การเพิ่มความถูกต้องในการทำนายค่าสารส้มในกระบวนการรวมตะกอนโดยใช้เหมืองข้อมูล กรณีศึกษา โรงผลิตน้ำประปาบางเขน, กรุงเทพมหานคร, ประเทศไทย

Enhancing Alum Dosage Prediction Accuracy in Coagulation Process by Data Mining Software: A Case Study of Bang khen Water Supply Plant, Bangkok, Thailand

อภิสิทธิ์ ทิพย์นาง

Aphisit Thipnang

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมสิ่งแวดล้อมและพลังงานเพื่อความยั่งยืน
คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ.2561

KMITL-2018-EN-M-167-137

การเพิ่มความถูกต้องในการทำนายค่าสารส้มในกระบวนการรวมตะกอนโดยใช้เหมืองข้อมูล กรณีศึกษา โรงผลิตน้ำประปาบางเขน, กรุงเทพมหานคร, ประเทศไทย

Enhancing Alum Dosage Prediction Accuracy in Coagulation Process by Data Mining Software: A Case Study of Bang khen Water Supply Plant, Bangkok, Thailand

อภิสิทธิ์ ทิพย์นาง

Aphisit Thipnang

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมสิ่งแวดล้อมและพลังงานเพื่อความยั่งยืน

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2561

KMITL-2018-EN-M-167-137

Enhancing Alum Dosage Prediction Accuracy in Coagulation Process by Data Mining Software: A Case Study of Bang khen Water Supply Plant, Bangkok, Thailand

Aphisit Thipnang

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF ENGINEERING IN ENVIRONMENTAL AND ENERGY ENGINEERING

FOR SUSTAINABILITY

FACULTY OF ENGINEERING

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2018

KMITL-2018-EN-M-167-137

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์	การเพิ่มความถูกต้องในการทำนายค่าสารส้มในกระบวนการรวมตะกอนโดยใช้เหมืองข้อมูล กรณีศึกษา โรงผลิตน้ำประปาบางเขน, กรุงเทพมหานคร, ประเทศไทย
Thesis Title	Enhancing Alum Dosage Prediction Accuracy in Coagulation Process by Data Mining Software: A Case Study of Bang khen Water Supply Plant, Bangkok, Thailand
ชื่อนักศึกษา	นายอภิสิทธิ์ ทิพย์นาง
รหัสประจำตัว	59601229
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมสิ่งแวดล้อมและพลังงานเพื่อความยั่งยืน
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผศ.ดร.ภาสกร ขันทองทิพย์
หมายเลขวิทยานิพนธ์	KMITL-2018-EN-M-167-137

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
พ.ต.ดร.เสกสรร	หมอยาดี	
ผศ.ดร.ชลิดา	อู่ตะเภา	
ผศ.ดร.วุฒิชัย	ชาติพัฒนานันท์	
ผศ.ดร.ภาสกร	ขันทองทิพย์	

วัน/เดือน/ปี ที่สอบ วันพฤหัสบดีที่ 26 กรกฎาคม พ.ศ. 2561 เวลา 11.00 - 13.00 น.

สถานที่สอบ ณ ห้องประชุม 4 ชั้น 5 อาคาร A

คณะวิศวกรรมศาสตร์ รับรองแล้ว

(รองศาสตราจารย์.ดร. คมสัน มาลิสี)

คณบดี คณะวิศวกรรมศาสตร์

หัวข้อวิทยานิพนธ์	การเพิ่มความถูกต้องในการทำนายค่าสารส้มในกระบวนการรวมตะกอนโดยใช้เหมืองข้อมูล กรณีศึกษา โรงผลิตน้ำประปาบางเขน, กรุงเทพมหานคร, ประเทศไทย
นักศึกษา	นายอภิสิทธิ์ ทิพย์นาง
รหัสประจำตัว	59601229
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมสิ่งแวดล้อมและพลังงานเพื่อความยั่งยืน
พ.ศ.	2561
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผศ.ดร.ภาสกร ขันทองทิพย์

บทคัดย่อ

งานวิจัยนี้นำเสนอการทำนายปริมาณสารส้มที่ใช้ในกระบวนการผลิตน้ำประปา โดยใช้โปรแกรม Weka version 3.6 เพื่อสร้างแบบจำลองทางคณิตศาสตร์ใช้ทำนายปริมาณสารส้ม ในงานวิจัยนี้ได้เปรียบเทียบผลลัพธ์การทำนายด้วยค่า Root Mean Square Error (RMSE), Correlation Coefficient (CC) และ R-square (R^2) จากแบบจำลองที่สร้างจาก REPTree, M5P, M5Rules และ Multilayer Perceptron (MLP) เพื่อหาว่าแบบจำลองใดสามารถทำนายได้แม่นยำใกล้เคียงค่าจริงที่สุด ตัวแปรต้นที่ใช้ในการศึกษามีจำนวน 3 พารามิเตอร์ ได้แก่ ค่าความขุ่น (Turbidity), ค่าปริมาณของแข็งแขวนลอย (Total Suspended Solids) และ พีเอช (pH) ตัวแปรตามคือ ปริมาณการเติมสารส้ม ข้อมูลที่ใช้ในการสร้างแบบจำลองเป็นข้อมูลน้ำดิบขาเข้า รวมรวมข้อมูลจากโรงผลิตน้ำประปาบางเขน กรุงเทพมหานคร ตั้งแต่วันที่ 1 มกราคม 2549 ถึงวันที่ 31 กรกฎาคม 2558 โดยข้อมูลทั้งหมดมีจำนวน 3,500 ชุด ได้ทำการแบ่งข้อมูลออกเป็น 3 ชุด คือ ฤดูร้อน ฤดูฝนและฤดูหนาว และแต่ในละชุดได้แบ่งการวิเคราะห์ออกเป็น 3 แบบ คือ Percentage Split 90%, Percentage Split 75% และ Percentage Split 50% ผลการวิเคราะห์แบบจำลองพบว่า แบบจำลองที่ให้ค่า RMSE ที่ต่ำที่สุด ให้ค่า CC และ R^2 สูงที่สุด คือ REPTree ชุดข้อมูลฤดูฝน โดยใช้ Percentage Split 75% ให้ค่า RMSE = 8.1006, CC = 0.8036 และ R^2 = 0.6457 แต่เมื่อนำไปประยุกต์ในการทำนายค่าการเติมสารปริมาณสารส้มในโรงผลิตน้ำประปาอีกแห่ง คือโรงผลิตน้ำประปาธนบุรี พบว่าแบบจำลองของแต่ละที่ ไม่สามารถใช้ร่วมกันได้อย่างมีประสิทธิภาพ โดยมีค่าดั้งนี้ RMSE = 23.0491, CC = - 0.3785 และ R^2 = 0.1433

Thesis Title	Enhancing Alum Dosage Prediction Accuracy in Coagulation Process by Data Mining Software: A Case Study of Bang khen Water Supply Plant, Bangkok, Thailand
Student	Mr. Aphisit Thipnang
Student ID.	59601229
Degree	Master of Engineering
Program	Environmental and Energy Engineering for Sustainability
Year	2018
Thesis Advisor	Asst. Prof. Dr.Passkorn Khanthongthip

ABSTRACT

This research presents an alum dosage prediction in coagulation process by using Weka Data Mining Software version 3.6. for construct a mathematical model to predict alum dosage by using Root Mean Square Error (RMSE), Correlation Coefficient (CC) และ R-square (R^2) to measure a precision from the model that created from REPTree, M5P, M5Rules and Multilayer Perceptron (MLP) to find out which model can predict precisely the closest actual value. The independent variables used in the study were 3 parameters: Turbidity, Total Suspended Solids and pH, the dependent variable: alum dosage. The input data used in the modeling is from Bangkhen Water Supply Plant, Bangkok from 1 January 2006 – 31 July 2015, data collected is 3,500 records. The data is divided to 3 groups: Summer, Rainy and Winter, and in each group, the analysis is divided into 3 types: Percentage Split 90%, Percentage Split 75% and Percentage Split 50%. The results of the model analysis revealed that model with the lowest RMSE values and highest CC and R^2 is REPTree Rainy group with Percentage Split 75% gave RMSE = 8.1006, CC = 0.8036 R^2 = 0.6457. However, when applied to predict alum dosage in another water supply plant: Thonburi Water Treatment Plant. Find out which model of each, cannot be used together effectively, it gave RMSE = 23.0491, CC = - 0.3785 and R^2 = 0.1433

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จได้เป็นอย่างดี ด้วยความช่วยเหลือ จาก ผศ.ดร.ภาสกร ขันทองทิพย์ และรศ.ดร.เพ็ชรพร เชาวกิจเจริญ ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ที่คอยให้คำแนะนำและคำปรึกษาในการทำงานมาโดยตลอด ข้าพเจ้ารู้สึกซาบซึ้งในความอนุเคราะห์จากท่านอาจารย์

