FORECASTING THE STOCK PRICES BY FEATURES ENGINEERING AND MACHINE LEARNING TECHNIQUE

RATCHAPON PARIYOTHAI

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Applied Mathematics
Suranaree University of Technology
Academic Year 2022

การทำนายราคาของหลักทรัพย์ด้วยวิศวกรรมคุณลักษณะและเทคนิคการเรียนรู้ของเครื่อง

นายรัชพล ปริโยทัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาคณิตศาสตร์ประยุกต์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2565

FORECASTING THE STOCK PRICES BY FEATURES ENGINEERING AND MACHINE LEARNING TECHNIQUE

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for a Master's Degree.

Thesis Examining Committee

(Assoc. Prof. Dr. Chairat Modnak)
Chairperson

Benjarvan Rodjanadid.

(Asst. Prof. Dr. Benjawan Rodjanadid)
Member (Thesis Advisor)

(Assoc. Prof. Dr. Eckart Eckart)
Member

Chatchai Jothit

(Assoc. Prof. Dr. Chatchai Jothityangkoon)
Vice Rector for Academic Affairs
and Quality Assurance

(Prof. Dr. Santi Maensiri)
Dean of Institute of Science

รัชพล ปริโยทัย : การทำนายราคาของหลักทรัพย์ด้วยวิศวกรรมคุณลักษณะและเทคนิคการเรียนรู้ของเครื่อง (FORECASTING THE STOCK PRICES BY FEATURES ENGINEERING AND MACHINE LEARNING TECHNIQUE) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์. ดร. เบญจวรรณ โรจนดิษฐ์, 68 หน้า.

คำสำคัญ : เทคนิคการเรียนรู้ของเครื่อง/การเรียนรู้เชิงลึก/วิศวกรรมคุณลักษณะ

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อหาคุณลักษณะและสร้างตัวแบบในการทำนายราคาของหลักทรัพย์จำนวน 10 หลักทรัพย์ ในตลาดหลักทรัพย์แห่งประเทศไทยซึ่งได้แก่ BANPU, BBL, GUNKUL, IRPC, KBANK, KTB, PTT, SUPER, KKP และ TTB โดยการหาคุณลักษณะจะใช้วิศวกรรมคุณลักษณะร่วมกับเทคนิคซัพพอร์ตเวกเตอร์สำหรับการถดถอย เทคนิคการเรียนรู้เชิงลึก เทคนิคป่าสุ่ม และเทคนิคเกรเดียนท์บูตทรี จากนั้นนำคุณลักษณะที่ได้ไปสร้างตัวแบบในการทำนายราคาของหลักทรัพย์ ซึ่งในการสร้างตัวแบบใช้เทคนิคการเรียนรู้เครื่อง 4 เทคนิค ซึ่งได้แก่ เทคนิคซัพพอร์ตเวกเตอร์สำหรับการถดถอย เทคนิคการเรียนรู้เชิงลึก เทคนิคป่าสุ่ม และเทคนิคเกรเดียนท์บูตทรี สำหรับโปรแกรมที่ใช้ในงานวิจัยนี้ได้แก่โปรแกรม Minitab Statistical Software version 20 , Microsoft Excel และโปรแกรม Rapidminer Studio version 10.1 (Education license)

ผลการศึกษาพบว่าตัวแบบที่มีประสิทธิภาพมากที่สุดในการทำนายราคาของหลักทรัพย์ BANPU, BBL, GUNKUL, IRPC, KBANK, KTB, PTT, SUPER, และ TTB คือตัวแบบที่ได้จากเทคนิคการเรียนรู้เชิงลึก และตัวแบบที่มีประสิทธิภาพมากที่สุดในการทำนายราคาของหลักทรัพย์ KKP คือตัวแบบที่สร้างจากเทคนิคซัพพอร์ตเวกเตอร์สำหรับการถดถอย

สาขาวิชาคณิตศาสตร์ ลายมือชื่อนักศึกษา รัชพล ปริโยทัย

ปีการศึกษา 2565 ลายมือชื่ออาจารย์ที่ปรึกษา Benjawan Rodjanadid.

RATCHAPON PARIYOTHAI : FORECASTING THE STOCK PRICES BY FEATURES ENGINEERING AND MACHINE LEARNING TECHNIQUE. THESIS ADVISOR : ASST. PROF. BENJAWAN RODJANADID, Ph.D. 68 PP.

Keyword : MACHINE LEARNING/DEEP LEARNING/FEATURE ENGINEERING

In this research, the primary objective was to identify and develop a predictive model for the prices of 10 securities in the Stock Exchange of Thailand. The securities included BANPU, BBL, GUNKUL, IRPC, KBANK, KTB, PTT, SUPER, KKP, and TTB. The identification of features involved a combination of feature engineering techniques, including support vector regression, deep learning, random forest, and gradient boosting. Subsequently, the identified features were used to construct a predictive model for the securities' prices. The model construction utilized four machine learning techniques: support vector regression, deep learning, random forest, and gradient boosting. The software employed for this research comprised Minitab Statistical Software version 20, Microsoft Excel, and Rapidminer Studio version 10.1 (Education license).

The study findings indicated that the most effective models for price prediction of the securities were obtained through deep learning. The models generated accurate price predictions for BANPU, BBL, GUNKUL, IRPC, KBANK, KTB, PTT, SUPER, and TTB securities. Additionally, the model constructed using support vector regression yielded the most accurate price predictions for KKP securities.

School of Mathematics Student's Signature รัชพล ปริโยทัย

Academic Year 2022 Advisor's Signature Benjawan Rodjanadid.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all those who have contributed to the completion of this project.

First and foremost, I am immensely thankful to my supervisor, Asst. Prof. Dr. Benjawan Rodjanadid, for their guidance, support, and valuable insights throughout the entire research process. Their expertise and encouragement have been instrumental in shaping this project.

Furthermore, I would like to give a million thank for the professional support received from Assoc. Prof. Dr. Eckart Schulz and all professors in the School of Mathematics, Institute of Science, Suranaree University of Technology (SUT). Likewise, many thanks to acknowledge the contributions of Mr. Jirakit Boonmunewai, Mr. Jakkrit Polrob, Mr. Saranchai Sinlapasorn and all of my friends at SUT for their technical support, assistance with data analysis, and valuable suggestions.

In addition, I am grateful to Development and Promotion of Science and Technology Talents Project (DPST) for a scholarship to study in the Master's degree program at SUT.

I would also like to express my heartfelt thanks to my friends and family for their unwavering support, understanding, and encouragement throughout this journey. Your belief in me has been a constant source of motivation.

Last but not least, I am grateful to the academic community and the institutions that have provided resources, literature, and references that have been integral to the successful completion of this project.

Although it is not possible to mention everyone individually, please know that your contributions have not gone unnoticed and are sincerely appreciated.

Thank you all for being part of this endeavor and for your invaluable contributions.

Ratchapon Pariyothai

CONTENTS

Page

CHAPTER

CONTENTS (Continued)

CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF FIGURES (Continued)

CHAPTER I

INTRODUCTION

Stock investment is significant and in the spotlight of both domestic and international investors at present. Many investment data analysts are trying to develop and improve on efficiency and accuracy of stock and asset value prediction models for fair value and trend performance analysis, and for data driven decision supports.

The Stock Exchange of Thailand, or SET, is one of the interesting investment options which are has received wide attention. They have developed and applied technology to make stock investment bettor accessible to investors. Moreover, investment data on stock has grown up year by year, so the deployment of information technology to enhance storage and quick data selection has increased the accessibility of big data to investors, and also shortened the duration of data preparation and storage. The technological application with machine learning has eliminated repetitive logical processes and analysis by humans and has applied artificial intelligence (AI) to investment decision making, helping the investor to contemplate, examine, and assess assets in desired investment patterns. (Tsai and Hsiao, 2010) However, the investment demands depend on the individuals' investing behaviors, and attention is required to include individual preferences for good decision making support. For example, some investors may be interested in asset value whereas other investors are interested in long-term profitability or trading volume of an asset, and what quantity of on asset is available in the market. These are significant decision parameters for matching an appropriate individual investing strategy. Asset value consideration for investors is necessary to comprehend parameters of fundamental analysis by considering financial variables and measuring a level of asset investment quality and value for decision. (Sattawat and Surachai, 2020)

For stock market capitalization predictions, the comprehensive assessment of asset value can be classified basically into 2 main methods, technical analysis and fundamental analysis. Firstly for technical analysis, as asset values and stock price fluctuations are generally impacted by multiple factors; the implementation of technical analysis which emphasizes stock pricing, moving average, and trading volume analysis for trend performance predictions has been difficult to implement as a single appropriate instrument or multiple techniques, because while technical analysis has a lot of instruments to apply, the accuracy of predictions of stock price and performance trend depend on individual investment experience. Secondly, fundamentals analysis emphasizes value evaluation of modern stock market capitalization by considering return on investment (ROI) and future selling price. The analysis result is taken as a decision criterion, that is to say, buy the stock if the price is less than the evaluated base value and sell if the price is higher than the base value. The fundamentals factor analyzes economics and political information at present time, relevant industry, company earnings, and also the financial condition of the company. (Surachai, Chayamin and Jeeranun, 2013)

Over the years, research on stock quotes and asset value predictions has been receiving more attention, has been widely studied, and has also constantly developed complicated methods. Machine learning is a method that has become interesting to researchers; as the amount and variety of data have increased, efficient performance data storage system are required. These are all potential factors of the prediction model which have helped to develop and implement possible quickly, and these factors are supported for complicated big data analysis. Many companies and investors have given considerable importance to these reasons; thus they have been trying to implement accurate models to increase the opportunity to make profit on investment with good risk management.

This research studies the prediction of stock daily prices by deploying a machine process to support an investor’s decision making. It uses Rapid Miner Studio 10.1, Minitab Statistical Software version 20, and Microsoft Excel for model creation and analysis.

1.1 Research Objective

To construct a model to predict the stock price in the Stock Exchange of Thailand by using machine learning, deep learning, and feature engineering.

1.2 Scope and Limitations

The stock prediction model utilizes the daily historical stock prices of 10 companies listed on the Stock Exchange of Thailand. To construct the prediction model, the approach involves feature engineering techniques and employs various algorithms such as Support Vector Regression (SVR), Deep Learning (DL), Random Forest (RF), and Gradient Boost Tree (GBT).

The stock price data consist of 10 companies, namely:

1. BANPU PUBLIC COMPANY LIMITED (BANPU)
2. BANGKOK BANK PUBLIC COMPANY LIMITED (BBL)
3. GUNKUL ENGINEERING PUBLIC COMPANY LIMITED (GUNKUL)
4. IRPC PUBLIC COMPANY LIMITED (IRPC)
5. KASIKORNBANK PUBLIC COMPANY LIMITED (KBANK)
6. KIATNAKIN PHATRA BANK PUBLIC COMPANY LIMITED (KKP)
7. KRUNG THAI BANK PUBLIC COMPANY LIMITED (KTB)
8. PTT PUBLIC COMPANY LIMITED (PTT)
9. SUPER ENERGY CORPORATION PUBLIC COMPANY LIMITED (SUPER)
10. TMBTHANACHART BANK PUBLIC COMPANY LIMITED (TTB)

1.3 Research Procedure

The research work proceeded as follows:

1. Study features engineering.
2. Study machine learning techniques, namely support vector machine, deep learning, random forest, and gradient boost tree.

3. Study the time series model.

4. Understand and prepare the daily stock data.

5. Analyze and construct the model to predict the stock closing price on the next day.

6. Compare the performance of each prediction model by using root mean square error, mean absolute error, mean absolute percentage error, and square error.

1.4 Results

Our obtained model, powered by feature engineering, machine learning, and deep learning, demonstrates good predictive capabilities in forecasting the closing price of stocks for the subsequent day.

CHAPTER II

LITERATURE REVIEW

This subject matter introduces the fundamental concepts of machine learning, deep learning, and feature engineering as they pertain to stock price prediction. The contents encompasses the key principles of machine learning, deep learning, and feature engineering that play a significant role in these research studies.

2.1 Stock Price Prediction

Stock price prediction is the process of forecasting the future price of a respective stock or group of stocks. The prediction dependent on a variety of factors including the stock's historical performance, current market trends, economic indicators, and other pertinent information. To maximize earnings, investors and traders use stock price prediction to make educated judgments about buying and selling stocks.

Nonetheless, it is crucial for investors to remember that stock price forecasting is advanced speculation amid too many uncertainties, and there is no underwriting that any prediction will be accurate. It is always to perform one's research, consult financial experts, and diversify one's investments to manage risk.

2.2 Information of Stock Price

Daily stock data is instrumental in helping investors make informed decisions regarding investments in stocks, both for long-term and short-term investment strategies, It consist of 5 attributes:

Table 2.1 The data used in this thesis.

Feature	Type	Description
Closing price [Close(T)]	real number	The price of the stock at the end of a trading day
Open price [Open]	real number	First price of the stock at the beginning of a trading day
Minimum Price [Min]	real number	The minimum price of the stock during the trading day
Maximum Price [Max]	real number	The maximum price of the stock during the trading day
Max-Min [Max-Min]	real number	maximum price - minimum price
Volume [Volume]	real number	Number of stocks traded for security in all the markets during a given time

2.3 Time Series Analysis Model

2.3.1 Moving Average Model

The moving average model (MA) is a widely used technical time series analysis method that determines the average price of an asset over a given time to help discover trends in data. The equation represents the MA model as follows:

$$Y_t = \frac{Y_{(t-1)} + Y_{(t-2)} + Y_{(t-3)} + ... + Y_{(t-p)}}{p} \tag{2.1}$$

where Y_t is the value of the variable at time t, $Y_{(t-1)}$ to $Y_{(t-p)}$ are lagged values of the variable at times $t-1$ to $t-p$, and p is a number of lagged.

The Moving Average model is primarily used for smoothing out fluctuations or noise in the data and identifying the underlying trend. By averaging out short-term variations, it provides a clearer representation of the overall pattern or direction of the time series. This can be helpful in making predictions or identifying turning points in the data

2.3.2 Autoregressive Model

The autoregressive model, often abbreviated as AR, is a statistical model that describes the relationship between an observation in a time series and a linear combination of its past observations. In other words, the autoregressive model assumes that the current value of a variable is influenced by its previous values.

