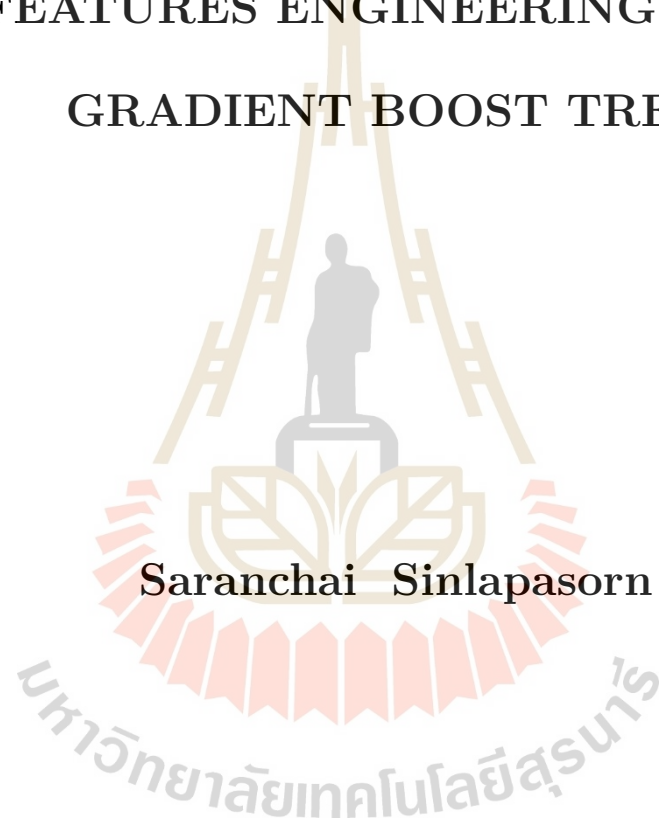


**MODELING TO PREDICT THE PATIENTS'
POSTOPERATIVE WOMAC SCORE AFTER
TOTAL KNEE REPLACEMENT BY
FEATURES ENGINEERING AND
GRADIENT BOOST TREE**



**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 2020

การสร้างตัวแบบที่ใช้ในการทำนายแบบประเมินข้อเข้าเสื่อม WOMAC
ของผู้ป่วยหลังการผ่าตัดเปลี่ยนข้อเข่าด้วยวิศวกรรมคุณลักษณะและ
เทคนิคเกรเดียนท์บูตทรี



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาคณิตศาสตร์ประยุกต์

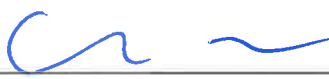
มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2563

**MODELING TO PREDICT THE PATIENTS'
POSTOPERATIVE WOMAC SCORE AFTER TOTAL
KNEE REPLACEMENT BY FEATURES
ENGINEERING AND GRADIENT BOOST TREE**


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.Eckart Schulz)


Chairperson



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



(Asst.Prof.Dr.Jessada Tanthanuch)
Member



(Assoc.Prof.Dr.Surattana Sungnul)

Member



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



(Assoc.Prof.Dr.Worawat Meevasana)
Dean of Institute of Science

ศรัณย์ชัย ศิลปสร : การสร้างตัวแบบที่ใช้ในการทำนายแบบประเมินข้อเข่าเสื่อม WOMAC ของผู้ป่วยหลังการผ่าตัดเปลี่ยนข้อเข่าด้วยวิศวกรรมคุณลักษณะและเทคนิคเกรเดียนท์บูตทรี (MODELING TO PREDICT THE PATIENTS' POSTOPERATIVE WOMAC SCORE AFTER TOTAL KNEE REPLACEMENT BY FEATURES ENGINEERING AND GRADIENT BOOST TREE) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.เบญจวรรณ โรจนดิษฐ์, 70 หน้า.

โรคข้อเข่าเสื่อม/แบบประเมินข้อเข่าเสื่อม WOMAC/วิศวกรรมคุณลักษณะ/เกรเดียนท์บูตทรี

งานวิจัยนี้มีจุดประสงค์เพื่อศึกษาปัจจัยที่มีอิทธิพลและสร้างตัวแบบในการทำนายคะแนนแบบประเมินข้อเข่าเสื่อม WOMAC ของผู้ป่วยหลังการผ่าตัดเปลี่ยนข้อเข่า สำหรับการหาปัจจัยที่มีอิทธิพลต่อการทำนายในงานวิจัยได้เลือกใช้เทคนิควิศวกรรมคุณลักษณะ โดยในขั้นตอนนี้ได้เลือกใช้เทคนิคเชิงเส้นวางนัยทั่วไป ซัพพอร์ตเวกเตอร์แมชชีน การเรียนรู้เชิงลึกและเกรเดียนท์บูตทรี จากนั้นนำปัจจัยที่ได้จากแต่ละเทคนิคไปสร้างตัวแบบในการทำนายคะแนนแบบประเมินข้อเข่าเสื่อม WOMAC ของผู้ป่วยหลังการผ่าตัดเปลี่ยนข้อเข่าด้วยเทคนิคเกรเดียนท์บูตทรี สำหรับโปรแกรมหลักที่ใช้ในการศึกษานี้คือโปรแกรม RapidMiner Studio รุ่น 9.9

ผลการศึกษาพบว่าการสร้างตัวแบบในการทำนายคะแนนแบบประเมินข้อเข่าเสื่อม WOMAC ของผู้ป่วยหลังการผ่าตัดเปลี่ยนข้อเข่าด้วยเทคนิคเกรเดียนท์บูตทรีและใช้คุณลักษณะที่ได้จากเทคนิควิศวกรรมคุณลักษณะในตัวแบบเกรเดียนท์บูตทรีมีประสิทธิภาพที่สุดโดยให้ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย ค่าคลาดเคลื่อนสัมบูรณ์และค่าคลาดเคลื่อนกำลังสองเป็น 5.311 ± 0.538 3.550 ± 0.376 และ 28.472 ± 5.881 ตามลำดับ

สาขาวิชาคณิตศาสตร์

ปีการศึกษา 2563

ลายมือชื่อนักศึกษา

ศรัณย์ชัย

ลายมือชื่ออาจารย์ที่ปรึกษา

Benjanan Radjanobol.

ACKNOWLEDGEMENTS

I would first like to express my sincere gratitude to my thesis supervisor Asst. Prof. Dr. Benjawan Rodjanadid for granting me the opportunity to do this research. She was nice and kind to guide, encourage and explain to me every time when I did not understand some content in this thesis work.

Furthermore, I would like to give a million thank for the professional support received from Asst. Prof. Dr. Jessada Tanthanuch, 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 Mr. Ratchapon Pariyothai, Mr. Jirakit Boonmunewai, Mr. Jakkrit Polrob and all of my friends at SUT for discussing and supporting my work.

In addition, I am grateful to SUT for a scholarship to study in the Master's degree program at SUT.

Finally, I am eternally grateful for encouragement from my family for believing in me, respect my way, and granting utmost encouragement to me on the days I felt desperate.

Sarachai Sinlapasorn

CONTENTS

	Page
ABSTRACT IN THAI	I
ABSTRACT IN ENGLISH	II
ACKNOWLEDGEMENTS	III
CONTENTS	IV
LIST OF TABLES	VIII
LIST OF FIGURES	X
CHAPTER	
I INTRODUCTION	1
1.1 Research objective	2
1.2 Scope and limitations	2
1.3 Research procedure	2
1.4 Results	3
II LITERATURE REVIEW	4
2.1 Questionnaire	4
2.1.1 WOMAC	4
2.1.2 KOOS	5
2.1.3 IKDC	5
2.1.4 OKS	5
2.2 Machine learning	6
2.2.1 Classification problem	6
2.2.2 Regression problem	6

CONTENTS (Continued)

	Page
2.3 Generalized linear model	7
2.4 Deep learning	7
2.5 Decision tree	8
2.5.1 Regression tree	8
2.6 Gradient boost machine	10
2.7 Support vector machine	11
2.7.1 Kernel	13
Linear kernel	14
Polynomial kernel	14
Gaussian radial basis function kernel	14
Anova	14
Epachnenikov	15
2.8 Features engineering	15
2.8.1 Feature selection	15
2.8.2 Feature extraction	15
2.8.3 Automatic feature engineering	16
2.9 K-fold cross validation	16
2.10 Performance evaluation	17
2.10.1 Root mean square error	17
2.10.2 Mean absolute deviation	18
2.10.3 Square error	18
2.10.4 Confusion matrix	18
2.10.5 Area under curve	19

CONTENTS (Continued)

	Page
2.11 Related researches	19
III RESEARCH METHODOLOGY	21
3.1 Data collection	22
3.2 Tool	22
3.3 Optimize parameters for feature engineering	23
3.4 Feature engineering	24
3.5 Creating a model by using the gradient boost tree	25
3.6 Predicting	25
3.7 Accuracy measurement of predicting model	25
IV RESULTS AND DISCUSSION	27
4.1 Data set	27
4.2 The novel attributes	35
4.2.1 Generalized linear model	35
4.2.2 Support vector machine	36
4.2.3 Deep learning	37
4.2.4 Gradient boost tree	38
4.3 Evaluation model	39
4.4 Performance of learning rate of gradient boost tree	42
V CONCLUSION AND RECOMMENDATION	46
REFERENCES	49
APPENDICES	

CONTENTS (Continued)

	Page
APPENDIX A QUESTIONNAIRE FOR OSTEOARTHRITIS	53
A.1 Western Ontario and McMaster Universities Arthritis Index	54
A.2 Knee injury and Osteoarthritis Outcome Score	55
A.3 International Knee Documentation Committee	59
APPENDIX B OPTIMIZE PARAMETERS FOR FEATURE ENGI- NEERING	61
B.1 Root mean square error of each model after op- timize the parameters	62
APPENDIX C FEATURE ENGINEERING	64
C.1 Optimized parameters for feature engineering	65
C.2 Feature engineering and optimize gradient boost tree	65
APPENDIX D OPTIMIZE GRADIENT BOOST TREE	67
D.1 RMSE of gradient boost tree with different at- tributes	68
CURRICULUM VITAE	70

LIST OF TABLES

Table		Page
3.1	Parameters of the Generalized Linear Model.	23
3.2	Parameters of Support Vector Machine.	23
3.3	Parameters of Deep Learning.	24
3.4	Parameters of Gradient Boost Tree.	24
3.5	Parameters of Gradient Boost Tree.	25
4.1	The attributes are used in this thesis.	28
4.2	Descriptive statistic of continuous data.	35
4.3	Features Engineering with Generalized Linear Model (FE+GLM). . .	36
4.4	Features Engineering with Support Vector Machine (FE+SVM). . .	37
4.5	Features Engineering with Deep Learning (FE+DL).	38
4.6	Features Engineering with Gradient Boost Tree (FE+GBT).	39
4.7	RMSE, MAD and SE of the models obtained.	40
4.8	Parameters Gradient Boost Tree.	43
4.9	RMSE, MAD and SE of the models obtained.	44
B.1	Generalized Linear Model.	62
B.2	Support Vector Machine.	62
B.3	Deep Learning.	63
B.4	Gradient Boost Tree.	63
D.1	Gradient Boost Tree with FE+GLM.	68
D.2	Gradient Boost Tree with FE+SVM.	68
D.3	Gradient Boost Tree with FE+DL.	69

LIST OF TABLES (Continued)

Table		Page
D.4	Gradient Boost Tree with FE+GBT.	69



LIST OF FIGURES

Figure		Page
2.1	Example of Decision Tree model.	8
2.2	Example of Regression Tree model.	9
2.3	The component of Gradient Boost Machine.	11
2.4	The Margin of hyperplane.	12
2.5	The penalty parameter.	12
2.6	k-fold Cross Validation method.	17
2.7	Confusion matrix.	18
3.1	Flow chart the process of the thesis.	21
4.1	Sex attribute distribution.	29
4.2	Age attribute distribution.	29
4.3	BMI attribute distribution.	30
4.4	Getting in or out of the car attribute distribution.	30
4.5	Lying on bed attribute distribution.	31
4.6	Meeting patient's expectations attribute distribution.	31
4.7	Patient ability to return to normal activity attribute distribution.	32
4.8	Sitting on the chair attribute distribution.	32
4.9	Walking on flat surface attribute distribution.	33
4.10	Walking without gait aid attribute distribution.	33
4.11	ROM after SX attribute distribution.	34
4.12	WOMAC score after surgery attribute distribution.	34
4.13	Predicting WOMAC Score by using FE+GLM.	40

LIST OF FIGURES (Continued)

Figure		Page
4.14	Predicting WOMAC Score by using FE+SVM.	41
4.15	Predicting WOMAC Score by using FE+DL.	41
4.16	Predicting WOMAC Score by using FE+GBT.	42
4.17	Tune Learning Rate in GBT.	42
4.18	Tune Learning Rate and Number of Trees in GBT.	43
4.19	Further Tune Learning Rate and Number of Trees in GBT.	44
A.1	Western Ontario and McMaster Universities Arthritis Index.	54
A.2	First page of Knee injury and Osteoarthritis Outcome Score.	55
A.3	Second page of Knee injury and Osteoarthritis Outcome Score.	56
A.4	Third page of Knee injury and Osteoarthritis Outcome Score.	57
A.5	Fourth page of Knee injury and Osteoarthritis Outcome Score.	58
A.6	First page of International Knee Documentation Committee.	59
A.7	Second page of International Knee Documentation Committee.	60
C.1	Overview process of optimized parameters.	65
C.2	Overview process of feature engineering and optimize gradient boost tree.	66

CHAPTER I

INTRODUCTION

Knee Osteoarthritis (KOA) is a common musculoskeletal disorder and the cause of disability in older adults which heavily affects a patient's daily life. The global prevalence of KOA was 22% or around 654 millions in individuals aged 40 and over in 2020 worldwide (Cui et al., 2020). Nowadays, 17.57% of Thailand's population is over 60 years old, divided into 9.82% women and 7.75% men (Department of Older Persons, 2020), and in the future these percentages may trend to increase.

At the present time, the diagnosis of knee osteoarthritis is done by several methodologies i.e., physical examination, clinical history, X-ray and MRI image if needed. For easy estimate of the level of knee osteoarthritis, some researchers have created questionnaires to evaluate osteoarthritis, one of which is the Western Ontario and McMaster Universities Arthritis Index (WOMAC) to evaluate Hip and Knee Osteoarthritis. WOMAC is usually used because it is easy to do an assessment and not waste time.

At this time, Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) are playing increasing roles in medicine for diagnosis or treatment plans. For example, Kuo-Ching Yuan et al. developed AI for early sepsis diagnosis in the intensive care unit, with an accuracy of 82% (Yuan et al., 2020). Most of the works of ML aim to predict or classify patients or evaluate patient's disease, e.g. Aleksei Tiulpin et al. predicted the potential need for total knee replacement using Gradient Boost Machine (GBM) and Convolutional Neural Network (CNN),

with an area under curve (AUC) accuracy of 0.79 (0.78-0.81) (Tiulpin et al., 2019). GBM and feature engineering are techniques that the data scientists usually use to improve efficiency of a model.

1.1 Research objective

The purpose of this research was to study the Gradient Boost Machine and Feature Engineering to construct the model to predict the patients' postoperative WOMAC score after total knee replacement.

1.2 Scope and limitations

The Techniques for solving the regression problem in this study consist of the Gradient boost Machine and Features Engineering using the data obtained from Asst. Prof. Lt. Col. MD.Bura Sindhupakorn who has collected information of patient's postoperative knee replacement for 12 years.