ขอขอบพระคุณเป็นอย่างสูง ขอขอบพระคุณคณาจารย์สาขาวิชาวิศวกรรมสิ่งแวดล้อมและพลังงานเพื่อความยั่งยืน คณะวิศวกรรมศาสลร์ สถาบันเทคโนโลยี พระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้ และถ่ายทอด ประสบการณ์ที่ดีให้แก่ข้าพเจ้า พร้อมทั้งได้ให้คำปรึกษาคำแนะนำ และข้อคิดเห็นต่างๆที่เป็น ประโยชน์ จนทำให้วิทยานิพนธ์นี้ได้บรรลุวัตถุประสงค์ได้ด้วยดี

ขอขอบคุณเพื่อน พี่ น้อง ทุกคนที่ให้คำแนะนำและคอยช่วยเหลือในการทำงาน พร้อมมอบกำลังใจให้เสมอมาจนสามารถฟันฝ่าอุปสรรคมาได้ด้วยดี

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณ บิดา มารดา ที่ให้การเลี้ยงดูอย่างดีตลอดมา และยังให้การสนับสนุน ทั้งกำลังทรัพย์ที่เป็นค่าใช้จ่ายในด้านล่างๆ พร้อมกับมอบกำลังใจที่ดี เมื่อยามท้อและเหน็ดเหนื่อย จนทำให้การทำงานสำเร็จไปได้ด้วยดีสำหรับคุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับผู้มีพระคุณทุกท่านซึ่งเป็นที่รักและเคารพยิ่ง

อภิสิทธิ์ ทิพย์นาง

สารบัญ

หน้า

บทคัดย่อภาษาไทย I
บทคัดย่อภาษาอังกฤษ II
กิตติกรรมประกาศ III
สารบัญ IV
สารบัญตาราง VII
สารบัญรูป IX
บทที่ 1 บทนำ 1
1.1 กล่าวนำ 1
1.2 ความสำคัญและที่มาของปัญหา 1
1.3 วัตถุประสงค์ของการศึกษา 2
1.4 ขอบเขตการศึกษา 2
1.5 ขั้นตอนการดำเนินงาน 2
1.6 ประโยชน์ที่คาดว่าจะได้รับ 3
บทที่ 2 ทฤษฎีที่เกี่ยวข้องและวรรณกรรมปริทัศน์ 4
2.1 กระบวนการสร้างตะกอน (Coagulation) และกระบวนการรวมตะกอน (Flocculation) 4
2.2 กลไกโคแอกกูเลชั่นด้วยสารส้ม 7
2.3 เหมืองข้อมูล (Data Mining) 10
2.4 งานวิจัยที่เกี่ยวข้อง 20
2.4.1 สรุปผลงานวิจัยที่เกี่ยวข้อง 36
บทที่ 3 วิธีดำเนินการ 37
3.1 แผนผังการดำเนินงานวิจัย 37
3.2 การเลือกโรงประปาที่เหมาะสม 38
3.3 แนวคิดในการคัดเลือกทฤษฎีที่นำมาใช้สังเคราะห์ข้อมูล 38
3.4 สังเคราะห์ข้อมูลจากทฤษฎีที่เลือกมาโดยใช้ Weka Software 40
3.5 การทดสอบความแม่นยำของแบบจำลองที่ได้ 48
บทที่ 4 ผลการทำลอง 52
4.1 กล่าวนำ 52
4.2 การปรับแต่งตัวแปรของทฤษฎีที่ใช้สร้างแบบจำลอง 52

สารบัญ (ต่อ)

หน้า

4.2.1 REPTree 52
4.2.2 M5P 54
4.2.3 M5Rules 55
4.2.4 Multilayer Perceptron(MLP) 57
4.3 การทดสอบแบบจำลองกับข้อมูลจริง 59
4.3.1 REPTree 60
4.3.1.1 ชุดฤดูฝน 60
4.3.1.2 ชุดฤดูหนาว 60
4.3.1.3 ชุดฤดูร้อน 60
4.3.2 M5P 61
4.3.2.1 ชุดฤดูฝน 61
4.3.2.2 ชุดฤดูหนาว 61
4.3.2.3 ชุดฤดูร้อน 62
4.3.3 M5Rules 62
4.3.3.1 ชุดฤดูฝน 62
4.3.3.2 ชุดฤดูหนาว 63
4.3.3.3 ชุดฤดูร้อน 63
4.3.4 Multilayer Perceptron 64
4.3.4.1 ชุดฤดูฝน 64
4.3.4.2 ชุดฤดูหนาว 64
4.3.4.3 ชุดฤดูร้อน 64
4.4 เปรียบเทียบแบบจำลองที่ได้จากการวิเคราะห์ของแต่ละทฤษฎี 65
4.5 นำแบบจำลองที่ดีที่สุดของแต่ละทฤษฎี ไปทำนายค่าการเติมปริมาณสารส้ม 66
4.5.1 การนำแบบจำลองทฤษฎี REPTree มาทำนายค่าสารส้มจริง 66
4.5.2 การนำแบบจำลองทฤษฎี M5P มาทำนายค่าสารส้มจริง 66
4.5.3 การนำแบบจำลองทฤษฎี M5Rules มาทำนายค่าสารส้มจริง 67
4.5.4 การนำแบบจำลองทฤษฎี MLP มาทำนายค่าสารส้มจริง 68
4.6 นำแบบจำลองที่ดีที่สุด ไปทำนายค่าการเติมปริมาณสารส้ม ในพื้นที่อื่น 68
4.7 การเพิ่มประสิทธิภาพ โดยใช้โมเดลจากซอฟแวร์ตัวอื่น 70

สารบัญ (ต่อ)

หน้า

บทที่ 5 วิเคราะห์ผลงานวิจัย 73

5.1 วิเคราะห์ข้อดีและข้อเสียของแต่ละทฤษฎี 73

5.2 วิเคราะห์ความแม่นยำจากการทำนายว่าสามารถนำมาใช้จริงได้หรือไม่ 75

5.3 วิเคราะห์การทำนายโดยใช้ ข้อมูลทดสอบ (Testing set) ที่มาจากโรงผลิตน้ำประปาประปาธนบุรี 76

5.4 วิเคราะห์ผลการเพิ่มประสิทธิภาพ โดยใช้โมเดลจากซอฟแวร์ตัวอื่น 76

บทที่ 6 สรุปผลงานวิจัยและข้อเสนอแนะ

6.1 สรุปผลการวิจัย 78

6.2 ข้อเสนอแนะ 78

บรรณานุกรม 80

ภาคผนวก 82

ภาคผนวก ก ข้อมูลผลการทดลอง 83

ภาคผนวก ข บทความงานวิจัยที่ได้รับตีพิมพ์เผยแพร่ 104

ประวัติผู้เขียน 109

สารบัญตาราง

ตารางที่		หน้า
2.1	แสดงค่า R-Square และ MAE ของแต่ละพารามิเตอร์	20
2.2	ตารางแสดงค่า RMSE ของปริการเติมสารส้มในแบบจำลอง	30
2.3	การเปรียบเทียบค่าจากการทดสอบแบบจำลอง	31
2.4	ตารางแสดงค่า จากการทำนายข้อมูลโดยใช้แบบจำลองจากข้อมูล กลุ่มที่ 2	34
2.5	การเปรียบเทียบค่าจากการทดสอบแบบจำลองของ 3 ทฤษฎี	36
4.1	แสดงผลการปรับแต่งตัวแปรเพื่อสร้างแบบจำลองของทฤษฎี REPTree	53
4.2	ผลจากการปรับแต่งตัวแปรจากทฤษฎี REPTree ที่ดีที่สุด	53
4.3	แสดงผลการปรับแต่งตัวแปรเพื่อสร้างแบบจำลองของทฤษฎี M5P	54
4.4	ผลจากการปรับแต่งตัวแปรจากทฤษฎี M5P ที่ดีที่สุด	55
4.5	แสดงผลการปรับแต่งตัวแปรเพื่อสร้างแบบจำลองของทฤษฎี M5P	56
4.6	ผลจากการปรับแต่งตัวแปรจากทฤษฎี M5Rules ที่ดีที่สุด	56
4.7	แสดงผลการปรับแต่งตัวแปรเพื่อสร้างแบบจำลองของทฤษฎี MLP	58
4.8	ผลจากการปรับแต่งตัวแปรจากทฤษฎี MLP ที่ดีที่สุด	59
4.9	ผลการสังเคราะห์ที่ดีที่สุดของแต่ละทฤษฎี	59
4.10	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูฝนจากแบบวิธี REPTree	60
4.11	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูหนาวจากวิธี REPTree	60
4.12	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูร้อนจากวิธี REPTree	60
4.13	ตารางเปรียบเทียบระหว่างแต่ละฤดูกาลของทฤษฎี REPTree	61
4.14	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูฝนจากแบบวิธี M5P	61
4.15	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูหนาวจากวิธี M5P	61
4.16	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูร้อนจากวิธี M5P	62
4.17	ตารางเปรียบเทียบระหว่างแต่ละฤดูกาลของทฤษฎี M5P	62
4.18	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลหน้าฝนจากแบบวิธี M5Rules	62
4.19	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลหน้าหนาวจากวิธี M5Rules	63
4.20	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลหน้าร้อนจากวิธี M5Rules	63
4.21	ตารางเปรียบเทียบระหว่างแต่ละฤดูกาลของทฤษฎี M5Rules	63
4.22	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลหน้าฝนจากแบบวิธี MLP	64
4.23	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลหน้าหนาวจากวิธี MLP	64
4.24	ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลหน้าร้อนจากวิธี MLP	64

