

# Super Learner for Predicting Stock Market Trends: A Case Study of Jakarta Islamic Index Stock Exchange

Gerry alfa Dito<sup>1</sup>, Bagus Sartono<sup>2\*</sup>, Annisa<sup>3</sup>  
{gerryalfadito@gmail.com<sup>1</sup>, bagusco@gmail.com<sup>2</sup>, annisa@apps.ipb.ac.id<sup>3</sup>}

<sup>1,2</sup> Statistics Department, IPB University, West Java, 16680, Indonesia

<sup>3</sup> Computer Science Department, IPB University, West Java, 16680, Indonesia,

**Abstract.** Predicting stock market trend is one of the challenging tasks over the years. It has diverse influencing factors which cause stock market trend is very dynamic and has high volatility. Forecasting model, which is a prevalent method to predict the stock market trend, has several difficulties with its characteristics. Although forecasting model is efficient, sometimes it has high forecasting error. Formulating forecasting problem into a classification problem might be considered an alternative approach to predict stock market trend. Several kinds of research have shown that machine learning is a suitable method for predicting stock market trend as a classification problem. This paper discusses applying one of powerful machine learning method, which is called Super learner, to predict stock market trends. Besides, this research employs several technical indicators as predictor variables. Results show that the Super Learner model is useful for predicting both the short-term and long-term trend.

**Keywords:** ensemble learning, machine learning, stock market trend, technical indicators.

## 1 Introduction

In the past several decades, predicting the stock market trends is one of primary interest for researchers and investors. They always investigate methods which can predict stock markets trend accurately. The most difficulty is the characteristics of the stock market price, which is very dynamic and has high volatility. The prevalent methods to predict stock market trends is a forecasting model which is part of a time series model. Some previous study, which is using the forecasting model are [1],[2], [3]. The difficulty to use forecasting model is sometimes this method produces high forecast error in short-term because the trends tremendously fluctuate. However, some stock becomes more stable in the long-term because it forms a linear trend. Formulating forecasting problem into a classification problem might be considered an alternative approach to predict the stock market trend.

Several kinds of research have shown that machine learning is a suitable method for predicting stock market trend as a classification problem. The first research using machine learning method which is known Support Vector Machine (SVM) using twelve technical indicators to predict stock price trend of Korea composite stock price index (KOSPI) and produced the accuracy approximately 65% for one days-ahead prediction [4]. Genetic-algorithm SVM first purposed for predicting the trend of stock price by using seven technical indicators in the Indian stock market and shown the accuracy of a model is about 60% for one

days-ahead prediction[5]. SVMs and artificial neural networks (ANNs) using twelve technical indicators for building the model used to predict future movement on the Istanbul Stock Exchange (ISE) National 100 index and shown the accuracy about 75% for one days-ahead prediction [6]. Combination of both fundamental and technical indicators is used by [7] and applied genetic algorithm SVM and ANN method to predict the stock market trend of 25 companies listed on the Bombay Stock Exchange and shown the accuracy about 80% for one days-ahead prediction. The study about discretized technical indicators by representing them as trend deterministic data and then trained predictive models using random forests, ANNs, naïve Bayes and SVM models on CNX Nifty, S&P BSE Sense indices Reliance Industries and Infosys Ltd companies on the National Stock Exchange of India and produced the accuracy around 89% [8]. Random Forest using six technical indicators to predict stock market trends in AAPL, MSFT, and Samsung and showed their accuracy in the range 85-95% for long term prediction [9].

Based on the previous study that has been discussed, it can be inferred that the ensemble learning methods still have a few numbers of research which applied to predict stock market trends. Moreover, the previous research has shown that Technical Indicators of the stock market are promising to become a predictor variable. This paper discusses the application of ensemble learning method, which is called Super learner, to predict stock market trends and employs 13 technical indicators as predictor variables. Concerning a few types of research exist in the literature to predict the stock market price index in Indonesia stock market, this paper uses Islamic Index Stock Exchange dataset to apply Super Learner method.

