Received: September 30, 2020. Revised: November 25, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Prediction of LQ45 Index in Indonesia Stock Exchange: A Comparative Study of Machine Learning Techniques

Abdul Syukur[1]* **Deden Istiawan[2]**

[1]*Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia*
[2]*Statistics Academy of Muhammadiyah, Semarang, Indonesia*
* Corresponding author's Email: abah.syukur01@gmail.com

Abstract: LQ45 is an Indonesia Stock Exchange Index (ISX) incorporate of 45 companies that meet certain criteria to target investors for selecting certain stocks. The prediction of stock price direction in the financial world is a major issue. The implementation of machine learning and other algorithms for market price analysis and forecasting is a very promising field. Different types of classification algorithms were used to predict the stock market. However, when individual studies are considered separately there is no clear consensus that algorithms work best. In this research, a comparison framework is proposed, which aims to benchmark the performance of a wide range of classification models and use them to predict the LQ45 index. The data in this research contains the transaction level and capitalization size are obtained from the Indonesian Stock Exchange (ISX). For analysis purposes, we set out 10 classifiers that can be used to build classification models and test their performance in the LQ45 dataset. The performance criterion chosen to measure this effect is accuracy, recall, and precision. The results showed that the random forest algorithm had the best performance for predicting the LQ45 index. Whilst the classification and regression trees, C4.5, support vector machine, and logistic regression algorithms also perform well. Besides, the models based on traditional statistical-based learners that are Naïve Bayes and linear discriminant analysis seem to underperform for predicting the LQ45 index. These results are not only beneficial to enrichment the machine learning techniques literature but also have a significant influence on the stock market prediction in terms of the ability to predict the LQ45 index.

Keywords: LQ45, Stock exchange, Stock price prediction, Machine learning.

1. Introduction

The stock market plays an important role in the development of the country. It offers investors an alternative investment and a source of corporate financing [1]. They have had a significant impact on many sectors, such as business, education, employment, technology, and thus on the economy [2-3]. Investment is the placement of the number of funds at a certain time in the hope of making profits in the future [4]. LQ45 is an Indonesia Stock Exchange Index (ISX) made up of 45 companies that meet certain criteria to target investors to select certain stocks [5]. LQ45 index is a calculation of 45 stocks, which are selected through several selection criteria. Apart from assessing liquidity, the selection of these shares considers market capitalization. The LQ45 index contains 45 stocks adjusted every six months at the beginning of February and August.
LQ45 index is one of the most popular and influential stock indices on the Indonesia Stock Exchange. LQ45 Index is a driving force for the Composite Stock Price Index, where the Composite Stock Price Index consists of all stocks on the Indonesian Stock Exchange. When the LQ45 index rises, the Composite Stock Price Index strengthens and vice versa. This is because the companies listed in the LQ45 index have good company performance. LQ45 index also contains blue-chip stocks, which refer to large companies that have a stable income. Blue-chip stocks are in great demand because they are considered safe and promising stocks. LQ45 index provides an objective and reliable tool for financial analysis, investment managers, investors as well as capital market observers in monitoring the price

movements of actively traded stocks. The main objective of investors investing in listed companies is to increase the wealth obtained through stock returns. Therefore, LQ45 index prediction can be an initial screening in investing to select stocks that have the potential to provide profit in the future. Investors can make it a reference for investing in shares in the Indonesian capital market.

Stock market prediction is an area of great interest due to the potential for very high returns on money invested in a very short period [6]. However, the analysis of stock market movements and price behavior is very challenging because of the dynamic, non-linear, non-stationary, non-parametric, noisy, and chaotic nature of markets [7]. Four approaches are widely used in stock market prediction techniques, which are statistics, pattern recognition, machine learning, and sentiment analysis [8].

Statistical techniques often assume linearity, stationarity, and normality provided a way to analyze and predict stock. Statistical techniques cannot be used to model the complexity and non-stationary nature of stock markets [9]. Pattern recognition techniques do pattern matching to identify future trends based on the historical template. Pattern recognition focus on the detection of pattern in data [10]. Patterns in stock markets are recurring sequences found in Open-High-Low-Close candlestick charts which traders have historically used as buy and sell signals. In general, pattern recognition techniques show promises but on their own do not give convincing results on stock prediction [11]. Sentiment analysis is another approach that has lately been used for stock market analysis [12]. It is the process of predicting stock trends via automatic analysis of text corpuses such as news feeds or tweets specific to stock markets and public companies. Sentiments can drive short-term market fluctuations which in turn causes a disconnect between the stock price and the true value of a company's share. Over long periods, however, the weighing machine kicks in as the fundamentals of a company ultimately cause the value and market price of its shares to converge [8].

Machine learning techniques are widely applied to stock market problems. They offer useful tools to predict noisy environments like stock markets, capturing their non-linear behavior [13]. Over the years, machine learning has played a vital role in predictions. Nowadays, machine learning techniques are used for predicting the stock market prediction. Particularly, for stock market prediction, the data size is huge, and also non-linear are used for stock market prediction. To deal with this variety of data efficient model is needed that can identify the hidden patterns and complex relations to deal with this variety of data efficient model is needed that can identify the hidden patterns and complex relations in this large data set. Machine learning techniques in this area have proved to improve efficiencies by 60-86 percent as compared to past methods [14].