สารบัญตาราง(ต่อ)

ตารางที่		หน้า
4.25	ตารางเปรียบเทียบระหว่างแต่ละฤดูกาลของทฤษฎี MLP	65
4.26	ตารางเปรียบเทียบแบบจำลองที่ดีที่สุดของแต่ละทฤษฎี	65
4.27	ตารางเปรียบเทียบค่าจากการทำนายปริมาณการเติมสารส้ม ของโรงผลิตน้ำประปาในพื้นที่อื่น	69
5.1	ผลการสังเคราะห์แบบจำลองจาก ทฤษฎี REPTree	73
5.2	ผลการสังเคราะห์แบบจำลองจาก ทฤษฎี M5P	74
5.3	ผลการสังเคราะห์แบบจำลองจาก ทฤษฎี M5Rules	74
5.4	ผลการสังเคราะห์แบบจำลองจาก ทฤษฎี Multilayer Perceptron	75
5.5	ค่าแสดงประสิทธิภาพของซอฟท์แวร์ตัวเสริม	76

สารบัญรูป

รูปที่		หน้า
2.1	การเปรียบเทียบปริมาณโคแอกกูแลนท์ ที่ใช้ในการทำลายเสถียรภาพของคอลลอยด์ด้วยกลไกแบบต่างๆ	5
2.2	เกณฑ์ที่เหมาะสมสำหรับการสร้างสัมผัสระหว่างอนุภาคต่างๆ ทั้ง 5 ประเภท	6
2.3	ความสัมพันธ์ระหว่างสารประกอบเชิงซ้อนสารส้ม และค่าพีเอช	8
2.4	กลไกในการสร้างโคแอกกูเลชั่นด้วยสารส้ม	9
2.5	ไดอะแกรมที่ใช้ในการออกแบบและควบคุมโคแอกกูเลชั่นด้วยสารส้ม	10
2.6	โครงสร้าง Neuro Network	16
2.7	การนำไปประยุกต์ใช้ ของ Data Mining	19
2.8	กราฟเปรียบเทียบระหว่าง ค่าจริง และค่าจากแบบจำลองของแต่ละพารามิเตอร์	21
2.9	โครงสร้างของ Neural Network	22
2.10	แบบแผนขั้นตอนการทำนาย	22
2.11	กราฟเปรียบเทียบค่า TSS จากการทำนาย และค่าที่ต้องการ	23
2.12	แผนภาพแสดงการทำงานของระบบ Adaptive Network-Based Fuzzy Inference System	23
2.13	กราฟระหว่างข้อมูลจากค่าจริงค่าจากการทำนายของ ANFIS	24
2.14	โครงสร้างของ RBFNN	24
2.15	กราฟเปรียบเทียบคุณภาพน้ำระหว่างค่าจริงกับค่าทำนาย	25
2.16	กราฟแสดงค่าความคลาดเคลื่อนของ BOD ระหว่างค่าจริงกับค่าจากการทำนาย	25
2.17	กราฟเปรียบเทียบความแม่นยำของค่าทำนายและค่าจริง	26
2.18	แผนภาพการดำเนินการของระบบเติมสารส้มอัตโนมัติ	26
2.19	กราฟข้อมูลจากค่าจริง และค่าจากการทำนายของ ANFIS	27
2.20	แผนภาพการดำเนินการสร้างแบบจำลองและทำนายผล (Pinar, 2014)	28
2.21	กราฟเปรียบเทียบพลังงานไฟฟ้าสูงสุดที่ผลิตได้ระหว่างค่าจริงกับค่าจากการทำนาย	29
2.22	กราฟระหว่างค่าสารส้มจริง กับค่าจากแบบจำลอง	29
2.23	กราฟระหว่างค่าใช้จ่ายจริง และค่าใช้จ่ายจากแบบจำลอง	30
2.24	กราฟเปรียบเทียบแบบจำลองโดยใช้ข้อมูล กลุ่มที่ 1	32
2.25	กราฟเปรียบเทียบแบบจำลองโดยใช้ข้อมูล กลุ่มที่ 2	32
2.26	กราฟเปรียบเทียบค่าจากการทำนายจริงโดยใช้แบบจำลองที่สร้างจากข้อมูลกลุ่มที่ 1	33

สารบัญรูป(ต่อ)

รูปที่		หน้า
2.27	กราฟแสดงผลการทำนาย ระหว่างค่าจริง และค่าจากแบบจำลอง ทั้ง 2 กลุ่ม	34
2.28	กราฟความสัมพันธ์สูงที่สุดระหว่างฝุ่นควันกับทิศทางของลม	35
2.29	กราฟความสัมพันธ์ต่ำที่สุดระหว่างฝุ่นควันกับอุณหภูมิ	35
3.1	แผนผังการดำเนินงานวิจัย	37
3.2	แบบจำลอง REPTree	38
3.3	แบบจำลอง M5P	39
3.4	แบบจำลอง Multilayer Perceptron	40
3.5	Weka data mining Software	40
3.6	ข้อมูลที่ไม่เกี่ยวข้องกับตัวแปร เช่นวันที่ในวงกลมสีแดง	42
3.7	หน้าโปรแกรมที่สามารถเปิดไฟล์ฐานข้อมูลได้	43
3.8	หน้าโปรแกรมหลังจากตัวแปรที่ไม่ต้องการถูก Remove ออกไปแล้ว	43
3.9	การเลือก Classifier ที่ต้องการ	44
3.10	การปรับแต่งแบบจำลองของวิธี Multilayer Perceptron	45
3.11	การปรับแต่งแบบจำลองของวิธี Multilayer Perceptron	46
3.12	การเปิดไฟล์ข้อมูลที่ใช้ทำนาย	46
3.13	เลือก Output prediction	47
3.14	เลือก Re-evaluate model on current test set	47
3.15	แสดงผลการทำนายค่า Alum Dosage	48
3.16	การจัดข้อมูลเพื่อหา R-square	49
3.17	การสร้างกราฟแบบ Scatter	50
3.18	การสร้าง Trendline	50
3.19	กราฟแสดงค่า R-square	51
4.1	หน้าต่างแสดงตัวแปรของทฤษฎี REPTree	52
4.2	หน้าต่างแสดงตัวแปรของทฤษฎี M5P	54
4.3	หน้าต่างแสดงตัวแปรของทฤษฎี M5Rules	55
4.4	หน้าต่างแสดงตัวแปรของทฤษฎี MLP	57
4.5	กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี REPTree	66
4.6	กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี M5P	67
4.7	กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี M5P	67

สารบัญรูป(ต่อ)

รูปที่		หน้า
4.8	กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี MLP	68
4.9	กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี REPTree ชุดข้อมูลฤดูฝน ทดสอบกับข้อมูลจาก โรงผลิตน้ำประปาธนบุรี	69
4.10	กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี M5P ชุดข้อมูลฤดูหนาว ทดสอบกับข้อมูลจาก โรงผลิตน้ำประปาธนบุรี	70
4.11	การทำนายค่าปริมาณการเติมสารส้มโดย Extra Tree	71
4.12	กราฟเปรียบเทียบค่าจริงกับค่าจากการทำนายโดย Extra Tree เพื่อหาค่า R-square	71
4.13	การทำนายค่าปริมาณการเติมสารส้มโดย REPTree	72
4.14	กราฟเปรียบเทียบค่าจริงกับค่าจากการทำนายโดย REPTree เพื่อหาค่า R-square	72

บทที่ 1
บทนำ

1.1 กล่าวนำ

ปัจจุบันประเทศไทยมีการบริโภคทรัพยากรธรรมชาติมากขึ้นทุกวัน ไม่ว่าจะเป็นทรัพยากรสัตว์ ทรัพยากรป่าไม้ โดยเฉพาะอย่างยิ่งทรัพยากรน้ำ ซึ่งได้มีการบริโภคอยู่ตลอดเวลาในกิจกรรมหลายชนิด เช่น การบริโภคในครัวเรือน การใช้ในอุตสาหกรรม น้ำเป็นสิ่งที่ขาดไม่ได้ในชีวิตประจำวันของมนุษย์ ไม่ว่าจะเป็นทางตรงหรือทางอ้อม ผู้ทำวิจัยได้เล็งเห็นความสำคัญของการบริโภคน้ำ ซึ่งมีอัตราการบริโภคมากขึ้นทุกวัน ดังนั้นในการผลิตน้ำจำเป็นต้องมีเทคโนโลยีในการผลิต ที่ทำให้การผลิตน้ำสะอาดมีอัตราที่เร็วขึ้น เพื่อผลผลิตที่มากขึ้นตามปริมาณการใช้ เทคโนโลยีที่นำมาใช้ในการผลิตน้ำสะอาดปัจจุบันนั้นมีหลากหลายรูปแบบโดยแต่ละรูปแบบขึ้นอยู่กับประเภทของน้ำดิบที่ต้องการจะผลิต ตัวอย่างเช่น