1.3 Research procedure

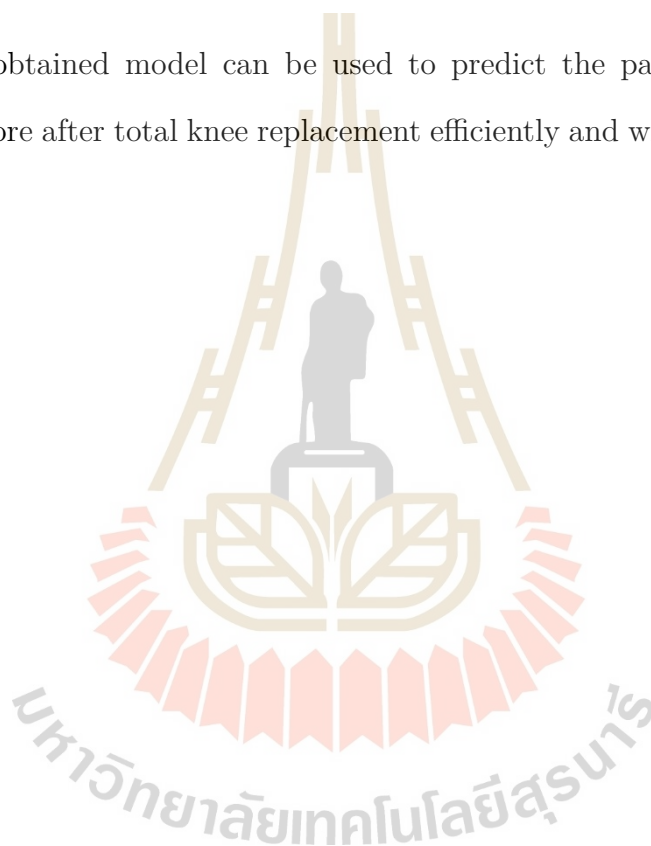
The research work proceeded as follows:

1. Study Features Engineering techniques, namely automatic feature engineering.
2. Study regression algorithm in data mining, namely decision tree, regression tree, and gradient boost tree.
3. Study the program Rapid Miner Studio Version 9.0.
4. Understand and prepare the data set of patients' postoperative knee replacement.

5. Construct the model for predict to postoperative total knee replacement patient's WOMAC score.
6. Measure the accuracy of the predicting model by using root mean square error, mean absolute deviation, and square error.

1.4 Results

Our obtained model can be used to predict the patients' postoperative WOMAC score after total knee replacement efficiently and with good accuracy.



CHAPTER II

LITERATURE REVIEW

In this section, the knowledge of basic mathematics and machine learning related with WOMAC is presented. The contents consists of the main idea of gradient boost machine and automatic feature engineering which are important in these studies.

2.1 Questionnaire

Nowadays, for comfortable to follow up symptom and diagnose osteoarthritis, there exists the scientist develop and create the questionnaire to evaluate osteoarthritis which has several questionnaires such as Western Ontario and McMaster Universities Arthritis Index (WOMAC), Knee injury and Osteoarthritis Outcome Score (KOOS), International Knee Documentation Committee (IKDC), and Oxford Knee Score (OKS).

2.1.1 WOMAC

WOMAC was developed in 1982, it is available in over 65 languages and has been linguistically validated, and is a self-administered questionnaire that takes approximately 12 minutes to complete to evaluate Hip and Knee Osteoarthritis and it has been used extensively in research studies. The WOMAC consisting of 24 items divided into three subsections which are Pain (5 items), Stiffness (2 items), and Physical Function (17 items), which fill out the form during 48 hours where the total WOMAC score is between 0 to 96 such that 0 being the best and

96 being the worst.

2.1.2 KOOS

KOOS was developed in the 1990s and used in the patient between 13 to 79 years old, which consists five subscales such as pain, other Symptoms, Function in daily living, Function in sport and recreation, and knee-related Quality of life which the score is a percentage score from 0 to 100, 0 and 100 represents extreme and no problems, respectively.

2.1.3 IKDC

IKDC was developed in 1987 and separable into 3 categories such as symptoms, sports activity, and knee function. The score is obtained by summation of each category which has a range between 0-100. There exist a modified version of the IKDC is The Pediatric International Knee Documentation Committee (Pedi-
IKDC) that was developed for use with children and adolescents between the ages of 10 and 18 years old.

2.1.4 OKS

OKS was developed in 1998 and used to measure pain and function after total knee replacement for use with individuals undergoing total knee replacement surgery, however, but it can be used to measure outcomes in pharmacological treatments, after surgery. The questionnaire consists of 12 questions about the activities of daily living which have been affected by pain for 4 weeks before doing the questionnaire.

2.2 Machine learning

Machine learning (ML) is the process by which a computer program adjusts parameters within the program from actual data, with the aim to detect relations within the data, so that. If there is new, not previously seen input into the computer, then the system computer can predict the output. Nowadays, ML is used in several fields such as medicine, marketing, sport and industry. ML is separated into 3 categories which are Supervised Learning, Unsupervised Learning and Reinforcement.

Supervised Learning is the most common method used in machine learning where the depended or predicted variables (label) depend on the data. Supervised learning can solve regression and classification problems.

2.2.1 Classification problem

The classification problem is the study relations between independent and dependent variables where the dependent variable is categorical, for example, to predict whether the weather is rainy, sunny, or cloudy. Classification models include logistic regression, random forest, decision tree, gradient boost tree, and Naive Bayes.

2.2.2 Regression problem

The regression problem is the study of relations between independent and dependent variables where the dependent variable takes discrete or continuous values, for example, to predict the number of students who study in the School of Mathematics in 2022. The regression problem can be solved by several methods e.g. linear regression model, gradient boost tree (GBT), or decision tree to

approximate the dependent value.

2.3 Generalized linear model

A generalized linear model (GLM) is a statistical modeling technique formulated by John Nelder and Robert Wedderburn in 1972. The model can build a linear relationship between the independent and dependent variables, even though the underlying data relationship is not linear. The model is popular for analyzing of data since this model can explain the influence of independent variables to a dependent variable. The model consists of 3 components: random component, systematic component, and link function. The solver of the generalized linear model is used to solve a certain optimization problem with an objective function that has several methods for using such as Iteratively reweighted least squares (IRLSM), Limited-memory BFGS (LBFGS), or Coordinate descent.

2.4 Deep learning

Deep learning is a sub-field of machine learning where the algorithms are inspired by the construction of neurons in humans called artificial neural network. The components of an artificial neural network consist of nodes (also called neurons), weights between nodes, layers and activation functions. The output of an artificial neural network is defined by the activation function which can often be categorized as binary function, linear activation function and non-linear activation function.

2.5 Decision tree

Decision tree is the most popular method for supervised learning for solving regression and classification problems. Since a decision tree can handle noisy data and many independent variables and which uses If-Else rule, a decision tree is easy to interpret.

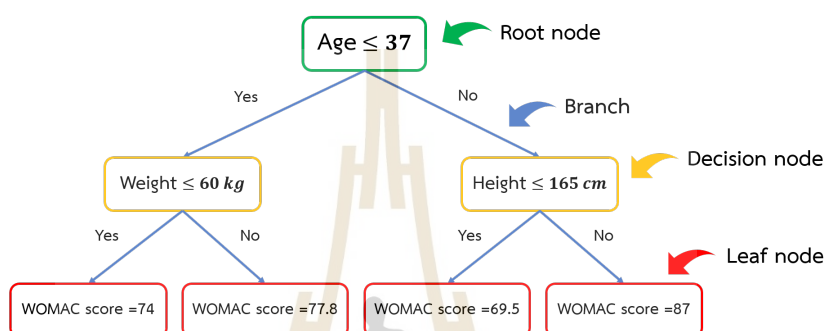


Figure 2.1 Example of Decision Tree model.

Figure 2.1 shows an example of a decision tree and shows the components of a decision tree such that each node contains a data attribute. The top node is called the root node and each branch represents the outcome of the node. A last node is called leaf node and the nodes between the root node and a leaf node are called decision nodes. The output of the model is defined in the form of a leaf node.

2.5.1 Regression tree

Regression Tree is one type of decision tree developed to estimate a continuous target 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.

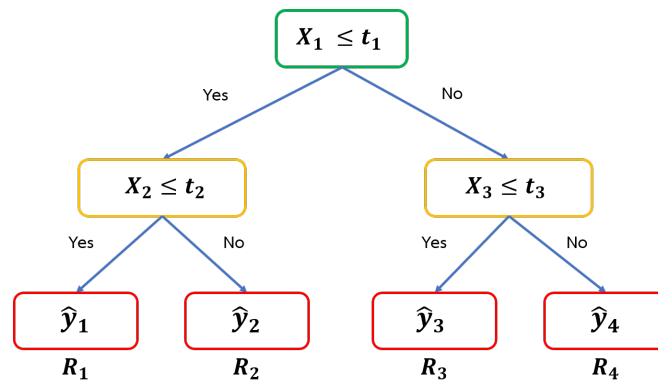


Figure 2.2 Example of Regression Tree model.

Let y_1, y_2, \dots, y_n be response variables depending on X_1, X_2, \dots, X_n that are predictor variables and implied that divided predictor space is the set of possible value for X_1, X_2, \dots, X_p into J -districts and non-overlapping regions, R_1, R_2, \dots, 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, \dots, R_J that minimize the Residual Sum of Squares (RSS), given by

$$RSS = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \quad (2.1)$$

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

$$\sum_{i: x_i \in R_1(j, s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j, s)} (y_i - \hat{y}_{R_2})^2, \quad (2.2)$$

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 (Gareth et al, 2017).

2.6 Gradient boost machine

Gradient Boost Machine (GBM) is one of the techniques of machine learning for classification and regression where by the model is an ensemble of several regression or classification tree models. The models are built sequentially and the error of the previous tree model determines the next tree model. The algorithm of GBM (Natekin and Knoll, 2013; Josh Starmer, 2019) is

Algorithm 1 Friedman's Gradient Boost algorithm.

Input:

- 1: Data input $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
- 2: Number of iterations M .
- 3: Differentiable **Loss function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma)$

Step 2: for $m = 1$ to M :

- 4: Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ where $F(x_i)$ is a predicted variable.
- 5: Fit a regression tree to the r_{im} values and create terminal regions $R_{j,m}$ for $j = 1, 2, \dots, J_m$
- 6: for $j = 1, 2, \dots, J_m$ compute $\gamma_{j,m} = \arg \min_{\gamma} \sum_{x_i \in R_{j,m}} L(y_i, F_{m-1}(x_i) + \gamma)$
- 7: Update $F_m(x) = F_{m-1} + v \sum_{j=1}^{J_m} \gamma_{j,m} I(x \in R_{j,m})$ where v and I are learning rate and indicator function, respectively.

Step 3: Output $F_m(x)$

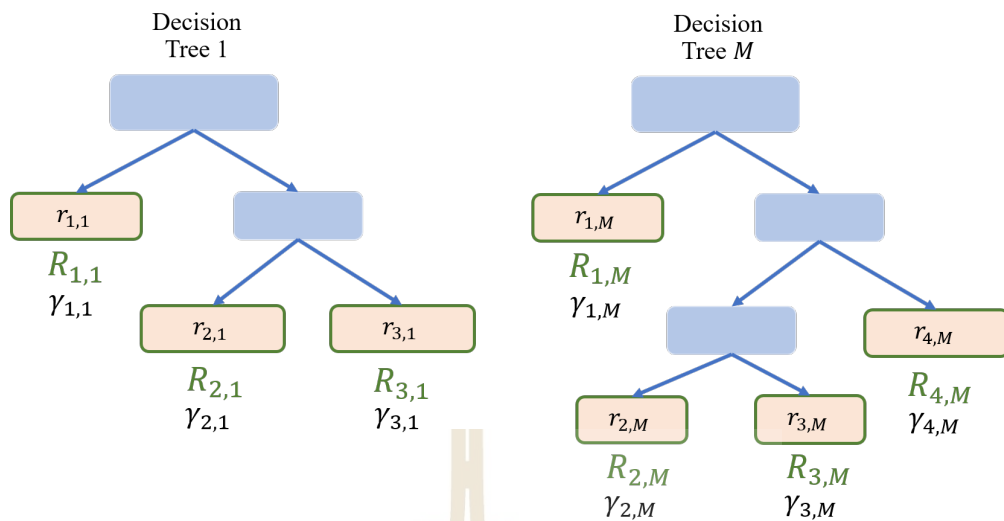


Figure 2.3 The component of Gradient Boost Machine.

2.7 Support vector machine

Support Vector Machine (SVM) is a supervised learning model for solving classification and regression problems. Support vector machines are suitable models for medium size of data with many attributes. To explain SVM, the simplest case of a 2 class problem where the two classes are linearly separable is presented. Let the data set D be given $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$ where $x_i \in \mathbb{R}^n$ for $i = 1, 2, \dots, l$ is the vector of attribute of training data set with label y_i for $i = 1, 2, \dots, l$ such that y_i can take one of two values, either -1 or 1 .

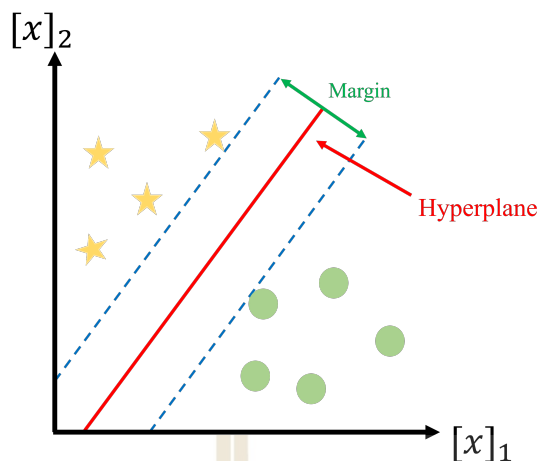


Figure 2.4 The Margin of hyperplane.

The ideal algorithm of a support vector machine is to find the maximum margin hyperplane as shown in Figure 2.4. The problem of optimal classification hyperplane is transformed into following optimization problem by

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i, \quad (2.3)$$

$$y_i(w \cdot x + b) - 1 + \xi_i \geq 0, \quad (2.4)$$

$$\xi_i \geq 0, i = 1, 2, \dots, l, \quad (2.5)$$

where $w \in \mathbb{R}^n, b \in \mathbb{R}$. In this model the data need not be totally linearly separable, C is the penalty parameter such that $C > 0$ and controls the degree of penalty for misclassification samples.

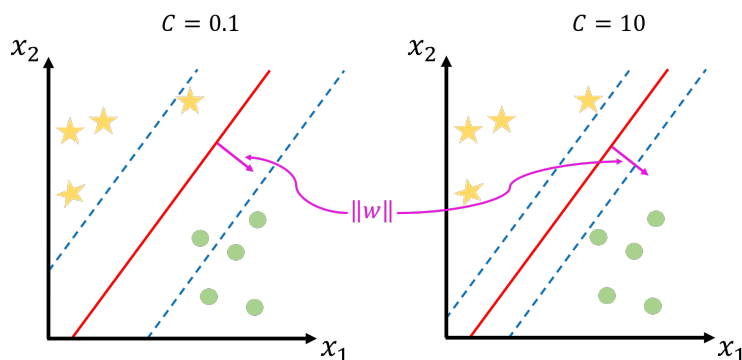


Figure 2.5 The penalty parameter.

The corresponding Lagrangian function is

$$L(w, b, \xi, \alpha) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i - \sum_{i=1}^l \alpha_i (y_i (w \cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^l \beta_i \xi_i, \quad (2.6)$$

where α_i, β_i are Lagrangian multipliers such that $\alpha_i, \beta_i > 0$ and (\cdot) denotes the inner product. We obtain the dual problem

$$\min \quad \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^l \alpha_i, \quad (2.7)$$

$$s.t. \quad \sum_{i=1}^l y_i \alpha_i = 0, \quad (2.8)$$

$$0 \leq \alpha_i \leq C, i = 1, 2, \dots, l. \quad (2.9)$$

2.7.1 Kernel

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 classified by SVM. 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 to the Hilbert space

$$\begin{aligned} \phi : R^n &\rightarrow \mathbb{H}, \\ x &\mapsto \phi(x), \end{aligned}$$

such that

$$K(x, x') = (\phi(x) \cdot \phi(x')), \quad (2.10)$$

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

Several types of kernel are typically used, for example:

Linear kernel

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:

$$K(x, x') = x^T x' + c. \quad (2.11)$$

Polynomial kernel

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

$$K(x, x') = ((x \cdot x') + 1)^d, \quad (2.12)$$

where d is a positive integer.

Gaussian radial basis function kernel

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

$$K(x, x') = \exp(-\gamma \|x - x'\|^2), \quad (2.13)$$

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

Anova

An anova kernel is a kernel function which function is similarly Gaussian kernel, the equation has a formula as following:

$$K(x, x') = \sum_{k=1}^n \exp(-\gamma(x^k - x'^k)^2)^d. \quad (2.14)$$

Epachnenikov

An epanechnikov kernel is another type of kernel aforementioned before, it is a kernel density estimation and used to estimate the probability density function of a random variable which the function is $\frac{3}{4}(1 - u^2)$ for $|u| \leq 1$ and the function is equal to zero where u outside that range.