In this study, we focus on machine learning techniques. Machine learning has been studied extensively for its potential in financial market prediction [15]. Several algorithms have been used in predicting the direction of stock prices. Simpler techniques such as decision tree [16], discriminant analysis [17], support vector machines [18-20], neural networks [21], and random forest [22-23].
While many studies individually report the comparative performance of the machine learning algorithms they have used, there is no clear agreement about which algorithm performs best when individual studies are looked at separately. Hence, the classification algorithms are widely used, and successful results obtained from the algorithms are also used for stock market prediction. Which classification algorithm to choose is a very important decision. There is no specific classification algorithm to solve the current problem. In other words, the best algorithm does not solve every problem in the best way. There are classification algorithms that give different results for different datasets or different problems. A classification algorithm that is considered to be the best solution to solve a problem may not work in another problem or dataset. For this reason, different classification algorithms for the given dataset must be compared before problem-solving. The algorithm that best solves the problem is the algorithm obtained by comparison with the specific statistical criteria. Thus, the algorithm to be used in problem-solving is determined. In this study, for determining the best algorithms for the current dataset, all classification algorithms were compared for the suitability of data. When choosing an algorithm that is known as giving good results without comparing it is performance to different algorithms may give misleading results.

The main contribution of this study is a comprehensive benchmark that compares the performance between the traditional statistical-based learners, distance-based learners, tree-based learners, artificial neural networks, and support vector machines. To build classification models and test their performance in the LQ45 dataset and then determined algorithms were compared with accuracy, precision, and recall. We believe that our findings obtained from a real application would contribute to facilitating the investors in determining the stock option. The results of this study are not only

Received: September 30, 2020. Revised: November 25, 2020.

beneficial to the literature but also have a significant influence on the stock market prediction in terms of the ability to predict the LQ45 index.

This research is organized as follows. In section 2, the proposed comparison framework is explained. The experimental results of the classification model's comparison are presented in section 3. Finally, our work in this paper is summarized in the last section.

2. Material and methods

2.1 Data description

The LQ45 index consists of 45 issuers with high liquidity, which were selected through several selection criteria. In addition to assessing liquidity, the selection of these issuers also considers market capitalization. The Indonesia Stock Exchange (ISE) regularly monitors developments in the performance of issuers included in the LQ45 index calculation. Every three months, an evaluation of the order of the shares is carried out. Share replacement will be carried out every six months, namely at the beginning of February and August.

The dataset of this study was issued from (https://www.idx.co.id) with a total of 261 companies approved in the ISE from 2015 to 2018 years. The number of companies included in LQ45 was 180 and 81 Non-LQ45. A brief description of the dataset of this study is presented in Table 1.

Table 1. Dataset description

No	Variable	Description
1	Volume	Volume is a measure of how much of a given financial asset has been traded in a given period.
2	Value	Value is a range of prices where the majority of trading volume took place on the prior trading day.
3	Frequency	Frequency is the number of trades executed in a specific time interval.
4	Days	The trading day or regular trading hours (RTH) is the period that a particular stock exchange is open.
5	Earnings per Share	Provides the profitability indication of a firm, and can be determined by dividing the firm's net income by its whole number of remaining stocks.
6	Book Values per Share	Market value ratio that weighs Stockholders' equity against shares outstanding. In other words, the value of all shares divided by the number of shares issued.
7	Debt to Assets Ratio	The leverage ratio measures the number of total assets that are financed by creditors instead of investors.
8	Debt to Equity Ratio	A financial, liquidity ratio that compares a company's total debt to total equity.
9	Return on Assets	This ratio signifies the proportion of earnings a firm earns about the firm's overall assets or resources. Thus, an indication of how profitable a firm is relative to the firm's total resources or assets.
10	Return on Equity	This ratio offers an overview of how well the shareholder's funds were used and the gain made out of its investment. When ROE is low, it implies that the shareholder's funds were not used properly.
11	Gross Profit Margin	Gross profit margin is a ratio that indicates the performance of a company's sales and production.
12	Operating Profit Margin	The earnings that a business generates from its operating activities.
13	Net Profit Margin	The percentage of revenue left after all expenses have been deducted from sales.
14	Payout Ratio	Shows the proportion of earnings paid out as dividends to shareholders.
15	Yield	Yield refers to the earnings generated and realized on an investment over a particular period.

2.2 Prediction models

This study aims to compare the performance of various machine learning techniques for the prediction of the LQ45 index. The ten classification algorithms have been chosen, which can be grouped into the categories of traditional statistical-based learners (logistic regression, linear discriminant analysis, and Naïve Bayes), distance-based learners (k-nearest neighbor and k-star), tree-based learners (classification and regression trees, C4.5, and random forest), artificial neural networks, and support vector machines. This selection aims to determine the best classification algorithm for LQ45 index prediction.

The proposed framework is shown in Fig. 1, while a brief explanation of each technique applied in this paper is presented below.