ก.กระบวนการสร้างตะกอน (Coagulation)

ข.กระบวนการรวมตะกอน (Flocculation)

ค.กระบวนการการตกตะกอน (Sedimentation)

ง.กระบวนการกรอง (Filtration)

จ.กระบวนการฆ่าเชื้อโรค (Disinfection)

แม้กระบวนการผลิตน้ำประปาในข้างต้นมีมากมาย หลายกระบวนการกว่าจะได้ผลิตภัณฑ์เป็นน้ำประปาออกมา แต่มีอยู่อย่างหนึ่งที่กระบวนการส่วนใหญ่มีเหมือนกันคือต้องมีการคำนวณ หรือทำการทดลอง เพื่อที่จะหาปริมาณสารเคมีที่ต้องเติมลงในน้ำ สำหรับการทำให้เกิดกระบวนการต่าง ๆ ในข้างต้นที่กล่าวมา ผู้วิจัยได้เล็งเห็นความสูญเสียในด้านนี้ เช่น เวลาที่เสียไปกับการทำการทดลอง และ สารเคมีที่สูญเสียในการทดลอง ซึ่งความสูญเสียที่กล่าวมามีการเกิดขึ้นทุกวัน ในโรงประปาทุกแห่ง ผู้วิจัยจึงมีความคิดที่จะนำ เทคโนโลยีที่เรียกว่า “การทำเหมืองข้อมูล (Data Mining)” เข้ามาช่วยในการวิเคราะห์ข้อมูลและสามารถทำนายปริมาณสารเคมีที่ต้องเติมลงไปในน้ำดิบในเวลาอันสั้น

1.2 ความสำคัญและที่มาของปัญหา

การใช้เทคโนโลยี การทำเหมืองข้อมูล (Data Mining) ได้ถูกนำไปใช้วิเคราะห์ข้อมูลในหลายด้าน ไม่ว่าจะเป็น ด้านการแพทย์ ด้านวิศวกรรม รวมไปถึงด้านสิ่งแวดล้อม งานวิจัยนี้ได้ทำการศึกษา กระบวนการสร้างตะกอน (Coagulation) เป็นหลัก ซึ่งในแต่ละวันของโรง

ผลิตน้ำประปาส่วนใหญ่นั้นได้มีการทำการทดลอง ที่เรียกว่า “จาร์เทสต์ (Jar-Test)” เพื่อหาปริมาณสารเคมีที่ต้องใช้ในการสร้างตะกอน โดยการเติมสารเคมีต่างความเข้มข้นลงในน้ำดิบในภาชนะแต่ละใบ แล้วกวนน้ำโดยเครื่อง จากนั้นสังเกตการตกตะกอน ตลอดจนขนาดของตะกอน ปล่อยน้ำทิ้งไว้ 1 ชั่วโมง จากนั้นวัดค่า ความขุ่น (Turbidity), ค่าของแข็งแขวนลอย (Suspended Solids) และ พีเอช (pH) เพื่อหาสภาวะที่เหมาะสม และได้ความเข้มข้นของสารส้มที่ต้องนำไปเติมในน้ำดิบจริง จะเห็นได้ว่าวิธีการทำ จาร์เทสต์ (Jar-Test) นั้นมีต้องใช้เวลามาก และสารเคมีเป็นจำนวนมาก

ดังนั้นถ้ามีเทคโนโลยีที่ช่วยให้การสูญเสียข้างต้นลดลง ก็จะเป็นประโยชน์ต่อการผลิตน้ำประปาเป็นอย่างมาก ซึ่งก็คือ การทำเหมืองข้อมูล (Data Mining) โดยจะทำการหาความสัมพันธ์ระหว่างตัวแปร จากข้อมูลน้ำดิบที่เก็บบันทึกในอดีต แล้วนำความสัมพันธ์มาสร้างเป็นแบบจำลอง (Model) เพื่อที่จะทำการทำนายค่าที่ต้องการจากความสัมพันธ์ของตัวแปร

1.3 วัตถุประสงค์ของการศึกษา

1. เพื่อทราบถึงความสัมพันธ์ของตัวแปรแต่ละตัวในน้ำดิบ ที่มีต่อปริมาณการเติมสารส้มในกระบวนการผลิตน้ำประปา
2. เพื่อพัฒนาการทำนายค่าสารส้มที่ต้องเติมลงไปเพื่อให้เกิดกระบวนการสร้างตะกอน (Coagulation) ให้มีความเป็นไปได้ในอนาคต

1.4 ขอบเขตการศึกษา

ในการศึกษาครั้งนี้จะทำการสร้างแบบจำลอง (Model) ของตัวแปรจากน้ำดิบ เพื่อทำนายปริมาณสารส้มที่เหมาะสมที่ต้องเติมในอนาคต โดยแบบจำลอง สร้างจากข้อมูลจากน้ำดิบ ของโรงประปาบางเขน กรุงเทพมหานคร โดยจะมีการนำแบบจำลองไปใช้ในการทำนายค่าการเติมปริมาณสารส้มในโรงผลิตน้ำประปาแห่งอื่นอีกด้วย

1.5 ขั้นตอนการดำเนินงาน

1. ศึกษาความสัมพันธ์ของตัวแปรในแต่ละตัว
2. ศึกษาหลักการทำงานของ การทำเหมืองข้อมูล (Data Mining)
3. เลือกโรงผลิตน้ำประปาที่เหมาะสม และทำการยื่นขอข้อมูล เพื่อทำการวิจัย
4. คัดกรองข้อมูล
5. ศึกษาทฤษฎีที่เหมาะสม เพื่อที่จะนำมาวิเคราะห์ ข้อมูลตัวแปรจากน้ำดิบ
6. วิเคราะห์ข้อมูล

7. เก็บรวบรวมผลการวิเคราะห์ข้อมูลที่ได้จากแต่ละทฤษฎี

8. เปรียบเทียบผล

9. นำแบบจำลองที่ดีที่สุดมาทำการทำนายผล

10. วิเคราะห์และสรุปผลการทดลอง

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. มีความเข้าใจในการทำเหมืองข้อมูล (Data Mining) จนสามารถนำไปประยุกต์ใช้จริงได้ในอนาคต

2. สามารถนำแบบจำลองของการสร้างตะกอน ไปทำนายค่าที่ต้องการได้ และมีความแม่นยำกว่างานวิจัยในอดีต

บทที่ 2
ทฤษฎีที่เกี่ยวข้องและวรรณกรรมปริทัศน์

2.1 กระบวนการสร้างตะกอน (Coagulation) และ กระบวนการรวมตะกอน (Flocculation)

อนุภาคขนาดเล็ก ซึ่งเรียกว่าอนุภาคคอลลอยด์ โดยทั่วไปมีขนาดอนุภาคอยู่ในช่วง 10^{-5} จนถึง 10^{-3} มิลลิเมตร เนื่องจากมีขนาดเล็กจึงไม่สามารถตกตะกอนได้ด้วยน้ำหนักของตัวมันเองในเวลาจำกัด นอกจากนี้อนุภาคคอลลอยด์เมื่ออยู่ในน้ำจะมีประจุประจำตัว โดยพวกที่ชอบน้ำ (Hydrophilic) จะมีประจุบวก ส่วนพวกที่ไม่ชอบน้ำ (Hydrophobic) มักจะเป็นประจุลบ และเนื่องจากอนุภาคดังกล่าวมีประจุที่ทำให้อนุภาคที่ชนิดเดียวกันเกิดแรงผลักระหว่างอนุภาค ทำให้อนุภาคเหล่านั้นมีเสถียรภาพสูง ดังนั้นการทำให้อนุภาคต่างๆรวมกันและจับตัวกันเป็นก้อนมีขั้นตอนดังนี้ (มั่นสิน ตัณฑุลเวศม์, 2537)

2.1.1 กลไกการลดความหนาของชั้นกระจาย (Diffuse Layer) โดยการเพิ่มประจุตรงกันข้ามกับ คอลลอยด์ในชั้นกระจายให้มากขึ้น ซึ่งจะทำให้ค่าศักดิ์ไฟฟ้า (Zeta Potential) ที่ ผิวนอกสุดของน้ำลดตามไปด้วย การทำลายเสถียรภาพโดยการลดความ หนาของชั้นกระจายด้วยการเติมสารละลายของเกลือต่างๆ มีดังนี้