The autoregressive model is denoted by $AR(p)$, where p represents the order of autoregression. The order p indicates how many lagged values of the variable are considered in the model. For example, $AR(1)$ represents a first-order autoregressive model, which includes the immediate lagged value of the variable as a predictor. $AR(2)$ includes the two most recent lagged values, and so on.

Mathematically, the autoregressive model can be represented as follows:

$$Y_t = \mu + \phi_1 Y_{(t-1)} + \phi_2 Y_{(t-2)} + \phi_3 Y_{(t-3)} + ... + \phi_p Y_{(t-p)} + \varepsilon_t \tag{2.2}$$

In this equation:

Y_t represents the value of the variable at time t.

μ is a constant term.

$\phi_1, \phi_2, ..., \phi_p$ are the autoregressive coefficients.

ε_t represents the random error term or noise component.

The autoregressive model captures the temporal dependencies and patterns present in the time series data. By estimating the autoregressive coefficients, the model can describe the behavior of the variable over time and make predictions for future values.

2.4 Machine Learning

At the basic knowledge level, ML indicates any type of computer algorithm that can learn on its own without having to be evidently programmed by a programmer. It is the procedure by which a computer algorithm optimizes parameters inside the program from input data; the purpose of this is to detect relationships within the input data, so that if there is unseen input into the computer, then the algorithm of the computer system can predict the output. At present, ML is being used in several fields, for example engineering, financial, medical, business, logistics, and industry. ML is divided into 3 types which are supervised learning, unsupervised learning, and reinforcement.

We model the stock price using different machine learning regression models based on supervised learning consisting of support vector regression, deep learning, random forest, and gradient boost tree. These models are explained below.

2.5 Regression Problem

In machine learning, a regression problem refers to a type of predictive modeling task where the goal is to predict a continuous numeric value or a real-valued output based on a set of input features or independent variables. In other words, the objective is to create a function that maps the input variables to a continuous output variable.

Regression problems are distinguished from classification problems, where the goal is to predict categorical or discrete labels. In regression, the predicted value can take on any numeric value within a range, making it suitable for tasks such as predicting prices, estimating quantities, or forecasting numerical values.

In a regression problem, the dataset consists of pairs of input features and corresponding target values. The model is trained on this data to learn the underlying patterns and relationships between the features and the target variable. Once trained, the model can then make predictions on new, unseen data.

There are various regression algorithms and techniques available in machine learning, including linear regression, polynomial regression, decision tree regression, support vector regression, random forest regression, and neural networks. The choice of algorithm depends on the characteristics of the data, the complexity of the problem, and the desired trade-offs between interpretability and predictive performance.

Overall, regression problems in machine learning focus on predicting continuous numerical values, enabling applications in fields such as finance, economics, healthcare, and many other domains.

2.6 Support Vector Regression (SVR)

Support vector regression is a type of support vector machine used to solve regression problems. This technique was developed by Vapnik (1995). Consider a data set T= $\{x_1, x_2, x_3, ..., x_n\}$ where $x_i \in \Re^n$ and $y_i \in \Re$ for $i = 1, 2, 3, ..., l$. Each x_i is the input vector for the response value, or output value y_i. A regression model is constructed from the data set T and is used to predict the response values of unseen input vectors. SVR is a nonlinear kernel-based regression method that attempts to construct a regression hyperplane with small fallibility in high-dimensional feature space. It possesses good function approximation and generalization capabilities.

Among the different types of support vector regression, the most commonly used is ε-SVR which constructs a regression hyperplane with an ε-band (Cristianini and Shawe-Taylor, 2000). To construct the model more efficiently, the input data is not required to be inside the ε-band. In case, where some of the input data situated outside the ε-band, penalty and slack variables are introduced to account for these data. For expedience, in the following, the term SVR is used to mean ε-SVR. The objective and constraint functions for an SVR are

$$\begin{aligned} &\text{min } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*); \\ &s.t.(\langle w, \phi(x_i)\rangle + b) - y_i \leq \varepsilon + \xi_i; \\ &y_i - (\langle w, \phi(x_i)\rangle + b) \leq \varepsilon + \xi_i^*; \\ &0 \leq \xi_i, \xi_i^*, i = 1, 2, 3, ..., l, \end{aligned} \tag{2.3}$$

where w is a coefficient vector, b is offset, l is the number of training data, C is a parameter which gives a trade off between model complexity and training error, ξ_i and ξ_i^* are slack variables for exceeding the response value by more than ε–band and for being below the response value by more than ε–band respectively. Notice that $\phi : X \to F$ is a nonlinear mapping function transforming the input space X to a feature space F. Also $\langle\cdot,\cdot\rangle$ designates the inner product of the involved arguments. The regression hyperplane to be derived is

$$f(x_i) = (\langle w, \phi(x_i)\rangle) + b \tag{2.4}$$

To solve Eq 2.3 we will use Lagrange Multipliers. The corresponding Lagrangian function is

$$\min_{\alpha^* \in \Re^{2l}} \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)k(x_i, x_j) + \varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i)$$

$$s.t. \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0$$

$$0 \leq \alpha_i^* \leq C, i = 1, 2, 3, ..., l \qquad (2.5)$$

where α_i^* and α_i are Lagrange multipliers. Notice that $k(x_i, x_j)$ is a kernel function.

2.6.1 Kernel Function

The kernel method is a technique used to deal with linearly inseparable data or non-linear data set. This method maps the data into a higher dimension space where it can be predict by SVR. The Kernel method is very powerful and the definition of the kernel is as follows.

Definition 2.1 A function $K(x, x')$ defined on $R^n \times R^n$ is called a kernel on $R^n \times R^n$ or kernel briefly if there exists a map ϕ from the space R^n some the Hilbert space

$$\phi : \quad R^n \to \mathbb{H},$$

$$x \mapsto \phi(x),$$

such that

$$K(x, x') = (\phi(x) \cdot \phi(x')), \qquad (2.6)$$

where $(\cdot)$ denotes the inner product of space $\mathbb{H}$.

Several types of kernel function are typically used, for example:

1. Linear Function

The linear or dot kernels are the simplest kernel function. It is given by the inner product between x and x' and then plus an optional constant c, the function is:

$$K(x, x') = x^T x' + c. \qquad (2.7)$$

2. Polynomial Function

A popular kernel used in SVR is a polynomial kernel; this method simply calculates the dot product of the data input. The form of a polynomial kernel is:

$$K(x,x') = ((x \cdot x') + 1)^d, \tag{2.8}$$

where d is a positive integer.

3. Gaussian Radial Basis Function

Gaussian radial basis function kernel of radial basis function is another kernel popularly used in SVR which the following format

$$K(x,x') = \text{exp}\left(-\gamma||x - x'||^2\right), \tag{2.9}$$

where $\gamma > 0$ is a parameter of radial basis function.

2.7 Deep Learning

Deep learning is a subfield of machine learning that focuses on training artificial neural networks with multiple layers, also known as deep neural networks. It is inspired by the structure and function of the human brain and aims to enable computers to learn and make intelligent decisions similar to how humans do. Deep learning algorithms learn to recognize patterns and extract features from raw data through a hierarchical representation of layers. Each layer in a deep neural network performs complex computations on the input data and passes the transformed information to the next layer. As the data propagates through these layers, the network learns to extract increasingly abstract and higher-level representations of the data. The key component of deep learning is the artificial neural network, which is composed of interconnected nodes, or artificial neurons, organized in layers. Each node applies a mathematical operation on its input and produces an output. Deep learning networks typically consist of an input layer, one or more hidden layers, and an output layer. The hidden layers, which can be numerous, allow for the network to learn intricate representations of the data. One of the main advantages of deep learning is its ability to automatically learn and extract relevant features from raw data, eliminating the need for manual feature engineering. This makes deep learning particularly powerful for tasks such as image and speech recognition, natural language processing, sentiment analysis, and many other areas where large amounts of data are available.

2.8 Decision Tree Regression

Decision tree regression is a machine learning algorithm used for regression analysis. It involves partitioning the input space into several regions and then fitting a simple model, usually a constant value, to each partition. The partitions are determined by recursively splitting the input space into smaller regions based on the value of a certain feature or input variable. The aim for regression problems is the prediction of a single output which takes continuous or discrete values by one more input variables. The output and input variables are known as respond and predictor variables, respectively.

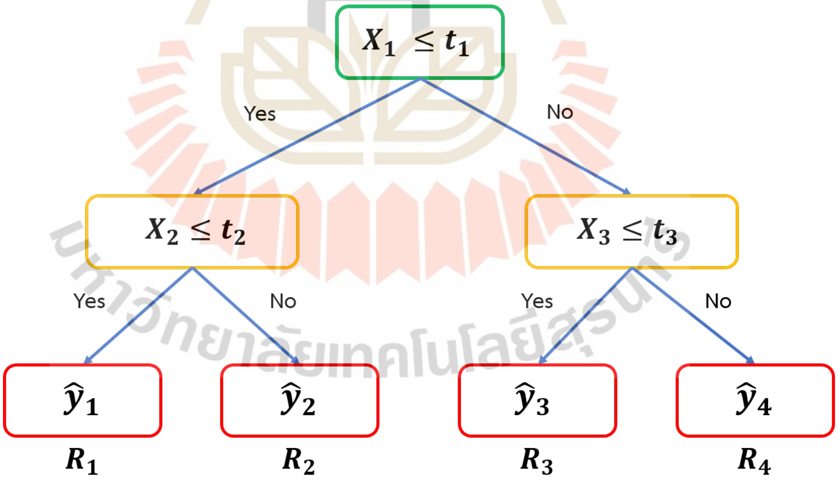

Figure 2.1 Example of Regression Tree model (Saranchai, 2022).

Let $y_1, y_2, \ldots, y_n \in \Re$ be output variables depending on $x_1, x_2, \ldots, x_n$ that are predictor variables and implied that divided predictor space is the set of possible value for $x_1, x_2, \ldots, x_p$ into $J-$districts and non-overlapping regions, $R_1, R_2, \ldots, R_J$, every observation at that R_i, it make the same response which implied the mean of observed in response value for training observation in R_i. The goal is to find boxes $R_1, \ldots, R_J$ that minimize the Residual Sum of Squares (RSS), given by

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \tag{2.10}$$

when y_i is a particular testing observation and $\hat{y}_{R_j}$ is the response mean of training observations within the $j - th$ box. In binary splitting, first we select the predictor x_j and cutpoint s which splits the predictor space into the regions $\{x|x_j < s\}$ and $\{x|x_j \geq s\}$ ($\{x|x_j < s\}$ means the region of predictor space in which x_j takes on a value less than s). We consider all predictors and all values of cutpoint where the resulting tree has the lowest RSS. We define

$$R_1(j,s) = \{x|x_j < s\} \text{ and } R_2(j,s) = \{x|x_j \geq s\},$$

we find the value of j and s by minimizing the equation

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2, \tag{2.11}$$

where $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are the means of response for the training observation in $R_1(j,s)$ and $R_2(j,s)$ respectively (Saranchai, 2023).

2.9 Random Forest

Random Forest is an algorithm of machine learning techniques. It is capable of solving both regression and classification problems. The concept is to combine multiple models of decision trees so to obtain the final output, instead of considering individual decision trees, in order to improves the efficiency of the model. In this work, newly constructed variables are provided for the training of each decision tree which in turn define the decisions at the nodes of the tree. It aims at minimizing the forecasting error by treating the stock market analysis as a regression problem, and depending on training variables, predict the next day's closing price of the stock for a particular company.

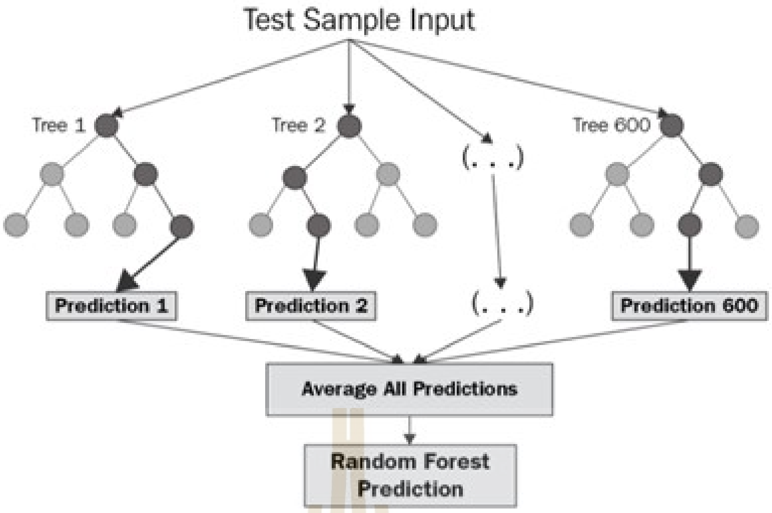

Figure 2.2 Example of Random Forest model (Afroz, 2019).

2.10 Gradient Boost Tree

Gradient Boost Tree is a sub-techniques of machine learning for solving regression and classification problems. The algorithm starts with a single decision tree, which is used to make predictions on the data. The errors made by the first tree are then used to train a second tree, which is added to the model. This process is repeated for a specified number of iterations, with each new tree learning from the errors made by the previous trees. The algorithm of GBT (Natekin and Knoll, 2013; Josh Starmer, 2019) is

Gradient Boost Tree algorithm.
Input $\{(x_i, y_i)\}_{i=1}^{n}$
1. Data input
2. Number of iterations M
3. Differentiable Loss function $L(y_i, F(x))$
Algorithm.
1. Initialize model with a constant value: $F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$
2. For $m = 1, 2, 3, \ldots, M$:
2.1 Compute so-called pseudo-residuals: $r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}; i = 1, 2, 3, \ldots, n$ where $F(x_i)$ is a predicted variable.
2.2 Fit a regression tree to the $r_{i,m}$ values and create terminal regions $\Re_{i,m}$ for $j = 1, 2, 3, \ldots, J_m$
2.3 Compute multiplier $\gamma_{j,m}$ for $j = 1, 2, 3, \ldots, J_m$ $\gamma_{j,m} = argmin_\gamma \sum_{x_i \in \Re_{i,j}} L(y_i, F_{m-1}(x_i) + \gamma)$
2.4 Update the model: $F_m(x) = F_{m-1} + \upsilon \sum_{j=1}^{j_m} \gamma_m I(x \in \Re_{j,m})$ where υ and I are learning rate and indicator function, respectively.
3. Output $F_m(x)$

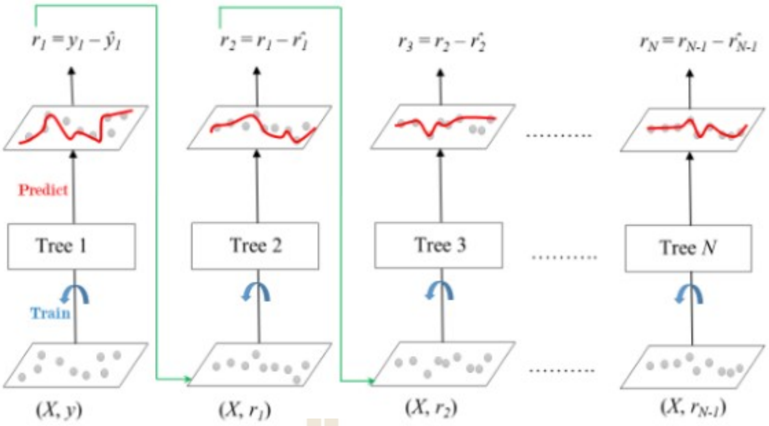

Figure 2.3 Example of the Gradient Boost Tree process (Neetika, 2020).