2.8 Features engineering

Feature engineering (FE) is a methodology to adjust and transform attributes to be suitable with machine learning. This technique is an important point of machine learning because only those attributes are used that extremely affect a model, thus reducing the time to build and execute the model (Zheng and Casari, 2018).

2.8.1 Feature selection

Feature selection is a technique to decrease the number of attributes and select attributes that most influence the dependent variables. This technique is important because the quantity of attributes affects the efficiency of the model. Example of feature selection method are the filter method, wrapper method and embedded method.

2.8.2 Feature extraction

This technique is used to reduce the quantity of attributes by generating new attributes that are more significant than the attributes in data. This technique is useful when the data has many attributes and one needs to reduce the quantity of attributes without losing important attributes. The most useful feature extraction

is Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA).

2.8.3 Automatic feature engineering

Automatic feature engineering (AFE) optimizes the process of building a model, increases accuracy and reduces complexity of the model. The basic algorithm behind the automated feature engineering method is called Deep Feature Synthesis (DFS).

DFS was invented by James Max Kanter and Kalyan Veeramachaneni which are researchers at Massachusetts Institute of Technology in 2014. The ideal of AFE is to select and generate new attributes which are more effective for a model than the model build by raw attributes. The new attributes are generated by mathematical operations from the raw attributes (Kanter and Veeramachaneni, 2015).

2.9 K-fold cross validation

The data set D is divided into k partitions (or fold) then $k - 1$ partitions are used for training and k^{th} is used for testing, this is repeated k times, with each partition used for testing only once.

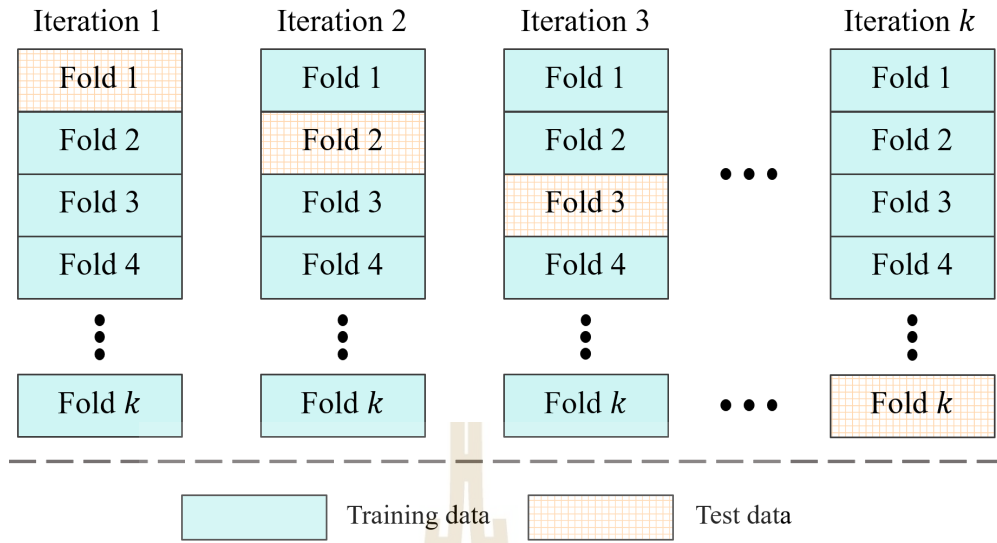


Figure 2.6 k-fold Cross Validation method.

2.10 Performance evaluation

In this study, solving regression problem, let

$$e_i = Y_i - \hat{Y}_i, i = 1, \dots, n, \quad (2.15)$$

where Y_i is an observed value, \hat{Y}_i is a predicted value and n is a number of data. In this thesis studies in regression problem, thus we considerate three measurement such as root mean square error, mean absolute deviation, and square error, and for classification problem usually use area under curve (AUC) and average precision (AP) for measurement the performance of model.

2.10.1 Root mean square error

Root mean square error (RMSE) is the square root of summation of square error divided by number of data

$$RMSE = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n}}. \quad (2.16)$$

2.10.2 Mean absolute deviation

Mean absolute deviation (MAD) is summation of absolute error divide by number of data

$$MAD = \frac{\sum_{i=1}^n |e_i|}{n}. \quad (2.17)$$

2.10.3 Square error

Square error (SE) is summation of square error

$$SE = \sum_{i=1}^n e_i^2. \quad (2.18)$$

2.10.4 Confusion matrix

The confusion matrix is used to measurement the performance of classification problem, the confusion matrix for binary classifier has the format as following

		Actual Values	
		True	False
Predicted Values	True	True Positive: TP	False Positive: FP
	False	False Negative: FN	True Negative: TN

Figure 2.7 Confusion matrix.

where TP is the case predicted yes and actual value was yes, TN is the case predicted no and actual value was no, FP is the case of predicted yes and actual values was no and FN is the case of predicted no and actual value was yes. Precision and recall can be calculate as:

$$Precision = \frac{TP}{TP + FP}, \quad (2.19)$$

$$Recall = \frac{TP}{TP + FN}. \quad (2.20)$$

Average precision is a measure that summation of recall and precision for achieved at each threshold, with the increase in recall from the previous threshold used as the weight:

$$AP = \sum_n (R_n - R_{n-1})P_n, \quad (2.21)$$

where R_n and P_n are recall and precision at n^{th} threshold, respectively.

2.10.5 Area under curve

The area under a receiver operating characteristic (ROC) curve, abbreviated as AUC is one of the tools to measures the performance of a binary classifier, such that the value of AUC is within 0.5 to 1 where the minimum and maximum value represents the performance of classifier as worst and perfection classifier, respectively.

2.11 Related researches

Kuo-Ching Yuan et al. (2020) developed AI for for early sepsis diagnosis in the intensive care unit established with pre-selected features and XGBoost and obtained an accuracy of 82%.

Aleksei Tiulpin et al. (2019) constructed the model using GBM cooperate with CNN to predict the increase of current KL-grade (KL-grade is a criterion for evaluating symptom of knee osteoarthritis which the level of the patient depends on the physician) and Potential need for total knee replacement within the next 7 years after baseline examination. The model yielded AUC of 0.79 (0.78-0.81) and average precision (AP) of 0.68 (0.66-0.70).

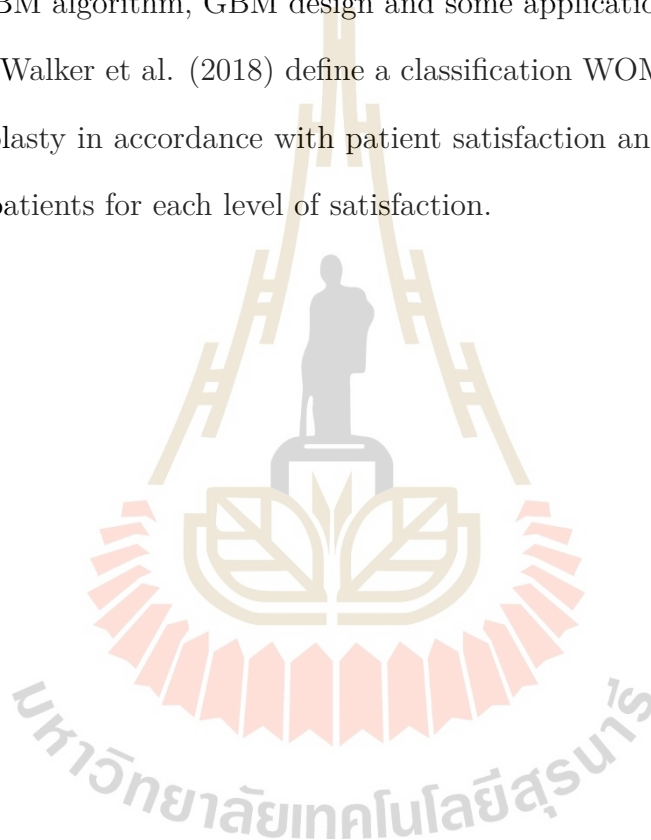
Christos Kokkotis et al. (2020) presented a review research to introduce the reader to key direction in machine learning for diagnosis and predictions of

knee osteoarthritis.

Afshin Jamshidi et al. (2019) informed guideline to diagnosis knee osteoarthritis early step of disease, collect factors from other researcher for knee osteoarthritis work and give examples of supervised learning algorithms for disease prediction and classification model.

Alexey Natekin and Alois Knoll (2013) introduced the Gradient Boost Machine i.e., GBM algorithm, GBM design and some application work with GBM.

Lucy Walker et al. (2018) define a classification WOMAC score after total knee arthroplasty in accordance with patient satisfaction and describe the demographics of patients for each level of satisfaction.



CHAPTER III

RESEARCH METHODOLOGY

The purpose of this chapter is shown process of this thesis which consists as following

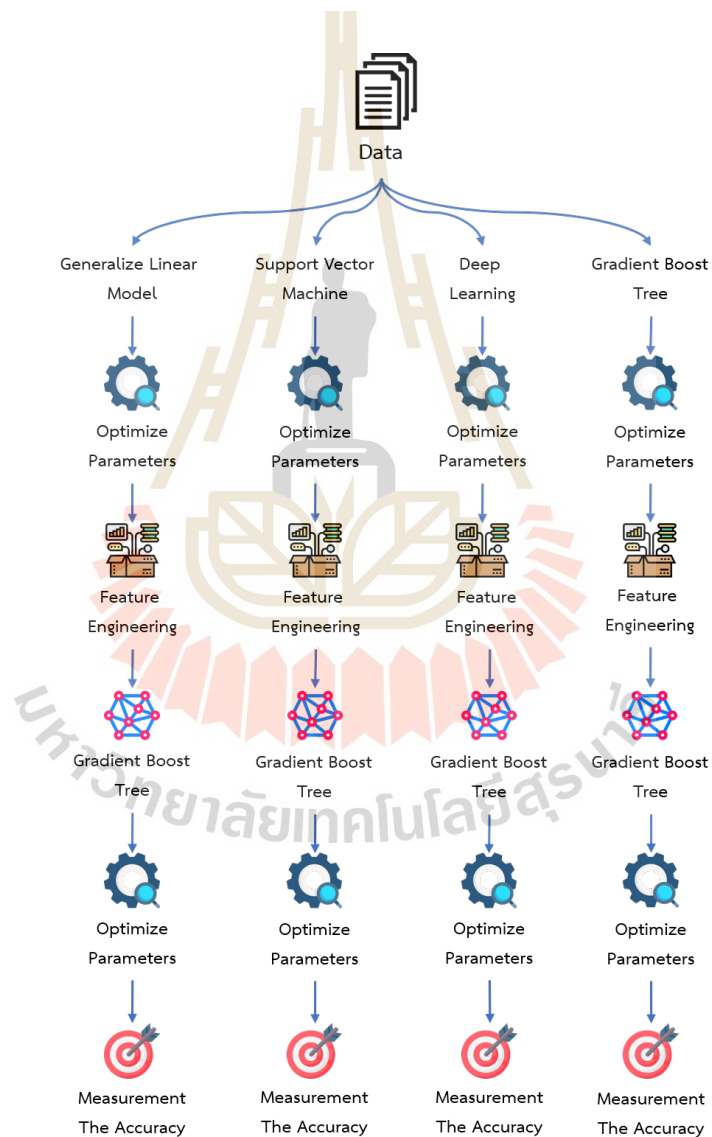


Figure 3.1 Flow chart the process of the thesis.

3.1 Data collection

The research used the questionnaire collected from the patient's post-operative total knee replacement, which Asst. Prof. Lt. Col. Bura Sindhupakorn at Suranaree university of Technology hospital has collected from 2007 to 2018. The data has 1250 instances and 21 attributes. In this research, we extracted 12 attributes and the dependent variable is the WOMAC score after surgery.

The attributes was used in this research as following

1. Sex,
2. Age,
3. Body mass index ("BMI"),
4. Flexibility of the knee after operation ("ROM after SX"),
5. Pain when lying with the back down on the bed ("Lying on bed"),
6. Pain when sitting on chair ("Sitting on chair"),
7. Number of the month to walk without a gait aid ("Walking without gait aid"),
8. Pain when walking on flat surface ("Walking on flat surface"),
9. Meeting patient's expectations,
10. Pain when getting in or out of the car ("Getting in or out of the car"),
11. Patient ability to return to normal activity,
12. WOMAC score after surgery.

3.2 Tool

In this work, we used the Rapidminer Studio version 9.9 (Education license) running on Microsoft Windows 10 operation system. Rapidminer is a software tool for data and text mining. It gathers the total data from data preparation to machine learning to predictive model deployment.

3.3 Optimize parameters for feature engineering

This step is used to optimize parameters of the generalized linear model (GLM), support vector machine (Philipp et al., 2019), deep learning (Koutsoukas et al., 2017), and gradient boost tree (Bentéjac et al., 2020) to set the parameters in feature engineering process for the purpose of selecting and generating the new attributes that are suitable with each model. The parameters of each model are shown in Tables 3.1 - 3.4. Some of the parameters in Table 3.2 apply to specific kernel only.

Table 3.1 Parameters of the Generalized Linear Model.

Parameter	Value
Family	Gaussian, Poisson, Gamma, Tweedie, Negativebinomial
Solver	IRLSM, LBFGS, Coordinate descent, Coordinate descent naive
Link function	Identity, Inverse, Logit

Table 3.2 Parameters of Support Vector Machine.

Parameter	Value
Kernel gamma	0, 1, 2, ..., 50
Kernel sigma1	0, 1, 2, 3, 4, 5
Degree	0,1,2,...,10
Penalty	-1, 0, 1, ..., 100
Epsilon	0.0001, 0.001, 0.01, 0.1, 0, 1, 2, 3
Kernel	Dot, Radial, Anova, Epachnenikov

Table 3.3 Parameters of Deep Learning.

Parameter	Value
Activation function	Rectifier, ExpRectifier, Tanh, Maxout
Loss function	Absolute, Quadratic, Quantile
Learning rate	10^{-7} , 10^{-6} , 10^{-5} , ..., 10^{-1}

Table 3.4 Parameters of Gradient Boost Tree.

Parameter	Value
Number of trees	100, 200, 300, 400, 500
Maximal depth	1, 2, 3, ..., 30
Learning rate	0.0001, 0.001, 0.01, 0.1

3.4 Feature engineering

This step is used to select and generate attributes by automatic feature engineering with 4 techniques such as generalized linear model, support vector machine, deep learning, and gradient boost tree. The parameters that were set in GLM are family, solver, and link function. The parameters that were set in SVM are kernel gamma, kernel sigma1, degree, penalty, epsilon, and kernel. The parameters that were set in deep learning are activation function, loss function, and learning rate. The parameters that were set in GBT is number of tree, maximal depth, and learning rate.

3.5 Creating a model by using the gradient boost tree

In this step we create 4 models by gradient boost tree; each model will use different attributes and each attribute comes from feature engineering on 4 techniques as described in section 3.4. The validation of each model used k -fold cross-validation with $k = 10$ and after that optimize every model which the parameters are used to optimize as following (Bentéjac et al., 2020).

Table 3.5 Parameters of Gradient Boost Tree.

Parameter	Value
Number of tree(NT)	100, 200, 300, 400, 500
Maximal of depth(MD)	1, 2, 3,..., 30
Learning rate(LR)	0.0001, 0.001, 0.01, 0.1

3.6 Predicting

The obtained model is used to predict the postoperative total knee replacement patient's WOMAC score.

3.7 Accuracy measurement of predicting model

Evaluation metrics for regression problems have many categories, e.g. mean absolute error, mean squared error, root mean squared error, root mean squared log error, adjusted R-squared and R-squared wherewith the evaluation metrics of this work are as follows:

- Root mean square error (RMSE)
- Mean absolute deviation (MAD)

- Square error (SE)



CHAPTER IV

RESULTS AND DISCUSSION

This chapter presents the results from processing using the Research Methodology in chapter III. The purpose is to show the computation results from Rapidminer Studio version 9.9 (Education license) i.e. the group attribute from feature engineering from generalized linear model, support vector machine, deep learning and gradient boost tree, the result of gradient boost tree with difference group attributes and performance learning rate parameter of gradient boost tree.

4.1 Data set

The data set has 1,250 instances and uses 12 attributes; the meaning of each attribute is shown in Table 4.1, while the distribution each attribute is shown in Figures 4.1-4.12

มหาวิทยาลัยเทคโนโลยีสุรนารี

Table 4.1 The attributes are used in this thesis.