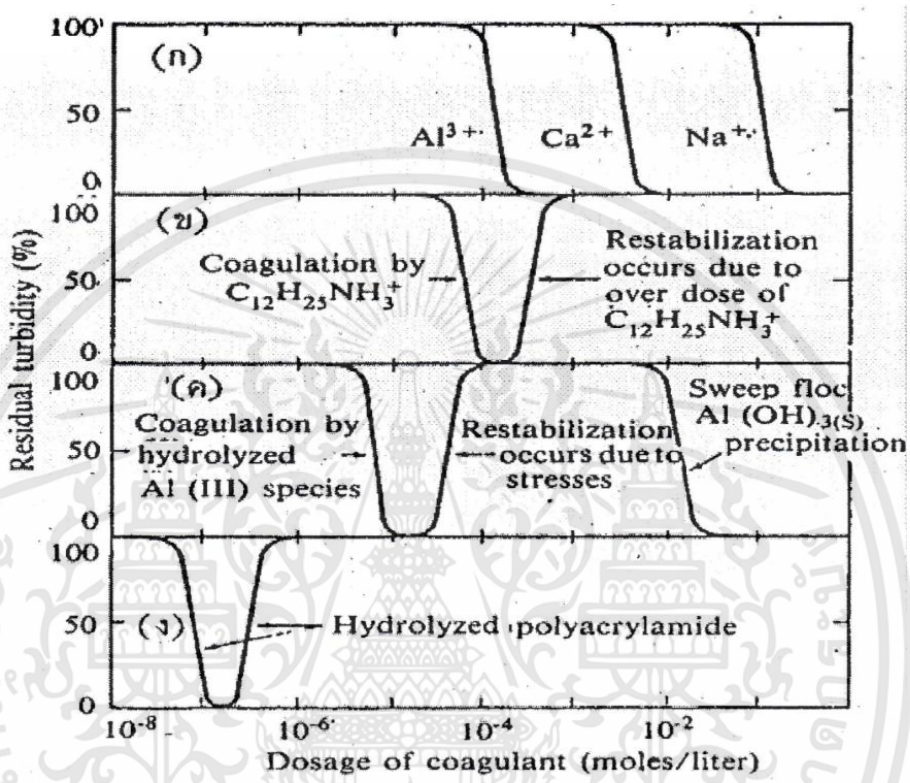
1. ปริมาณสารตัวนำไฟฟ้า (อิอออนประจุบวก) ที่เติมเพื่อทำลายเสถียรภาพของคอลลอยด์ด้วยวิธีลดความหนาของชั้นกระจายไม่ขึ้นอยู่กับ ความเข้มข้นของคอลลอยด์
2. ไม่สามารถทำให้คอลลอยด์เปลี่ยนประจุไฟฟ้าจากลบเป็นบวก

2.1.2 กลไกดูดติดผิวและทำลายประจุของอนุภาคคอลลอยด์ (Adsorption and Charge Neutralization)โดยใส่สารเคมีบางหมู่ที่มีความสามารถให้ประจุตรง
กันข้ามกับอนุภาคคอลลอยด์และดูดติดผิวได้ซึ่งจะมีผลในการลดศักย์ไฟฟ้าของคอลลอยด์ ซึ่ง เป็นการทำลายเสถียรภาพนั่นเอง

2.1.3 กลไกการสร้างผลึกขึ้นมาเพื่อให้อนุภาคคอลลอยด์มาเกาะจับ (Sweep Coagulation) เช่น การใส่สารส้มให้เกิดผลึก $Al(OH)_3$ เหมือนวุ้นสีขาว เพื่อให้อนุภาคมาเกาะแล้ว รวมกันเป็นปล็อคได้ กลไกการใช้ผลึกสาร อินทรีย์ในการทำลายเสถียรภาพของ คอลลอยด์มีลักษณะทีแตกต่างจากกลไก 2 แบบแรกคือ ปริมาณโคแอกกูแลนท์ที่เหมาะสม (Optimum Dosage) แปรผกผันกับความเข้มข้นของคอลลอยด์ กล่าวคือ นำที่มีความขุ่นห้อยต้องใช้โคแอกกูแลนด์น้อยกว่า เหตุผลคือน้ำที่มีความขุ่นต่ำจะมี โอกาสสัมผัสระหว่างอนุภาคห้อย ดังนั่นแม้ว่าการทำลายเสถียรภาพของคอลลอยด์จะ เกิดขึ้นแล้วก็ตาม โคแอกกูเลชั่นอาจไม่เกิดได้ดีเท่าที่ควร การใช้โคแอกกูแลนด์ ปริมาณสูงก็เพื่อสร้างผลึกจำนวนมากๆ สำหรับเป็นสารเป้าสัมผัสให้ กับอนุภาค คอลลอยด์ แต่ในกรณีที่น้ำมีความขุ่น

สูง โอกาสสัมผัสย่อมมีมาก จึงไม่จำเป็นต้อง อาศัยเป้าสัมผัสจากภาคนอกมากเท่ากับกรณีแรก

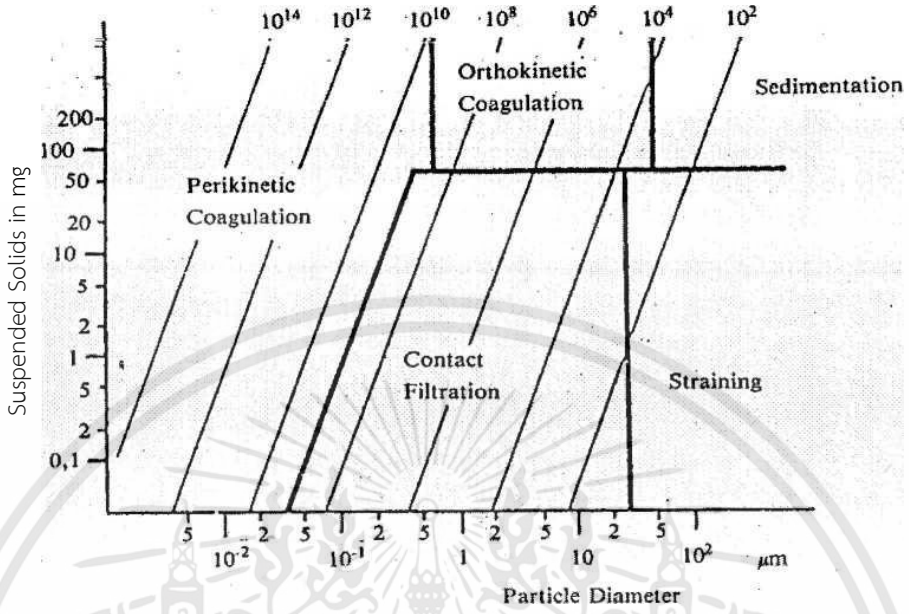
2.1.4 กลไกสร้างสะพานเชื่อมต่ออนุภาคคอลลอยด์โดยใช้สารโพลีเมอร์ ที่มีโมเลกุลขนาดใหญ่ เมื่อใส่ลงในน้ำจะให้อิออนเป็นจำนวนมากเพื่อ เกาะจับกับอนุภาคคอลลอยด์และยังมีแขนเชื่อมติดกับอนุภาคคอลลอยด์ตัว อื่นๆ เพื่อทำให้เกิดฟล็อค



รูปที่ 2.1 การเปรียบเทียบปริมาณโคแอกกูแลนท์ ที่ใช้ในการทำลายเสถียรภาพของคอลลอยด์ด้วยกลไกแบบต่างๆ (มั่นสิน ตัณฑุลเวศม์, 2537)

2.1.5 ทำให้อนุภาคคอลลอยด์ที่หมดเสถียรภาพแล้วเคลื่อนที่มาสัมผัสและเกาะจับ กันเป็นกลุ่มหรือฟล็อคคูเลชั่น (Flocculation) (พรศักดิ์ สมรไกรสรกิจ, 2550) วิธีการสร้างสัมผัสให้อนุภาคมีหลายวิธีดังแสดงในรูปที 2.1 ดังนั้นจะเห็นว่าแบบ (ก) ซึ่งเป็นการลดความหนาของชั้นกระจาย ด้วย Al^{+3}, Ca^{+2} และ Na^{+} ต้องการสารเคมีมากที่สุด ส่วนแบบ (ง) ซึ่งเป็นการใช้โพลีเมอร์ เป็นตัวเชื่อมโยง (สะพาน) ให้อนุภาคคอลลอยด์มารวมตัวกัน มีความต้องการสารโคแอกกูแลนท์ที่น้อยที่สุดทำให้อนุภาคคอลลอยด์เคลื่อนที่ไปมาในนำจนกว่าจะมีการสร้างสัมผัสเกิดขึ้น วิธีปฏิบัติเป็นที่นิยมมากที่สุด คือ กวนน้ำให้เคลื่อนที่ในลักษณะที่ส่วนต่างๆ ของน้ำมีอัตราเร็ว ในการไหลแตกต่างกัน เป็นเหตุให้อนุภาคต่างๆ มีอัตราเร็วในการเคลื่อนที่ไม่เท่ากันจึงมีกาว สัมผัสเกิดขึ้นการเคลื่อนที่ของนำต้องไม่รวดเร็วจนเกินไป มิฉะนั้นแล้วฟล็อคที่เกิดขึ้นอาจ แตกหรือหลุดออกจากกันได้ วิธีนี้เป็นวิธีธรรมดาที่นิยมใช้กันทั่วไป ซึ่งอุปกรณ์ในการสร้าง สัมผัสหรือสร้างฟล็อคคูเลชั่นเรียกว่า ถังกวนช้า และวิธีการสร้างสัมผัสแบบนี้มีชื่อเทคนิคว่า Ortho kinetic Flocculation อนุภาคคอลลอยด์ที่มีฟล็อคคูเลชั้น