## **2 Methodology**

### **2.1 Research Data**

The dataset, which is used in this research is the Islamic Index Stock Exchange. It is collected from 14 April 2013 to 12 March 2019 in day's format from Yahoo Finance website. This dataset contains 55 missing values, and then these values are removed for further analysis.

### **2.2 Variable Extraction**

Technical Indicators are used as predictor variables which might give benefit to predict stock market trends. Technical Indicators are derived from historical stock market data. The purpose of technical indicators is used by investors to forecast stock market direction. The technical indicators description based on [10] (except German and Klass, and Roger and Satchell) which can be seen in Table 1. These indicators are used as predictor variables, but some of the indicators have output more than one output. The list of technical indicators which have more than one output can be seen below:

1. MACD has two outputs; those are price oscillator and a moving average of the oscillator.
2. Stochastic Oscillator has five outputs; those are Stochastic Fast %K, Stochastic Fast %D, Stochastic Slow %D, Stochastic Momentum Index (SMI), and Stochastic Momentum Index signal line.

3. ADX has four outputs; those are positive direction index, negative direction index, direction index, and average direction index.
4. Bollinger Band has four outputs; those are lower Bollinger Band, middle moving average Bollinger Band, upper Bollinger Band, and %B.
5. Average True Range has four outputs; those are True Range series, Average True range, True High and True Low.

### 2.3 The model

The statistical model, which is used in this paper is Super Learner. Super Learner is a model which uses cross-validation to integrate multiple statistical models. These statistical models are called base learners. Super Learner uses constrained regression model to combine base models, which is called meta-learner. The original idea about Super Learner was proposed by [11], but in that time, this model was called Stacked Generalization. Then, [12] used cross-validation to combine base models, which is called Stacking. However, the mathematical reasons for why stacking worked were unknown until [13] proved that the stacking ensemble represents an asymptotically optimal for modeling. Super Learner can be performed by following procedures from Figure 3.

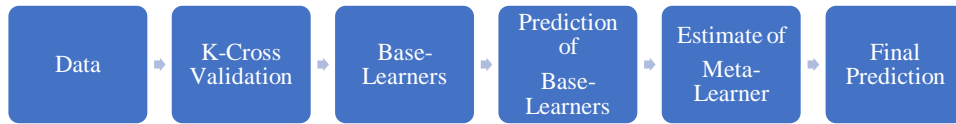
**Table 1.** Technical Indicators.

| No | Names  | Short Description  |
|----|--|--|
| 1  | Parabolic SAR                                | A chart overlay showing reversal points below prices in an uptrend and above prices in a Downtrend                         |
| 2  | Moving Average Convergence/Divergence (MACD) | A momentum oscillator based on the difference between two EMAs   |
| 3  | Relative Strength Index (RSI)                | Shows how strongly a stock is moving in its current direction  |
| 4  | Stochastic Oscillator                        | shows how a stock's price is doing relative to past movements. Fast, Slow and Full Stochastics are explained               |
| 5  | Averages Directional Index (ADX)             | Shows whether a stock is trending or oscillating   |
| 6  | Rate of Change (ROC)                         | Shows the speed at which a stock's price is changing   |
| 7  | Williams %R                                  | Uses Stochastics to determine overbought and oversold levels   |
| 8  | On Balance Volume                            | Combines price and volume in a very simple way to show how money may be flowing into or out of a stock                     |
| 9  | Chaikin Accumulation                         | Combines price and volume to show how money may be flowing into or out of a stock. Based on Accumulation/Distribution Line |
| 10 | Bollinger Bands                              | A chart overlay showing the upper and lower limits of "normal" price movements based on                                    |

|    |                     |  |
|----|---------------------|--|
|    |                     | the Standard Deviation of prices   |
| 11 | Avarage True Range  | Measures a stock's volatility  |
| 12 | German and Klass    | Measures a stock's volatility which assumes Brownian motion with zero drift and no opening jumps |
| 13 | Rogers and Satchell | Measures a stock's volatility which allows for non-zero drift, but assumed no opening jump       |