Attribute	Type	Description
Sex	Independent	1:Male, 2:Female
Age	Independent	Integer (47-85)
BMI	Independent	Real numbers (18.91-41.02)
Walking without gait aid	Independent	Real numbers (0.46-24)
Getting in or out of the car	Independent	1:Uncertain, 2:Unsatisfied, 3:Satisfied, 4:Very satisfied
Lying on bed	Independent	1:Uncertain, 2:Unsatisfied, 3:Satisfied, 4:Very satisfied
Meeting patient's expectations	Independent	1:Uncertain, 2:Unsatisfied, 3:Satisfied, 4:Very satisfied
Patient ability to return to normal activity	Independent	1:Uncertain, 2:Unsatisfied, 3:Satisfied, 4:Very satisfied
Sitting on chair	Independent	1:Uncertain, 2:Unsatisfied, 3:Satisfied, 4:Very satisfied
Walking on flat surface	Independent	1:Uncertain, 2:Unsatisfied, 3:Satisfied, 4:Very satisfied
ROM after SX	Independent	1:0°-45°, 2:46°-90°, 3:91°-135°, 4:more than 135°
WOMAC score after surgery	Dependent	Score (15.2-100)

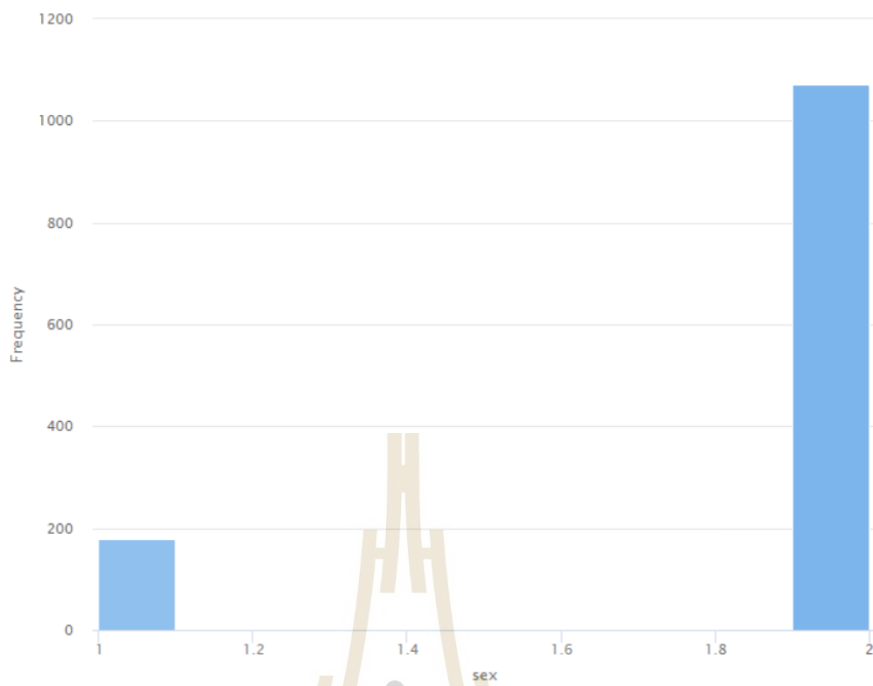


Figure 4.1 Sex attribute distribution.

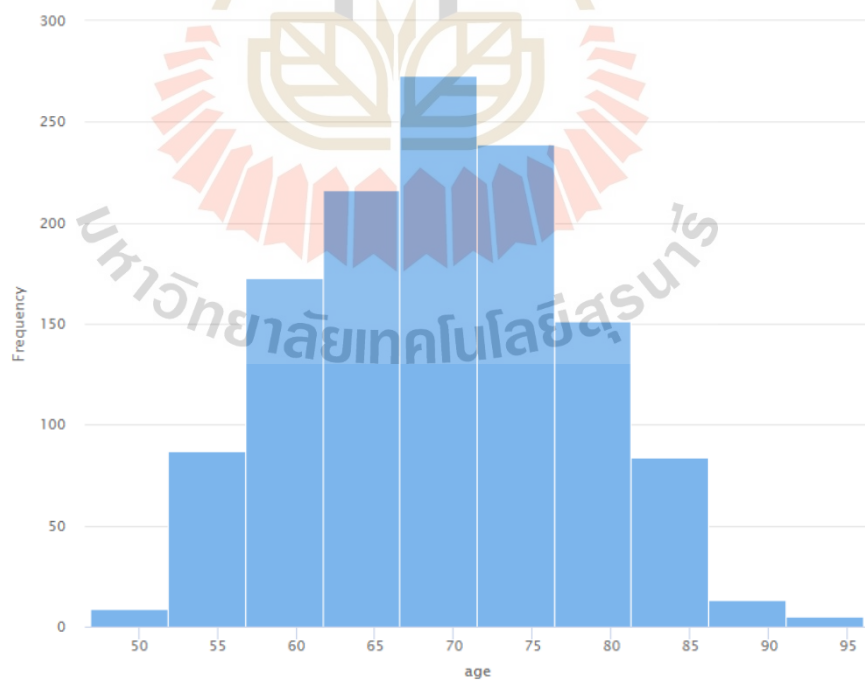


Figure 4.2 Age attribute distribution.

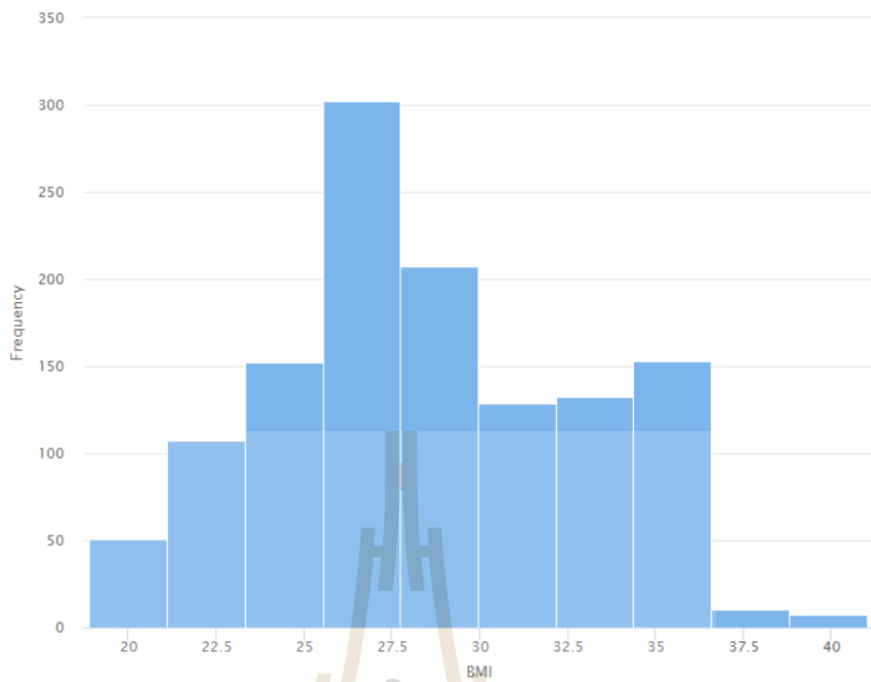


Figure 4.3 BMI attribute distribution.

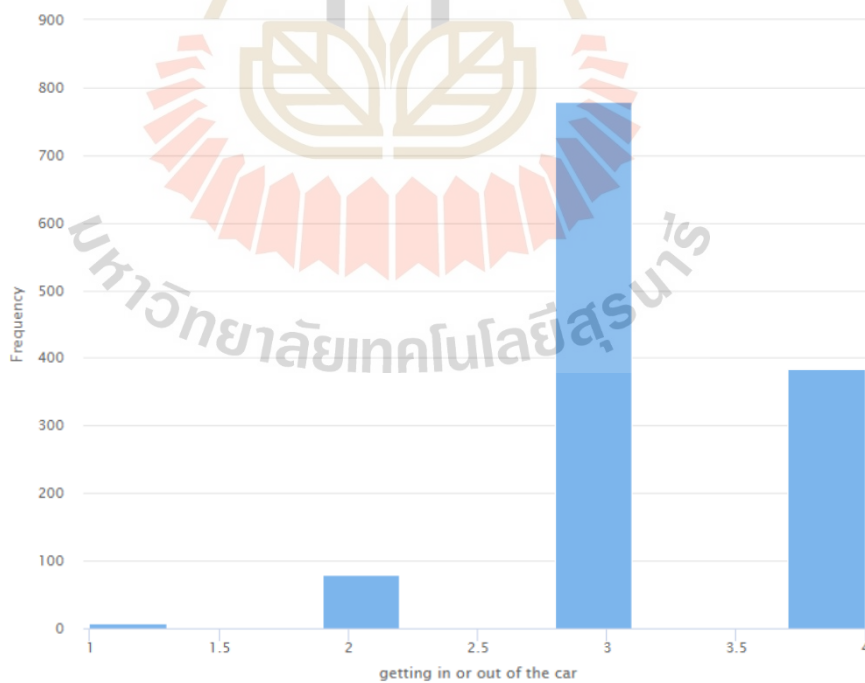


Figure 4.4 Getting in or out of the car attribute distribution.

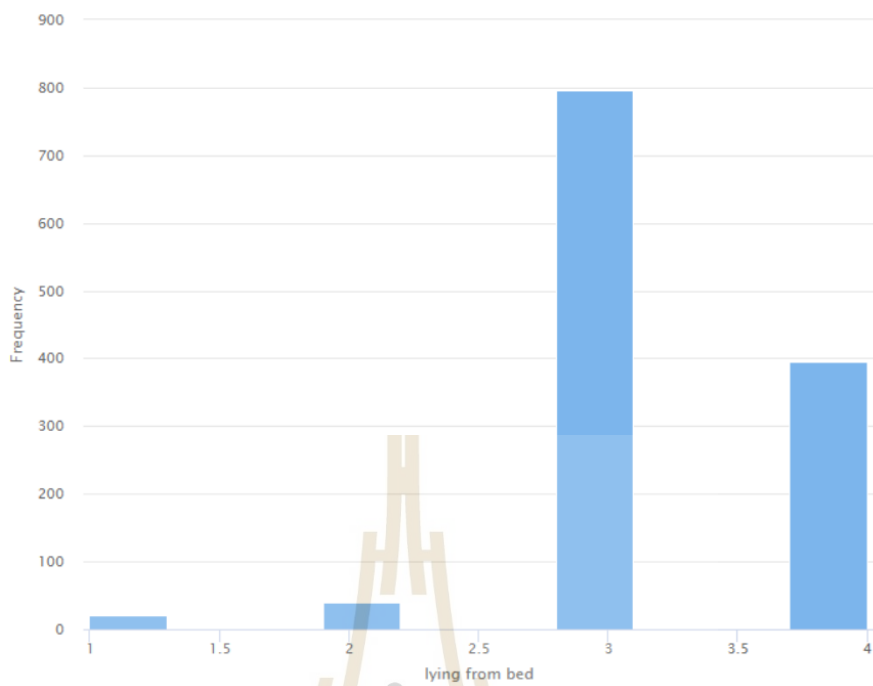


Figure 4.5 Lying on bed attribute distribution.

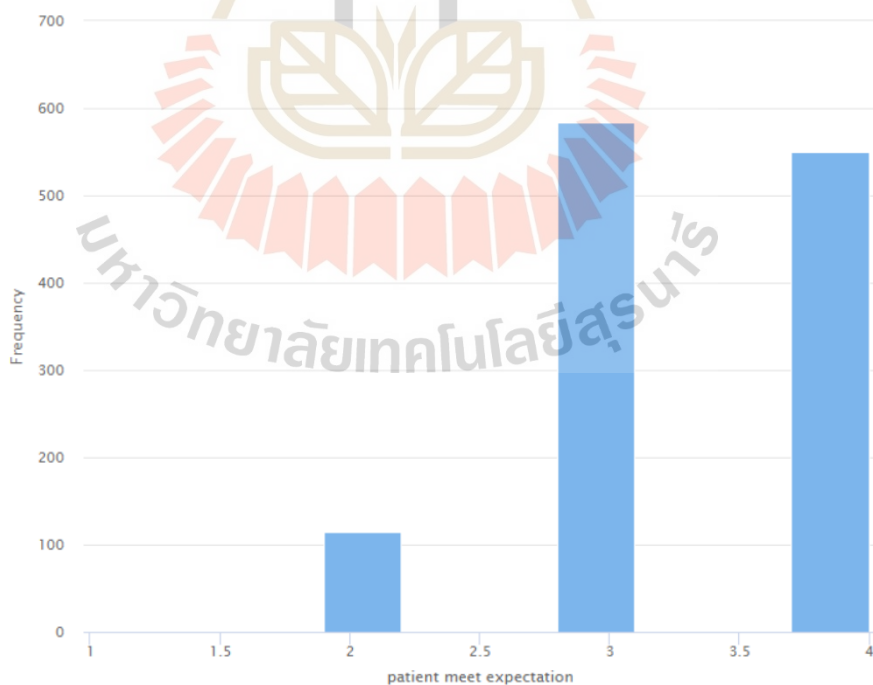


Figure 4.6 Meeting patient's expectations attribute distribution.

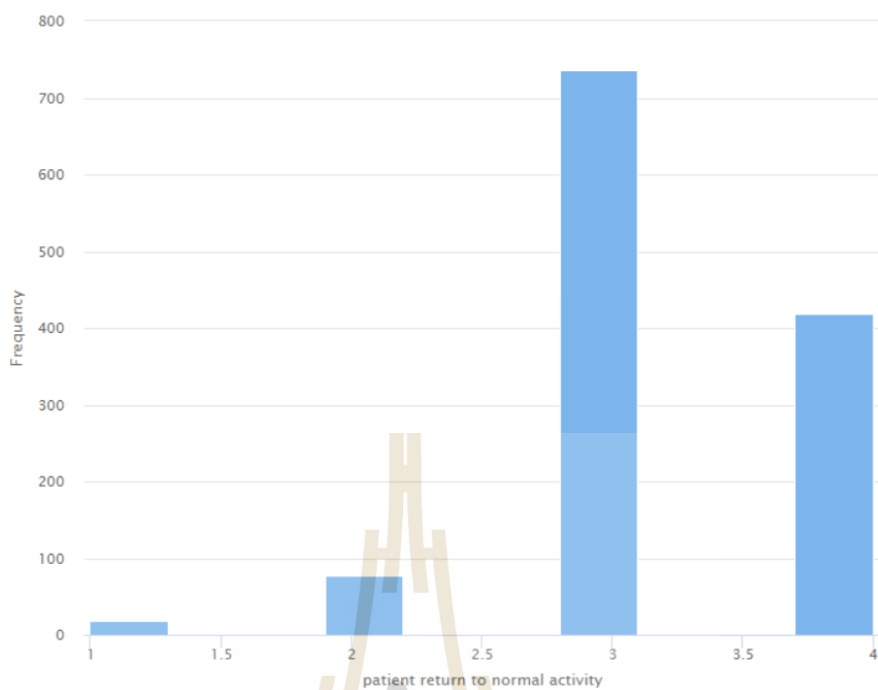


Figure 4.7 Patient ability to return to normal activity attribute distribution.

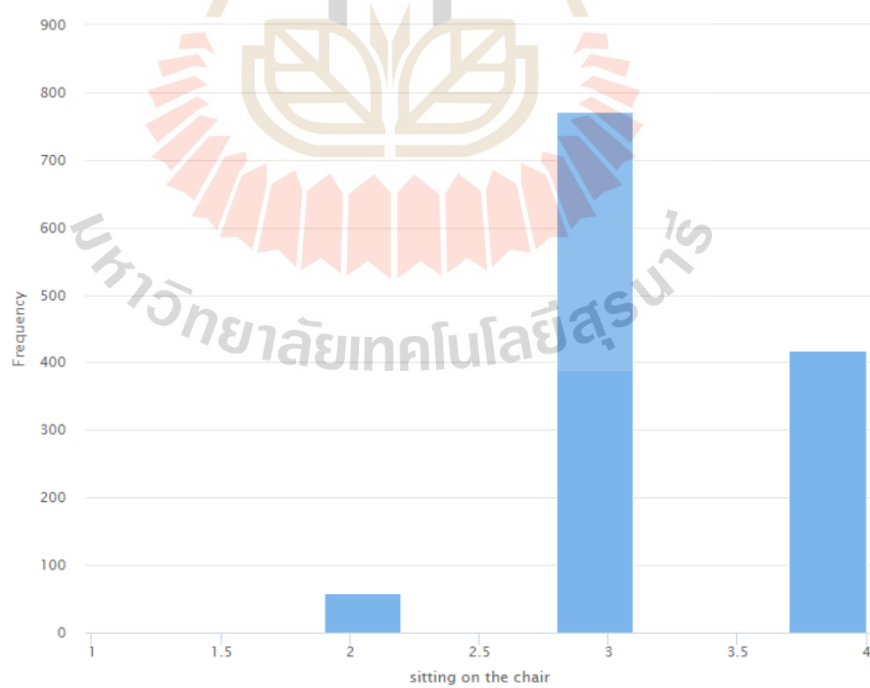


Figure 4.8 Sitting on the chair attribute distribution.

