economic, industrial and commercial.

Data mining can be defined as preparing, visualizing and exploring massive databases (Parr & Vaudrevange, 2020), whereas the techniques for discovering patterns from these databases are based on the knowledge to be mined, such as descriptive, estimation, prediction, classification, clustering and association (Alsultanny, 2013).

The COVID-19 pandemic has inflicted heavy human and economic losses and confused social and health systems around the world. Understanding and counteracting the pandemic requires recognizing its properties and attributes by collecting and analyzing relevant big data. Consequently, big data analytics tools are an essential requirement for those needing to make decisions and establish precautionary measures (Almuhaideb, 2021). Also, Tariq *et al.* (2022) emphasized that software automation plays an important role in e-health and generally improves health care services for individuals by bringing efficiency to the systems. Moreover, many countries around the world are trying to smooth out the curve of the epidemic with the help of smartphone apps. Because of this, it is very important to carefully carry out strategies to protect data privacy when utilizing big data. Improvements in technology can offer several benefits, but they can also pose a risk of breaching privacy. Governments and companies in many industries use big data as a basis for automating processing and extracting important insights to aid decision-making. While big data has been confirmed as being useful in analysis and prediction, it is important to implement security procedures for maintaining confidential data on big data systems (Haafza *et al.*, 2021; Rafiq *et al.*, 2022).

The researcher found that the effective research tool for the analysis of big data is data mining, which is also known as data analytics and predictive analytics (Murray & Scime, 2015). Predictive analytics can be defined as building and evaluating a model that is aimed at creating empirical predictions, including an empirical predictive model, which is a statistical model, along with other methods, such as data mining algorithms designed for predicting new or future observations, or events and methods for evaluating and estimating the quality of the predictive power of a model (Shmueli & Koppius, 2011).

Big data is used in the forecasting process, particularly in the financial market, as forecasting is used in the stock market to create an automatic prediction of volatility in share prices. The main purpose of forecasting by data mining in the stock market is to discover knowledge that can assist decision-makers. It is important that companies use data mining with utmost care to improve their business by increasing revenue and reducing costs (Ahmed, 2004). For example, Amazon encourages its customers to use the Amazon Price Check Mobile App to collect a database of market price data. Another example is when Google collaborates with the Center for Disease Control (CDC) (Petersen, 2016).

Recently, many researchers such as Gupta, Bhatia, Dave, and Jain, (2019), Kohli, Zargar, Arora, and Gupta, (2019) and Petrova, Pauwels, Svidt, and Jensen, (2019) have used data mining techniques to predict the stock market, which contains rich, valuable and considerable data that need careful analysis before good decisions can be made. The Kuwait Stock Exchange (KSE) is one of the oldest stock markets in the Gulf Cooperation Council (GCC) as it was inaugurated in 1952, and it has used the automated trading system since 1995 (Bley & Chen, 2006). Electronic trading led to an accumulation of data from many sources, such as market index, sector index and the index for each company. The Kuwait stock market has 173 companies distributed for three markets – premier market, main market and auction market – where there are trading shares and investment funds (Boursakuwait, 2019).

Thus, the current study will focus on predicting the stock returns in the KSE by using data mining techniques, namely, regression, support vector machine, decision tree and random forest.

Background

Previous research has used several data mining techniques to predict future results and trends. The support vector machine method was used to predict the direction of the stock market by Chen and Hao (2017), Lai and Liu (2010) and Yu, Wang, and Lai, (2005) from China, and Usmani, Adil, Raza and Ali, (2016) from Pakistan. All agreed that the SVM strongly forecasts the performance of the stock markets.

Two studies used the decision tree technique: Tsai and Hsiao (2010) from Taiwan, and Al-Radaideh, Assaf, and Alnagi (2013) from Jordan. The Taiwan study indicated that the movement of the stock market could be forecast by the decision tree model, but Al-Radaideh *et al.*'s data from Jordan suggested that the accuracy of the decision tree model is low. In addition, Tsai and Hsiao presented 85 variables as important factors affecting stocks, including the observation that the US stock market has a leading effect on the Taiwan stock market.

Ou and Wang (2009) attempted to find the ability of the ten data mining techniques to predict the movement of the Hang Seng Index in Hong Kong by using tree-based classification, the logistical regression model and SVM. The results of the study showed that the SVM had a better level of prediction than the decision tree and the logistical regression model. Also, Imandoust and Bolandraftar (2014) from Iran predicted the stock trend based on the decision tree and random forest, and found the performance of the decision tree model to be better than the random forest. Awan *et al.* (2021a) in *Social media and stock market prediction: a big data approach* predicted future pricing and sales of products by using linear regression and random forest, and found that linear regression gives higher accuracy than random forest.