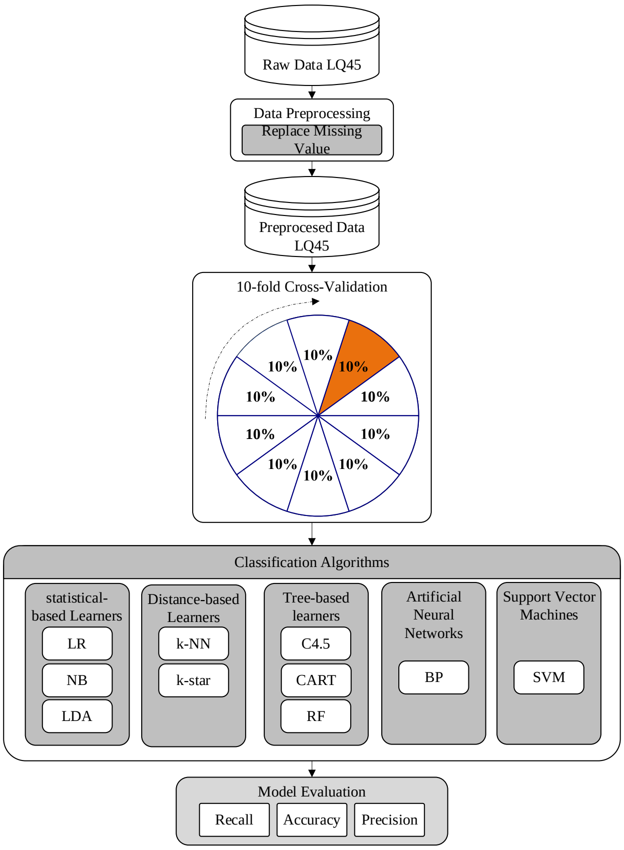

Figure. 1 Proposed comparison framework of classification models for prediction of the LQ45 index

2.2.1. Logistic regression

Logistic regression (LR) is an extension of linear regression. This technique is used when the response variable or class is discrete, which means that linear regression cannot be used directly for data modeling [24]. In case the response variable is binomial, the output of the logistic regression model is expressed in probabilistic terms, where a probability value close to 0 indicates a low probability of occurrence and a probability value close to 1 indicates a high probability of occurrence [25]. The logistic regression model then takes the form in Eq. (1) as follows:

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta^T \boldsymbol{x} \qquad (1)$$

where α is the intercept parameter and β^T contains the variable coefficients [26].

2.2.2. Linear discriminant analysis

Linear discriminant analysis (LDA) is a dependency statistical analysis technique that has the use of classifying several groups of objects. This grouping with discriminant analysis occurs because there is the influence of one or more other variables that are independent. The linear combination of these variables will form a discriminant function [27]. A linear discriminant function is formulated in Eq. (2) as follows:

$$z = w_1x_1 + w_2x_2 + \cdots + w_kx_k \qquad (2)$$

where $x_1, x_2, \ldots, x_k$ are independent variables. The quantity z is called the discriminant score, and $w_1, w_2, \ldots, w_k$ are called weights [28].

2.2.3. Naïve bayes

Naive Bayes (NB) is an algorithm in machine learning that applies the Bayes theory of classification. The main characteristic of the NB classifier is a very strong assumption (naive) of the independence of each condition or event [29]. Given a training dataset, $D = (X_1, X_2, \ldots, X_n)$, each data record is represented as, $X_i = (X_1, X_2, \ldots, X_n)$. D contains the following attribute values $(A_1, A_2, \ldots, A_n)$ and also contains a set of classes $C = (C_1, C_2, \ldots, C_m)$. For test instance X the classifier will predict that X belongs to the class with the highest posterior probability conditioned on X (as formulated in Eq. (3), if and only if $P(C_i|X) > P\left(C_j\middle|X\right)$ for $1 \leq j \leq m, j \neq i$. The class C_i for which $P(C_i|X)$ is maximized called the maximum posterior hypothesis [30].

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (3)$$

As $P(X)$ is constant for all classes, only $P(X|C_i)P(C_i)$ needs to be maximized.

2.2.4. *K*-Nearest Neighbour

The K-Nearest Neighbour (K-NN) classification algorithm is a supervised learning algorithm, which means that this algorithm uses existing data and the output is known. The K-NN is a machine learning algorithm that is used to classify objects based on the training data that is closest to the object. The purpose of the K-NN algorithm is to classify new objects based on attributes and samples from training data. Suppose that there are j training categories, as $C_1, C_2, \ldots, C_j$ and the sum of the training samples is N. Also, class X is the same feature vector as all the training samples. When d_i is one of the neighbors in the training set, $y(d_i, C_j) \in (0,1)$ indicates whether d_i belongs to class C_j and $sim(X, d_i)$ is the similarity function for X and d_i. Then, the probability density function $P(X, C_j)$ for the feature data X, given class

C_j can be written in Eq. (4) as follows:

$$P(X, C_j) = \sum_{d_i \in k-NN} sim(X, d_i) y(d_i, C_j) \quad (4)$$

The $sim(X, d_i)$ can be calculated using the Euclidean distance, cosine, and correlation methods. In this study, the Euclidean distance method was selected because it is often used as the distance metric [31].