2.11 Features Engineering

Feature engineering is a process to modulate and transform features to be appropriate for a machine learning model. This methodology is an important process of a machine learning model, because only those attributes are used that exceedingly influence the model, thus reducing the time to construct and analyze the model (Zheng and Casari, 2018).

2.11.1 Feature Selection

Feature selection, also known as variable selection, is the process of selecting a subset of relevant features or variables from a larger set of available features in a dataset. It is an important step in machine learning and data analysis, as it aims to identify the most informative and significant features that contribute to the predictive performance of a model.

The goal of feature selection is to improve the model's performance and interpretability by reducing the dimensionality of the dataset. By selecting a subset of relevant features, unnecessary or redundant information is eliminated, which can lead to more efficient and accurate models. Feature selection can also help mitigate the risk of overfitting by reducing noise and focusing on the most informative features.

Feature selection techniques can be broadly categorized into 3 types:

1. Filter methods: These methods assess the statistical properties of the features independently of the machine learning model. They measure the correlation, mutual information, or statistical significance between each feature and the target variable. Features are selected based on predetermined criteria or rankings derived from statistical tests.

2. Wrapper methods: These methods evaluate the performance of a machine learning model by iteratively selecting different subsets of features and measuring their impact on the model's performance. The selection process is guided by a specific machine learning algorithm and performance metric, such as accuracy or error rate. Wrapper methods can be computationally expensive but generally result in more accurate feature subsets.

3. Embedded methods: These methods incorporate feature selection as part of the model training process itself. Certain machine learning algorithms have built-in feature selection mechanisms that assess the relevance of features during the training phase. These methods automatically select the most informative features while training the model.

The choice of feature selection technique depends on various factors, including the dataset size, dimensionality, the relationship between features and the target variable, and the specific goals of the analysis. It is often recommended to combine multiple feature selection methods to obtain a robust and reliable subset of features. By performing feature selection, the model complexity is reduced, training time is decreased, and the model's performance can be improved. Additionally, feature selection aids in interpreting the model by focusing on the most important features and providing insights into the underlying relationships in the data.

2.11.2 Feature Extraction

Feature extraction is a process in machine learning and data analysis that involves transforming raw or high-dimensional data into a reduced set of representative features, capturing the most relevant information for a given task. It aims to extract meaningful features that can effectively represent the data and facilitate subsequent analysis or modeling. Feature extraction is particularly useful when dealing with complex data, such as images, audio, text, or sensor data, where the original data may contain a large number of irrelevant or redundant features. By reducing the dimensionality and extracting relevant features, the computational complexity can be reduced, and the performance of subsequent tasks, such as classification or clustering, can be improved.

There are various techniques for feature extraction, depending on the type of data and the specific requirements of the task. Some commonly used techniques include:

1. **Principal Component Analysis (PCA):** PCA is a statistical technique that transforms high-dimensional data into a lower-dimensional space while maximizing the variance of the data. It identifies the principal components that capture the most significant information in the data.
2. **Linear Discriminant Analysis (LDA):** LDA is a supervised feature extraction technique that aims to find a linear combination of features that maximizes class separability. It is commonly used for dimensionality reduction in classification problems.
3. **Wavelet Transform:** The wavelet transform decomposes data into different frequency bands, capturing both local and global patterns. It is often used for feature extraction in signal processing tasks.
4. **Bag-of-Words (BoW):** BoW is a technique commonly used in natural language processing to extract features from text data. It represents text documents based on the frequency or presence of specific words or n-grams.
5. **Convolutional Neural Networks (CNN):** CNNs are deep learning models commonly used for feature extraction from images. They automatically learn hierarchical representations of images, extracting features at different levels of abstraction.
6. **Mel-Frequency Cepstral Coefficients (MFCC):** MFCC is a feature extraction technique commonly used in speech and audio processing. It captures the spectral characteristics of audio signals by analyzing the frequency content over time.

The choice of feature extraction technique depends on the specific task, the nature of the data, and the available domain knowledge. The extracted features are then used as inputs for subsequent analysis or modeling tasks, such as classification, regression, or clustering. Feature extraction plays a crucial role in improving the efficiency and effectiveness of machine learning algorithms by reducing dimensionality and focusing on the most relevant information in the data.

2.11.3 Automatic Feature Engineering

Automatic feature engineering, also known as automated feature engineering or feature synthesis, refers to the process of automatically generating new features from existing data without explicit manual intervention. It leverages machine learning algorithms, statistical techniques, and domain knowledge to derive new features that can enhance the performance of predictive models. Traditionally, feature engineering has been a manual and time-consuming process, requiring domain expertise and extensive trial and error. However, with the advent of automated feature engineering techniques, the process has become more efficient and less reliant on human effort.

Automatic feature engineering methods typically involve the following steps:

1. **Feature Generation:** Various techniques are used to generate a large set of potential features from the available data. This can include mathematical transformations, statistical aggregations, interaction terms, time-based features, or combining information from multiple sources.
2. **Feature Selection:** To handle the large number of generated features, selection methods are applied to identify the most relevant and informative ones. This can be done based on statistical tests, feature importance rankings, or machine learning algorithms.
3. **Feature Transformation:** Once the selected features are identified, transformations may be applied to improve their representation or normalize their distributions. This can include scaling, normalization, log transformations, or encoding categorical variables.
4. **Validation and Evaluation:** The generated and transformed features are evaluated using cross-validation or other validation techniques to assess their impact on the model's performance. This helps identify the most effective set of features for the given task.

The main advantages of automatic feature engineering include:

1. **Efficiency:** It reduces the manual effort required for feature engineering, allowing for quicker iteration and exploration of feature spaces.
2. **Exploration of Complex Relationships:** Automated methods can identify complex interactions and patterns in the data that may be difficult for humans to detect.
3. **Generalization:** Automatically generated features can capture general patterns and re-

lationships in the data, potentially improving the model's performance on unseen data.

4. Reduction of Bias: Manual feature engineering may introduce bias or subjectivity, whereas automated methods can minimize such biases by relying on data-driven techniques.

However, it is important to note that automatic feature engineering is not a one-size-fits-all solution. The quality and effectiveness of the generated features still depend on the quality of the input data, the appropriateness of the feature generation techniques, and the domain knowledge incorporated into the process. Overall, automatic feature engineering provides a valuable approach to discovering new and relevant features from data, potentially enhancing the performance of machine learning models and saving time and effort in the feature engineering process.

2.12 K-fold Cross Validation

Cross-validation is a resampling process used to estimate machine learning models on finite data. It is commonly used in practical machine learning models to compare and select a model for classification or regression problems because it is easy to understand, easy to actualize, and resultant in skill estimates that generally have a lower bias than other methods.

The data set T is separated into K sections (or folds), then $K-1$ sections are used for the training model and the remaining section is used for testing; this is repeated K times.

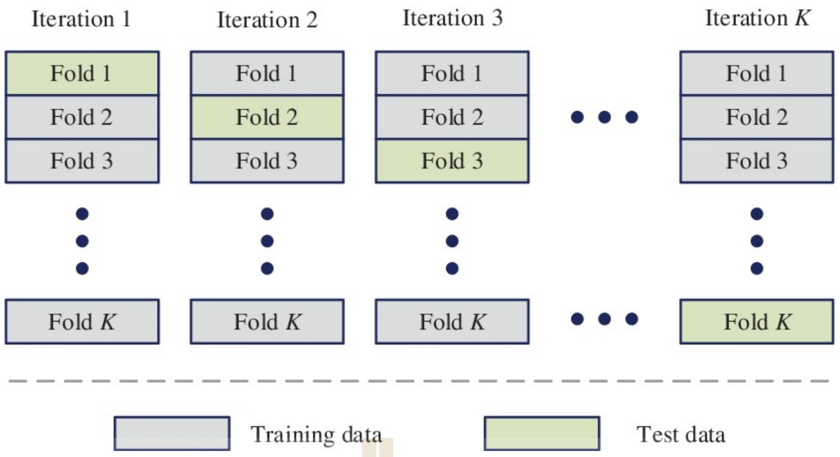

Figure 2.4 Example of K-Fold Cross Validation process (Ren, Li and Han, 2019).

2.13 Performance Metrics

As this dissertation studies in regression problems, thus we compare the performance of each model by considering three measurements names, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Square Error (SE).

$$\begin{aligned} RMSE &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}} \\ MAE &= \frac{\sum_{i=1}^{n}|y_i - \widehat{y_i}|}{n} \\ MAPE &= \frac{1}{n}\sum_{i=1}^{n}(\frac{|y_i - \widehat{y_i}|}{y_i})100 \\ SE &= \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \end{aligned}$$

where y_i is the actual data, $\widehat{y_i}$ is the predicted data and n is the number of data.

2.14 Related Researches

Wu et al. (2008) have proposed an ensemble prediction model using support vector regression and an artificial neural network for predicting stock price. The model performance of the ensemble model was compared with those of the support vector regression and the artificial neural network model alone. It was observed by the researcher that the ensemble model presented better results than all single models.

Senol et al. (2008) worked on stock price direction prediction using an Artificial Neural Network (ANN) approach: the case of Turkey. The proposed ANN can be used to predict stock prices and their direction of changes. The result was promising with a forecast accuracy of 78.47% on average.

Phua et al. (2000) carried out a study on stock price movements prediction using neural networks with genetic algorithms in the Singapore Stock Exchange. The performance of the prediction model was found to be 81% on the test data set and showed that the model was moderately proficient in its prediction.

Hanias et al. (2012) carried out a study to predict the daily stock exchange price index of the Athens Stock Exchange (ASE) using a Neural Network (NN) with backpropagation. The NN was used to conduct a prediction for nine days ahead. The performance of the prediction model obtained a mean square error of 0.0024.

Nusrat Rouf et al. (2021) presented a decade survey on methodologies, recent developments, and future directions on stock market prediction using machine learning techniques.

Mehar et al. (2019) worked on stock closing price prediction using machine learning techniques. This research used ANN and RF for predicting the next day's closing price for five companies. It was observed by the researcher that the ANN model presented more efficient results than RF.

CHAPTER III

RESEARCH METHODOLOGY

The main content of this chapter presents the process of this thesis which consists as following.

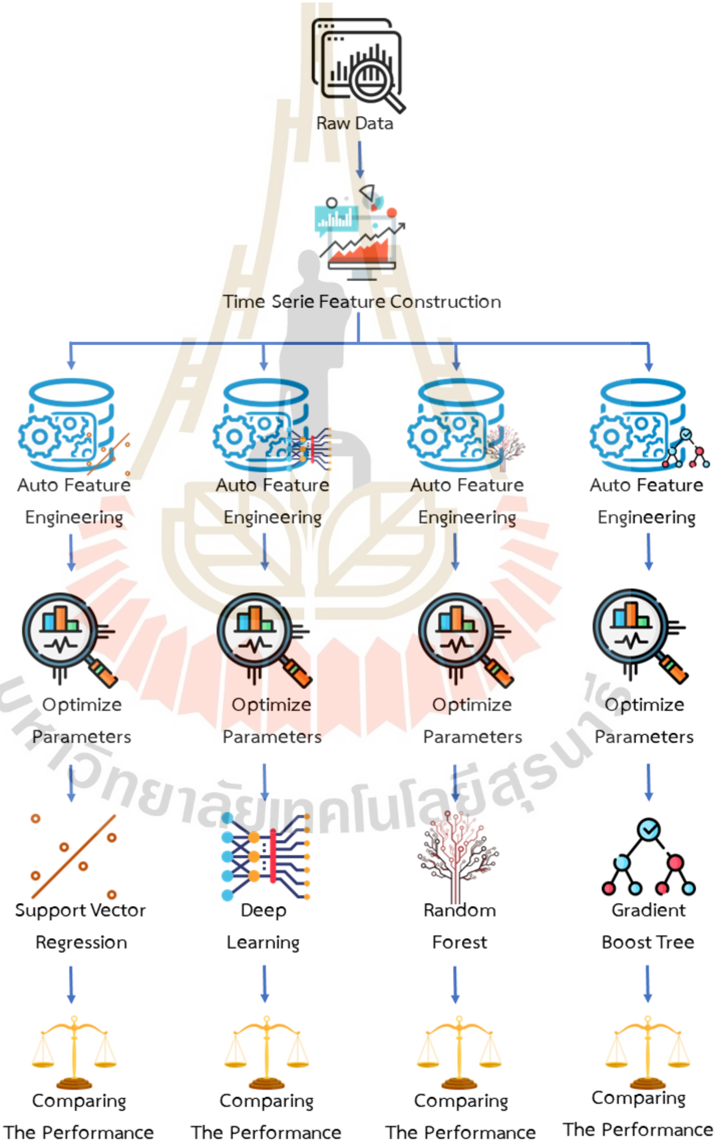

Figure 3.1 Flow chart the process of the thesis.