There are also studies using the four methods, including *A big data approach to Black Friday sales* by Awan *et al.* (2021b) which used linear regression, generalized linear regression, random forest and decision tree to predict market trends, and found that linear regression, random forest and generalized linear regression provide an accuracy of 80%–98%, while the decision tree did not perform as well. Shashaank, Sruthi, Vijayalakshimi and Garcia, (2015) also used a full mix of classification algorithms – random forest, decision tree, support vector machine and multinomial logistic regression – to predict the stock price. The results of this Indian study showed that random forest had the best prediction performance, followed by decision tree, then SVM and, lastly, multinomial logistic regression.

Methodology

The methodology used is quantitative techniques, which are mathematical and statistical models that describe a various array of the relationships of variables for assisting managers to use these techniques in order to provide insight into problems and facilitate daily decision-making. The statistics algorithms are the processes of collecting a sample, organizing, analyzing and interpreting data; and the numeric values in characteristics analyzed in this process to help with problem-solving and decision-making (Devi & Devaki, 2019).

Quantitative methods used to predict the direction of stock market returns aim to assist decision-makers in taking action to buy or sell stocks at the best possible time. Prediction is one of the data mining techniques adopted in this research to achieve the research objective of using data analysis tools – regression, support vector machine, decision tree and random forest – to extract knowledge. The predictive approach is a technique of data mining that forecasts predictions based on historical data or on aggregate indicators, such as key

performance indicators, so that potential problems can be detected ahead of time and thereby managed and mitigated (Metzger *et al.*, 2014).

Limitations

Some of the limitations identified for this study are as follows:

(1) Location limitation: KSE.

(2) Time limitation: The amount of time available to accomplish the study, where the period was completed within the academic year 2019-2020 and the academic year 2020-2021. During 2020, the coronavirus pandemic (COVID-19), which was a major obstacle, occurred during data collection and analysis.

(3) Data limitation: The KSE data were collected from May 2019 to March 2020, while the factors affecting the stock exchange data were collected in July 2020 due to the corona pandemic.

The research framework of this study is summarized in Figure 1. As depicted there, the process starts with data collection from various sources. The data are then pre-processed and converted to a proper format, ready for analysis. The next step is to perform some analysis (EDA) to understand the data before the final step, which is developing the model for prediction. A more detailed description of the steps in this framework follows.

Data collection

The data were collected from two sources:

(1) Stock market data were collected for companies in Boursa Kuwait from January 6, 2015 to November 26, 2019.

(2) Based on the literature review, the available fundamental and economic variables are selected in related work, such as oil price, gold price, the exchange rate of Kuwaiti dinar (KWD) to US dollar (USD), money supply, interest rate, earnings per share (EPS), dividends per share (DPS) and Gulf stock market index namely, Kuwait, Oman, Saudi Arabia, Bahrain and Dubai (Chatzis, Siakoulis, Petropoulos, Stavroulakis, & Vlachogiannakis, 2018; Cheng *et al.*, 2021; Kim, 2021; Ou & Wang, 2009; Yu *et al.*, 2005; Zhong & Enke, 2017).

- The oil price, gold price and exchange rates of KWD to USD were collected from investing websites.
- The money supply and interest rate were collected from the Central Bank of Kuwait website.

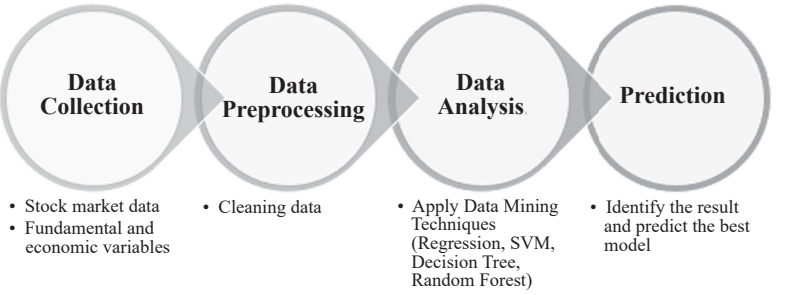

Figure 1.
Research framework design

- Data related to companies such as EPS and DPS were collected from Boursa Kuwait website.
- The Gulf stock market index was collected from Kuwait's financial brokerage firms.

Drawing up the charts helped us decide to choose the banking sector and telecommunication sector to analyze. Figure 2 shows the market capitalization for all Kuwait stock market sectors, with the banking and telecommunication sectors representing three-quarters of the stock market.

Data preprocessing

This step uses the historical data to convert the raw data into an understandable form (CSV). The CSV format file is a record of data in a tabular format, which is easy to handle by researchers (Mao *et al.*, 2018).