## 2.4 The method

In this paper, analyzing the Jakarta Islamic Index by using Super Learner has the following procedures:



**Fig.1.** Super Learner model builder.

- Extracting technical Indicators in Table 1 from the Jakarta Islamic Index datasets based on
- Generating return of stock price index  $y_{ret}$ , by performing equation (1) and transform to binary predictor variable,  $y_{bin}$  based on equation (2). The result of step 2 is  $y_{bin,k}$  according k days-head, which  $isk = 1,5,10,30,60$ .

$$y_{ret,k} = y_{t-k} - y_t \quad (1)$$

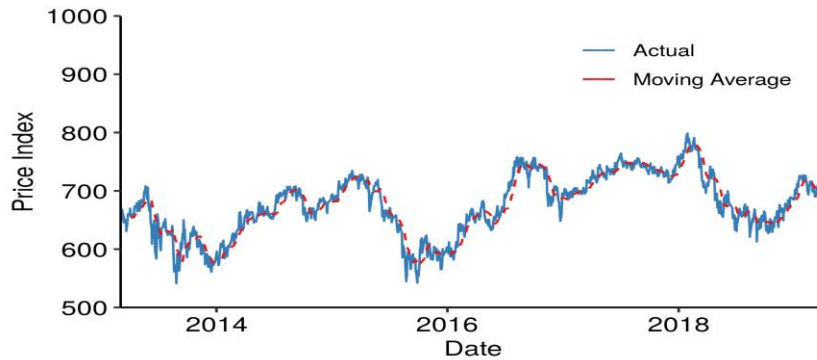
$$y_{bin,k} = \begin{cases} down, & y_{ret} < 0 \\ up, & y_{ret} \geq 0 \end{cases} \quad (2)$$

- Stacking dataset from step 1 and 2 to get 5 final datasets, that is one day ahead, five days ahead, ten days ahead, 30 days ahead and 60 days ahead final dataset.
- Splitting final dataset into a training and testing, according to a 10-fold cross-validation scheme
- Setting 5 base learners of Super Learner methods as shown in Table 2.
- Fitting Super Learner in 10-fold cross-validation dataset and calculating accuracy in each fold.
- Averaging accuracy in each fold to get final accuracy.

**Table 2.** Base Learner.

| No | Base Learners              | Short Description  |
|----|----------------------------|--|
| 1  | Conditional Inference Tree | Conditional inference trees estimate a regression relationship by binary recursive partitioning in a conditional inference framework |
| 2  | C50 Tree                   | Classification tree models using Quinlan's C5 algorithm.   |
| 3  | J48 Tree                   | Classification tree models using Quinlan's C45 algorithm.  |
| 4  | Recursive Partitioning     | Classification tree models using Recursive Partitioning  |

|   |                        |   |
|---|------------------------|---|
|   | Tree                   | algorithm   |
| 5 | Naïve Bayes            | Model based on applying Bayes' theorem with independence assumptions between the variables.   |
| 6 | Neural Network         | Model that are inspired by the biological neural networks.  |
| 7 | One Rule               | Classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule". |
| 8 | Instance Base Learning | Simple instance-based learner that uses the class of the nearest k training instances for the class of the test instances.                              |



**Fig.2.**Jakarta Islamic Index Stock Exchange historical price plot with Moving Average (MA) prediction

### 3. Result and Discussion

Figure 2 shows a glimpse of the data. It shows that the Price Index movement slightly fluctuates over time(days). It causes 1-day ahead prediction might not as good as the rest prediction because its random movement is difficult for modelling. In Figure 3, which tells about differencing in a different setting, shows the increasing lag in differencing might cause steady Price Index movement. Consequently, the forecasting result of 60 days ahead might be better than 30 days ahead, 30 days ahead might be better than ten days ahead, and so on. Then, the return of the price index is transformed into two classes, which are Down and Up. The frequency of each class (as shown in Table 3) has almost the same number in each setting. It caused the classification problem might not have an issue to give accurate result in one class only.