2.2.5. K-Star

K-star is an instance-based classifier. The class of a test instance is based on the training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function [32]. Given a set of infinite points and a set of transformations predefined T. Let t be a value of set T. This t will map $t: I \rightarrow I$. To map instances with itself σ is used in $T(\sigma(\alpha) = \alpha)$. σ terminates all prefix codes from T^*. Set T^* consists member which defines transformation one to one on I. The function t is formulated in Eq. (5) as follows:

$$t(\alpha) = t_n(t_n - 1(\ldots t_1(\alpha) \ldots)) \quad (5)$$

where $t = t_1, t_2, \ldots, t_n$. Calculation of total probability of all paths from instance α to an instance b. The P^* follows the probability of all paths from instance α to instance b:

$$P^*\left(\frac{b}{a}\right) = \sum_{t \in p; t(\alpha)=b} P(t) \quad (6)$$

The K^* is then formulated as Eq. (7) as follows:

$$K^*\left(\frac{b}{a}\right) = -log_2 P^*\left(\frac{b}{a}\right) \quad (7)$$

The K^* is not exactly a distance function, and the point to be an underscore $K^*(a|a)$ is usually non-zero also is not a symmetric function [33].

2.2.6. C4.5

C4.5 algorithm was introduced firstly by Quinland (1996) which is a development of the ID3 algorithm. The ID3 algorithm can only be used on features with categorical types, while numeric types cannot be used. The improvement that differentiates the C4.5 algorithm from ID3 is that it can handle features with numeric types and pruning decision trees [34]. C4.5 algorithm uses the gain ratio in determining the features that break the nodes in the induced tree [35] which is formulated by Eq. (8) as follows:

$$Gain\ Ratio(s, j) = \frac{Gain\ (s, j)}{Split\ Info\ (s, j)} \quad (8)$$

In Eq. (8), the value of the gain ratio is $j - th$ feature with $Split\ Info\ (s, j)$ is obtained from Eq. (9) below.

$$Split\ Info\ (s, j) = -\sum_{i=1}^{k} p(v_i|s) log_2\, p(v_i|s) \quad (9)$$

2.2.7. Classification and regression tree

The classification and regression trees (CART) is a binary recursive partitioning procedure capable of processing continuous and nominal attributes both as targets and predictors [36]. The CART algorithm implements the least squared deviation (LSD) impurity to determine the splitting rules and goodness of fit [37]. The LSD measure ($R(t)$) can be simply calculated in Eq. (10), Eq. (11), and Eq. (12) as follows:

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} \omega_i f_i (y_i - \bar{y}))^2 \quad (10)$$

$$\bar{y} = \frac{1}{N_w(t)} \sum_{i \in t} \omega_i f_i y_i \quad (11)$$

$$N_w(t) = \sum_{i \in t} \omega_i f_i \quad (12)$$

where $N_w(t)$is the weighted number of samples in node t, ω_i is the value of the weighting response for record i (if any), f_i is the value of the record response (if any), y_i is the response value, and $\bar{y}_i$ is the mean value of response values. To split s at node t, the LSD uses the following criterion in Eq. (13) as follows:

$$Q(s, t) = R(t) - R(t_L) - R(t_R) \quad (13)$$

where t_R and t_L are the left and right child nodes of node t, respectively. The split s is determined to maximize the $Q(s, t)$.

2.2.8. Random forest

Random Forest (RF) is an ensemble classifier combining a random selection of features with bagging. Multiple decision trees are created using a random subset of the attributes. For classification, all trees are applied and make a prediction [38]. First, the bagging method is used to extract k number of training subsets from the set of training samples marked $Di = (i = 1,2, \ldots, k)$ and ensure that the sample size of each subset is identical to that of the training set. Second, the random subspace method is applied to build k number of decision trees for the subsets $\{h(X,\Theta_i), i = 1,2,\ldots,k\}$. X is the eigenvector used to determine the classification, and Θ_i is an independent and identically distributed random variable. Finally, a simple majority vote is adopted for the classification results of the k decision trees to acquire the final classification results [39]. The results can be expressed in Eq. (14) as follows:

$$H(x) = \underset{y}{\operatorname{argmax}} \sum_{i=1}^{k} I(h_i(x) = Y) \qquad (14)$$

where $H(x)$ represents the combined classification model, h_i is a single decision-tree classification model, Y is the output variable, and I is an indicator function

2.2.9. Artificial neural network

Artificial Neural Network (ANN) is a knowledge engineering concept in the field of artificial intelligence which is designed by adopting the human nervous system [20]. In general, the ANN design consists of an input vector with some values or features, which are given as input values to the ANN, each input value passes through a w weighted relationship, then all values are combined. The combined value is then processed by the activation function to produce a signal as output. The activation function uses a threshold value to limit the output value to always be within the specified threshold value [40].

2.2.10. Support vector machine

The Support Vector Machine (SVM) algorithm originates from a statistical learning theory [41] whose results are very promising to provide better results than other algorithms. SVM can be applied to high-dimensional data, even SVM that uses kernel techniques must map the original data from its original dimension to the relatively higher dimension. The basic idea of SVM is to maximize the best hyperplane boundary that functions as a separator of two data classes in the input space by measuring the hyperplane margin and finding its maximum point. Margin is the distance between the hyperplane and the closest data from each class. This closest data is known as a support vector [42].