Two processes are applied in this study for using the data in forecasting the stock market. They are:

(1) Preparing the data for forecasting by modifying the data in tables and using it in Excel software in one sheet (.XLS).

(2) Converting the dataset to (.XLSX), which is more secure and faster. The (.XLS) format is limited to 65,000 rows, but the (.XLSX) format transacts up to a million rows (Fairhurst, 2021; Gerth, Sieverling, & Trognitz, 2017).

The database for the regression test, the decision tree test, the support vector machine test and the random forest test is split and partitioned into subsets according to 75% and 25%, where the 14 variables of 16,982 observations divide into 12,736 observed training data and

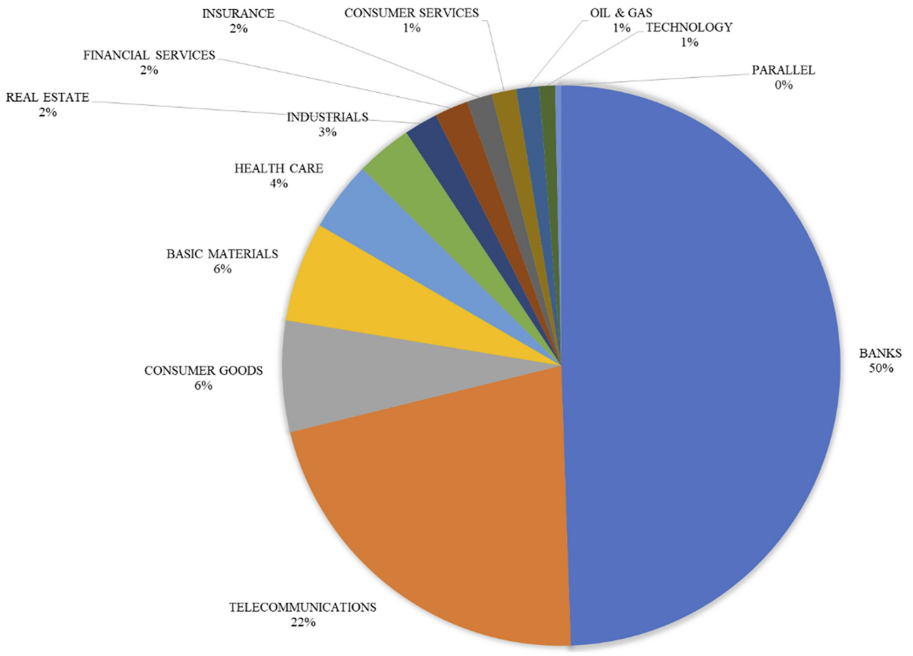

Figure 2.
The market capitalization for all Kuwait stock market sectors

4,246 observed testing data. In order to know the effect of the rate of return in Boursa Kuwait based on factors such as oil price, gold price, exchange rate of KWD to USD, money supply, interest rate, EPS, DPS and the five indices of the Gulf stock markets, a new feature has been added to the table, which is the rate of return.

The following mathematical equation was used (Dayananda *et al.*, 2002):

$$\text{Rate of returns} = \ln(\text{close price}_n) - \ln(\text{close price}_{n-1})$$

This feature was then analyzed based on its direction – downward, upward or stable.

Data analysis

In these steps, the data were converted to information by using different data mining techniques such as regression, SVM, decision tree and random forest on a sample of Kuwait stock market data, along with variables that affected the Kuwaiti stock market. Then we used RStudio software to build the models for each technique, and subsequently compared the models based on the accuracy rate. RStudio is free and open-source software for data science, scientific research and the technical community, which uses the *R* language. *R* is a free language and environment for statistical computing and graphics, which provides a wide variety of statistical and graphical techniques (Misra, 2020; RStudio, 2020). RStudio software was used to analyze the collected data for this study, and techniques such as regression, support vector machine, decision tree and random forest were used.

Results and discussion

Regression test

This study used the multinomial logistic regression analysis test, which is a type of regression that predicts the probabilities of more than two possible outcomes of a categorically distributed dependent variable based on independent variables. So, the regression test will present the effect of selected variables on the direction rate of return for the Kuwait stock market in three categories – upward, downward and stable – as well as providing an equation for making predictions about the rate of return based on selected variables. The accuracy of the multinomial logic regression test is 54.24.