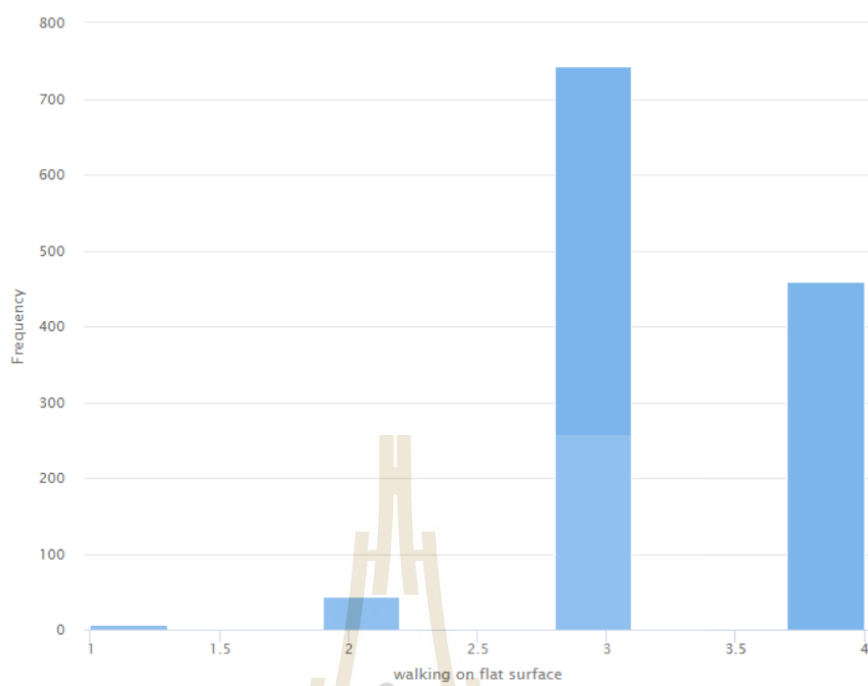


Figure 4.9 Walking on flat surface attribute distribution.

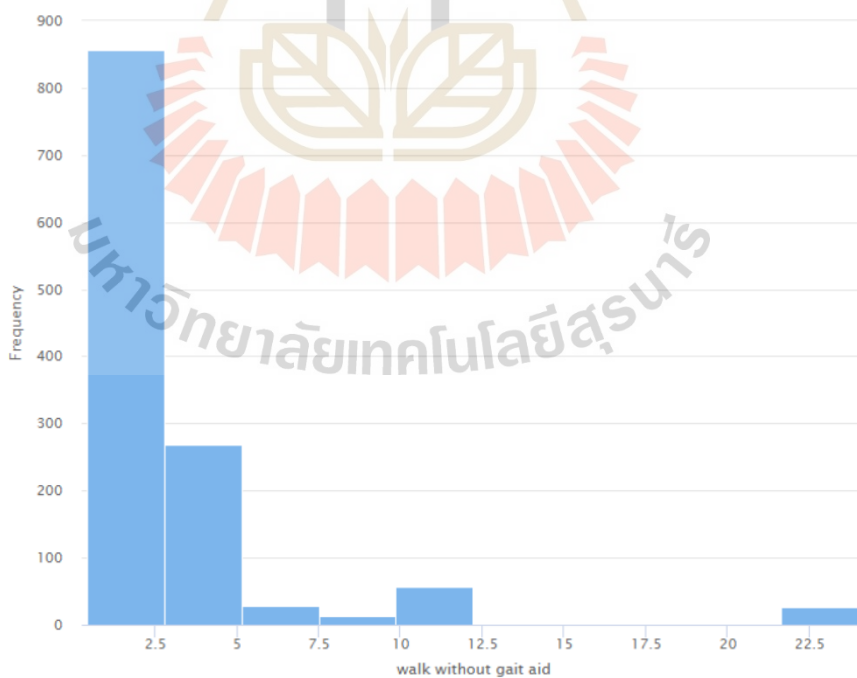


Figure 4.10 Walking without gait aid attribute distribution.

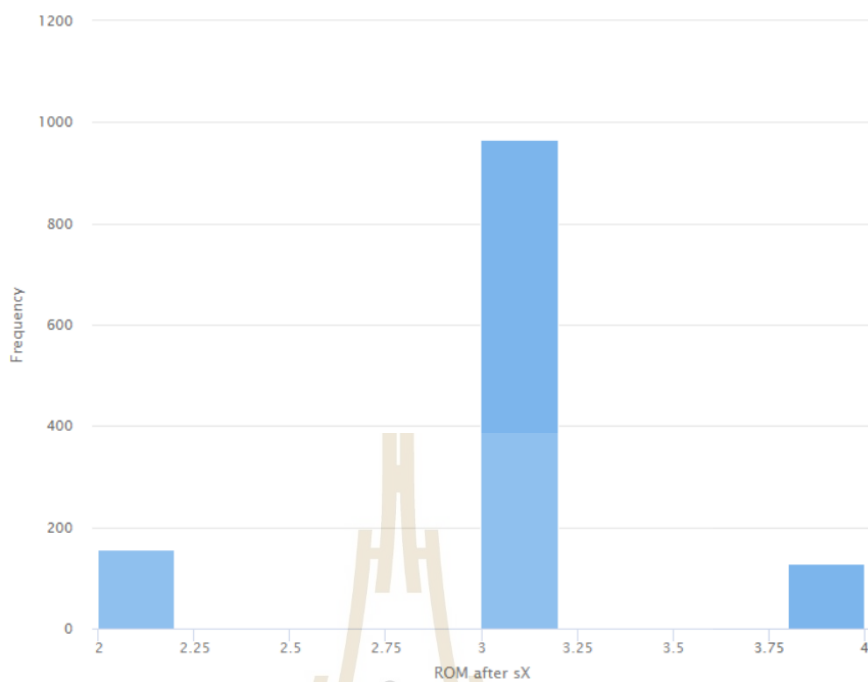


Figure 4.11 ROM after SX attribute distribution.

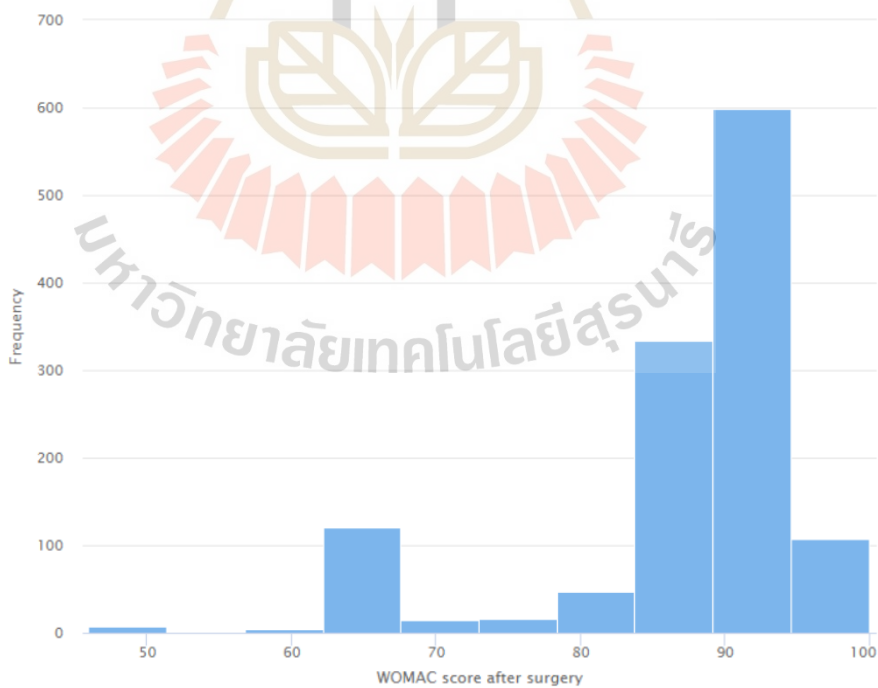


Figure 4.12 WOMAC score after surgery attribute distribution.

Table 4.2 Descriptive statistic of continuous data.

Attribute	Min	Max	Mean	Variance	SD
Age	47	96	68.98	73.57	8.58
BMI	18.91	41.02	28.44	20.01	4.48
WOMAC score after surgery	46	100	87.86	90.04	9.51

4.2 The novel attributes

The process of this part was to select and generate new attributes by automatic feature engineering on GLM, SVM, DL, and GBT. The following new attributes were obtained:

4.2.1 Generalized linear model

The parameters of GLM are set on the automatic feature engineering, including

- Family: Gaussian
- Link function: Identity
- Solver: IRLSM.

Then the group attribute is obtained by the automatic feature engineering on GLM as following Table 4.3

Table 4.3 Features Engineering with Generalized Linear Model (FE+GLM).

Attribute	Description
G_1	BMI
G_2	Walking without gait aid
G_3	Meeting patient's expectations
G_4	Patient ability to return to normal activity
G_5	$\log(\max(\text{Meeting patient's expectations}, \text{Walking without gait aid}))$

where the attribute G_5 means the value logarithm of maximum value between “Meeting patient's expectations” and “Walking without gait aid”.

4.2.2 Support vector machine

The parameters of SVM are set on the automatic feature engineering, including

- Kernel: Radial
- Kernel Gamma: 3
- Penalty: 38
- Epsilon: 2

Then the group attribute is obtained by the automatic feature engineering on SVM as following Table 4.4

Table 4.4 Features Engineering with Support Vector Machine (FE+SVM).

Attribute	Description
S_1	Age
S_2	BMI
S_3	ROM after SX
S_4	Walking without gait aid
S_5	Meeting patient's expectations
S_6	Patient ability to return to normal activity

4.2.3 Deep learning

The parameters of DL are set on the automatic feature engineering, including

- Activation function: Rectifier
- Loss function: Quadratic
- Learning rate: 0.01

Then the group attribute is obtained by the automatic feature engineering on DL as following Table 4.5

Table 4.5 Features Engineering with Deep Learning (FE+DL).

Attribute	Description
D_1	Age
D_2	BMI
D_3	ROM after SX
D_4	Walking without gait aid
D_5	Meeting patient's expectations
D_6	Getting in or out of the car
D_7	Patient ability to return to normal activity
D_8	Lying on bed \div Patient Meeting patient's expectations

where D_8 means value of “Lying on bed” divided by value of “Meeting patient's expectations”.

4.2.4 Gradient boost tree

The parameters of GBT are set on the automatic feature engineering, including

- Number of tree: 200
- Maximal of depth: 11
- Learning rate: 0.1

Then the group attribute is obtained by the automatic feature engineering on GBT as following Table 4.6

Table 4.6 Features Engineering with Gradient Boost Tree (FE+GBT).

Attribute	Description
T_1	Age
T_2	BMI
T_3	ROM after SX
T_4	Walking on flat surface
T_5	Meeting patient's expectations
T_6	Getting in or out of the car
T_7	Patient ability to return to normal activity
T_8	Patient ability to return to normal activity – ROM after SX
T_9	BMI \times Walking without gait aid

where T_8 and T_9 mean value of “Patient ability to return to normal activity” minus the value of “ROM after SX”, and the value of “BMI” multiplied by “Walking without gait aid”, respectively.

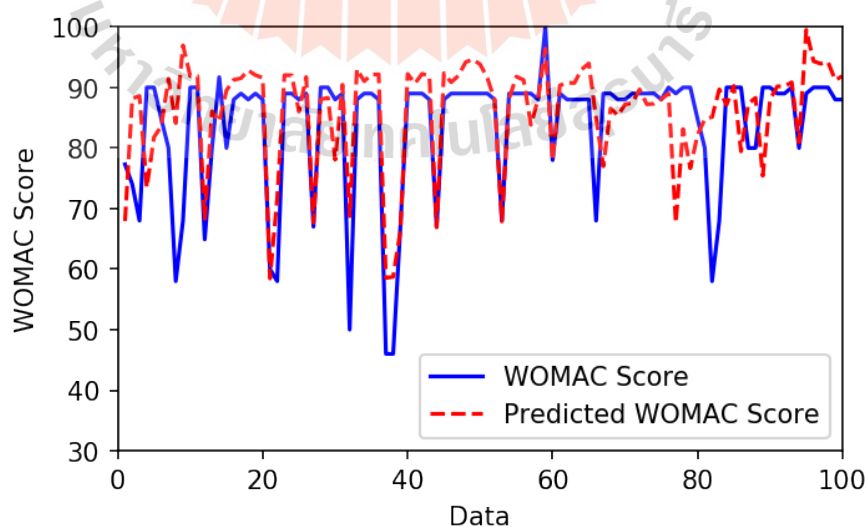
4.3 Evaluation model

This part demonstrates the efficiency of gradient boost tree with other attribute where 10-fold cross validation is used to validate the model. The model can be evaluated by RMSE, MAD, and SE metric where a lower value of each metric is better than a high value. The values of the metrics are demonstrated in Table 4.7.

Table 4.7 RMSE, MAD and SE of the models obtained.

Model	Attributes	NT	MD	LR	RMSE	MAD	SE
GBT	FE+GLM	200	15	0.1	5.819 ± 0.801	3.738 ± 0.481	34.443 ± 9.233
GBT	FE+SVM	200	10	0.1	5.517 ± 0.699	3.660 ± 0.442	30.880 ± 7.781
GBT	FE+DL	200	11	0.1	5.544 ± 0.652	3.632 ± 0.414	31.113 ± 7.288
GBT	FE+GBT	200	11	0.1	5.316 ± 0.539	3.529 ± 0.370	28.525 ± 5.828

where NT, MD, and LR are parameter values of tree, maximal depth, and learning rate, respectively. Table 4.7 demonstrates the performance of the gradient boost tree with different attributes with 10-fold cross validation. The best model is the gradient boost tree with attributes originating by automatic feature engineering on gradient boost tree (or group attribute 4) where RMSE, MAD and SE are 5.316 ± 0.539 , 3.529 ± 0.370 , and 28.525 ± 5.828 , respectively. Figures 4.13-4.16 show the predictions of the patients' postoperative WOMAC score after total knee replacement first 100 data created with the model of gradient boost tree together with FE+GLM, FE+SVM, FE+DL, and FE+GBT, respectively.

**Figure 4.13** Predicting WOMAC Score by using FE+GLM.

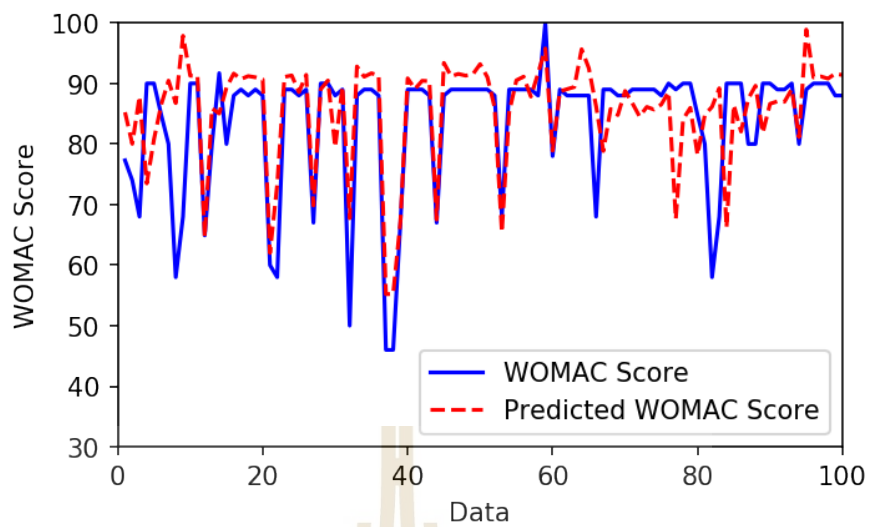


Figure 4.14 Predicting WOMAC Score by using FE+SVM.