แบบนี้ควรมีขนาดใหญ่กว่า 0.1-1 ไมครอนและมีความเข้มข้นไม่น้อยกว่า 50 มิลลิกรัม /ลิตร



รูปที่ 2.2 เกณฑ์ที่เหมาะสมสำหรับการสร้างสัมผัสระหว่างอนุภาคต่างๆ 5 ประเภท (มั่นสิน ตัณฑุลเวศม์, 2537)

2.1.6 การสัมผัสของอนุภาคคอลลอยด์ อาจเกิดขึ้นได้เองโดยอาศัยการเคลื่อนที่แบบบราวเนียนซึ่งเกิดขึ้นเนื่องจากอนุภาคคอลลอยด์กระทบกันเองหรือถูกชนโดยโมเลกุลของน้ำเนื่องจากการเคลื่อนที่ของ โมเลกุลของน้ำขึ้นอยู่กับอุณหภูมิ การสัมผัสแบบนี้จึงขึ้นอยู่กับอุณหภูมิด้วย จึงอาจกล่าวได้ว่าการเคลื่อนที่ แบบบราวเนียน เรียกว่า Perikinetic Flocculation.

2.1.7 การสัมผัสระหว่างอนุภาคเกิดขึ้น เนื่องจากการตกตะกอนที่มีอัตราไม่เท่ากันของอนุภาคต่างๆ ฟล็อคคูเลชันด้วยวิธีนี้เกิดขึ้นพร้อมๆกับการตกตะกอน ทำให้สามารถกำจัดอนุภาคคอลลอยด์ออกจากนำได้เลย อนุภาคทีสามารถสร้างฟล็อคคูเลชนแบบนีได้ต้องมีขนาด ใหญ่กว่า 5 ไมครอน และมีความเข้มข้นไม่น้อยกว่า 50 มก./ล. ในทางปฏิบัติอนุภาคที่มี ขนาดดังกล่าวอาจเกิดฟล็อคคูเลชันมาก่อนแล้วครั้งหนึ่ง เมือมาถึงการตกตะกอนจึงเกิดฟ ล็อคคูเลชั่นอีกในขณะที่มีการตกตะกอน

2.1.8 ในกรณีที่อนุภาคคอลลอยด์มีขนาดใหญ่กว่า 0.1-1 ไมครอน แต่เล็กกว่า 5 ไมครอน และ มีความเข้มข้นน้อยกว่า 50 มก./ล. ฟล็อคคูเลชั่นอาจเกิดขึ้นโดยการสร้างสัมผัสแบบ Ortho kinetic Flocculation แต่อาจเกิดขึ้นช้าเนื่องจากโอกาสสัมผัสน้อย วิธีแก้ไขอาจ กระทำดังนี้

1. ใช้ถังกรองทรายแบบกรองเร็วหรือถังกรองแบบ 2 ชั้น ชั้นกรองช่วยเพิ่มอัตราสัมผัสให้และยังบังคับให้อนุภาคต่างๆเคลื่อนที่เข้ามาชิดกันด้วย การใช้ถังกรอง ช่วยสร้างฟลือคคูเลชั่นเช่นนี้ เรียกว่า กรองสัมผัส (Contact Filtration) แต่ เนื่องจากช่องว่างในชั้น

กรองมีจำกัด วิธีนี้จึงใช้ได้กับอนุภาคที่มีความเข้มข้นไม่เกิน 50 มก./ล. การใช้กรวดทรายขนาดเล็กแทนทรายอาจเพิ่มปริมาตรช่องว่าง ได้ แต่เป็นการลดพื้นที่สัมผัส ดังนั้นจึงอาจได้ผลในทางฟล็อคคุเลชั่นไม่ดีเท่าชั้น ท่ราย

2. ใช้อนุภาคที่จับตัวกันเป็นฟล็อคแล้วเป็นเป้าสัมผัสให้กับอนุภาคใหม่ ในทางปฏิบัติสามารถกระทำได้ 2 วิธีคือ ทำให้ฟล็อคจับตัวคันเป็นชั้นสลัดจ์ (Sludge Blanket) และบังคับให้อนุภาคคอลลอยด์เคลื่อนที่ผ่านชั้นสลัดจ์ อีกวิธีหนึ่งคือ นำเอาฟล็อคกลับคืนมาผสมกับอนุภาคคอลลอยด์ จากนั้นจึงสร้างสัมผัสตามแบบ Ortho kinetic Flocculation ไปตามปกติ การใช้ถังตกตะกอนแบบ Solids Contact Clarifier ก็ใช้หลักนี้

2.1.9 ในกรณีที่อนุภาคคอลลอยด์มีขนาดใหญ่กว่า 3 ไมครอนแต่มีความเข้มข้นตามการสร้างสัมผัส อาจใช้วิธีกรองได้เช่นกันแต่สารกรองที่ใช้ควรมีขนาดใหญ่กว่าทราย

2.2 กลไกโคแอกกูเลชั่นด้วยสารส้ม

สารส้มเป็นโคแอกกูแลนท์ที่นิยมใช้กันมากที่สุดในประเทศไทย เนื่องจากสามารถใช้ได้ดีกับน้ำดิบจากแหล่งต่างๆ และหาซื้อได้ง่ายในราคาที่ไม่แพงมากนัก สารส้ม (อะลูมิเนียมซัลเฟต) มีสูตรโมเลกุล $A1_2(S0_4)_3 * H_20$ ซึ่งโดยปกติ * มีค่าเท่ากับ 14.3 หรือ 18 เมื่อเติมสารส้มลงในน้ำ จะแตกตัวให้อิออน บวกและลบ ดังปฏิกิริยา

$Al_2(S0_4)_3 \rightarrow 2Al^{3+} + 3S0_4^{2-}$ (1)

เมื่อเติมสารส้มในน้ำ อลูมิเนียมไอออนจาก $Al_2(S0_4)_3$ จะถูกล้อมรอบด้วยโมเลกุลของ น้ำได้ $Al((H_20)_6)^{3+}$ หรือ Al^{+3} ไฮดรอลิค (Hydrolysis) ของ Al^{+3} จะเกิดขึ้นทันทีโดยไลแกนด์ (Ligands) ชนิดต่างๆที่อยู่ในน้ำ โดยเฉพาะอย่างยิ่ง OH^- จะเข้าแทนที่โมเลกุลของน้ำ เกิดเป็น สารประกอบเชิงซ้อน (Complex substance) ระหว่างอลูมิเนียมกับไฮดรอกไซด์ไอออน (Hannah, 1967) ดังสมการต่อไปนี้

$Al^{3+} + H_20 \rightarrow Al(OH)^{2+} + H^+$ *(2)*

$Al^{+3} + 2H_20 \rightarrow Al(O)^{2+} + 2H^+$ *(3)*

$7Al^{+3} + 17H_20 \rightarrow Al_7((OH)_{17})^{4+} + 17H^+$ (4)

ในกรณีที่ความเข้มข้นของสารส้มสูงกว่าความเข้มข้นที่จุดอิ่มตัว (Saturation Point) ไฮโดรไลซีสจะดำเนินต่อไปจนได้ผลของปฏิกิริยาสุดท้ายเป็นผลึก

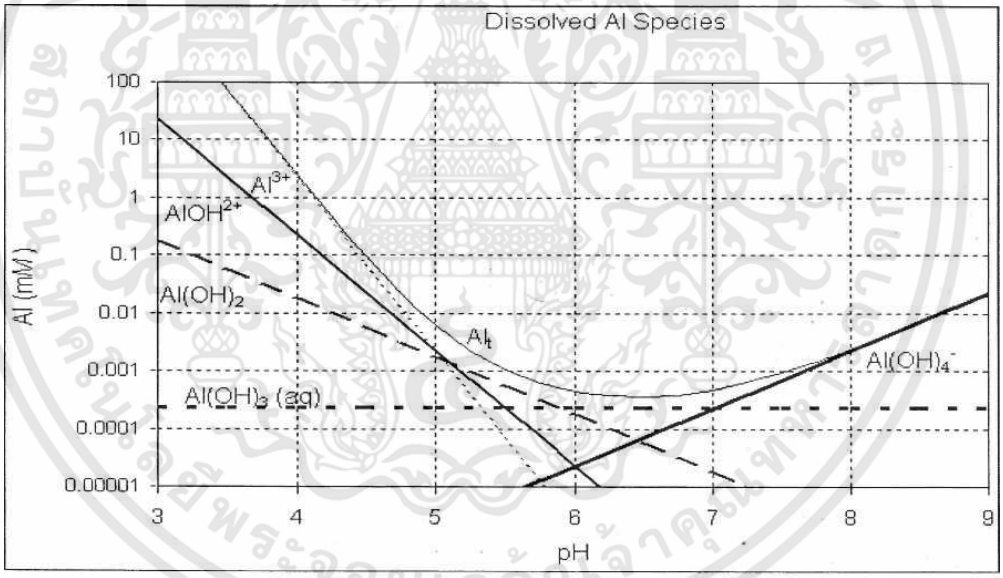
$Al^{3+} + 3H_20 \rightarrow Al(OH)_3(s) + 3H^+$ (5)