The results of the multinomial logic regression test. All variables are statistically significant at the 5% level except gold price and oil price. Also, the variables that do not have an influence in the direction rate of return of Boursa Kuwait are money supply and gold price. The regression equation of the logistic model is based on this test, with three categories having two logit functions: the first logit function is for the probability of a stable direction relative to the probability of a downward direction; the second logit function is for the probability of an upward direction relative to the probability of a downward direction; and the equations based on this test are given as:

$$\begin{aligned}\ln(P_0/P_{-1}) = {} & 3.98 - 0.76\,(\text{ALMUTAHED}) - 1.12\,(\text{AUB}) - 0.43\,(\text{BOUBYAN}) \\ & - 1.04\,(\text{BURG}) + 0.42\,(\text{CBK}) - 1.1\,(\text{GBK}) - 1.64\,(\text{KFH}) - 0.93\,(\text{KIB}) \\ & - 1.6\,(\text{NBK}) - 1.13\,(\text{OOREDOO}) - 1.58\,(\text{VIVA}) - 1.4\,(\text{Warba}) \\ & - 1.68\,(\text{ZAIN}) - 0.02\,(\text{Oil price}) - 1.03\,(\text{Exch. rate}) - 117.16\,(\text{Int. rate}) \\ & + 3.85\,(\text{EPS}) - 5.16\,(\text{DPS}) + 57.52\,(\text{Kuwait}) + 12.13\,(\text{Oman}) - 6.5\,(\text{KSA}) \\ & + 16.6\,(\text{Bahrain}) + 9.03\,(\text{Dubai})\end{aligned}$$

$$\ln(P_1/P_{-1}) = -3.88 + 0.17\,(\text{ALMUTAHED}) - 0.03\,(\text{AUB}) - 0.04\,(\text{BOUBYAN}) - 0.09\,(\text{BURG}) + 0.34\,(\text{CBK}) + 0.09\,(\text{GBK}) + 0.02\,(\text{KFH}) + 0.08\,(\text{KIB}) + 0.14\,(\text{NBK}) - 0.13\,(\text{OOREDOO}) - 0.49\,(\text{VIVA}) - 0.08\,(\text{ZAIN}) - 0.01\,(\text{Oil price}) + 0.43\,(\text{Exch.Rate}) + 0.17\,(\text{Int. Rate}) + 5.78\,(\text{EPS}) - 3.47\,(\text{DPS}) + 128.15\,(\text{Kuwait}) + 24.80\,(\text{Oman}) - 2.56\,(\text{KSA}) + 1.75\,(\text{Bahrain}) + 3.62\,(\text{Dubai})$$

Table 1 shows the confusion matrix of multinomial logistic regression for training data and testing data. As we see, the highest value in the confusion matrix for both training and testing data is when the actual direction rate of return is stable and predicts stability. The lowest value in the confusion matrix for both training and testing data is when the actual direction rate of return is stable and predicts a downward direction. Also, the accuracy scores for both training data and testing data are nearly the same.

Support vector machine test

The support vector machine test is built by RStudio software. Table 2 displays the confusion matrix of the SVM, where the parameters used are the kernel function polynomial basis, and the regularization parameter (C) is 10. Seemingly, the highest value in the confusion matrix is when the actual direction rate of return is stable and predicts stability, and the lowest value is when the actual direction rate of return is stable and predicts an upward direction. The accuracy of the confusion matrix of the polynomial kernel function of SVM is 52.73.

Table 3 indicates the class of the direction of rate of return of the polynomial kernel function of the SVM; the highest-class accuracy is the stable class, and the positive and negative classes are nearly the same.

Table 4 displays the confusion matrix of the SVM, where the parameters used are the kernel function radial basis and the regularization parameter (C) is 10.

Table 5 indicates the class of the direction of the rate of return of the radial kernel function of the SVM; the highest-class accuracy is the stable class, and the classes of upward and downward direction are very close.

	Training data			Testing data		
Prediction	−1	0	1	−1	0	1
−1	1608	663	739	543	213	268
0	1458	3239	1377	446	1111	437
1	1022	781	1849	320	259	649
Accuracy		47.42			45.76	

Table 1. The confusion matrix of multinomial logistic regression

	Reference		
Prediction	−1	0	1
−1	535	208	280
0	561	1223	593
1	213	152	481

Table 2. The confusion matrix of polynomial kernel function of SVM

Decision trees test

The decision tree is used to predict the direction of the rate of return in Boursa Kuwait and is constructed by RStudio software. Figure 3 shows the results of implementing the decision tree. The size of this tree is 9. The leaves of the tree explain the decision tree's prediction rules.

Also, it is visible that if the condition is less than the interest rate of 0.019, then the predicted direction is stable (0). Furthermore, the stability overwhelms the results where it is predicted three times: one time, the rules predict a downward direction, and another time, an upward direction. The accuracy of the decision tree is 53.56.

Table 6 presents the confusion matrix of the decision tree, where it seems that the highest value in the confusion matrix is when the actual direction of the rate of return is stable and predicts stability.