3.1 Data Collection

This thesis utilized historical daily stock price data from ten securities listed on the Stock Exchange of Thailand. The securities included in the analysis are BANPU, BBL, GUNKUL, IRPC, KBANK, KKP, KTB, PTT, SUPER, and TTB. The data covers the period from January 1, 2019, to December 31, 2021, and comprises a total of 727 instances. The dataset consists of five attributes, namely:

1. Closing Price
2. Open Price
3. Maximum Price
4. Minimum Price
5. Volume

3.2 Tool

For this research, Minitab Statistical Software version 20 and Microsoft Excel were utilized to construct and manage new features. Rapidminer Studio version 10.1 (Education license), running on the Microsoft Windows 10 operating system, was employed to analyze the data.

Minitab Statistical Software is a widely used tool for data analysis, statistical analysis, and process improvement. Organizations around the world rely on this software to enhance quality and decrease costs.

Rapidminer is a software tool designed for data and text mining. It covers the entire data analysis process, including data preparation, machine learning, and the deployment of predictive models.

3.3 Time series Feature Construction

In this study, new features were constructed using time series model analysis, which involved considering Moving Average periods from 1 to 7 days and Auto Regressive periods from 1 to 5 days. These analyses allowed for the creation of informative and relevant features for further investigation.

3.4 Features Collection

The features employed in this research are as follows:

1. Closing Price
2. Open Price
3. Maximum Price
4. Minimum Price
5. Volume
6. Maximum-Minimum
7. Autoregressive of closing price, period 1-5 days
8. Move average of closing price, period 1-7 days

3.5 Feature Engineering

In this research, an automatic feature engineering process was employed to select and generate new features based on four techniques: support vector regression model, deep learning, random forest, and gradient boost tree. These techniques were utilized to extract valuable information from the data and create enhanced features for further analysis.

3.6 Optimize Parameters for Construct Machine Learning Model

In this procedure, we employ parameter optimization techniques to enhance the performance of Support Vector Regression (SVR), Deep Learning (DL), Random Forest (RF), and Gradient Boost Tree (GBT) models during the construction process. The model parameters for each approach can be found in Tables 3.1 - 3.4. It is worth noting that certain parameters listed in Table 3.1 are applicable only to specific kernel types.

Table 3.1 Parameters of Support Vector Regression.

Parameter	Value
Kernel gamma	0.10, 0.15, 0.20, 0.25, ..., 0.9, 0.95, 1
Epsilon	0.0001, 0.001, 0.01, 0.1, 0, 1, 2, 3
C	1, 2, 3, ..., 100
Kernel	linear, polynomial, Radial basis function

Table 3.2 Parameters of Deep Learning.

Parameter	Value
Activation function	Rectifier, ExpRectifier, Tanh, Maxout
Loss function	Absolute, Quadratic, Quantile, Huber
Distribution function	Bernoulli, multinomial, Gaussian, poisson, huber
Hidden layer	50
Hidden node	50

Table 3.3 Parameters of Random Forest.

Parameter	Value
Number of Tree	1, 2, 3, ..., 100
Maximum Depth	1, 2, 3, ...,10

Table 3.4 Parameters of Gradient Boost Tree.

Parameter	Value
Number of Tree	1, 2, 3, ..., 100
Maximum Depth	1, 2, 3, ...,10
Learning Rate	0.001, 0.01, 0.1, 0.2, 0.3, 1

3.7 Creating a Model by Machine Learning

In this stage, we develop four prediction models, namely SVR, RF, DL, and GBT, each utilizing distinct sets of features. These features are derived through feature engineering techniques outlined in section 3.6. The validation of each model, we employ K-fold cross-validation with a value of Kequal to 10.

3.8 Predicting

The machine learning model we have obtained is employed to forecast the closing price of the stock for the following day.

3.9 Accuracy Measurement of Predicting Model

For regression problems, there are several commonly used accuracy measurements, including mean absolute error, mean squared error, root mean squared error, root mean squared log error, adjusted R-squared, and R-squared. In this research, the evaluation metrics employed are as follows:

1. Root mean squared error
2. Mean absolute error
3. Mean absolute percentage error
4. Squared error

CHAPTER IV

RESULTS AND DISCUSSION

This chapter presents the outcomes derived from the processing of the research methodology discussed in Chapter III. The aim is to showcase the results obtained using Rapidminer Studio version 10.1 (Education license). These results encompass the feature groups generated through support vector regression, deep learning, random forest, and gradient boost tree techniques. Additionally, the chapter includes the parameters specific to each model, the predictions made by each model using different feature groups, and an evaluation of the performance of each model for prediction purposes.

4.1 Data Set

The data set utilized in this study consists of historical daily stock prices for 10 companies. Each security within the data set comprises 727 instances (trading days) and is characterized by 18 features. The specific interpretation of each feature can be found in Table 4.1.

Table 4.1 The features are used in this thesis.

Feature	Type	Description
Closing price [Close(T)]	real number	The price of the stock at the end of a trading day
Open price [Open]	real number	First price of the stock at the beginning of a trading day
Minimum Price [Min]	real number	The minimum price of the stock during the trading day
Maximum Price [Max]	real number	The maximum price of the stock during the trading day
Max-Min [Max-Min]	real number	maximum price - minimum price
Volume [Volume]	real number	Number of stocks traded for security in all the markets during a given time
AR(1-5) [AR(1)-AR(5)]	real number	Stock price's period one - five days auto regressive
MA(1-7) [MA(1)-MA(7)]	real number	Stock price's period one - seven days moving average

4.2 Descriptive Statistic of Data

Table 4.2 Descriptive statistic of data.

Securities	Statistic	Data				
		Close(T)	Open	Max	Min	Volume
BANPU	Max	17.000	17.000	17.100	16.8	776.110
	Min	4.010	3.790	4.200	3.540	5.320
	Average	10.079	10.091	10.247	9.914	83.835

Continued on next page

Table 4.2 Descriptive statistic of data. (Continued)

Securities	Statistic	Data				
		Close(T)	Open	Max	Min	Volume
	S.D.	3.618	3.628	3.647	3.605	87.783
	Variance	13.089	13.160	13.304	12.996	7705.785
BBL	Max	215.000	215.000	216.000	214.000	83.910
	Min	88.000	90.250	90.750	88.000	0.840
	Average	140.755	140.943	142.255	139.478	9.581
	S.D.	38.012	38.098	38.046	38.049	8.151
	Variance	1444.879	1451.487	1447.528	1447.710	66.440
GUNKUL	Max	5.700	5.750	5.850	5.600	1840.000
	Min	1.910	1.910	1.960	1.840	2.160
	Average	3.202	3.204	3.255	3.158	96.262
	S.D.	0.921	0.920	0.941	0.906	179.694
	Variance	0.848	0.847	0.886	0.820	32289.799
IRPC	Max	6.150	6.050	6.200	5.950	1600.000
	Min	1.880	1.900	1.970	1.760	16.000
	Average	3.748	3.756	3.811	3.699	148.851
	S.D.	1.060	1.063	1.066	1.056	150.475
	Variance	1.124	1.130	1.136	1.114	22642.691
KBANK	Max	202.000	203.000	205.000	201.000	172.610
	Min	70.750	71.750	73.000	70.000	1.980
	Average	134.887	135.072	136.493	133.465	18.717
	S.D.	35.846	35.930	35.770	35.902	17.073
	Variance	1284.952	1290.940	1279.465	1288.939	291.477
KKP	Max	73.250	73.500	74.000	72.750	26.330
	Min	33.250	32.000	34.250	32.000	0.318
	Average	57.438	57.456	58.061	56.893	4.422

Continued on next page

Table 4.2 Descriptive statistic of data. (Continued)

Securities	**Statistic**	**Data**				
		Close(T)	Open	Max	Min	Volume
	S.D.	11.081	11.091	11.008	11.154	3.271
	Variance	122.792	123.011	121.173	124.423	10.697
KTB	Max	20.200	20.100	20.200	20.000	245.060
	Min	8.400	8.400	8.600	8.350	3.570
	Average	13.706	13.713	13.860	13.566	33.660
	S.D.	3.653	3.656	3.650	3.660	24.854
	Variance	13.345	13.364	13.320	13.394	617.705
PTT	Max	49.750	50.000	50.250	49.500	643.780
	Min	25.750	24.200	26.000	23.200	9.210
	Average	41.026	41.064	41.514	40.615	62.914
	S.D.	5.040	5.078	4.972	5.124	44.520
	Variance	25.401	25.785	24.721	26.257	1982.025
SUPER	Max	1.060	1.070	1.100	1.050	5290.000
	Min	0.310	0.320	0.330	0.280	7.270
	Average	0.77	1 0.773	0.785	0.759	271.265
	S.D.	0.194	0.195	0.196	0.193	462.761
	Variance	0.038	0.038	0.038	0.037	214147.745
TTB	Max	2.120	2.140	2.160	2.100	3070.000
	Min	0.680	0.700	0.740	0.600	24.340
	Average	1.311	1.312	1.330	1.294	334.941
	S.D.	0.352	0.353	0.354	0.352	326.059
	Variance	0.124	0.125	0.125	0.124	106314.644

where S.D. is standard deviation.

4.3 The Novel Features

This section highlights the outcomes of the automatic feature engineering process on SVR, DL, RF, and GBT models. The result includes the selection and generation of new attributes. The following are the newly obtained features:

Table 4.3 New feature by automatic feature engineering.

Securities	Model	Feature
BANPU	SVR	Max+Close(T)
	DL	Close(T)
	RF	Close(T), Min, MA(5), Close(T)-Max+Close(T), Max+Close(T), Max+Close(T)-MA(4)
	GBT	Close(T), AR(4)
BBL	SVR	Close(T), Max * Max, exp(Close(T))
	DL	Close(T)
	RF	Close(T), Max, MA(7), Close(T)*Close(T)/MA(1)
	GBT	Max, Max-Min, MA(3)
GUNUL	SVR	Close(T)
	DL	Close(T)
	RF	min(Volume,Min), exp(1/exp(Close(T)))
	GBT	Close(T), Min, MA(5), AR(1), AR(4), AR(5,) AR(5)/AR(4)
IRPC	SVR	Closs(T), Min, Volume, Closs(T)+Max, abs(Volume)
	DL	Close(T)
	RF	Closs(T), AR(3), Closs(T)+Max-AR(4)
	GBT	Close(T), Open, Max, Min, Max-Min, Volume, MA(1), MA(2), MA(3), MA(4), MA(5), MA(6), MA(7), AR(1), AR(2), AR(3), AR(4), AR(5)
KBANK	SVR	Close(T), abs(Close(T))

Continued on next page

Table 4.3 New feature by automatic feature engineering. (Continued)

Securities	Model	Feature
	DL	Close(T)
	RF	Max, MA(5), 1/exp(Close(T)), exp(Close(T))-(Min*Close(T)), Min*Close(T), 1/exp(Close(T))*MA(5)
	GBT	Close(T), Open, MA(7), AR(2), exp(Close(T))
KKP	SVR	Close(T), Open, Max, Min, Max-Min, Volume, MA(1), MA(2), MA(3), MA(4), MA(5), MA(6), MA(7), AR(1), AR(2), AR(3), AR(4), AR(5)
	DL	Close(T)
	RF	Close(T), Min, 1/exp(Close(T))
	GBT	Close(T), MA2
KTB	SVR	Close(T), Open, Max, Min, Max-Min, Volume, MA(1), MA(2), MA(3), MA(4), MA(5), MA(6), MA(7), AR(1), AR(2), AR(3), AR(4), AR(5)
	DL	Close(T)
	RF	Max, Min, Close(T)*Close(T)
	GBT	Close(T), Volume, MA(2)
PTT	SVR	Close(T)+Close(T)+Min
	DL	Close(T)
	RF	Close(T), Max, MA(1), AR(3)
	GBT	Close(T), Open, Max, Min, Max-Min, Volume, MA(1), MA(2), MA(3), MA(4), MA(5), MA(6), MA(7), AR(1), AR(2), AR(3), AR(4), AR(5)
SUPER	SVR	Close(T)+Close(T)+Min
	DL	Close(T)

Continued on next page

Table 4.3 New feature by automatic feature engineering. (Continued)

Securities	Model	Feature
	RF	abs(Close(T))*Close(T)/Max, Close(T)+Open, 1/1/exp(Close(T))
	GBT	Close(T), Open, Max, Min, Max-Min, Volume, MA(1), MA(2), MA(3), MA(4), MA(5), MA(6), MA(7), AR(1), AR(2), AR(3), AR(4), AR(5)
TTB	SVR	Max+Min+Max, Max+Close(T)
	DL	Close(T)
	RF	Close(T), Open, Max - MA(6) +Close(T)
	GBT	Close(T), MA(4), MA(6), exp(Close(T))

4.4 Parameter for Construct of each Model

This section presents the results of the parameter optimization process for the SVR, DL, RF, and GBT models. The following are the obtained parameters for each model:

Table 4.4 Parameter of BANPU.

Model											
SVR				DL			RF		GBT		
EP	KF	DG	C	AF	LF	DF	NT	MD	NT	MD	LR
0.01	poly	0	72	Rectifier	Quantile	Gaussian	93	6	57	2	0.01

Table 4.5 Parameter of BBL.

Model											
SVR				DL			RF		GBT		
EP	KF	GM	C	AF	LF	DF	NT	MD	NT	MD	LR
1	lin.	0	27	Tanh	Huber	huber	29	6	64	2	0.01

Table 4.6 Parameter of GUNKUL.

Model											
SVR				DL			RF		GBT		
EP	KF	GM	C	AF	LF	DF	NT	MD	NT	MD	LR
0.01	rbf	0.7	0	Rectifier	Quadratic	Gaussian	45	6	82	2	0.01

Table 4.7 Parameter of IRPC.

Model											
SVR				DL			RF		GBT		
EP	KF	GM	C	AF	LF	DF	NT	MD	NT	MD	LR
0.01	rbf	0.05	2	Rectifier	Absolute	Gaussian	57	6	73	2	0.1

Table 4.8 Parameter of KBANK.

Model											
SVR				DL			RF		GBT		
EP	KF	GM	C	AF	LF	DF	NT	MD	NT	MD	LR
1	lin	0	30	Tanh	Quadratic	Gaussian	76	6	67	2	0.1

Table 4.9 Parameter of KKP.