Table 3Frequency of Price Index Movement (Down and Up) in each setting

| Class | 1 day ahead | 5 days ahead | 10 days ahead | 30 days ahead | 60 days ahead |
|-------|-------------|--------------|---------------|---------------|---------------|
| Down  | 687         | 676          | 678           | 664           | 631           |
| Up    | 750         | 757          | 750           | 744           | 747           |

The result of the Super Learner Model consists of two things, which are the regression coefficient of each base learner and prediction performance. Coefficient of each base learner states about the contribution of each base learner in Super Learner. Higher coefficient implies the higher contribution, which is given by base learners. For example, Conditional Tree has the highest contribution for modelling 1 day ahead setting (as shown in table 4). Another essential characteristic of coefficients is they summed to one. Based on Table 4, Neural Network and Naïve Bayes have a small contribution in all setting.

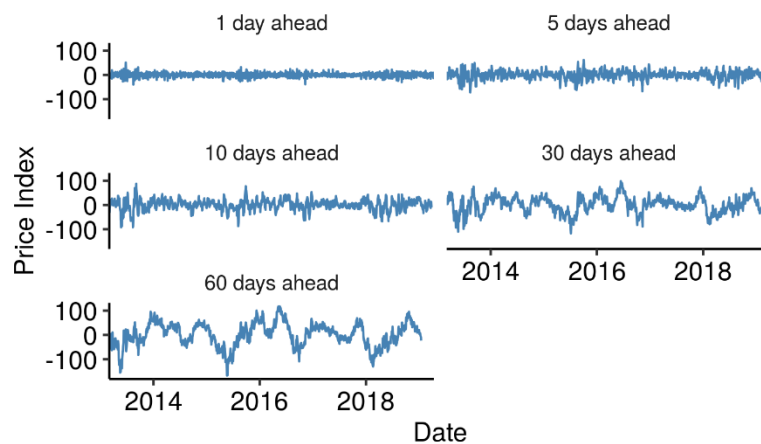
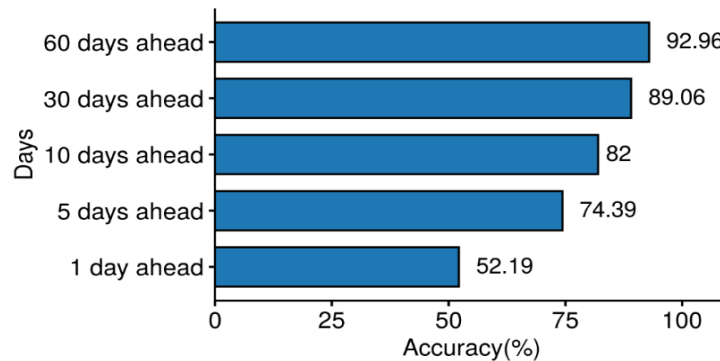


Fig. 3. Jakarta Islamic Index Stock Exchange historical price return

Table 4 Coefficient of each Base Learner in Super Learner

| Base learner     | 1 days ahead | 5 days ahead | 10 days ahead | 30 days ahead | 60 days ahead |
|------------------|--------------|--------------|---------------|---------------|---------------|
| Conditional Tree | 0.408        | 0.157        | 0.085         | 0.076         | 0.133         |
| C50              | 0.343        | 0.098        | 0.168         | 0.171         | 0.202         |
| J48              | 0.029        | 0.098        | 0.15          | 0.117         | 0.193         |
| CARRT            | 0.07         | 0.121        | 0.107         | 0.13          | 0.055         |
| Naïve Bayes      | 0.048        | 0.016        | 0.02          | 0             | 0.005         |
| Neural Network   | 0.081        | 0.028        | 0.022         | 0             | 0             |
| OneR             | 0.01         | 0.07         | 0.018         | 0.04          | 0.023         |
| Ibk              | 0.013        | 0.412        | 0.431         | 0.465         | 0.389         |