2.3 Model validation

We use stratified k -fold cross-validation for learning and testing data. In k-fold cross-validation, the available data are split into mutually exclusive subsets. Each subset is used as a validation set one time while the model is constructed using the remaining subsets [43]. The 10-fold cross-validation breaks data into 10 sets of size $N/10$. It trains the classifier on nine datasets and tests it using the remaining one dataset. This repeats 10 times and we take a mean accuracy rate. For classification, the accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of instances in the initial dataset. We employ the stratified 10-fold cross-validation because this method has become the standard and state-of-the-art validation method in practical terms. Some tests have also shown that the use of stratification improves results slightly [44].

2.4 Model evaluation

Evaluation of the results of the experiment is a measuring tool that can be used to assess or measure how well the proposed method against other methods and whether the proposed method has a significant difference in the results of other models. In this study, the evaluation models used are accuracy, recall, and precision. The three statistical criteria are explained as follows:

Accuracy value (AC) is calculated by taking the correct prediction percentage from the whole data. According to the confusion matrix, it can be calculated using Eq. (15) as follows:

$$AC = \frac{TP+TN}{TP+FP+TN+FN} \qquad (15)$$

where TN is the true negative, TP is the true positive, FP is the false positive, and FN: false negative.

Sensitivity or recall (R) in the field of information search measures the proportion of original positives that are correctly predicted as positive. Sensitivity is related to the ability of testing to identify positive results from actually positive, which is calculated using Eq. (16) as follows:

$$R = \frac{TP}{TP+FN} \qquad (16)$$

where TP is the true positive and FN: false negative

Precision or positive predictive value (P) is a matrix to measure system performance in getting relevant data. Precision is the amount of data that is true positive divided by the amount of data that is recognized as positive, which is calculated using Eq. (17) as follows:

$$P = \frac{TP}{TP+FP} \qquad (17)$$

Table 2. Confusion matrix

Predicted	Actual	
	A	B
A	TP	FP
B	FN	TN

where TP and FP are the true positive and false positive values, respectively.

The values of the statistical criteria that are compared to classification algorithms are calculated by using a confusion matrix. The confusion matrix is shown in Table 2.

3. Experimental results

The experiments were conducted using a computing platform based on Intel Core i5 1.6 GHz CPU, 8 GB RAM, and Microsoft Windows 10 Home 64-bit. The development environment is the RapidMiner 9.2 library with the default parameter. For the artificial neural network, we tested the hidden layer: 1, training cycles: 200, learning rate: 0.01, momentum: 0.9, shuffle: yes, error epsilon: 1.0E-4. For support vector machine we tested the kernel type: dot, kernel cache: 200, C: 0, convergence epsilon: 0.001, max iteration:100000, L Pos: 1.0, L Neg: 1.0 and epsilon: 0.0. Since the prediction results of these classifiers are sensitive to data split, cross-validation is used to judge the ability of generalization. Besides, about 10% of the data are used as a test data set, while to simulate this partition, the 10-fold cross-validation method to evaluate models.

In this study, 10 classification algorithms are built and compared to each other using their predictive accuracy. Using the 10-fold cross-validation, the RF produced the best results with an overall prediction rate of 93.49%, and the SVM came out as the runner up with an overall prediction rate of 90.06%, as shown in Fig. 2.

On the other hand, models based on traditional statistical-based learners that are NB and LDA seem to underperform. The result confirmed the RF as the top algorithm for stock market prediction [22-23]. The other study also shows that RF performs significantly more accurately than C4.5, ELM, LR, NN [45]. Besides, the experimental study comparing the use of 9 classifiers (repeated incremental decision trees, RF, KNN, NB, LR, SVM, and ANN) shows that RF has remarkable performance in all datasets [46].

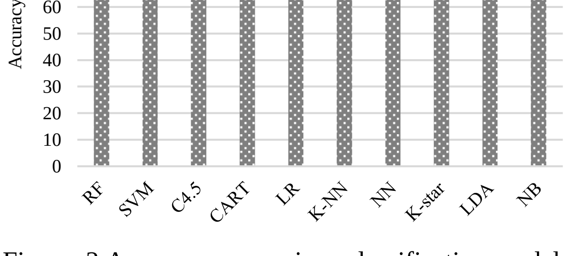

Figure. 2 Accuracy comparison classification model

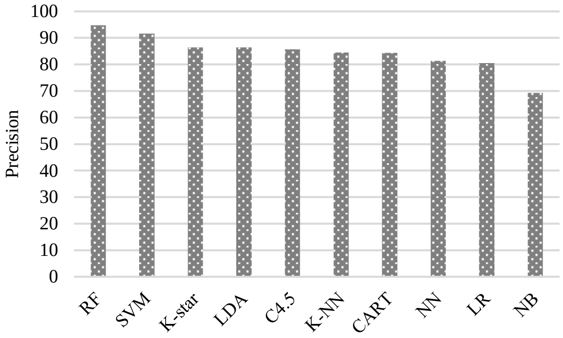

Figure. 3 Precision comparison classification model

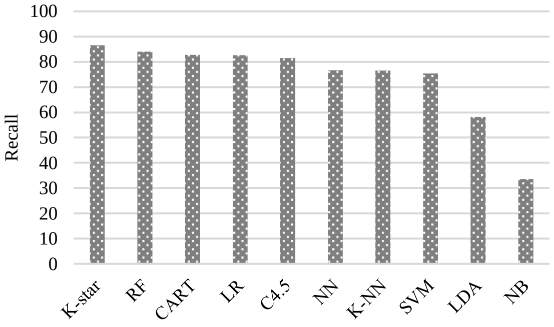

Figure. 4 Recall comparison classification model

Random forest involves the subdivision of the data into subsets separated by the values of the input variables until the basic classification unit is obtained by the training samples. The consensual classification of the most accurate trees is combined into a single one, comprising the RF algorithm. The combination of decision trees in the RF technique can be used in regressions or classification, leading to good results for financial market prediction, as demonstrated by [2, 47-50].