Model											
SVR				DL			RF		GBT		
EP	KF	GM	C	AF	LF	DF	NT	MD	NT	MD	LR
0.001	lin.	0	4	Tanh	Huber	Gaussian	33	6	63	2	0.1

Table 4.10 Parameter of KTB.

Model											
SVR				DL			RF		GBT		
EP	KF	GM	C	AF	LF	DF	NT	MD	NT	MD	LR
0.1	rbf	0	56	Rectifier	Huber	Gaussian	70	7	42	4	0.1

Table 4.11 Parameter of PTT.

Model											
SVR				DL			RF		GBT		
EP	KF	GM	C	AF	LF	DF	NT	MD	NT	MD	LR
0.1	rbf	0	27	ExpRectifier	Huber	Gaussian	50	6	44	4	0.2

Table 4.12 Parameter of SUPER.

Model											
SVR				DL			RF		GBT		
EP	KF	GM	C	AF	LF	DF	NT	MD	NT	MD	LR
0.1	rbf	0.1	0	Tanh	Quadratic	Gaussian	35	8	36	3	0.2

Table 4.13 Parameter of TTB.

Model											
SVR				DL			RF		GBT		
EP	KF	GM	C	AF	LF	DF	NT	MD	NT	MD	LR
0.001	lin	0	0	Rectifier	Huber	Gaussian	50	6	27	3	0.2

where EP, KF, GM, DG, AF, LF, DF, NT, MD, and LR are epsilon, kernel function, gamma, degree of polynomial, activation function, loss function, distribution function, number of tree, maximum depth, and learning rate respectively.

4.5 Result of Model Prediction

This section presents the predictions of stock prices for each security using the SVR, DL, RF, and GBT models. Furthermore, it provides an evaluation of the performance of each model’s predictions. The following are the obtained prediction results:

4.5.1 Performance of Predicting Model

Table 4.14 RMSE, MAE, MAPE and SE of the models.

Securities	Model	RMSE	MAE	MAPE(%)	SE
BANPU	SVR	2.286	2.044	16.891	758.011
	DL	0.381	0.269	2.290	21.097
	RF	0.514	0.389	3.261	38.356
	GBT	0.480	0.358	2.976	33.356
BBL	SVR	1.720	1.242	1.073	429.027
	DL	1.712	1.235	1.073	425.125
	RF	1.936	1.531	1.318	543.684
	GBT	2.272	1.767	1.522	748.611

Continued on next page

Table 4.14 RMSE, MAE, MAPE and SE of the models. (Continued)

Securities	Model	RMSE	MAE	MAPE(%)	SE
GUNKUL	SVR	0.632	0.447	8.638	57.837
	DL	0.527	0.475	9.444	40.299
	RF	0.948	0.865	17.243	130.263
	GBT	0.954	0.881	17.582	131.996
IRPC	SVR	0.077	0.061	1.519	0.849
	DL	0.066	0.051	1.288	0.628
	RF	0.093	0.067	1.662	1.255
	GBT	0.097	0.068	1.674	1.355
KBANK	SVR	2.511	1.833	1.450	914.011
	DL	2.507	1.825	1.444	911.507
	RF	2.791	1.970	1.561	1129.868
	GBT	2.950	2.114	1.682	1262.212
KKP	SVR	0.778	0.587	1.041	87.742
	DL	0.788	0.608	1.081	90.038
	RF	0.945	0.686	1.215	129.392
	GBT	0.873	0.688	1.222	110.431
KTB	SVR	0.224	0.157	1.380	7.276
	DL	0.175	0.124	1.109	4.427
	RF	0.250	0.182	1.584	9.044
	GBT	0.205	0.147	1.316	6.120
PTT	SVR	0.521	0.403	1.058	39.409
	DL	0.517	0.400	1.055	38.823
	RF	0.610	0.490	1.289	53.943
	GBT	0.784	0.617	1.623	89.112
SUPER	SVR	0.015	0.010	0.753	0.033
	DL	0.014	0.009	0.672	0.028

Continued on next page

Table 4.14 RMSE, MAE, MAPE and SE of the models. (Continued)

Securities	Model	RMSE	MAE	MAPE(%)	SE
	RF	0.019	0.012	1.286	0.051
	GBT	0.028	0.022	2.270	0.111
TTB	SVR	0.027	0.021	1.818	0.102
	DL	0.021	0.017	1.467	0.065
	RF	0.029	0.020	1.696	0.122
	GBT	0.026	0.018	1.562	0.096

Table 4.14 presents the performance of various models in predicting the closing prices of different securities using a 10-fold cross-validation technique. The following table summarizes the best models for each security and their corresponding evaluation metrics:

Table 4.15 The best models for each security.

Security	Best Model	RMSE	MAE	MAPE	SE
BANPU	Deep Learning	0.381	0.26	2.290	21.09
BBL	Deep Learning	1.712	1.235	1.073	25.125
GUNKUL	Deep Learning	0.632	0.447	8.638	57.837
IRPC	Deep Learning	0.077	0.061	1.519	0.849
KBANK	Deep Learning	2.507	1.825	1.444	911.507
KKP	Support Vector Regression	0.778	0.587	1.041	87.742
KTB	Deep Learning	0.175	0.124	1.109	4.427
PTT	Deep Learning	0.517	0.400	1.055	38.823
SUPER	Deep Learning	0.014	0.009	0.672	0.028
TTB	Deep Learning	0.021	0.017	1.467	0.065

The table 4.15 show cases that the deep learning model consistently performs well across most securities, except for KKP, where the support vector regression model stands out as the best performer.

4.5.2 Stock Price Prediction Results

In this section present the stock price prediction of each securities based on SVR, DL, RF and GBT model.

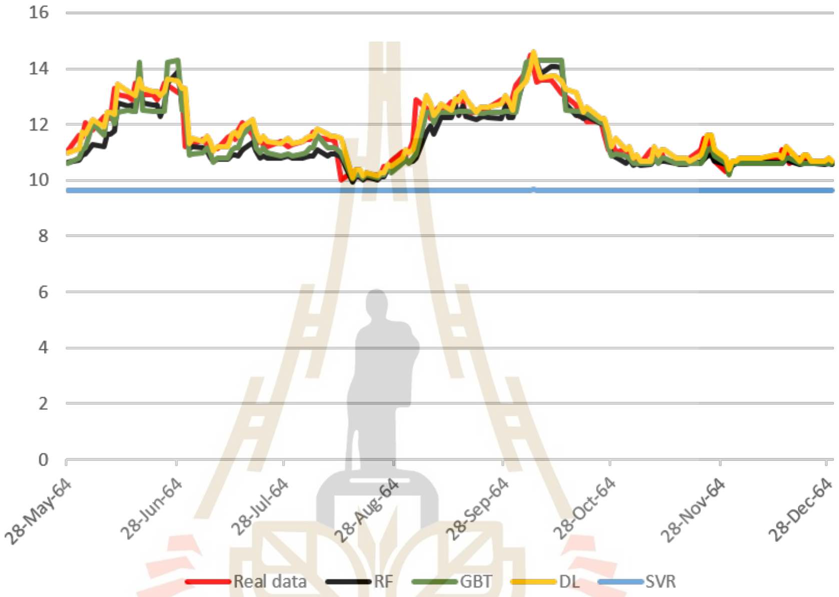

Figure 4.1 Price predicting of BANPU.

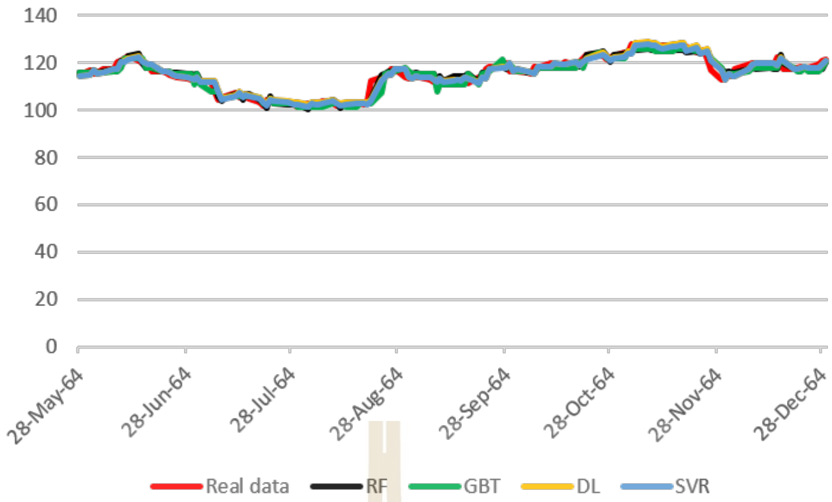

Figure 4.2 Price predicting of BBL.

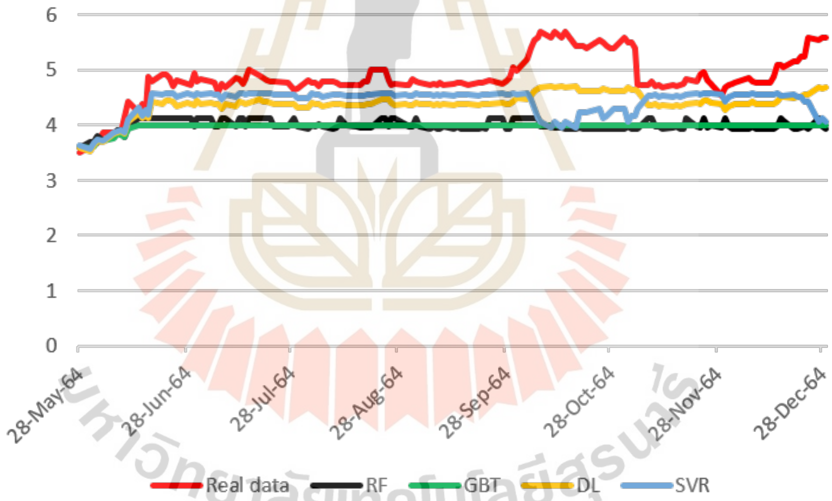

Figure 4.3 Price predicting of GUNKUL.

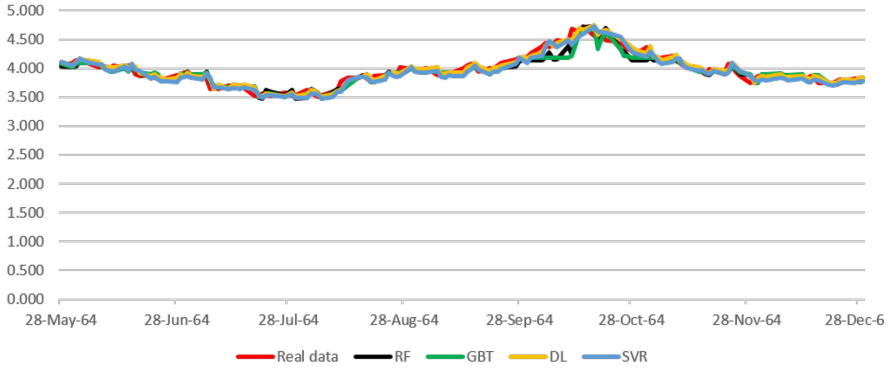

Figure 4.4 Price predicting of IRPC.

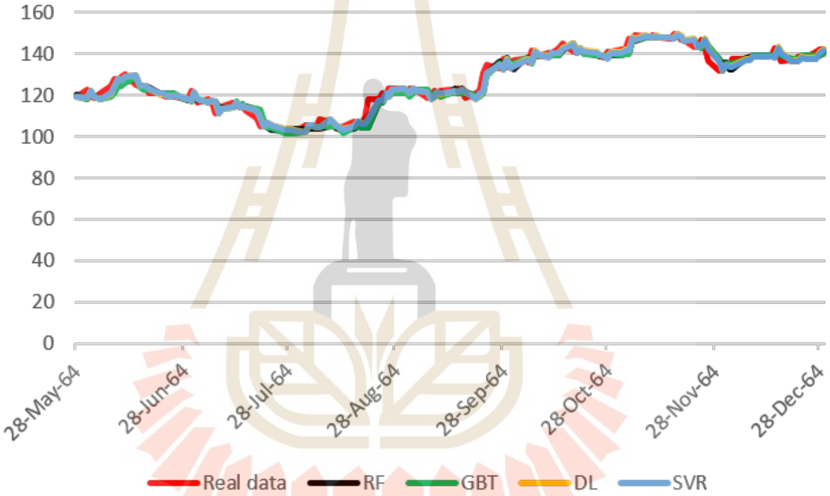

Figure 4.5 Price predicting of KBANK.

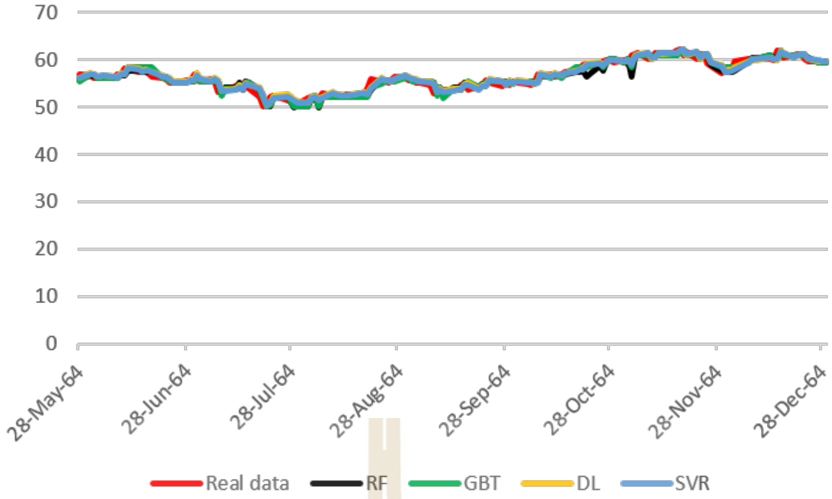

Figure 4.6 Price predicting of KKP.