Table 7 indicates the class of the direction of the rate of return of the decision tree; the highest-class accuracy is the stable class, and the positive and negative classes are nearly the same.

Random forest test

First, the random forest test searches for the best number of variables available for splitting at each tree node from 2 to 10 based on accuracy. Table 8 shows the accuracy for each variable, and the highest accuracy for the number 8 is 53.1.

Table 9 displays the confusion matrix of the random forest, where the parameters used are the number of variables for splitting at each tree node (5) and the number of trees to grow (100). The result of random forest accuracy is 53.04.

Table 10 indicates the class of the direction of the rate of return of the random forest; the highest-class accuracy is the stable class, then the positive class, followed by the negative class.

Table 3.
Statistics by classes of the direction of rate of return of the polynomial kernel function of SVM

	Class: −1	Class: 0	Class: 1
Sensitivity	40.87	77.26	35.52
Specificity	83.38	56.67	87.38
Positive predictive value	52.3	51.45	56.86
Negative predictive value	75.99	80.74	74.32
Balanced accuracy	62.13	66.96	61.45

Table 4.
The confusion matrix of radial kernel function of SVM

		Reference	
Prediction	−1	0	1
−1	615	262	350
0	411	1093	434
1	283	228	570

Table 5.
Statistics by classes of the direction of rate of return of the radial kernel function of SVM

	Class: −1	Class: 0	Class: 1
Sensitivity	46.98	69.05	42.1
Specificity	79.16	68.27	82.33
Positive predictive value	50.12	56.4	52.73
Negative predictive value	77.01	78.77	75.23
Balanced accuracy	63.07	68.66	62.21

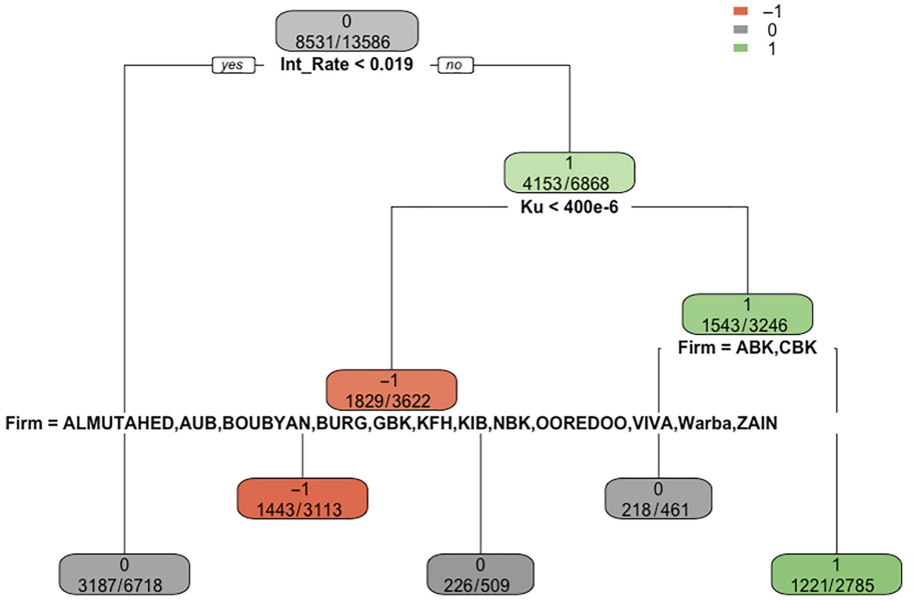

Figure 3.
Decision tree for prediction the direction of the rate of return

Prediction	Reference −1	Reference 0	Reference 1
−1	512	182	309
0	568	1271	554
1	229	130	491

Table 6.
The confusion matrix of decision tree for prediction direction rate of return based on Kuwait index

	Class: −1	Class: 0	Class: 1
Sensitivity	39.11	80.29	36.26
Specificity	83.28	57.87	87.59
Positive predictive value	51.05	53.11	57.76
Negative predictive value	75.42	83.16	74.59
Balanced accuracy	61.2	69.08	61.92

Table 7.
Statistics by classes of the direction of rate of return of the decision tree

While the random forest model is popular for its predictive performance, it also provides the feature of being a fully non-parametric measure of variable importance (VIMP), which supplies insight into a system by identifying which variables play a key role in prediction (Ishwaran & Lu, 2019).

Figure 4 illustrates the variable importance for our developed random forest model. As shown in Figure 4, the three highest variables that affect the direction of the rate of return in Boursa Kuwait are the firm's variable, the Kuwait index and the EPS. We can also observe that money supply has the lowest effect on the direction of the rate of return in Boursa Kuwait.