ผลของปฏิกิริยาที่จะเกิดการดูดติดผิวอนุภาคคอลลอยด์คือสารคอมเพล็กซ์ ซึ่งเกิดขึ้นใน ระหว่างไฮโดรไลซิสจาก Al^{3+} ถึง $Al(OH)_3$ สารคอมเพล็กซ์อาจมีประจุลบหรือบวกก็ได้ ทั้งนี้ ขึ้นอยู่กับพีเอชของน้ำ กล่าวคือ ถ้าพีเอชของน้ำสูงกว่าจุดสะเทินทางไฟฟ้า (Zero Point of Charge) ของ

$Al(OH)_3(s)$ จะเกิดสารคอมเพล็กซ์ประจุลบ เช่น $Al(OH)_4^-$, $Al(OH)_5^{2-}$ ถ้าพีเอชของน้ำต่ำกว่า จะสะเทินทางไฟฟ้าของ $Al(OH)_3(S)$ ซึ่งเป็นลักษณะที่เกิดขึ้นโดยทั่วไปในกระบวนการโคแอกกูเลชั่นจะเกิดสารคอมเพล็กซ์ประจุบวก เช่น

ผลของปฏิกิริยาที่จะเกิดการดูดติดผิวอนุภาคคอลลอยด์คือสารคอมเพล็กซ์ ซึ่งเกิดขึ้นใน ระหว่างไฮโดรไลซิสจาก Al^{3+} ถึง $Al(OH)_3$ สารคอมเพล็กซ์อาจมีประจุลบหรือบวกก็ได้ ทั้งนี้ ขึนอยู่กับพีเอชของน้ำ กล่าวคือ ถ้าพีเอชของน้ำสูงกว่าจุดสะเทินทางไฟฟ้า (Zero Point of Charge) ของ $Al(OH)_3(s)$ จะเกิดสารคอมเพล็กซ์ประจุลบ เช่น $Al(OH)_4^-$, $Al(OH)_5^{2-}$ ถ้าพีเอชของน้ำต่ำกว่า จะสะเทินทางไฟฟ้าของ $Al(OH)_3(S)$ ซึ่งเป็นลักษณะที่เกิดขึ้นโดยทั่วไปในกระบวนการโคแอกกูเลชั่นจะเกิดสารคอมเพล็กซ์ประจุบวก เช่น

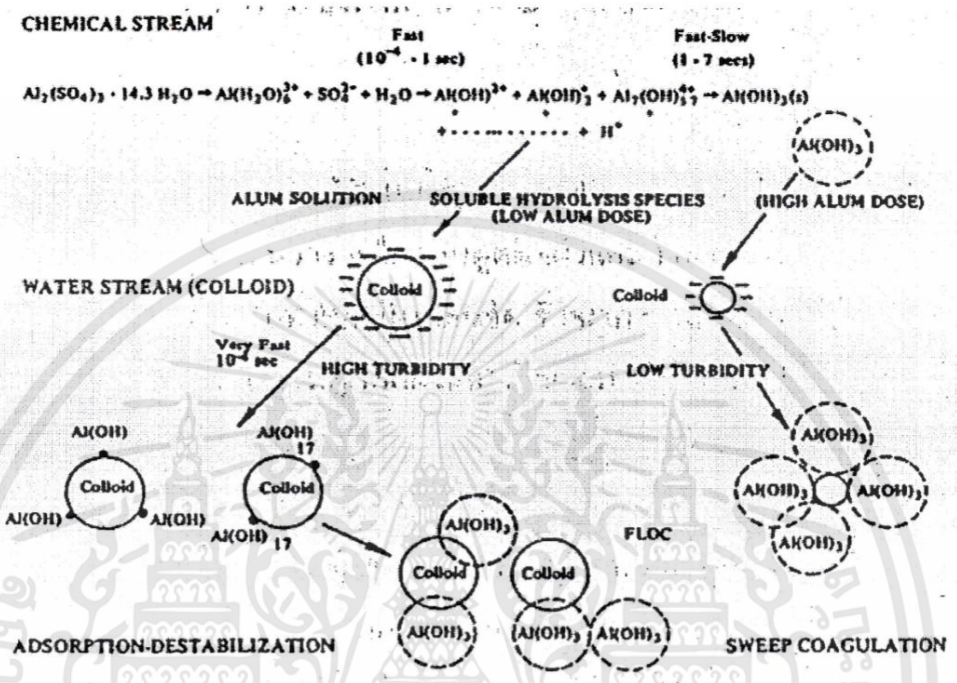
$Al(OH)^{2+}$, $Al((OH)_2)^+$, $Al_7((OH)_{17})^{+4}$, $Al_{13}((OH)_{34})^{+5}$ (รูป 2.3)
สารส้มที่เติมลงในน้ำจะเกิดการทำลายเสถียรภาพของอนุภาคคอลลอยด์ ด้วยกลไกหลัก ดังนี้ (รูป 2.4) (Amirtharjah และ Mill, 1*982)*



รูปที่ 2.3 ความสัมพันธ์ระหว่างสารประกอบเชิงซ้อนสารส้ม และค่าพีเอช (Driscol และ Shecher, 1990)

1. กล ไกแบบดูดติดผิวและทำลายประจุ (Adsorption and Charge Neutralization) เกิดจาก สารประกอบเชิงซ้อนสารส้มที่มีประจุบวก ทำลายเสถียรภาพของคอลลอยด์ซึ่งมักมีประจุเป็นลบให้เป็นกลาง (Neutralization) เป็นการสร้างโอกาสสัมผัสให้อนุภาครวมตัวกันจนมีขนาด ใหญ่และสามารถตกตะกอนด้วยน้ำหนักของอนุภาคเพียงลำพัง กลไกนี้มีช่วงความเหมาะสมที่แคบ ซึ่งจะควบคุมการทำงานให้ดีนั้นยาก เพราะสารประกอบเชิงซ้อนที่เกิดขึ้นต้องพอเหมาะเท่านั้น ถ้าหากมีปริมาณต่ำเกินไป โคแอกกูเลชั่นจะไม่

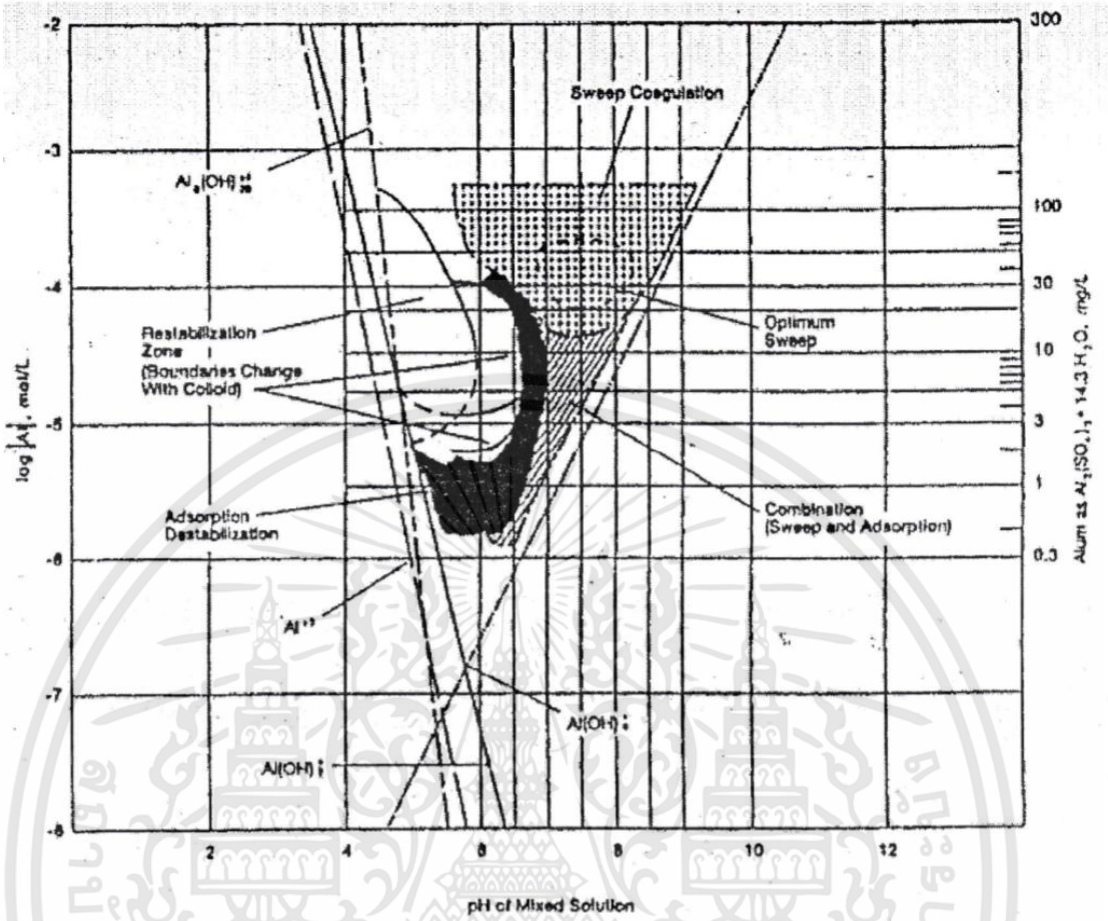
เกิด แต่ถ้าสูงเกินไปสารประกอบ เชิงซ้อนจะดูดติดผิวอนุภาคมากทำให้อนุภาคเปลี่ยนเป็นประจุบวกและเกิดเสถียรภาพขึ้นอีกแต่ตะกอนที่เกิดจากกลไกนี้สามารถแยกออกจากน้ำได้ง่าย ทำให้ประหยัดค่าใช้จ่าย