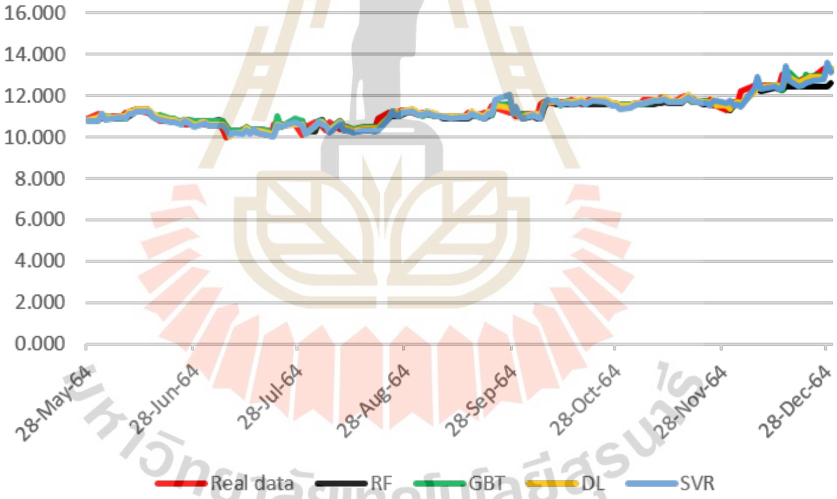

Figure 4.7 Price predicting of KTB.

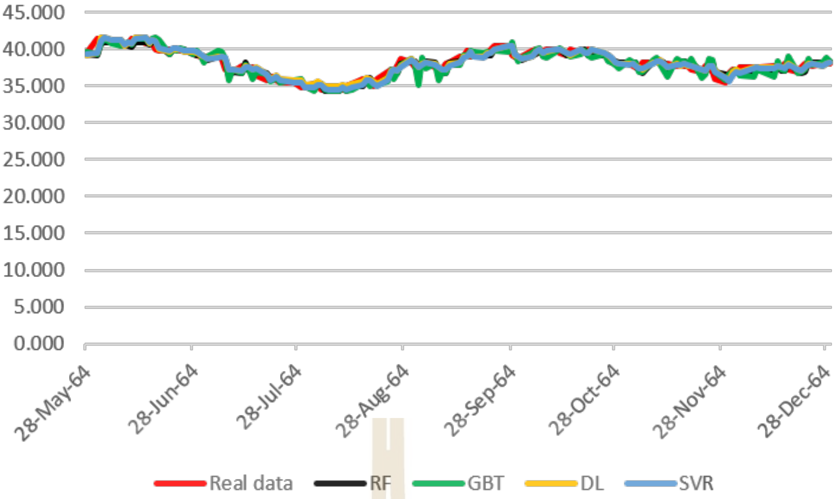

Figure 4.8 Price predicting of PTT.

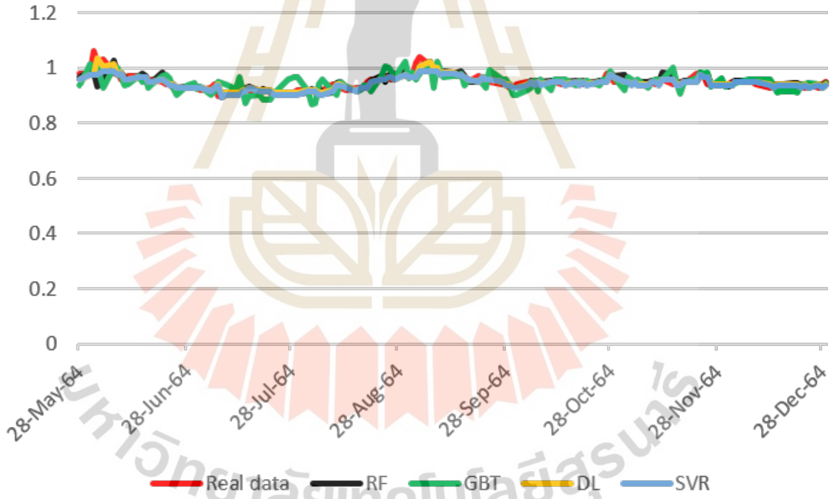

Figure 4.9 Price predicting of SUPER.

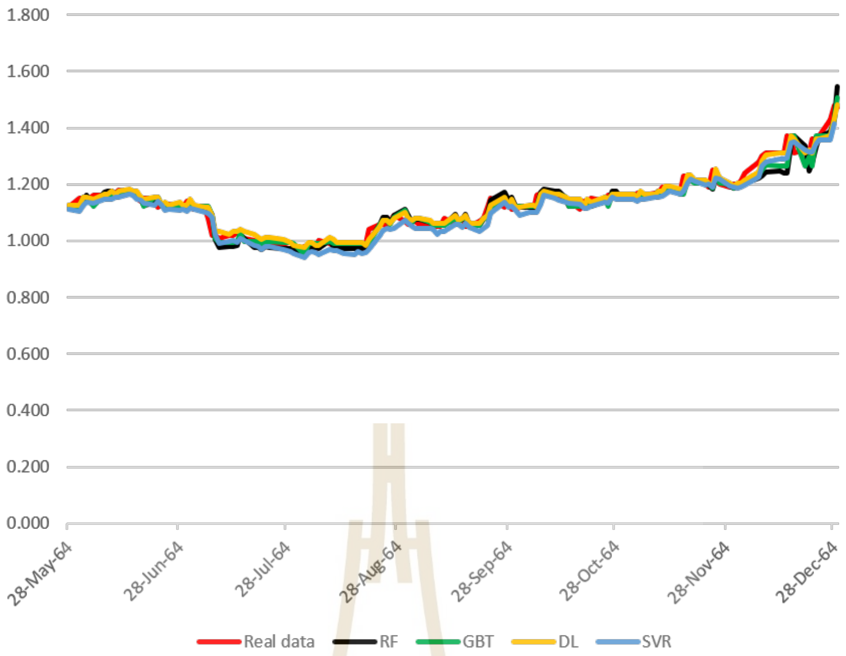

Figure 4.10 Price predicting of TTB.

Figures 4.1 - 4.10 illustrate the stock price predictions generated by obtained model.

CHAPTER V

RESULTS AND DISCUSSION

In this thesis, the objective was to predict the closing price of stocks on the following day using historical daily stock price data from ten companies listed on the Stock Exchange of Thailand. The prediction models employed in this study included regression with support vector regression, deep learning, random forest, and gradient boost tree. The implementation of these models was done using Rapidminer Studio 10.1, utilizing an Education license.

The analysis was divided into three main sections. The first section focused on selecting and generating new group features through automatic feature engineering. This process was applied to support vector regression, deep learning, random forest, and gradient boost tree models. Consequently, group attributes were obtained from each model, which served as the outcome of this section.

Moving on to the second section, the goal was to optimize the model parameters. This involved constructing the support vector regression, deep learning, random forest, and gradient boost tree models to predict the stock closing price on the following day. The objective was to obtain the parameter values for each model that would yield the best performance.

In the third part of the analysis, the group features obtained in the first section were combined with the optimized parameters from the second section. This integration allowed for the construction of a prediction model using the support vector regression, deep learning, random forest, and gradient boost tree, incorporating 10-fold cross-validation to enhance the reliability of the predictions.

The model was constructed by evaluating different group features using RMSE, MADE, MAPE, and SE, as presented in Table 4.14. The results in Table 4.14, indicate that the deep learning model performed the best in predicting the prices of nine securities, namely BANPU, BBL, GUNKUL, IRPC, KBANK, KTB, PTT, SUPPER, and TTB. On the other hand, the support vector regression model demonstrated the highest performance in predicting the prices of KKP securities. Figures 4.1 - 4.10 depict the price predictions for each security based on the testing data.

In the future, the model developed in this study holds potential for application and further development in the field of artificial intelligence. For instance, it can be utilized to construct an application that assists in analyzing securities prices, enabling investors to leverage the information for making short-term trading decisions. Daily stock price data can be collected and used for automated data analysis.

Researchers may also be interested in applying the findings and the automatic feature engineering method to other models in their future work. The approach of auto feature engineering and machine learning can be adapted and applied to analyze data of various types. It is anticipated that researchers will find the outcomes of this study beneficial in their own research

REFERENCES

REFERENCES

Afroz, C. (2019). *Random Forest Regression*. Retrieved from https://medium.com/swlh/random-forest-and-its-implementation- 71824ced454f.

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge.

Hanias, M., Curtis, P., and Thalassinos, J. (2012). *Time Series Prediction with Neural Networks for the Athens Stock Exchange Indicator. European Research Studies, 5*(2), 23-31.

Kanter, J. M., and Veeramachaneni, K. (2015). *Deep feature synthesis: Towards automating data science endeavors.* 2015 IEEE International Conference on Data Science and dvanced Analytics (DSAA).

Mehtab, S., and Sen, J. (2020). *A time series analysis-based stock price prediction using machine learning and deep learning models. International Journal of Business Forecasting and Marketing Intelligence (IJBFMI), 6*(4), 272 - 335.

Natekin, A., and Knoll, A. (2013). *Gradient boosting machines, a tutorial.* Frontiers in Neurorobotics, *7*(21).

Neetika, K. (2020). *ML – Gradient Boosting.* Retrieved from https://www.geeksforgeeks.org/ml–gradient–boosting/

Patel, J., Shah, S., Thakkar, P., and Kotecha, K. (2015). *Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42*(1), 259-268.

Phua, P. K. H., Ming, D., and Lin, W. (2000). *Neural network with genetic algorithms for stock prediction.* 5th Conference of the Association of Asian-Pacific Operations Research Societies, Singapore.

Ren, Q., Li, M., and Han, S. (2019). *Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. Big Earth Data, 3*(1), 8–25.

Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., and Kim, H.-C. (2021). *Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions.* Electronics, *10*(21).

Saranchai, S., Benjawan, R., Jessada, T., Bura, S., and Arjuna, C. (2023). *Modeling to Predict the Patients' Postoperative WOMAC Score by Features Engineering and Gradient Boost Tree. Suranaree Journal of Science and Technology. 30*(3), 030107(1-8).

Sattawat, B., and Surachai, C. (2020). *Trading strategy and portfolio optimization using support vector machine : An empirical study on the stock exchange of thailand. KKU Research Journal of Humanities and Social Sciences (Graduate Studies), 8*(2).

Senol, D., and Ozturan, M. (2008). *Stock price direction prediction using artificial neural network approach: the case of Turkey. Journal of Artificial Intelligence, 1*(2), 70-77.

Starmer, J. (2019). *Gradient Boost Part 2 (of 4) : Regression Details.* Retrieved from https://www.youtube.com/watchv=2xudPOBz–vst=492sab_channel=StatQuestwithJosh Starmer.

Surachai, C., Chayamin, C., and Jeeranun, K. (2013). *Forecast stock price using neuro-fuzzy. Journal of Management Scienc, 30*(2).

The Stock Exchange of Thailand. (2022). *History of the Stock Exchange of Thailand.* Retrieved from https://www.set.or.th/set/mainpage.dolanguage=thcountry=TH.

Tsai, C. F., and 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.

Vijh, M., Chandola, D., Tikkiwal, V. A., and Kumar, A. (2020). *Stock closing price prediction using machine learning techniques. Procedia Computer Science, 167*, 599 –606.

Wu, Q., Chen, Y., and Liu, Z. (2008). *Ensemble model of intelligent paradigms for stock market forecasting.* Proceedings of the IEEE 1st International Workshop on Knowledge Discovery and Data Mining, Washington, DC, USA, 205 – 208.

Yang, J., Zhao, C., Yu, H., and Chen, H. (2020). *Use GBDT to Predict the Stock Market. Procedia Computer Science, 174*, 161-171.

Zheng, A., and Casari, A. (2018). *Feature Engineering for Machine Learning.* USA, O'Reilly Media.

ปรเมษฐ์ ธันวานนท์ม, ชัยกรณ์ ยิ่งเสรี, วรพล พงษ์เพ็ชร, และ ธนภัทร ฆังคะจิตร. (2560). การประยุกต์ใช้โมเดลการเรียนรู้แบบรวมกลุ่มเพื่อพยากรณ์แนวโน้มราคาของหลักทรัพย์ในตลาดหลักทรัพย์แห่งประเทศไทย. *Journal of Information Science and Technology, 2*(1).

APPENDICES

APPENDIX A

PROCESS OF AUTOMATIC FEATURE ENGINEERING

A.1 Process of Automatic Feature Engineering in the Rapidminer Studio program

In this section we present the process of automatic feature engineering in the Rapidminer Studio program, with validation by 10-fold cross validation. We obtained the process of each model as follows:

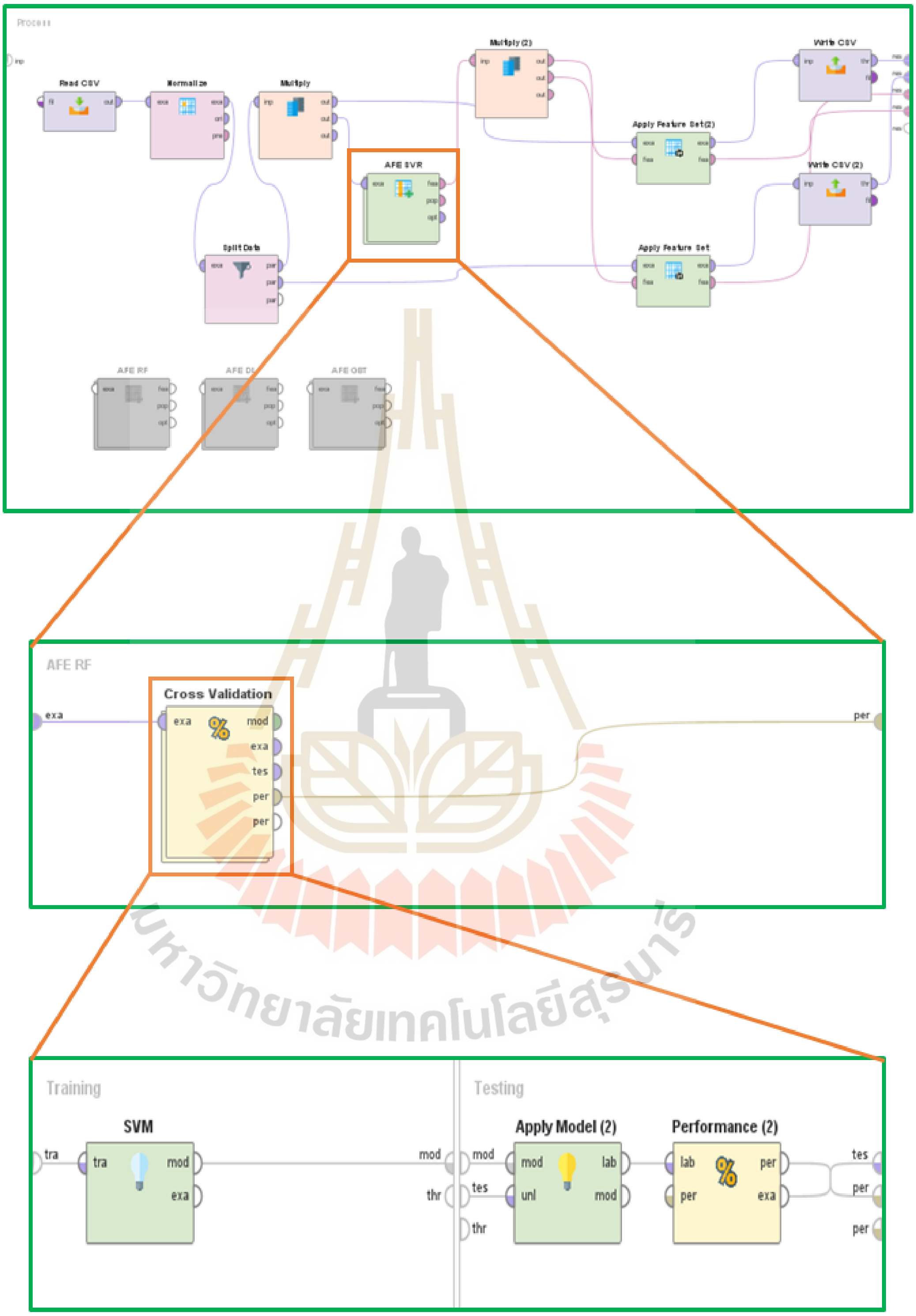

Figure A.1 Process of automatic feature engineering base on SVR.