Results and discussion

This research study was prepared to predict the direction of the rate of return for the Kuwait stock market by using data mining techniques. The data were collected from two sources: the

Table 8.
The accuracy scores for random variables of trees

No.	Variables	Accuracy
1	2	52.7
2	3	53.1
3	4	52.7
4	5	52.7
5	6	53
6	7	52.8
7	8	53.4
8	9	53.1
9	10	53.2

Table 9.
The confusion matrix of random forest

	Reference		
Prediction	−1	0	1
−1	616	282	307
0	397	973	348
1	368	292	663

Table 10.
Statistics by classes of the direction of rate of return of the random forest

	Class: −1	Class: 0	Class: 1
Sensitivity	44.61	62.9	50.3
Specificity	79.44	72.4	77.46
Positive predictive value	51.12	56.64	50.11
Negative predictive value	74.84	77.29	77.59
Balanced accuracy	62.02	67.65	63.88

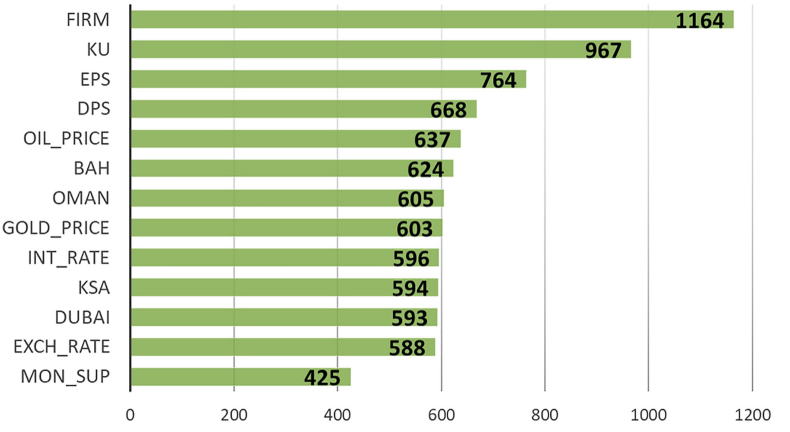

Figure 4.
Scale of variable importance in random forest

data about Boursa Kuwait, and the variables that affect the direction of the rate of return in the Kuwait stock market, such as company data, EPS, DPS, oil price, gold price, exchange rate of KWD to USD, interest rate, money supply and the Gulf stock market index. Consequently, it consists of 16,982 readings from January 6, 2015 to November 26, 2019. The data were analyzed using four methods: the multinomial logistic regression test, the support vector machine test, the decision tree test and the random forest test.

The multinomial logistic regression test was used to measure the correlation and the influence of the variables on the direction of the rate of return. There are 14 variables that have significance on the direction of the rate of return, and the Kuwait index has the highest coefficient, equal to 99.460. The results of this test were applied to predict the value of the direction of the rate of return based on equations.

In the support vector machine test, the radial kernel function has better accuracy than the polynomial kernel function. Also, the stable class performs best based on the accuracy of other classes.

The decision tree test has good accuracy with a value of 53.56 and, based on the confusion matrix, its best prediction is in stability. Furthermore, the stable class performs best based on sensitivity, with a value of 80.29 and an accuracy of 69.08 compared to other classes.

The random forest test has good accuracy and, based on the confusion matrix, its best prediction is in stability. Moreover, the stable class performs best based on sensitivity with a value of 62.9 and an accuracy of 67.65 compared to other classes. Additionally, the height score of the variable that affects the direction of the rate of return is the firms.

Table 11 indicates the accuracy of the overall performance of the four models. Accuracy is the parameter for evaluating the performance of a model; all the tests have around 50% accuracy scores, which means that the models are only moderately effective. Therefore, the accuracy will be arranged as follows: multinomial logistic regression, radial kernel function in SVM, decision tree, polynomial kernel function in SVM and, finally, random forest.

Conclusion

The main purpose of this study was to discover the relation of the probabilities of more than two possible directions of Boursa Kuwait based on other variables, and also to determine which class has highly accurate prediction of the directions of the rate of return of Boursa Kuwait. Furthermore, this study assists the decision-making process by mapping out different potential outcomes of directions of the rate of return of Boursa Kuwait by the decision trees, identifies the variables that affect the directions of the rate of return of Boursa Kuwait and recognizes suitable methods of analyzing the big data of the Boursa market.

The multinomial logistic regression analysis is used in this study to indicate the effect of selected variables on the direction and the rate of return for the Kuwait stock market. So, the variable that has more effect in the direction rate of return of Boursa Kuwait is the Kuwait index, and the variables that have no influence in the direction rate of return of Boursa Kuwait are money supply and gold price.