In terms of precision performance, Fig. 3 shows the comparison of the precision of the 10 classification algorithms. Random Forest has the largest precision value then SVM is in the second position. Precision is the ratio of positive true predictions compared to the overall positive predicted results. Precision answers the question of what percentage of companies are right in the LQ45 category from all companies predicted to be in the LQ45 category.

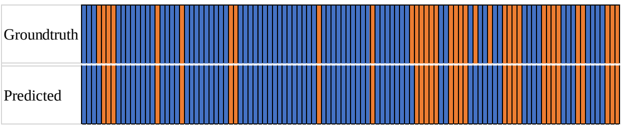

Figure. 5 The prediction results using radom forest

Table 3. Comparison of classification algorithm

Algorithm	Accuracy	Precision	Recall
LR	88.12	80.54	82.64
LDA	83.92	86.38	58.19
NB	74.74	69.23	33.61
K-NN	87.74	84.48	76.53
K-star	86.59	86.40	**86.60**
C4.5	89.66	85.66	81.53
CART	88.87	84.32	82.78
RF	**93.49**	**94.78**	84.03
NN	86.61	81.40	76.67
SVM	90.06	91.61	75.42

In terms of recall performance, Fig. 4 shows the recall comparison of 10 classification algorithms. K-Star algorithm has the best performance with a recall value of 86.60, while RF has a recall value of 84.03. In this study, NB has the worst performance compared to other algorithms. This is due to the assumption of independence that makes the performance of the NB algorithm decrease. The recall is the ratio of true positive predictions to the overall true positive data. Recall answers the question of what percentage of companies are predicted to be in the LQ45 category compared to all companies that are actually in the LQ45 category. While the performance of SVM is statistically similar to RF.

A large number of classification algorithms have been proposed in the long history of machine learning, some of which have been recognized as highly accurate, in particular, SVM and RF (RF) [51].

We also visualize the prediction results in Fig. 5. In Fig. 5, the first row of each example represents the true tag value, and the second row represents the predicted value, where the blue box indicates LQ45 and the orange box indicates Non-LQ45. Each column in the instances corresponds to the actual value and the predicted value at a specific moment. The prediction is correct if two bars in the same column have the same color, otherwise, the prediction is incorrect.

4. Conclusion

This study set out to benchmark the performance of traditional statistical-based learners (logistic regression, linear discriminant analysis, and naïve Bayes), distance-based learners (k-Nearest Neighbour and K-star), tree-based learners (C4.5, classification and regression trees, random forest), artificial neural networks and support vector machines in predicting LQ45 index.

A comparison framework is proposed for comparing the performance of classification algorithms in the prediction of the LQ45 index in the Indonesia stock exchange. The framework is comprised of 10 classification algorithms, 10-fold cross-validation models, and accuracy indicators. The experimental results show that the RF performs best in the prediction of the LQ45 index. C.45, CART, SVM, and LR also perform well. Traditional statistical-based learners tend to underperform, as well as NB and LDA.

The random forest has the best performance to predict the LQ45 index because RF is an ensemble machine learning technique. It is capable of performing both regression and classification tasks. The idea is to combine multiple decision trees to determine the final output instead of relying on individual decision trees to reduce the variance in the model. The noise in stock market data is usually quite high because of its huge size and can cause the trees to grow in a completely different manner as compared to the expected growth. It aims at minimizing forecasting error by treating the stock market analysis as a classification problem [52].

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization and methodology, A. Syukur; formal analysis and investigation, A. Syukur and D. Istiawan; resources and data curation, D. Istiawan; writing—original draft preparation, D. Istiawan; writing—review and editing, A. Syukur and D. Istiawan; visualization, D. Istiawan; supervision, A. Syukur; project administration, D. Istiawan. All authors have read and approved the final manuscript.

References

[1] Y. Andrianto and A. R. Mirza, "A Testing of Efficient Markets Hypothesis in Indonesia Stock Market", *Procedia - Social and Behavioral Sciences*, Vol. 219, pp. 99-103, 2016.

[2] A. Syukur and A. Marjuni, "Stock Price Forecasting Using Univariate Singular Spectral Analysis through Hadamard Transform", *International Journal of Intelligent Engineering and Systems*, Vol. 13, No. 2, pp. 96-107, 2020.

[3] M. Hiransha, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "NSE Stock Market Prediction Using Deep-Learning Models", *Procedia Computer Science*, Vol. 132, pp. 1351-1362, 2018.