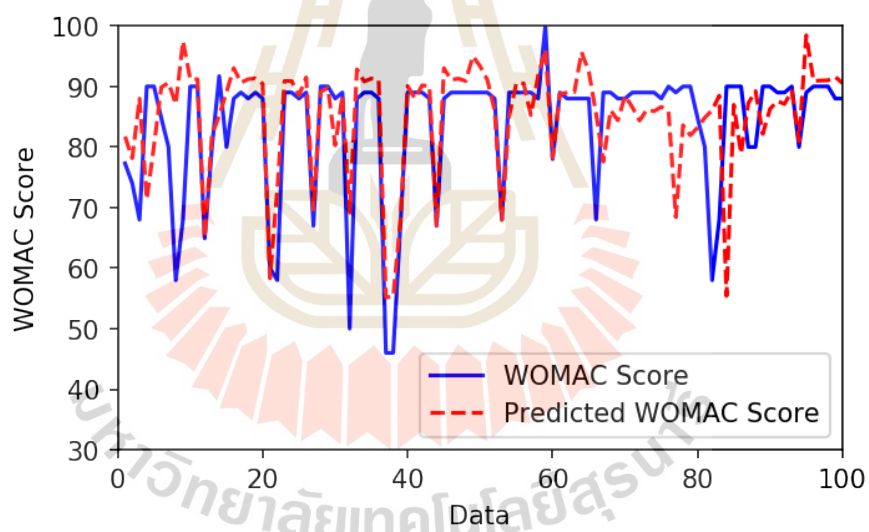


Figure 4.15 Predicting WOMAC Score by using FE+DL.

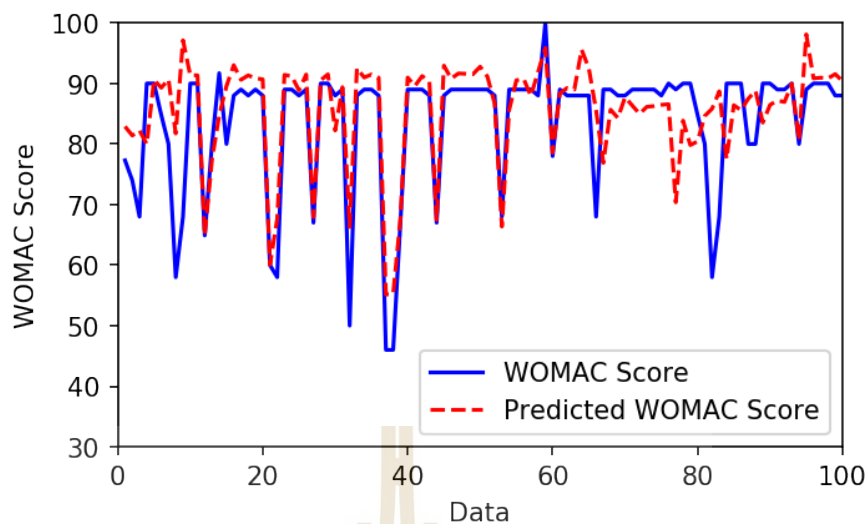


Figure 4.16 Predicting WOMAC Score by using FE+GBT.

4.4 Performance of learning rate of gradient boost tree

Table 4.7 demonstrates that the model built by group attributes 4 is the best of all, and subsequently, only the learning rate parameter is considered. Thus, the parameters of the model remain as before; however the learning rate parameter is adjusted only.

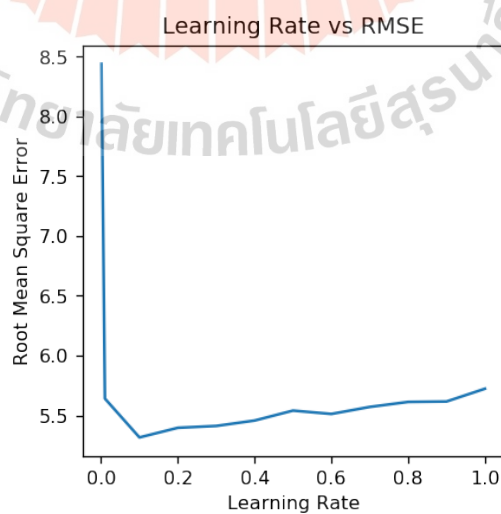


Figure 4.17 Tune Learning Rate in GBT.

Figure 4.17 shows the RMSE for different values of the learning rate in range 0,0.1,0.2,...,1. A learning rate 0.1 gives the least RMSE. After this, the model is further adjusted, modifying two parameters which are learning rate and number of trees. The following parameters were chosen:

Table 4.8 Parameters Gradient Boost Tree.

Parameter	Value
Number of tree	100, 200, 300, 400, 500
Learning rate	0.0001, 0.001, 0.01, 0.1

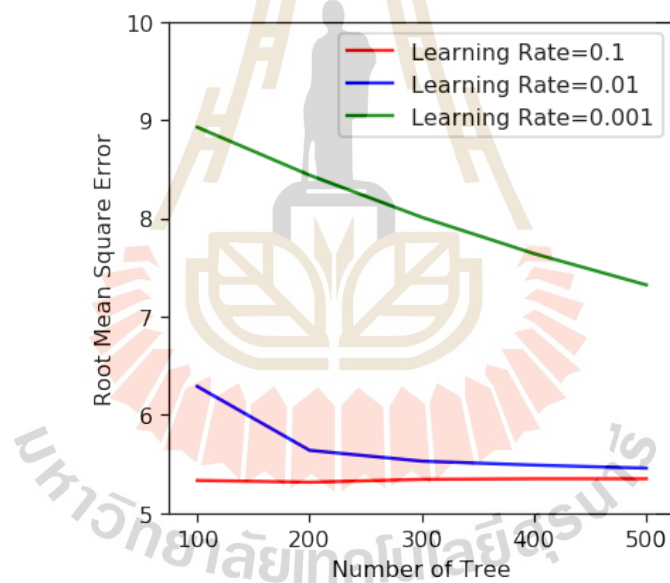


Figure 4.18 Tune Learning Rate and Number of Trees in GBT.

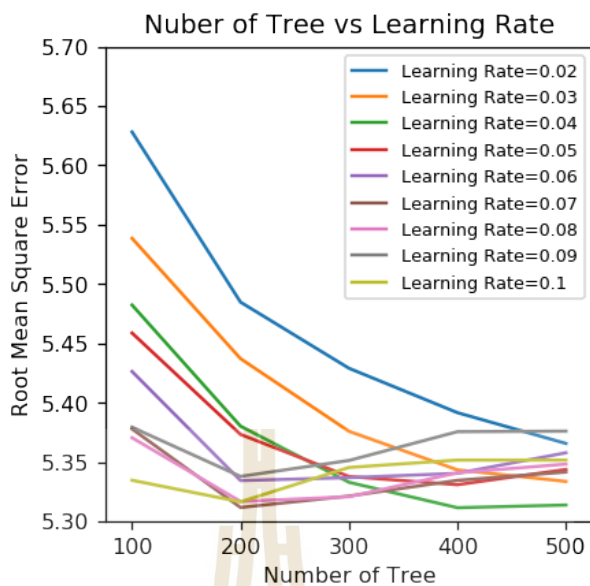


Figure 4.19 Further Tune Learning Rate and Number of Trees in GBT.

The results shown in Figure 4.18 demonstrate whether for every number of trees, the learning rate of 0.1 is better than the other learning rates.

Table 4.9 RMSE, MAD and SE of the models obtained.

Model	Attributes	NT	MD	LR	RMSE	MAD	SE
GBT	FE+GBT	500	11	0.01	5.459 ± 0.654	3.678 ± 0.427	30.190 ± 7.289
GBT	FE+GBT	500	11	0.02	5.366 ± 0.563	3.594 ± 0.379	29.075 ± 6.212
GBT	FE+GBT	500	11	0.03	5.334 ± 0.544	3.549 ± 0.367	28.713 ± 5.934
GBT	FE+GBT	400	11	0.04	5.311 ± 0.538	3.550 ± 0.376	28.472 ± 5.881
GBT	FE+GBT	400	11	0.05	5.331 ± 0.517	3.536 ± 0.365	28.659 ± 5.606
GBT	FE+GBT	200	11	0.06	5.334 ± 0.566	3.561 ± 0.392	28.743 ± 6.199
GBT	FE+GBT	200	11	0.07	5.312 ± 0.565	3.532 ± 0.391	28.501 ± 6.177
GBT	FE+GBT	200	11	0.08	5.317 ± 0.544	3.540 ± 0.352	28.533 ± 5.922
GBT	FE+GBT	200	11	0.09	5.338 ± 0.508	3.543 ± 0.356	28.724 ± 5.552
GBT	FE+GBT	200	11	0.1	5.316 ± 0.539	3.529 ± 0.370	28.525 ± 5.828

After that, further tuning of learning rate and number of trees gave the data of Figure 4.19 and Table 4.9, which shows that a learning rate of 0.07 and number of trees of 200 has RMSE of 5.312 ± 0.565 . A learning rate of 0.04 and number of trees of 400 has RMSE of 5.311 ± 0.538 . So that both models have an RMSE lower than the other models of Table 4.7. This shows that the learning rate is a parameter that affects the gradient boost tree.



CHAPTER V

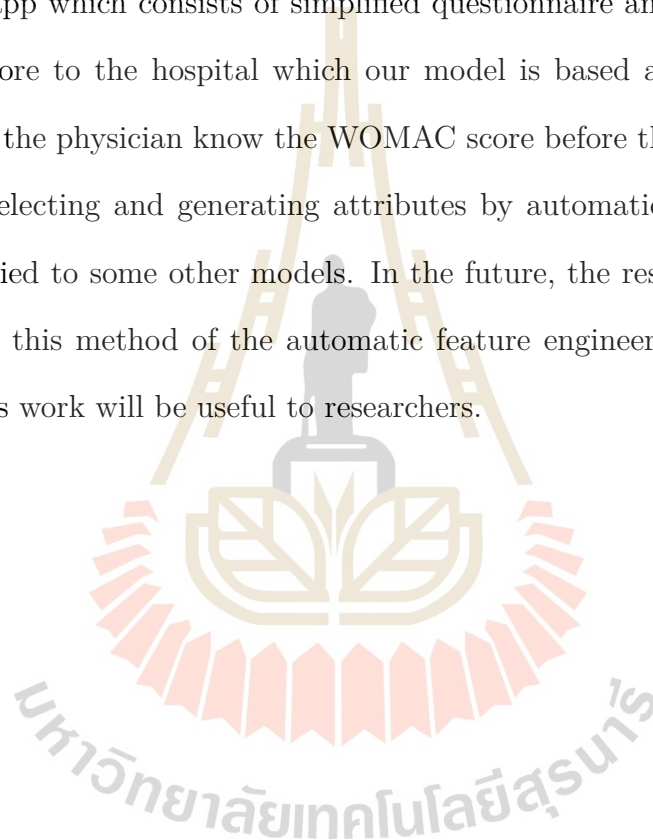
CONCLUSION AND RECOMMENDATION

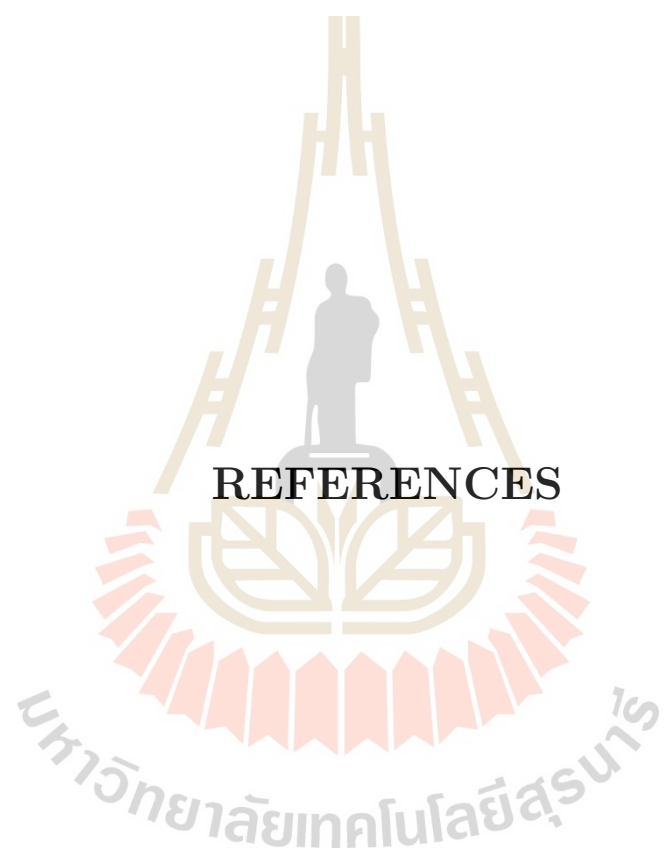
This thesis has studied the modeling to predict patient's postoperative WOMAC score after total knee replacement, using the data set obtained from Asst. Prof. Lt. Col. Bura Sibdhupakron, as mentioned before. Since the WOMAC score is an integer, the model for solving chosen was a regression with a gradient boost tree. Furthermore, the model was implemented in Rapidminer Studio 9.9 (Education license). This analysis was divided into three parts. The first part is optimizing the model, including a generalized linear model, support vector machine, deep learning, and gradient boost tree to predict the patient's postoperative WOMAC score after total knee replacement to receive the value of parameters of each model. The second part is to select and generate new group attributes by using the automatic feature engineering on a generalized linear model, support vector machine, deep learning, and gradient boost tree such that each model is set the value of parameter obtained from the first part, the result of this part obtain the group attributes from each model. The third part used each group's attributes from the second part to construct the prediction model for predicting by using the gradient boost tree with 10-fold cross-validation, followed by optimizing three parameters of gradient boost tree, namely the number of trees, maximal of depth, and learning rate.

The model is created in which different group attributes were measured by RMSE, MAD, and SE in Table 4.7. The result of Table 4.7 show that the gradient boost tree using group attributes received from automatic feature engineering on

gradient boost tree is the best. Second, third, and fourth best performance are group attributes from auto feature engineering on deep learning, support vector machine, and generalized linear model. Figure 4.17-4.19, and table 4.9 show that the learning rate of the gradient boost tree affects the efficiency to the model.

In the future, the model from this work can be used and developed in artificial intelligence to a great variety of applications. For example one may develop an app which consists of simplified questionnaire and send the predicted WOMAC score to the hospital which our model is based algorithm behind the app, so that the physician know the WOMAC score before the appointment. The method of selecting and generating attributes by automatic feature engineering may be applied to some other models. In the future, the researcher interested in this work or this method of the automatic feature engineering to other models. We hope this work will be useful to researchers.





REFERENCES

REFERENCES

- Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. **Artificial Intelligence Review**. <https://doi:10.1007/s10462-020-09896-5>
- Botchkarev, A. (2018). Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio. **SSRN Electronic Journal**. <https://doi:10.2139/ssrn.3177507>
- Cui1, A., Li, H., Wang, D., Zhong, J., Chen, Y. and Lu H. (2020). Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. **EClinicalMedicine**. <https://doi.org/10.1016/j.eclinm.2020.100587>
- The Data WOMAC Score after TKR data set is available on <https://www.kaggle.com/saranchaisinlapasorn/data-womac-score-after-tkr>
- Department of Older Persons, (2020), **Situation of the Thai Elderly**. Retrieved from <http://www.dop.go.th/th/know/1>
- Gareth, J., Daniela W., Trevor H. and Rob T. (2017). **An Introduction to Statistical Learning**. New York: Springer.
- Jamshidi, A., Pelletier, J.-P. and Martel-Pelletier, J. (2018). Machine-learning-based patient-specific prediction models for knee osteoarthritis. **Nature Reviews Rheumatology**. <https://doi:10.1038/s41584-018-0130-5>
- Kanter, J. M. and Veeramachaneni, K. (2015). Deep feature synthesis: Towards automating data science endeavors. **2015 IEEE International**

Conference on Data Science and Advanced Analytics (DSAA).