The support vector machine test and the random forest test are used to identify which class can more accurately predict the directions of the rate of return of Boursa Kuwait, and both tests agree that the highest-class accuracy is the stable class.

Table 11. The accuracy of the overall performance of the models' results

		SVM			
	Multinomial logistic regression	Radial Kernel function	Polynomial kernel function	Decision tree	Random forest
Accuracy	54.24	53.65	52.73	53.56	52.4

The decision tree test is used in this study to identify the direction of the rate of return of Boursa Kuwait based on independent variables; therefore, the decision tree is constructed based on the interest rate. The other variables that affect the direction of the rate of return are the Kuwait index and firms.

The random forest test provides the feature of a non-parametric measure of variable importance (VIMP) that can identify which variables play a key role in prediction. The highest variables that affect the direction of the rate of returns in Boursa Kuwait is the firm's variable, followed by the Kuwait index and then EPS.

Based on the accuracy scores provided by the models used in this study, all tests showed very similar accuracy, which was moderately effective; therefore, the accuracy will be arranged as follows: multinomial logistic regression, radial kernel function in SVM, decision tree, polynomial kernel function in SVM and, finally, random forest.

Recommendations

Based on the results of this study, the following recommendations are suggested:

(1) In the case of data mining results, accuracy depends on the quality of the used data, so it is paramount that an effort is made to verify and preprocess the data.

(2) The three highest variables that affect the direction of the rate of returns in Boursa Kuwait are firms, the Kuwait index and EPS.

(3) Employ data mining techniques in the stock market in order to provide more considered findings, which will lead to an increase in the quality of decisions.

(4) Encourage the decision-makers to utilize data mining techniques within their analytical and strategic planning efforts.

(5) Conduct further studies on utilizing more data mining techniques and tools to support decisions in the stock market.

Future works

For further works, the following are a few suggestions:

(1) Improve the models that are used in this study, such as multinomial logistic regression, SVM, decision tree and random forest, by applying the models to all the companies listed in the Kuwait stock market.

(2) Elaborate on the knowledge of Kuwait stock market returns by using further data mining techniques, such as Apriori, k-means, C4.5, AdaBoost, naïve Bayes, k-nearest neighbors (k-NN) and expectation-maximization (EM).

(3) Reconsider the factors that affect the Kuwait stock market return, such as trading volume, financial news, political news, global indicators and Morgan Stanley Capital International (MSCI).

(4) Finally, employ these data mining techniques on other stock markets, such as GCC, Middle East countries and global markets.

References

Ahmed, S. R. (2004). Applications of data mining in retail business. *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.*

Al-Radaideh, Q. A., Assaf, A. A., & Alnagi, E. (2013). Predicting stock prices using data mining techniques. *The International Arab Conference on Information Technology (ACIT'2013)*.

Almuhaideb, A. (2021). Big data analytics in COVID-19. *Encyclopedia*.

Alsultanny, Y. A. (2013). Labor market forecasting by using data mining. *Procedia Computer Science*, *18*, 1700–1709. doi: 10.1016/j.procs.2013.05.338.

Awan, M. J., Mohd Rahim, M. S., Nobanee, H., Munawar, A., Yasin, A., & Zain, A. M. (2021a). Social media and stock market prediction: A big data approach. *Computers, Materials and Continua*, *67*(2), 2569–2583.

Awan, M. J., Mohd Rahim, M. S., Nobanee, H., Yasin, A., & Khalaf, O. I. (2021b). A big data approach to black friday sales. *Intelligent Automation and Soft Computing*, *27*(3), 785–797.

Bley, J., & Chen, K. H. (2006). Gulf cooperation council (GCC) stock markets: The dawn of a new era. *Global Finance Journal*, *17*(1), 75–91.

Boursakuwait (2019). *About Boursa Kuwait*. Boursakuwait, available from: https://www.boursakuwait.com.kw/page/39.

Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications*, *112*, 353–371.

Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Systems with Applications*, *80*, 340–355.

Cheng, K.-C., Huang, M.-J., Fu, C.-K., Wang, K.-H., Wang, H.-M., & Lin, L.-H. (2021). Establishing a multiple-criteria decision-making model for stock investment decisions using data mining techniques. *Sustainability*, *13*(6). doi: 10.3390/su13063100.

Dayananda, D., Irons, R., Harrison, S., Herbohn, J., & Rowland, P. (2002). *Capital budgeting: Financial appraisal of investment projects*. Cambridge: Cambridge University Press.

Devi, A. M., & Devaki, A. (2019). Applications of quantitative techniques in decision making of business organisation. *International Journal of Trend in Scientific Research and Development*, *3*(3), 568–571.