**Fig.4.** Accuracy of Super Learner

Performance of Super Learner prediction can be evaluated by using classification accuracy. The accuracy in Figure 4 is the mean of accuracy from 10-folds cross-validations. The model accuracy is improving when the number of days ahead is also increasing. The reason for this result might be that the stock price is more fluctuate for a small number of days. Prediction of 1 day ahead might not be considered for deciding because it is only about 50%. However, five days ahead and ten days ahead has prediction above 74%, which means Super Learner might help short-term decision takers to choose whether or not invest in this stock market. In 30 and 60 days-ahead predictions, the model has high accuracy over 85%, which means Super Learner is reliable for predicting the long-term behavior of this stock markets.

#### 4. Conclusion

The paper used the Super Learner method to predict short-term and long-term Jakarta Islamic IndexStock exchange. It has shown achieve high accuracy for both short-term and long-term prediction expect for case one day ahead, which only get around 50% accuracy. Future research should consider optimizing hyperparameter of the Super Learner method to achieve higher accuracy. The use of an algorithm which can select technical indicators automatically also should be an issue for future research to explore.

#### References

- [1] Gençay, R.: Linear, non-linear and essential foreign exchange rate prediction with simple technical trading rules. *J. Int. Econ.* vol. 47, no. 1, pp. 91–107 (1999)
- [2] Timmermann, A. and Granger, C. W. J.: Efficient market hypothesis and forecasting. *Int. J. Forecast.*, vol. 20, no. 1, pp. 15–27 (2004)
- [3] Bao, D and Yang, Z.: Intelligent stock trading system by turning point confirming and probabilistic reasoning. *Expert Syst. Appl.*, vol. 34, no. 1, pp. 620–627 (2008)
- [4] Kim, K.: Financial time series forecasting using support vector machines. vol. 55, pp. 307–319 (2003)
- [5] Choudhry, R. and Garg, K.: A Hybrid Machine Learning System for Stock Market Forecasting. vol.2, no. 3, pp. 689–692(2008)

- [6] Kara, Y. Acar, M. and Kaan, Ö.:Expert Systems with Applications Predicting direction of stock price index movement using artificial neural networks and support vector machines : The sample of the Istanbul Stock Exchange.*Expert Syst. Appl.*, vol. 38, no. 5, pp. 5311–5319 (2011)
- [7] Chandwani, D. and Singh SalujaM.:Stock Direction Forecasting Techniques: An Empirical Study Combining Machine Learning System with Market Indicators in the Indian Context.*Int. J. Comput. Appl.*, vol. 92, no. 11, pp. 8–17(2014)
- [8] Patel, J. Shah, S. Thakkar, P. and Kotecha, K.: Expert Systems with Applications Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques.*Expert Syst. Appl.*, vol. 42, no. 1, pp. 259–268( 2015)
- [9] Khaidem,L. and Dey, S.R.: Predicting the direction of stock market prices using random forest. vol. 00, no. 00, pp. 1–20(2016)
- [10] Edwards RD, Magee J, Bassetti WHC. 2018. Technical Analysis Of Stock Trends. Eleventh Edition.Revised Edition of the Authors’ Technical Analysis of Stock Trends, C2013. CRC Press, US (2018)
- [11] Wolpert,D. H.:Stacked Generalization. vol. 5, pp. 241–259(1991)
- [12] Breiman, L.: Stacked regressions.*Mach. Learn.*, vol. 24, no. 1, pp. 49–64, Jul (1996)
- [13] van der Laan, M. J. Polley, E. C and Hubbard, A. E.: Super Learner.*Stat. Appl. Genet. Mol. Biol.*, vol. 6, no. 1, pp. 321–332 (2007)