[4] D. Devianto, Maiyastri, Randy, M. Hamidi, S. Maryati, and A. W. Ahmad, "Efficiency Analysis of Optimal Portfolio Selection for Stocks in LQ45 Index", In: *Proc. of International Conf. on Applied Information Technology and Innovation (ICAITI)*, Padang, Indonesia, pp. 78-83, 2018.

[5] A. Z. R. Langi, S. W. Pitara, and Kuspriyanto, "Stock Prices Trends Analysis using Wavelet Transform", In: *Proc. of International Conf. on Cloud Computing and Social Networking (ICCCSN)*, Bandung, West Java, Indonesia, pp. 2-5, 2012.

[6] B. B. Nair, N. M. Dharini, and V. P. Mohandas, "A Stock Market Trend Prediction System using a Hybrid Decision Tree-Neuro-Fuzzy System", In: *Proc. of International Conf. on Advances in Recent Technologies in Communication and Computing (ICARTCC)*, Kottayam, India, pp. 381–385, 2010.

[7] Y. S. A.-Mostafa and A. F. Atiya, "Introduction to Financial Forecasting", *Applied Intelligence*, Vol. 6, No. 3, pp. 205-213, 1996.

[8] D. Shah, H. Isah, and F. Zulkernine, "Stock market analysis: A Review And Taxonomy of Prediction Techniques", *International Journal of Financial Studies*, Vol. 7, No. 2, pp. 1-22, 2019.

[9] G. Kumar, S. Jain, and U. P. Singh, "Stock Market Forecasting Using Computational Intelligence: A Survey", *Archives of* Computational *Methods in Engineering*, pp. 1–33, 2020.

[10] J. L. Wang and S. H. Chan, "Stock market trading rule discovery using pattern recognition and technical analysis", *Expert Systems with Applications*, Vol. 33, No. 2, pp. 304–315, 2007.

[11] M. Velay and F. Daniel, "Stock Chart Pattern recognition with Deep Learning", *arXiv*, pp. 1–6, 2018.

[12] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market", *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8, 2011.

[13] T. Anbalagan and S. U. Maheswari, "*Classification* and Prediction of Stock Market Index Based on Fuzzy Meta graph", *Procedia Computer Science*, Vol. 47, pp. 214–221, 2015.

[14] L. Li, Y. Wu, Y. Ou, Q. Li, Y. Zhou, and D. Chen, "Research on Machine Learning Algorithms and Feature Extraction for Time Series", *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–5, 2017.

[15] S. Shen, H. Jiang, and T. Zhang, "Stock Market Forecasting Using Machine Learning Algorithms", *Department of Electrical Engineering, Stanford University*, pp. 1-5, 2012.

[16] M. C. Wu, S. Y. Lin, and C. H. Lin, "An Effective Application of Decision Tree to Stock Trading", *Expert System with Applications.*, Vol. 31, No. 2, pp. 270-274, 2006.

[17] P. Ou and H. Wang, "Prediction of Stock Market Index Movement by Ten Data Mining Techniques", *Modern Applied Science*, Vol. 3, No. 12, pp. 28-42, 2009.

[18] W. Huang, Y. Nakamori, and S. Y. Wang, "Forecasting Stock Market Movement Direction with Support Vector Machine", *Computer & Operational Research*, Vol. 32, No. 10, pp. 2513-2522, 2005.

[19] K. Kim, "Financial Time Series Forecasting using Support Vector Machines", *Neurocomputing*, Vol. 55, No. 1-2, pp. 307-319, 2003.

[20] M. C. Lee, "Using Support Vector Machine with a Hybrid Feature Selection Method to the Stock Trend Prediction", *Expert System with Applications*, Vol. 36, No. 8, pp. 10896-10904, 2009.

[21] S. H. Kim and S. H. Chun, "Graded Forecasting using an Array of Bipolar Predictions: Application of Probabilistic Neural Networks to a Stock Market Index", *International Journal of Forecasting*, Vol. 14, No. 3, pp. 323-337, 1998.

[22] M. Ballings, D. V. D. Poel, N. Hespeels, and R. Gryp, "Evaluating Multiple Classifiers for Stock Price Direction Prediction", *Expert System with Applications*, Vol. 42, No. 20, pp. 7046-7056, 2015.

[23] N. Milosevic, "Equity Forecast: Predicting Long Term Stock Price Movement using Machine Learning", *arXiv:1603.00751*, pp. 1-9, 2016.

[24] D. Delen, "A Comparative Analysis of Machine Learning Techniques for Student Retention Management", *Decision Support System*, Vol. 49, No. 4, pp. 498-506, 2010.

[25] D. G. Kleinbaum, L. L. Kupper, K. E. Muller, and A. Nizam, *Logistic Regression Analysis. Applied Regression Analysis and Other Multivariate Methods*, 4th ed. Cole, Belmont: Thomson Brooks, 2008.

[26] D. T. Larose, *Data Mining Methods and Models*. Hoboken, New Jersey: John Wiley & Sons, 2006.

[27] M. Khashei, A. Z. Hamadani, and M. Bijari, "A

Novel Hybrid Classification Model of Artificial Neural Networks and Multiple Linear Regression Models", *Expert System with Applications*, Vol. 39, No. 3, pp. 2606-2620, 2012.