<https://doi:10.1109/dsaa.2015.7344858>

Kokkotis, C., Moustakidis, S., Papageorgiou, E., Giakas, G. and Tsaopoulos, D. E. (2020), Machine learning in knee osteoarthritis: A review. **Osteoarthritis and Cartilage Open**, Volume 2, Issue 3, Article 100069.

<https://doi.org/10.1016/j.ocarto.2020.100069>

Koutsoukas, A., Monaghan, K. J., Li, X. and Huan, J. (2017). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. **Journal of Cheminformatics**, 9(1). <https://doi:10.1186/s13321-017-0226-y>

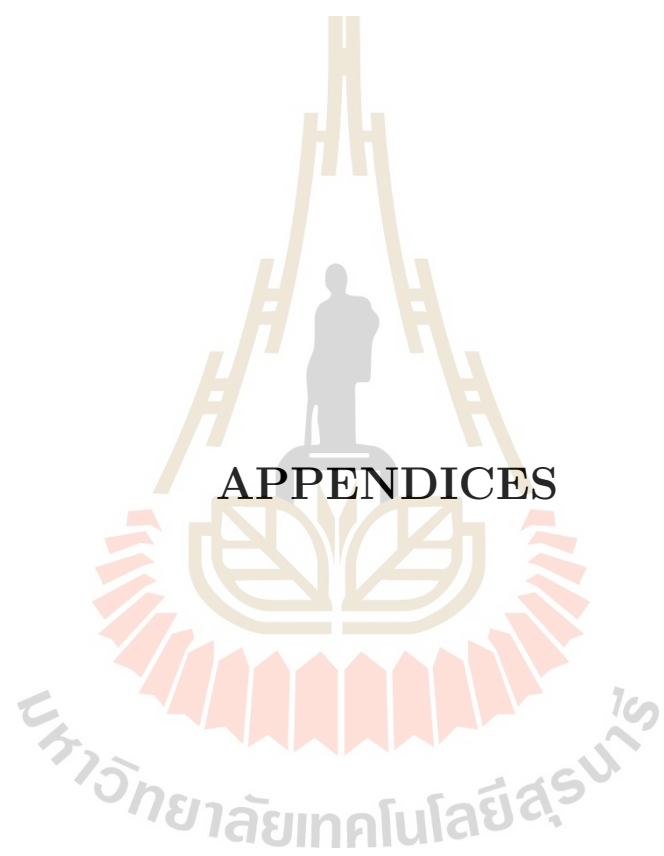
Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. **Frontiers in Neurorobotics**, 7. <https://doi:10.3389/fnbot.2013.00021>

Philipp P., Anne-Laure, B. and Bernd B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. **Journal of Machine Learning Research**, Volume 20, Retrieved from <http://jmlr.org/papers/v20/18-444.html>

Starmer, J. (2019, April 2), **Gradient Boost Part 2 (of 4): Regression Details** [Video file], Retrieved from https://www.youtube.com/watch?v=2xudPOBz-vs&t=492s&ab_channel=StatQuestwithJoshStarmer

Starmer, J. (2019, April 20), **Regression Trees, Clearly Explained!!!** [Video file], Retrieved from https://www.youtube.com/watch?v=g9c66TUy1Z4&t=254s&ab_channel=StatQuestwithJoshStarmer

- Tiulpin, A., Klein, S., Bierma-Zeinstra, S. M. A., Thevenot, J., Rahtu, E., Meurs, J. Van, Oei, E. H. G. and Saarakkala, S. (2019). Multi-modal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. **Scientific Reports**, 9(1). <https://doi:10.1038/s41598-019-56527-3>
- Walker, L. C., Clement, N. D., Bardgett, M., Weir, D., Holland, J., Gerrand, C. and Deehan, D. J. (2018). The WOMAC score can be reliably used to classify patient satisfaction after total knee arthroplasty. **Knee Surgery, Sports Traumatology, Arthroscopy**. doi:10.1007/s00167-018-4879-5
- Yuan, K.-C., Tsai, L.-W., Lee, K.-H., Cheng, Y.-W., Hsu, S.-C., Lo, Y.-S. and Chen, R.-J. (2020). The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. **International Journal of Medical Informatics**, 141, Article 104176. <https://doi:10.1016/j.ijmedinf.2020.104176>
- Zheng, A. and Casari, A. (2018). **Feature Engineering for Machine Learning** (1st ed.). USA: O'Reilly Media.






APPENDIX A
QUESTIONNAIRE FOR OSTEOARTHRITIS

มหาวิทยาลัยเทคโนโลยีสุรนารี

A.1 Western Ontario and McMaster Universities Arthritis Index

	PATIENT NAME	DOB				
WESTERN ONTARIO AND MCMASTER OSTEOARTHRITIS INDEX (WOMAC) Please circle the appropriate rating for each item.						
RATE YOUR PAIN WHEN...	NONE	SLIGHT	MODERATE	SEVERE	EXTREME	HOSPITAL USE ONLY
Walking	0	1	2	3	4	
Climbing stairs	0	1	2	3	4	
Sleeping at night	0	1	2	3	4	
Resting	0	1	2	3	4	
Standing	0	1	2	3	4	
RATE YOUR STIFFNESS IN THE...	NONE	SLIGHT	MODERATE	SEVERE	EXTREME	HOSPITAL USE ONLY
Morning	0	1	2	3	4	
Evening	0	1	2	3	4	
RATE YOUR DIFFICULTY WHEN...	NONE	SLIGHT	MODERATE	SEVERE	EXTREME	HOSPITAL USE ONLY
Descending stairs	0	1	2	3	4	
Ascending stairs	0	1	2	3	4	
Rising from sitting	0	1	2	3	4	
Standing	0	1	2	3	4	
Bending to floor	0	1	2	3	4	
Walking on even floor	0	1	2	3	4	
Getting in/out of car	0	1	2	3	4	
Going shopping	0	1	2	3	4	
Putting on socks	0	1	2	3	4	
Rising from bed	0	1	2	3	4	
Taking off socks	0	1	2	3	4	
Lying in bed	0	1	2	3	4	
Getting in/out of bath	0	1	2	3	4	
Sitting	0	1	2	3	4	
Getting on/off toilet	0	1	2	3	4	
Doing light domestic duties (cooking, dusting)	0	1	2	3	4	
Doing heavy domestic duties (moving furniture)	0	1	2	3	4	
PATIENT SIGNATURE		DATE				
REVIEWED BY PHYSICAL THERAPIST		DATE				WOMAC TOTAL SCORE /96

YAVAPAI REGIONAL MEDICAL CENTER
 PHYSICAL REHABILITATION SERVICES

WOMAC OSTEOARTHRITIS INDEX QUESTIONNAIRE

REHABILITATION SERVICES
 PT THA/TKA WOMAC QUESTIONNAIRE
 MR-1433 (11/15)

Figure A.1 Western Ontario and McMaster Universities Arthritis Index.

A.2 Knee injury and Osteoarthritis Outcome Score

Knee Injury and Osteoarthritis Outcome Score (KOOS)

Source: Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)—development of a self-administered outcome measure. *J Orthop Sports Phys Ther.* 1998 Aug;28(2):88-96.

The Knee Injury and Osteoarthritis Outcome Score (KOOS) is a questionnaire designed to assess short and long-term patient-relevant outcomes following knee injury. The KOOS is self-administered and assesses five outcomes: pain, symptoms, activities of daily living, sport and recreation function, and knee-related quality of life. The KOOS meets basic criteria of outcome measures and can be used to evaluate the course of knee injury and treatment outcome. KOOS is patient-administered, the format is user-friendly and it takes about 10 minutes to fill out.

Scoring instructions

The KOOS's five patient-relevant dimensions are scored separately: Pain (nine items); Symptoms (seven items); ADL Function (17 items); Sport and Recreation Function (five items); Quality of Life (four items). A Likert scale is used and all items have five possible answer options scored from 0 (No problems) to 4 (Extreme problems) and each of the five scores is calculated as the sum of the items included.

Interpretation of scores

Scores are transformed to a 0–100 scale, with zero representing extreme knee problems and 100 representing no knee problems as common in orthopaedic scales and generic measures. Scores between 0 and 100 represent the percentage of total possible score achieved.

Figure A.2 First page of Knee injury and Osteoarthritis Outcome Score.

Knee Injury and Osteoarthritis Outcome Score (KOOS)

Pain

P1	How often is your knee painful?	<input type="checkbox"/> Never	<input type="checkbox"/> Monthly	<input type="checkbox"/> Weekly	<input type="checkbox"/> Daily	<input type="checkbox"/> Always
What degree of pain have you experienced the last week when...?						
P2	Twisting/pivoting on your knee	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
P3	Straightening knee fully	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
P4	Bending knee fully	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
P5	Walking on flat surface	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
P6	Going up or down stairs	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
P7	At night while in bed	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
P8	Sitting or lying	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
P9	Standing upright	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme

Symptoms

Sy1	How severe is your knee stiffness after first wakening in the morning?	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
Sy2	How severe is your knee stiffness after sitting, lying, or resting later in the day?	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
Sy3	Do you have swelling in your knee?	<input type="checkbox"/> Never	<input type="checkbox"/> Rarely	<input type="checkbox"/> Sometimes	<input type="checkbox"/> Often	<input type="checkbox"/> Always
Sy4	Do you feel grinding, hear clicking or any other type of noise when your knee moves?	<input type="checkbox"/> Never	<input type="checkbox"/> Rarely	<input type="checkbox"/> Sometimes	<input type="checkbox"/> Often	<input type="checkbox"/> Always
Sy5	Does your knee catch or hang up when moving?	<input type="checkbox"/> Never	<input type="checkbox"/> Rarely	<input type="checkbox"/> Sometimes	<input type="checkbox"/> Often	<input type="checkbox"/> Always
Sy6	Can you straighten your knee fully?	<input type="checkbox"/> Always	<input type="checkbox"/> Often	<input type="checkbox"/> Sometimes	<input type="checkbox"/> Rarely	<input type="checkbox"/> Never
Sy7	Can you bend your knee fully?	<input type="checkbox"/> Always	<input type="checkbox"/> Often	<input type="checkbox"/> Sometimes	<input type="checkbox"/> Rarely	<input type="checkbox"/> Never

Figure A.3 Second page of Knee injury and Osteoarthritis Outcome Score.

Activities of daily living

What difficulty have you experienced the last week...?

A1 Descending	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A2 Ascending stairs	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A3 Rising from sitting	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A4 Standing	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A5 Bending to floor/picking up an object	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A6 Walking on flat surface	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A7 Getting in/out of car	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A8 Going shopping	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A9 Putting on socks/stockings	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A10 Rising from bed	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A11 Taking off socks/stockings	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A12 Lying in bed (turning over, maintaining knee position)	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A13 Getting in/out of bath	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A14 Sitting	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A15 Getting on/off toilet	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A16 Heavy domestic duties (shovelling, scrubbing floors, etc)	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
A17 Light domestic duties (cooking, dusting, etc)	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme

Sport and recreation function

What difficulty have you experienced the last week...?

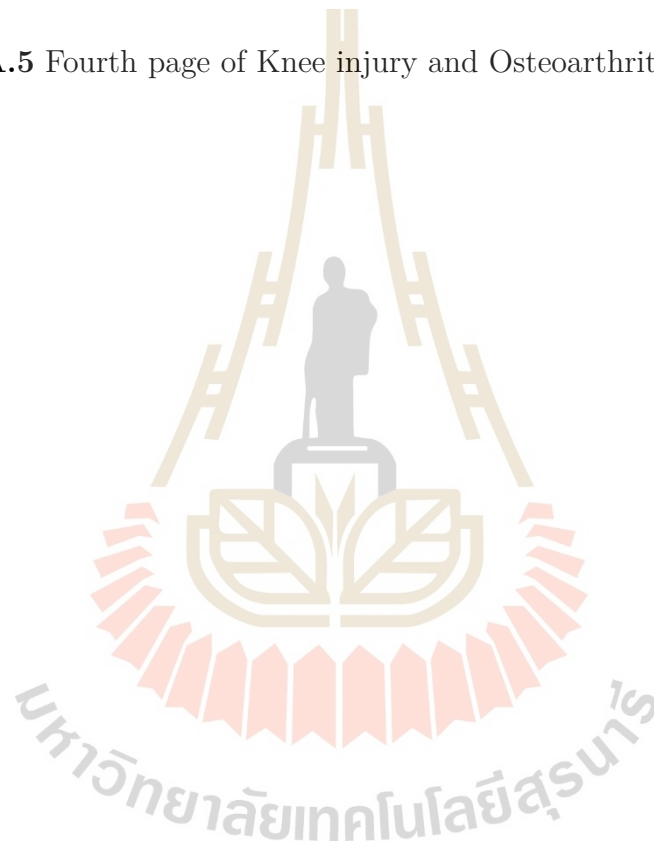
Sp1 Squatting	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
Sp2 Running	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
Sp3 Jumping	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
Sp4 Turning/twisting on your injured knee	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme
Sp5 Kneeling	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme

Figure A.4 Third page of Knee injury and Osteoarthritis Outcome Score.

Knee-related quality of life

Q1 How often are you aware of your knee problems?	<input type="checkbox"/> Never	<input type="checkbox"/> Monthly	<input type="checkbox"/> Weekly	<input type="checkbox"/> Daily	<input type="checkbox"/> Always
Q2 Have you modified your lifestyle to avoid potentially damaging activities to your knee?	<input type="checkbox"/> Not at all	<input type="checkbox"/> Mildly	<input type="checkbox"/> Moderately	<input type="checkbox"/> Severely	<input type="checkbox"/> Totally
Q3 How troubled are you with lack of confidence in your knee?	<input type="checkbox"/> Not at all	<input type="checkbox"/> Mildly	<input type="checkbox"/> Moderately	<input type="checkbox"/> Severely	<input type="checkbox"/> Totally
Q4 In general, how much difficulty do you have with your knee?	<input type="checkbox"/> None	<input type="checkbox"/> Mild	<input type="checkbox"/> Moderate	<input type="checkbox"/> Severe	<input type="checkbox"/> Extreme

Figure A.5 Fourth page of Knee injury and Osteoarthritis Outcome Score.



A.3 International Knee Documentation Committee

IKDC Subjective Knee Evaluation

SYMPTOMS*:

*Grade symptoms at the highest activity level at which you think you could function without significant symptoms, even if you are not actually performing activities at this level.

1. What is the highest level of activity that you can perform without significant knee pain?
 - 4 Very strenuous activities like jumping or pivoting as in gymnastics or football
 - 3 Strenuous activities like heavy physical work, skiing or tennis
 - 2 Moderate activities like moderate physical work, running or jogging
 - 1 Light activities like walking, housework or gardening
 - 0 Unable to perform any of the above activities due to knee pain

2. During the past 4 weeks, or since your injury, how often have you had pain?

Never	0	1	2	3	4	5	6	7	8	9	10	Constant
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

3. If you have pain, how severe is it?

No pain	0	1	2	3	4	5	6	7	8	9	10	Worst pain imaginable
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

4. During the past 4 weeks, or since your injury, how stiff or swollen has your knee been?
 - 4 Not at all
 - 3 Mildly
 - 2 Moderately
 - 1 Very
 - 0 Extremely

5. What is the highest level of activity you can perform without significant swelling in your knee?
 - 4 Very strenuous activities like jumping or pivoting as in gymnastics or football
 - 3 Strenuous activities like heavy physical work, skiing or tennis
 - 2 Moderate activities like moderate physical work, running or jogging
 - 1 Light activities like walking, housework or gardening
 - 0 Unable to perform any of the above activities due to knee swelling

6. During the past 4 weeks, or since your injury, has your knee locked or caught?
 - 0 Yes
 - 1 No

7. What is the highest level of activity you can perform without significant giving way in your knee?
 - 4 Very strenuous activities like jumping or pivoting as in gymnastics or football
 - 3 Strenuous activities like heavy physical work, skiing or tennis
 - 2 Moderate activities like moderate physical work, running or jogging
 - 1 Light activities like walking, housework or gardening
 - 0 Unable to perform any of the above activities due to giving way of the knee

Figure A.6 First page of International Knee Documentation Committee.