Fairhurst, D. S. (2021). *Financial modeling in excel for dummies*. John Wiley & Sons.

Gerth, P., Sieverling, A., & Trognitz, M. (2017). Data curation: How and why. A showcase with re-use scenarios. *Studies in Digital Heritage*, *1*(2), 182–193.

Gupta, A., Bhatia, P., Dave, K., & Jain, P. (2019). Stock market prediction using data mining techniques. *2nd International Conference on Advances in Science and Technology (ICAST)*.

Haafza, L. A., Awan, M. J., Abid, A., Yasin, A., Nobanee, H., & Farooq, M. S. (2021). Big data COVID-19 systematic literature review: Pandemic crisis. *Electronics*, *10*(24). doi: 10.3390/electronics10243125.

Imandoust, S. B., & Bolandraftar, M. (2014). Forecasting the direction of stock market index movement using three data mining techniques: The case of Tehran stock exchange. *International Journal of Engineering Research and Applications*, *4*(6), 106–117.

Ishwaran, H., & Lu, M. (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine*, *38*(4), 558–582.

Kim, M. (2021). A data mining framework for financial prediction. *Expert Systems with Applications*, *173*. doi: 10.1016/j.eswa.2021.114651.

Kohli, P. P. S., Zargar, S., Arora, S., & Gupta, P. (2019). Stock prediction using machine learning algorithms. In *Applications of Artificial Intelligence Techniques in Engineering* (pp. 405–414). Springer.

Lai, L. K., & Liu, J. N. (2010). Stock forecasting using support vector machine. *2010 International Conference on Machine Learning and Cybernetics*.

Mao, S., Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., & Huang, G.-B. (2018). An automatic identification system (AIS) database for maritime trajectory prediction and data mining. *Proceedings of ELM-2016* (pp. 241–257). Springer.

Metzger, A., Leitner, P., Ivanović, D., Schmieders, E., Franklin, R., Carro, M., . . ., & Pohl, K. (2014). Comparing and combining predictive business process monitoring techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *45*(2), 276–290.

Misra, B. B. (2020). Open-source software tools, databases, and resources for single-cell and single-cell-type metabolomics. In *Single Cell Metabolism* (pp. 191–217). Springer.

Murray, G., & Scime, A. (2015). Data mining. In R. Scott, & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences : an interdisciplinary, searchable, and linkable resource* (pp. 1–15). John Wiley & Sons. doi: 10.1002/9781118900772.etrds0071.

Ou, P., & Wang, H. (2009). Prediction of stock market index movement by ten data mining techniques. *Modern Applied Science*, *3*(12), 28–42.

Parr, E., & Vaudrevange, P. K. (2020). Contrast data mining for the MSSM from strings. *Nuclear Physics B*, *952*, 114922.

Petersen, R. (2016). 20 companies do data mining and make their business better. BarnRaisers, LLC. available from: https://barnraisersllc.com/2016/11/07/companies-data-mining-business-better/ [accessed 7 November 2016].

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R. L. (2019). In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data. In *Advances in informatics and computing in civil and construction engineering* (pp. 19–26). Springer.

Rafiq, F., Awan, M. J., Yasin, A., Nobanee, H., Zain, A. M., & Bahaj, S. A. (2022). Privacy prevention of big data applications: A systematic literature review. *SAGE Open*, *12*(2). doi: 10.1177/21582440221096445.

RStudio (2020). What makes RStudio different?. available from: https://rstudio.com/about/what-makes-rstudio-different/.

Shashaank, D., Sruthi, V., Vijayalakshimi, M., & Garcia, J. S. (2015). Turnover prediction of shares using data mining techniques: A case study. *arXiv*, preprint arXiv:1508.00088.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572.

Tariq, A., Awan, M. J., Alshudukhi, J., Alam, T. M., Alhamazani, K. T., Meraf, Z., & Velmurugan, P. (2022). Software measurement by using artificial intelligence. *Journal of Nanomaterials*, 1–10. doi: 10.1155/2022/7283171.

Tsai, C.-F., & Hsiao, Y.-C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, *50*(1), 258–269.

Usmani, M., Adil, S. H., Raza, K., & Ali, S. S. A. (2016). Stock market prediction using machine learning techniques. *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*.

Yu, L., Wang, S., & Lai, K. K. (2005). Mining stock market tendency using GA-based support vector machines. *International workshop on Internet and network economics*.

Zhong, X., & Enke, D. (2017). A comprehensive cluster and classification mining procedure for daily stock market return forecasting. *Neurocomputing*, *267*, 152–168.

Corresponding author
Bedour M. Alshammari can be contacted at: bedourmfs@agu.edu.bh