[28] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. Hoboken, New Jersey: John Wiley & Sons, 2011.

[29] L. Jiang, Z. Cai, and D. Wang, "Improving Naive Bayes for Classification", *International Journal of Computers and Applications*, Vol. 32, No. 3, pp. 328-332, 2010.

[30] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, "Hybrid Decision Tree and Naïve Bayes Classifiers for Multi-Class Classification Tasks", *Expert System with Applications*, Vol. 41, No. 4, pp. 1937-1946, 2014.

[31] K. S. Kim, H. H. Choi, C. S. Moon, and C. W. Mun, "Comparison of K-Nearest Neighbor, Quadratic Discriminant and Linear Discriminant Analysis in Classification of Electromyogram Signals Based on the Wrist-Motion Directions", *Current Applied Physics*, Vol. 11, No. 3, pp. 740-745, 2011.

[32] D. Y. Mahmood and M. A. Hussein, "Intrusion Detection System Based on K-Star Classifier and Feature Set Reduction", *IOSR Journal of Computer Engineering*, Vol. 15, No. 5, pp. 107-112, 2013.

[33] S. Painuli, M. Elangovan, and V. Sugumaran, "Tool Condition Monitoring using K-Star Algorithm", *Expert System with Applications*, Vol. 41, No. 6, pp. 2638-2643, 2014.

[34] C. J. Mantas and J. Abellán, "Credal-C4.5: Decision Tree Based on Imprecise Probabilities to Classify Noisy Data", *Expert System with Applications*, Vol. 41, No. 10, pp. 4625-4637, 2014.

[35] I. Brown and C. Mues, "An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets", *Expert System with Applications*, Vol. 39, No. 3, pp. 3446-3453, 2012.

[36] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "*Top 10 Algorithms in Data Mining*", *Knowledge and Information Systems*, Vol. 14, No. 1, pp. 1-37, 2008.

[37] S. M. H.-Ziabari and T. Bakhshpoori, "Improving The Prediction of Ground Motion Parameters Based on an Efficient Bagging Ensemble Model of M5 and CART Algorithms", *Applied Soft Computing*, Vol. 68, pp. 147-161, 2018.

[38] M. Reif, F. Shafait, and A. Dengel, "Meta-Learning for Evolutionary Parameter Optimization of Classifiers", *Machine Learning*, Vol. 87, No. 3, pp. 357-380, 2012.

[39] L. Tang, F. Cai, and Y. Ouyang, "Applying a Nonparametric Random Forest Algorithm to Assess the Credit Risk of the Energy Industry in China", *Technological Forecasting and Social Change*, Vol. 144, 2019, pp. 563-572, 2019.

[40] S. Huo, Z. He, J. Su, B. Xi, and C. Zhu, "Using Artificial Neural Network Models for Eutrophication Prediction", *Procedia Environmental Sciences*, Vol. 18, pp. 310–316, 2013.

[41] V. N. Vapnik, *The Nature of Statical Learning Theory*. 2000.

[42] S. Dreiseitl and L. O.-Machado, "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review", *Journal of Biomedical Informatics*, Vol. 35, No. 5-6, pp. 352-359, 2002.

[43] T. W. Schiller, Y. Chen, I. E. Naqa, and J. O. Deasy, "Modeling Radiation-Induced Lung Injury Risk with an Ensemble of Support Vector Machines", *Neurocomputing*, Vol. 73, No. 10-12, pp. 1861-1867, 2010.

[44] T. J. McCabe, "A Complexity Measure", *IEEE Transaction of Software Engineering*, Vol. SE-2, No. 4, pp. 308-320, 1976.

[45] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research", *European Journal of Operational Research*, Vol. 247, No. 1, pp. 124-136, 2015.

[46] A. C. Lorena, L. F. O. Jacintho, M. F. Siqueirab, R. D. Giovanni, L. G. Lohmann, A. C. P. L. F. de Carvalho, M. Yamamoto, "Comparing Machine Learning Classifiers in Potential Distribution Modeling", *Expert System with Applications*, Vol. 38, No. 5, pp. 5268-5275, 2011.

[47] M. *Kumar* and M. Thenmozhi, "Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models", *International Journal of Banking, Accounting, and Finance*, Vol. 5, No. 3, pp. 284–308, 2014.

[48] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques", *Expert Systems with Applications*, Vol. 42, No. 1, pp. 259–268, 2015.

[49] D. Kumar, S. S. Meghwani, and M. Thakur, “Proximal *support* vector machine-based hybrid prediction models for trend forecasting in financial markets”, *Journal of Computational Science*, Vol. 17, pp. 1–13, 2016.

[50] C. Krauss, X. A. Do, and N. Huck, “Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500”, *European Journal of Operational Research*, Vol. 259, No. 2, pp. 689–702, 2017.

[51] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, “An Up-To-Date Comparison of State-of-the-Art Classification Algorithms”, *Expert System with Applications*, Vol. 82, pp. 128-150, 2017.

[52] M. Vijh, D. *Chandola*, V. A. Tikkiwal, and A. Kumar, “Stock Closing Price Prediction using Machine Learning Techniques”, *Procedia Computer Science*, Vol. 167, No. 2019.