SPORT ACTIVITIES:

8. What is the highest level of activity you can participate in on a regular basis?
- 4 Very strenuous activities like jumping or pivoting as in gymnastics or football
 3 Strenuous activities like heavy physical work, skiing or tennis
 2 Moderate activities like moderate physical work, running or jogging
 1 Light activities like walking, housework or gardening
 0 Unable to perform any of the above activities due to knee
9. How does your knee affect your ability to:
- | | Not difficult
at all | Minimally
difficult | Moderately
Difficult | Extremely
difficult | Unable to
do |
|---------------------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| a. Go up stairs | 4 <input type="checkbox"/> | 3 <input type="checkbox"/> | 2 <input type="checkbox"/> | 1 <input type="checkbox"/> | 0 <input type="checkbox"/> |
| b. Go down stairs | 4 <input type="checkbox"/> | 3 <input type="checkbox"/> | 2 <input type="checkbox"/> | 1 <input type="checkbox"/> | 0 <input type="checkbox"/> |
| c. Kneel on the front of your knee | 4 <input type="checkbox"/> | 3 <input type="checkbox"/> | 2 <input type="checkbox"/> | 1 <input type="checkbox"/> | 0 <input type="checkbox"/> |
| d. Squat | 4 <input type="checkbox"/> | 3 <input type="checkbox"/> | 2 <input type="checkbox"/> | 1 <input type="checkbox"/> | 0 <input type="checkbox"/> |
| e. Sit with your knee bent | 4 <input type="checkbox"/> | 3 <input type="checkbox"/> | 2 <input type="checkbox"/> | 1 <input type="checkbox"/> | 0 <input type="checkbox"/> |
| f. Rise from a chair | 4 <input type="checkbox"/> | 3 <input type="checkbox"/> | 2 <input type="checkbox"/> | 1 <input type="checkbox"/> | 0 <input type="checkbox"/> |
| g. Run straight ahead | 4 <input type="checkbox"/> | 3 <input type="checkbox"/> | 2 <input type="checkbox"/> | 1 <input type="checkbox"/> | 0 <input type="checkbox"/> |
| h. Jump and land on your involved leg | 4 <input type="checkbox"/> | 3 <input type="checkbox"/> | 2 <input type="checkbox"/> | 1 <input type="checkbox"/> | 0 <input type="checkbox"/> |
| i. Stop and start quickly | 4 <input type="checkbox"/> | 3 <input type="checkbox"/> | 2 <input type="checkbox"/> | 1 <input type="checkbox"/> | 0 <input type="checkbox"/> |

FUNCTION:

10. How would you rate the function of your knee on a scale of 0 to 10 with 10 being normal, excellent function and 0 being the inability to perform any of your usual daily activities which may include sport?

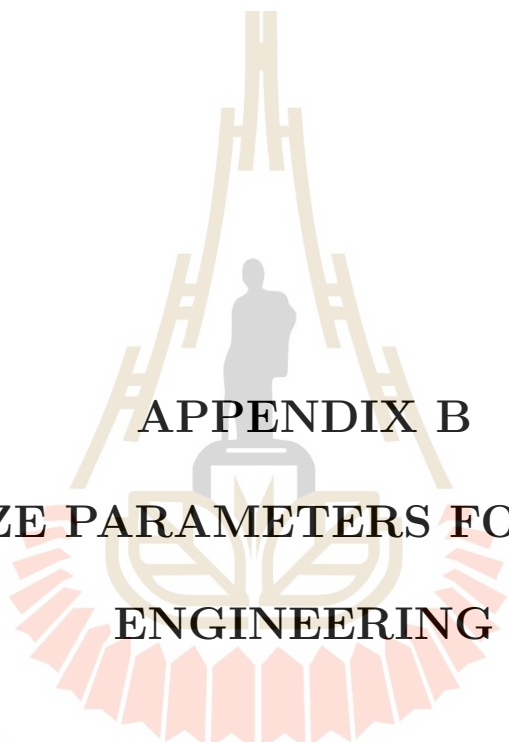
FUNCTION PRIOR TO YOUR KNEE INJURY:

Couldn't perform daily activities	0	1	2	3	4	5	6	7	8	9	10	No limitation in daily activities
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

CURRENT FUNCTION OF YOUR KNEE:

Cannot perform daily activities	0	1	2	3	4	5	6	7	8	9	10	No limitation in daily activities
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

Figure A.7 Second page of International Knee Documentation Committee.



APPENDIX B
OPTIMIZE PARAMETERS FOR FEATURE
ENGINEERING

มหาวิทยาลัยเทคโนโลยีสุรนารี

B.1 Root mean square error of each model after optimize the parameters

In this chapter we show the RMSE of each model, using the optimize parameters of each model in the Rapidminer Studio program, with validation by 10-fold cross validation. We obtained the performance of each model as follows:

Table B.1 Generalized Linear Model.

Model	Family	Link function	Solver	RMSE
GLM	Negativebinomial	-	IRLSM	7.646 ± 0.479
GLM	Gaussian	Logit	Coordinate descent	7.649 ± 0.468
GLM	Poisson	Logit	IRLSM	7.654 ± 0.464
GLM	Gamma	Identity	IRLSM	7.659 ± 0.486
GLM	Gaussian	Identity	IRLSM	7.667 ± 0.462
GLM	Tweedie	-	Coordinate descent naive	7.674 ± 0.459

Table B.2 Support Vector Machine.

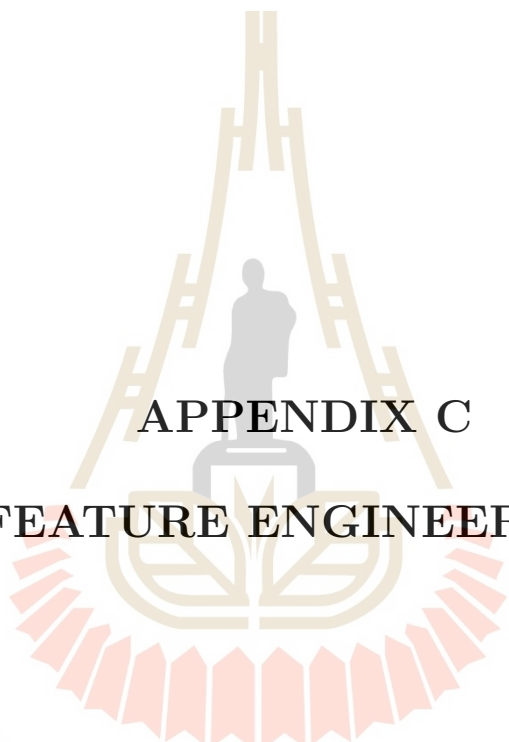
Model	Kernel	Kernel gamma	Kernel sigma	Degree	Penalty	Epsilon	RMSE
SVM	radial	3	-	-	38	2	5.865 ± 0.418
SVM	epachnenikov	-	2	7	44	2	5.868 ± 0.434
SVM	anova	5	-	2	2	0.1	5.969 ± 0.858
SVM	dot	-	-	-	0	7	7.676 ± 0.461

Table B.3 Deep Learning.

Model	Activation function	Learning rate	Loss function	RMSE
DL	Rectifier	0.01	Quadratic	6.310 ± 0.732
DL	Tanh	0.01	Quadratic	6.492 ± 0.757
DL	Maxout	0.001	Quadratic	6.604 ± 0.623
DL	ExpRectifier	0.001	Quadratic	7.215 ± 0.542

Table B.4 Gradient Boost Tree.

Model	Number of tree	Maximal depth	Learning rate	RMSE
GBT	200	11	0.1	5.466 ± 0.595
GBT	500	13	0.01	5.575 ± 0.624
GBT	500	18	0.001	7.359 ± 0.497
GBT	500	16	0.0001	9.205 ± 0.719



APPENDIX C
FEATURE ENGINEERING

มหาวิทยาลัยเทคโนโลยีสุรนารี

This chapter demonstrate the process feature engineering in Rapidminer Studio.

C.1 Optimized parameters for feature engineering

In this part, we show the process to optimize parameters of generalized linear model, support vector machine, deep learning, and gradient boost tree in Rapidminer Studio program which validation model by 10-fold cross-validation.

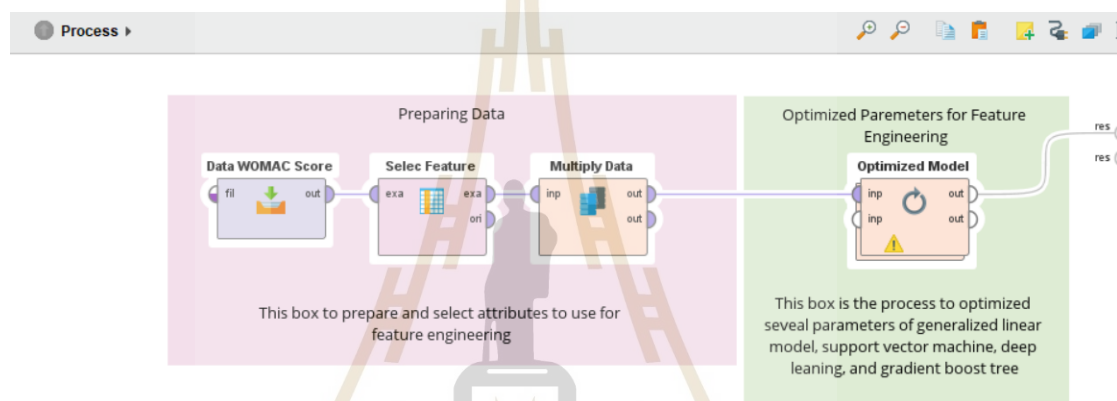


Figure C.1 Overview process of optimized parameters.

C.2 Feature engineering and optimize gradient boost tree

In this part, we show the process to feature engineering such that using automatic feature engineering to select and generate attributes with several techniques such as generalized linear model, support vector machine, deep learning, and gradient boost tree in Rapidminer Studio program which validation each technique by 10-fold cross-validation and process of optimizing gradient boost tree.

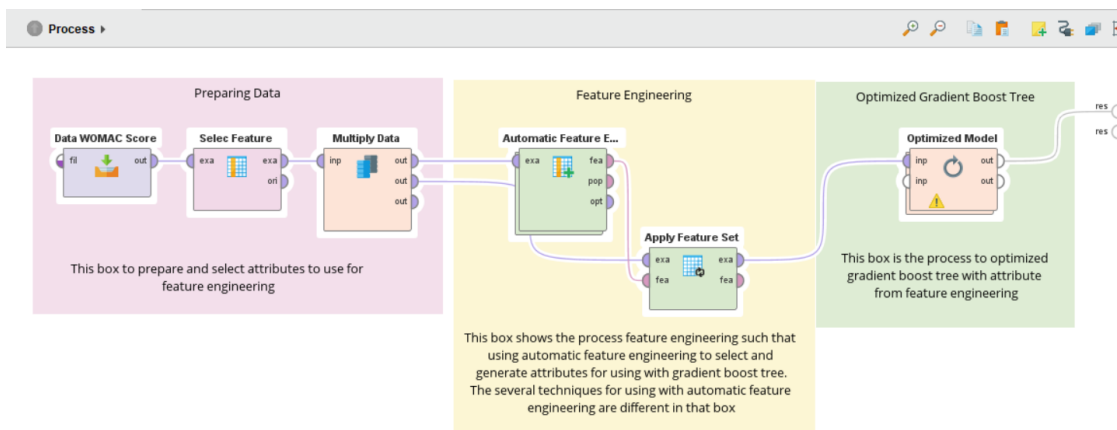


Figure C.2 Overview process of feature engineering and optimize gradient boost tree.





APPENDIX D
OPTIMIZE GRADIENT BOOST TREE

มหาวิทยาลัยเทคโนโลยีสุรนารี

This chapter demonstrate the RMSE of gradient boost tree for different learning rate.

D.1 RMSE of gradient boost tree with different attributes

In this chapter we show RMSE of gradient boost tree with different attributes in Rapidminer Studio program which validation model by 10-fold cross validation then we obtained the performance of each model as following

Table D.1 Gradient Boost Tree with FE+GLM.

Model	Number of tree	Maximal depth	Learning rate	RMSE
GBT	200	15	0.1	5.819 ± 0.801
GBT	500	16	0.01	5.870 ± 0.798
GBT	500	16	0.001	7.468 ± 0.502
GBT	500	16	0.0001	9.214 ± 0.711

Table D.2 Gradient Boost Tree with FE+SVM.

Model	Number of tree	Maximal depth	Learning rate	RMSE
GBT	200	10	0.1	5.517 ± 0.699
GBT	500	21	0.01	5.605 ± 0.716
GBT	500	16	0.001	7.392 ± 0.495
GBT	500	16	0.0001	9.204 ± 0.718

Table D.3 Gradient Boost Tree with FE+DL.

Model	Number of tree	Maximal depth	Learning rate	RMSE
GBT	200	11	0.1	5.544 ± 0.652
GBT	500	20	0.01	5.611 ± 0.621
GBT	500	18	0.001	7.383 ± 0.481
GBT	500	16	0.0001	9.203 ± 0.717

Table D.4 Gradient Boost Tree with FE+GBT.

Model	Number of tree	Maximal depth	Learning rate	RMSE
GBT	200	11	0.1	5.316 ± 0.539
GBT	500	13	0.01	5.429 ± 0.650
GBT	500	17	0.001	7.307 ± 0.520
GBT	500	15	0.0001	9.196 ± 0.720

CURRICULUM VITAE

NAME : Saranchai Sinlapasorn

GENDER : Male

EDUCATION BACKGROUND:

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

SCHOLARSHIP:

- Kittibandit Scholarship for graduate honor student of Suranaree University of Technology.

CONFERENCE:

- The 25th Annual Meeting in Mathematics, King Mongkut's Institute of Technology Ladkrabang, Bangkok, May 27th-29th, 2021.

EXPERIENCE:

- Teaching assistant in Suranaree University of Technology, Term 2019-2020.

CURRICULUM VITAE

NAME : Saranchai Sinlapasorn

GENDER : Male

EDUCATION BACKGROUND:

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

SCHOLARSHIP:

- Kittibandit Scholarship for graduate honor student of Suranaree University of Technology.

CONFERENCE:

- The 25th Annual Meeting in Mathematics, King Mongkut's Institute of Technology Ladkrabang, Bangkok, May 27th-29th, 2021.

EXPERIENCE:

- Teaching assistant in Suranaree University of Technology, Term 2019-2020.

CURRICULUM VITAE

NAME : Saranchai Sinlapasorn

GENDER : Male

EDUCATION BACKGROUND:

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

SCHOLARSHIP:

- Kittibandit Scholarship for graduate honor student of Suranaree University of Technology.

CONFERENCE:

- The 25th Annual Meeting in Mathematics, King Mongkut's Institute of Technology Ladkrabang, Bangkok, May 27th-29th, 2021.

EXPERIENCE:

- Teaching assistant in Suranaree University of Technology, Term 2019-2020.

CURRICULUM VITAE

NAME : Saranchai Sinlapasorn

GENDER : Male

EDUCATION BACKGROUND:

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

SCHOLARSHIP:

- Kittibandit Scholarship for graduate honor student of Suranaree University of Technology.

CONFERENCE:

- The 25th Annual Meeting in Mathematics, King Mongkut's Institute of Technology Ladkrabang, Bangkok, May 27th-29th, 2021.

EXPERIENCE:

- Teaching assistant in Suranaree University of Technology, Term 2019-2020.

CURRICULUM VITAE

NAME : Saranchai Sinlapasorn

GENDER : Male

EDUCATION BACKGROUND:

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

SCHOLARSHIP:

- Kittibandit Scholarship for graduate honor student of Suranaree University of Technology.

CONFERENCE:

- The 25th Annual Meeting in Mathematics, King Mongkut's Institute of Technology Ladkrabang, Bangkok, May 27th-29th, 2021.

EXPERIENCE:

- Teaching assistant in Suranaree University of Technology, Term 2019-2020.

CURRICULUM VITAE

NAME : Saranchai Sinlapasorn

GENDER : Male

EDUCATION BACKGROUND:

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

SCHOLARSHIP:

- Kittibandit Scholarship for graduate honor student of Suranaree University of Technology.

CONFERENCE:

- The 25th Annual Meeting in Mathematics, King Mongkut's Institute of Technology Ladkrabang, Bangkok, May 27th-29th, 2021.

EXPERIENCE:

- Teaching assistant in Suranaree University of Technology, Term 2019-2020.