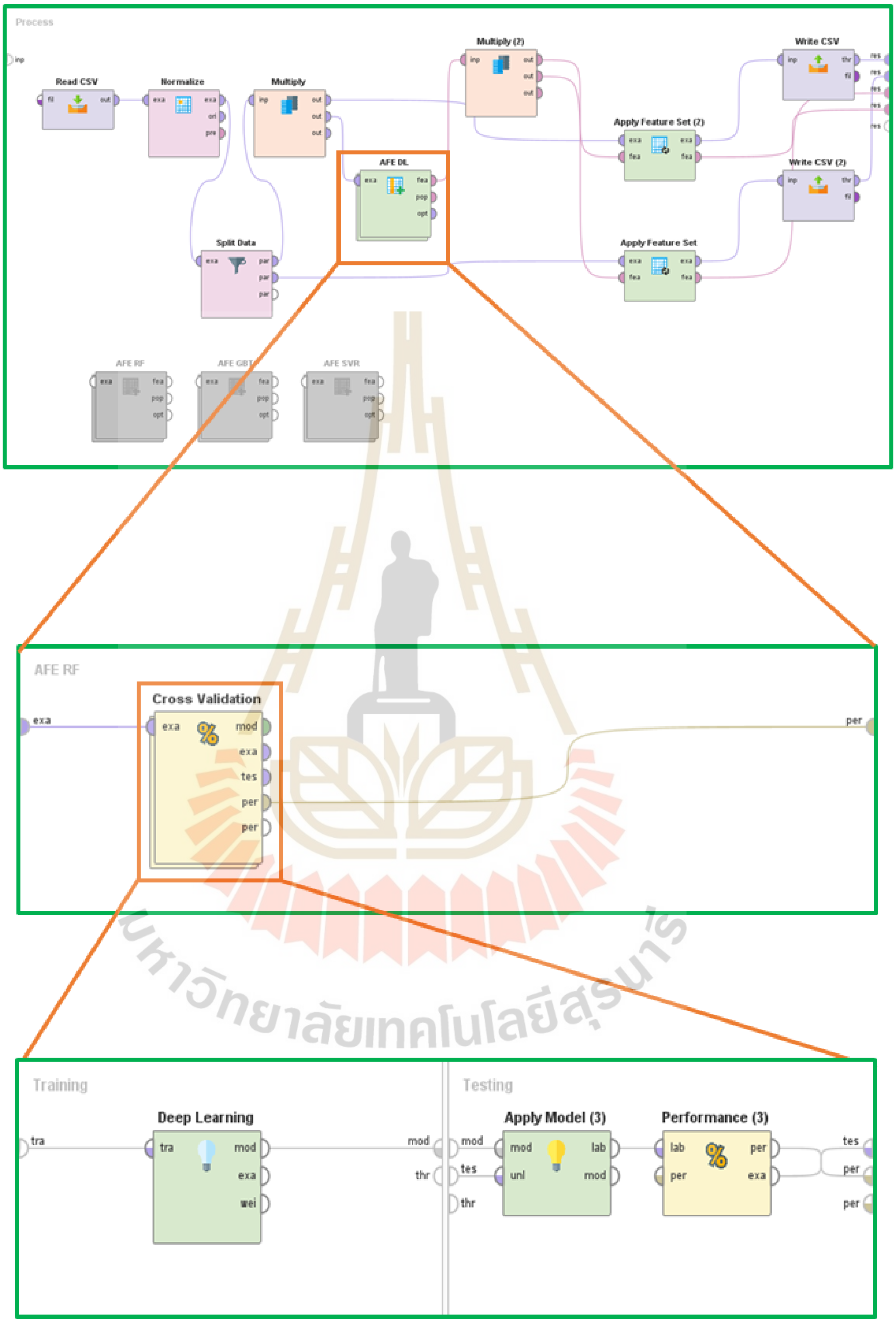

Figure A.2 Process of automatic feature engineering base on DL.

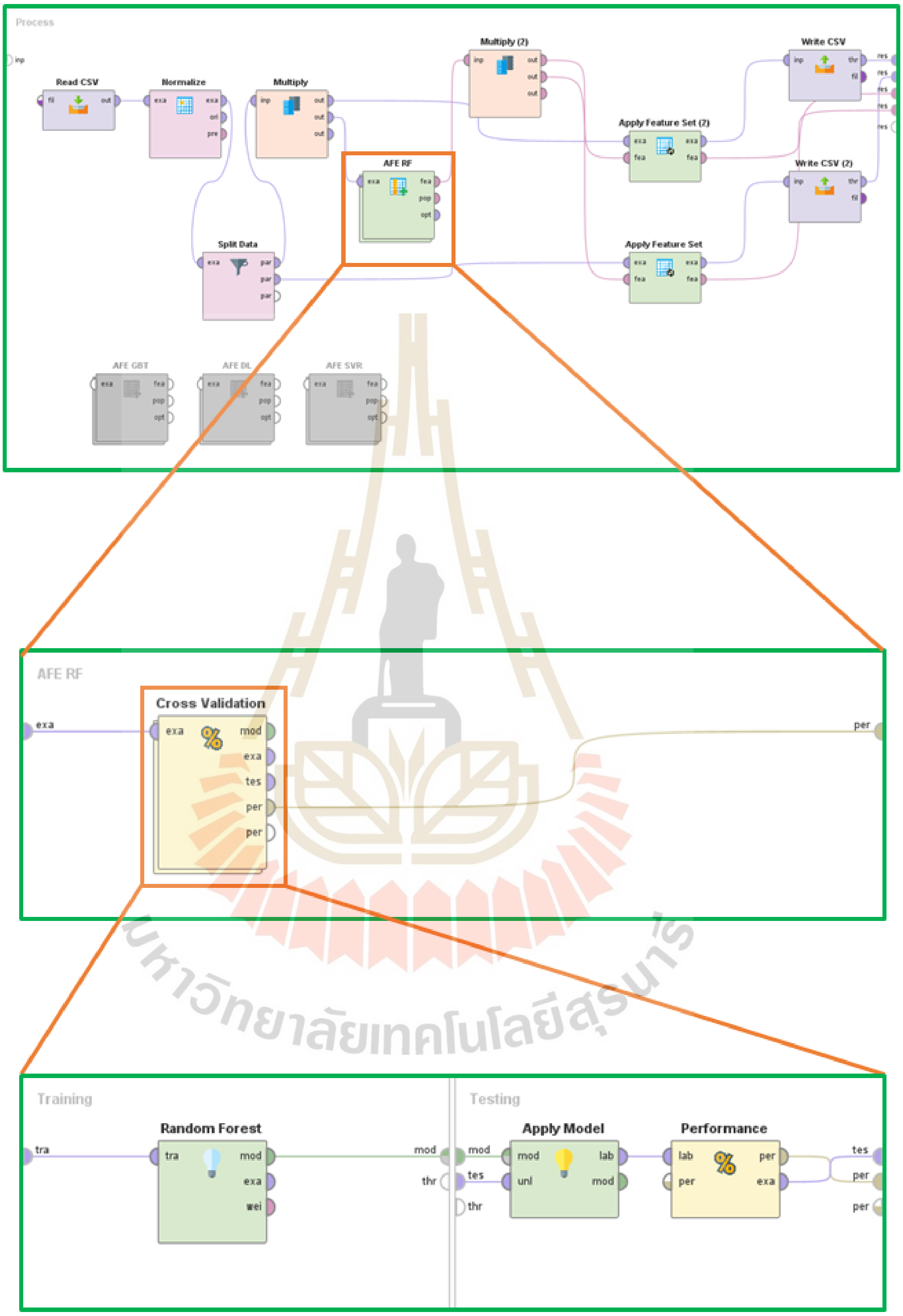

Figure A.3 Process of automatic feature engineering base on RF.

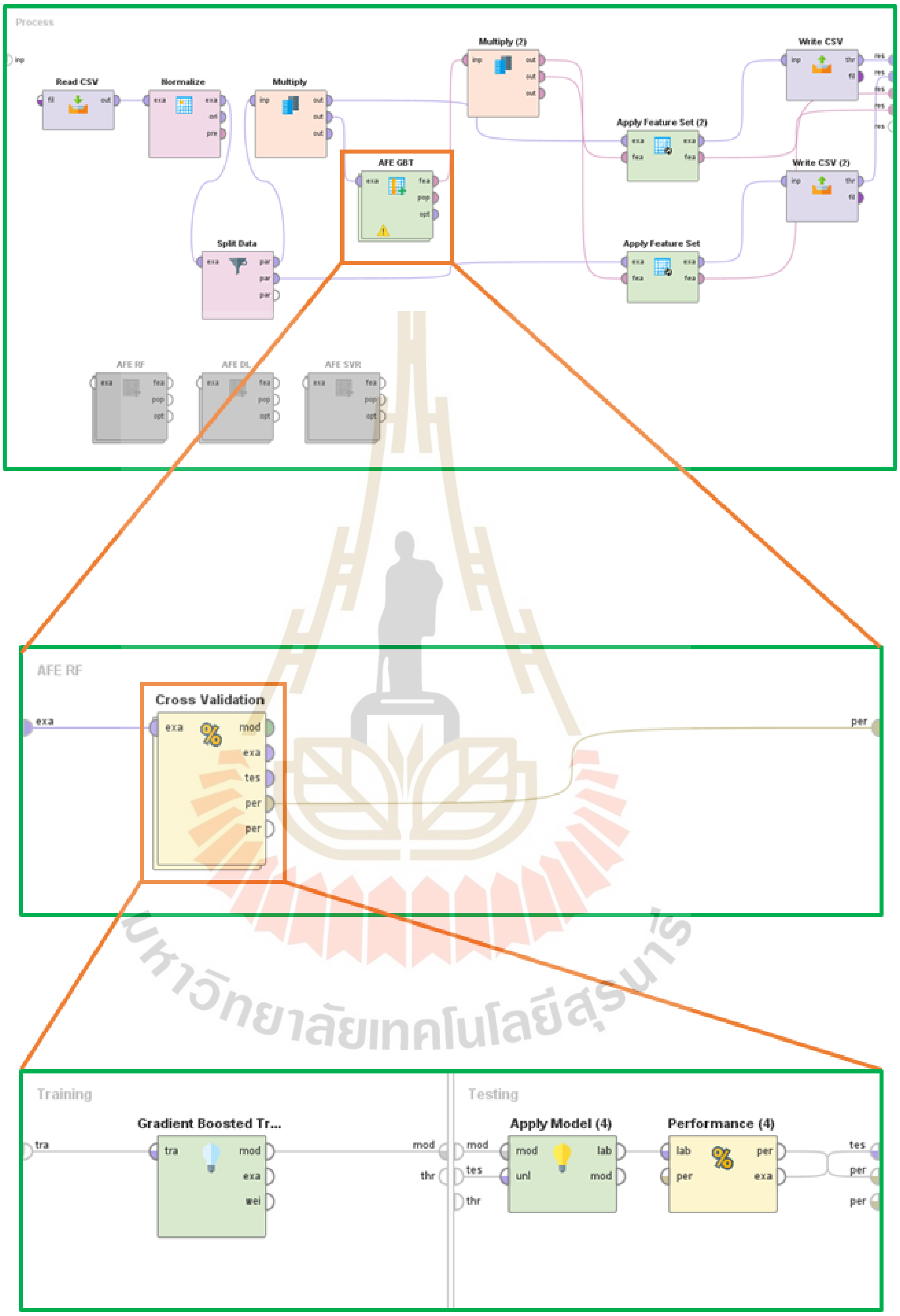

Figure A.4 Process of automatic feature engineering base on GBT.

APPENDIX B

OPTIMIZE PARAMETER FOR CONSTRUCT MODEL PREDICTION

B.1 Optimize parameter for construct model prediction

In this section, we present the process to optimize parameters of support vector regression, deep learning, random forest, and gradient boost tree for construct model prediction in Rapidminer Studio program which validation model by 10-fold cross-validation.

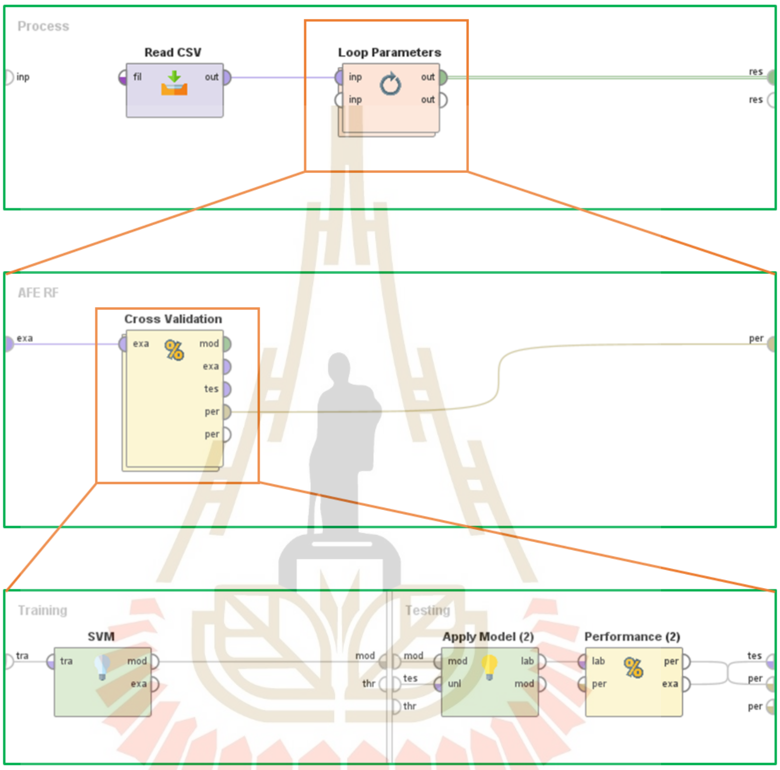

Figure B.1 Process of optimize parameter of SVR.

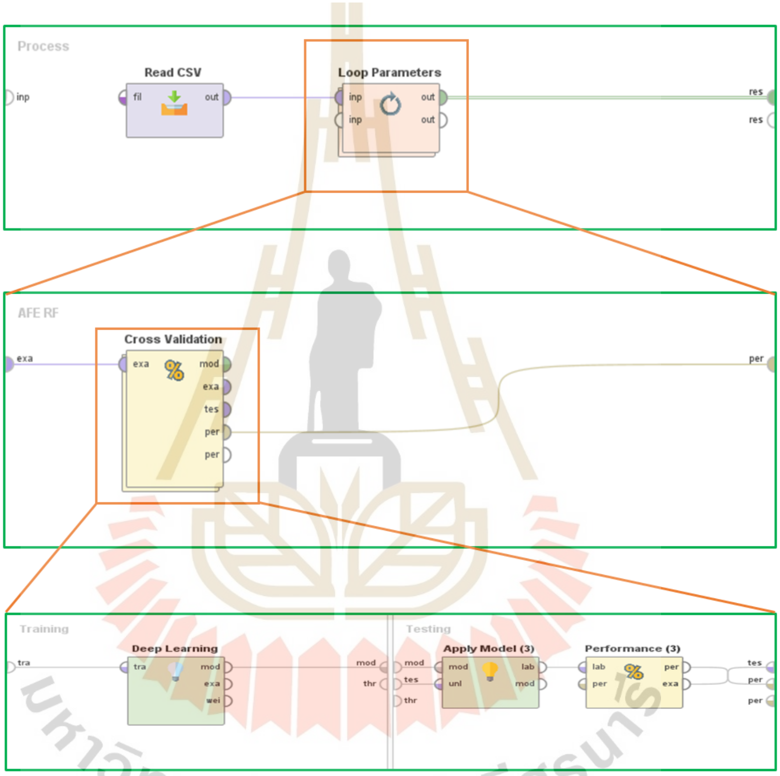

Figure B.2 Process of optimize parameter of DL.

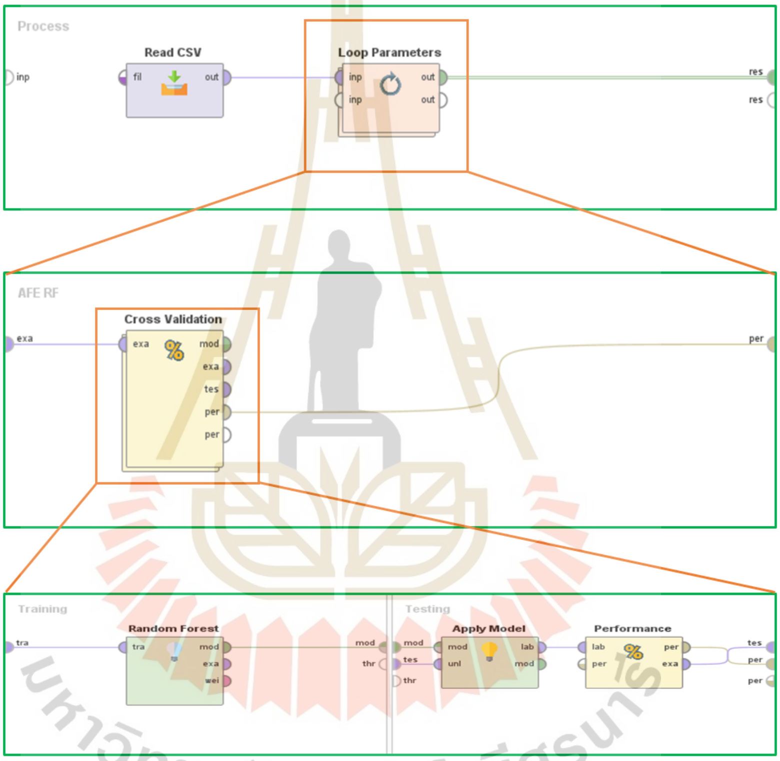

Figure B.3 Process of optimize parameter ofn RF.

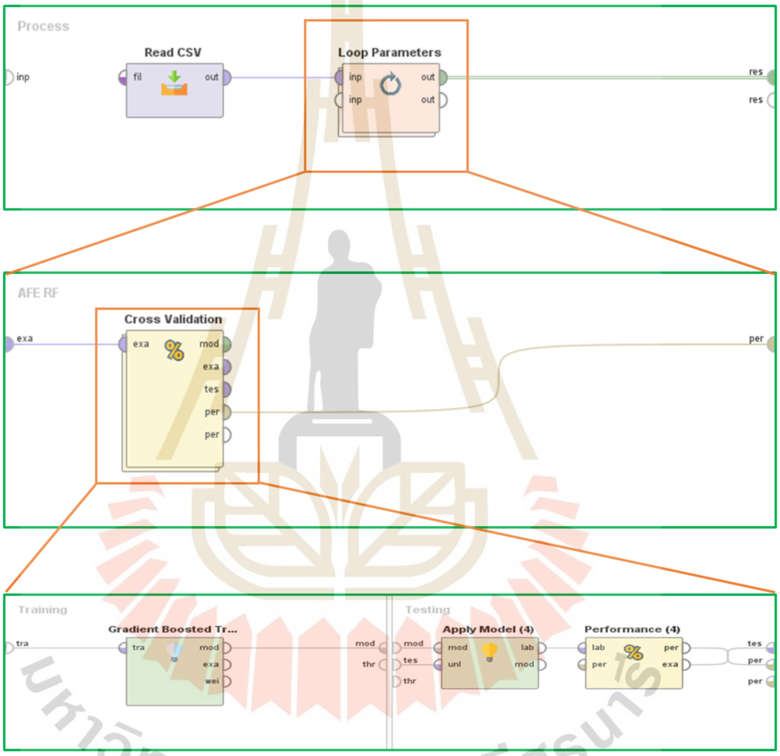

Figure B.4 Process of optimize parameter of GBT.

APPENDIX C

CONSTRUCTING MODEL PREDICTION

C.1 Constructing model prediction

In this section, we present the process to methodology for creating a Model for stock Price prediction consist of support vector regression, deep learning, random forest, and gradient boost tree for construct model prediction in Rapidminer Studio program which validation model by 10-fold cross-validation.

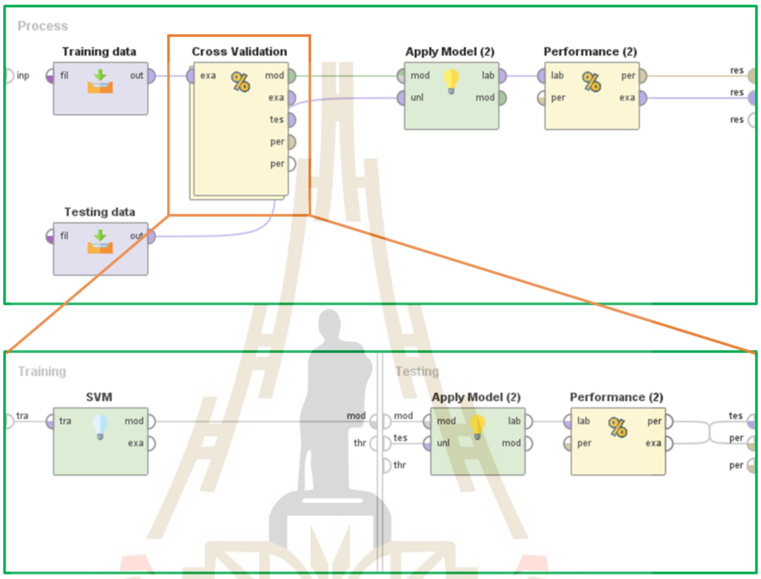

Figure C.1 Process for constructing an SVR model.

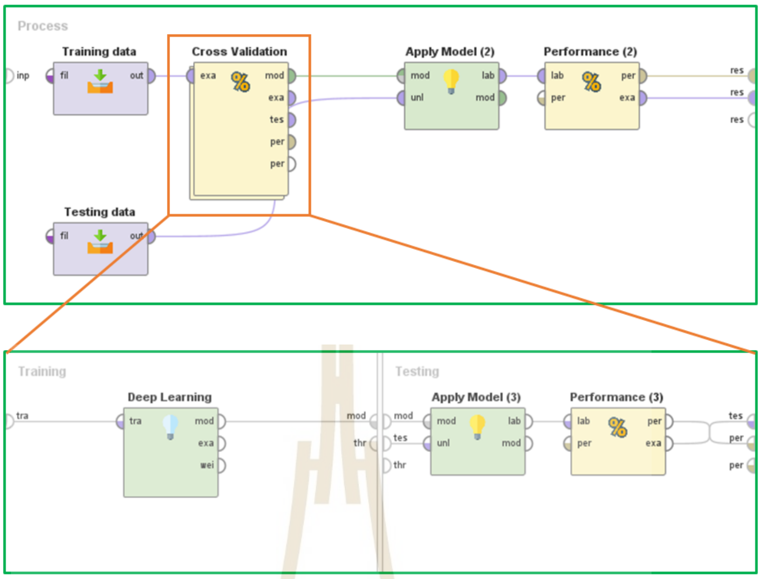

Figure C.2 Process for constructing a DL model.

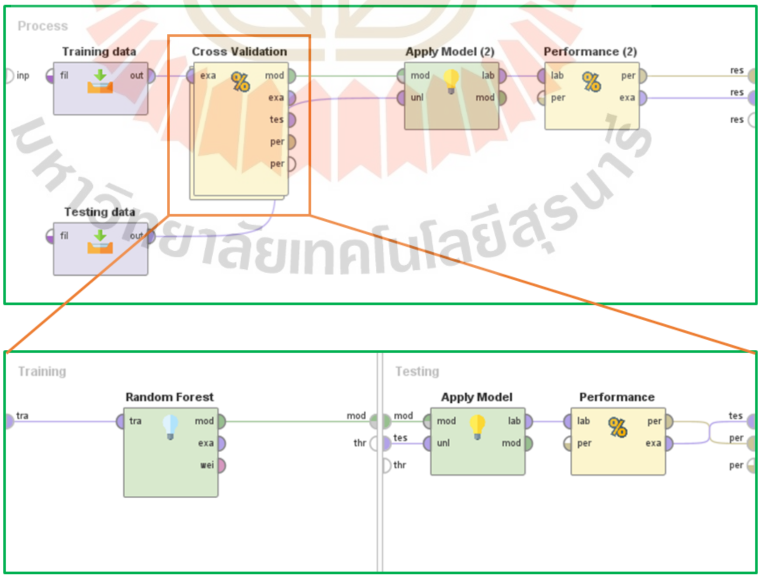

Figure C.3 Process for constructing a RF model.

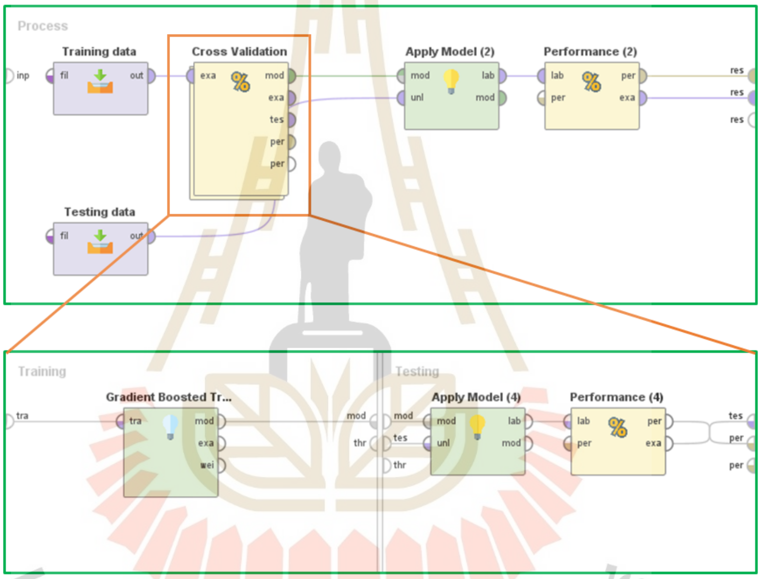

Figure C.4 Process for constructing a GBT model.

CURRICULUM VITAE

NAME : Ratchapon Pariyothai **GENDER :** Male

EDUCATION BACKGROUND:

- Bachelor of Science (Mathematics), Honors Program (First class honors), Suranaree University of Technology, Thailand, 2019.

SCHOLARSHIP:

- Development and Promotion of Science and Technology Talents Project (DPST) scholarship for graduate honor student of Suranaree University of Technology.

CONFERENCE:

- The 26th Annual Meeting in Mathematics 2022 (AMM 2022) and The 1st International Annual Meeting in Mathematics 2022, Suranaree University of Technology, May 18th - 20th, 2022.

EXPERIENCE:

- Working on U2T for BCG in Technopolis Suranaree University of Technology, 2022.