รูปที่ 2.4 กลไกในการสร้างโคแอกกูเลชั่นด้วยสารส้ม (Amirtharajah และ Mill, 1982)

2. *กลไกแบบกวาด* (Sweep Coagulation) ในกรณีความเข้มข้นของสารส้มเกินพอจนปฏิกิริยา ดำเนินต่อไปจนได้ $Al(OH)_3$ดังสมการที่ (4) การทำลายเสถียรภาพของอนุภาคคอลลอยด์ด้วย กลไกนีจะเกิดขึ้นเมื่อมีการเติมสารส้มเป็นจำนวนมากพอ จนมีความเข้มข้นเกินจุดอิ่มตัว ซึ่งทำให้ผลึกของ $Al(0H)_3$ ซึ่งมีลักษณะเหนียวสามารถห่อหุ้มอนุภาคและทำให้ผิวของอนุภาคม ความเหนียว ไม่แสดงอิทธิพลทางประจุไฟฟ้า จึงทำหน้าที่สร้างสัมผัสอนุภาคคอลลอยด์จนมีขนาดใหญ่และสามารถยึดเกาะลำพัง

3. กลไกโคแอกกูเลชั่นแบบร่วม (Combine Coagulation) เป็นการทำลายเสถียรภาพอนุภาค คอลลอยด์ร่วมกันระหว่างกลไกแบบดูดติดผิว และทำลายประจุ และแบบกวาด โดยที่ความแตกต่างระหว่างอิทธิพลของกลไกทั้งสองมีไม่เด่นชัด ซึ่งจะเกิดขึ้นเมื่อมีการใช้ปริมาณสารส้ม เพิ่มสูงขึ้นกว่ากลไกการทำลายเสถียรภาพแบบดูดติดผิวและทำลายประจุ แต่จะใช้ปริมาณ สารส้มต่ำกว่ากลไกแบบกวาด Amirtharajah และ Mill (1982) ได้รวบรวมผลการวิจัยเกี่ยวกับโคแอกคูเลชั่น ด้วยสารส้มและ นำมาวิเคราะห์จึงเสนอหลักการออกแบบและควบคุมโคแอกกเลชั่นด้วยสารส้ม ดังแสดงในรูปที่ 2.5 ซึ่งจากภาพแสดงให้เห็นว่าโคแอกกูเลชั่น จะได้ผลดีที่สุดที่พีเอช 6.8 ถึง 8.2



รูปที่ 2.5 ไดอะแกรมที่ใช้ในการออกแบบและควบคุมโคแอกกูเลชั่นด้วยสารส้ม (Amirtharajah และ Mill, 1982)

2.3 เหมืองข้อมูล (Data Mining)

Data Mining คือการค้นหาความสัมพันธ์และรูปแบบ (Pattern) ทั้งหมดซึ่งมีอยู่ จริงในฐานข้อมูล แต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก Data Mining จะทำการสำรวจและวิเคราะห์ปริมาณข้อมูลจำนวน มากอย่างอัตโนมัติหรือกึ่งอัตโนมัติ ให้อยู่ในรูปแบบความสัมพันธ์ และอยู่ในรูปของกฎ (Rule) โดยความสัมพันธ์เหล่านี้ แสดงให้เห็นถึง ความรู้ต่างๆ ที่มีประโยชน์ แต่ถูกซ่อนไว้ในฐานข้อมูล

Data Mining คือการสังเคราะห์ข้อมูลอย่างละเอียดจากฐานข้อมูลขนาดใหญ่ หรืออาจวิเคราะห์มาจากรายการ Transaction โดยเรียนรู้ข้อมูลจากอดีต หรือปัจจุบัน ผลลัพธ์ที่ได้จากการสังเคราะห์ของ Data Mining อาจจะเป็นข้อมูลแบบ Unknown, Valid หรือ Actionable ซึ่งความหมายของข้อมูลทั้ง 3 ประเภทนี้ มีดังนี้

1. ข้อมูลแบบ Unknown เป็นข้อมูลที่ผู้ใช้งานไม่เคยรู้มาก่อนไม่ชัดเจนไม่สามารถตั้งสมมติฐานล่วงหน้าได้ว่าจะเป็นแบบใด เช่น

- ห้างสรรพสินค้าแห่งหนึ่งค้นพบพฤติกรรมของผู้บริโภค ที่พ่อบ้านมักซื้อเบียร์และผ้าอ้อมในวันศุกร์ตอนเย็น ดังมันเป็นสัญญาณให้เจ้าของกิจการควรจะ เตรียมสินค้าไว้เพื่อจำหน่าย ในขณะที่ห้างคู่แข่งอาจจะไม่รู้ข้อมูลเหล่านี้

2. ข้อมูลแบบ Valid คือ การที่ผู้ใช้เริ่มใช้เทคนิคของ Data Mining ค้นพบสิ่งที่น่าสนใจและพิจารณาด้วยว่าสิ่งนั้นถูกต้อง หรือไม่ เช่น

- ความสัมพันธ์ของการซื้อสินค้า 2 อย่าง เมื่อจำนวนความหลากหลายของสินค้ามากขึ้น แต่ไม่ได้หมายความว่าจะต้องให้ห้างสรรพสินค้า เก็บสินค้าใน คลังมากขึ้น เพราะข้อมูลที่ได้อาจเกิดความคลาดเคลื่อน เพราะฉะนั้นจะต้อง ทำการตรวจสอบความถูกต้อง (Validation and Checking) ของข้อมูลและ วิเคราะห์ความถูกต้องอีกครั้ง

3. ข้อมูลแบบ Actionable ข้อมูลจะต้องถูกแปลงออกมาและนำมาตัดสินใจ เพื่อ สร้างความได้เปรียบในเชิงธุรกิจ บางครั้งข้อมูลที่เราค้นพบเป็นสิ่งที่คู่แข่งได้ทำไปแล้ว หรืออาจผิดกฎหมาย ซึ่งจะต้องมีวิจารณญาณในการใช้ด้วย และในบางที ข้อมูลดังกล่าวอาจไม่มีประโยชน์อะไร

2.3.1 วิวัฒนาการของ Data Mining

1. ปี ค.ศ. 1960 : Data Collection มีการนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่น่าเชื่อถือ เพื่อป้องกันการสูญหายได้เป็นอย่างดี

2. ปี ค.ศ. 1980 : Data Access มีการนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์ระหว่างกันเพื่อนำไปวิเคราะห์ และตัดสินใจอย่างมีประสิทธิภาพ

3. ปี ค.ศ. 1990 : Data Warehouse and Decision Support มีการนำข้อมูลมาเก็บลง ในฐานข้อมูลขนาดใหญ่ ครอบคลุมการใช้งานที่ หมดขององค์กร เพื่อช่วยสนับสมุน การตัดสินใจ

4. ปี ค.ศ. 2000 : Data Mining นำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผลโครงสร้างแบบจำลองและความสัมพันธ์ทางสถิติ

2.3.2 วัตถุประสงค์ในการใช้ Data Mining

1. เพื่อการค้นพบองค์ความรู้ใหม่ในฐานข้อมูล
2. เพื่อการสกัดองค์ความรู้ที่ซ่อนเร้นอยู่
3. เพื่อจัดการกับข้อมูลในอดีต
4. เพื่อสำรวจข้อมูล
5. เพื่อค้นหา Pattern ของข้อมูลที่ซ่อนอยู่
6. เพื่อใช้ขุดเจาะข้อมูล
7. เพื่อเก็บเกี่ยวผลประโยชน์ให้ได้มาซึ่งสารสนเทศที่มีประโยชน์

2.3.3 เป้าหมายหลักของ Data Mining

คุณลักษณะและเป้าหมายหลักของ Data Mining คือ ใช้การสกัดหรือค้นหา Pattern ของ ข้อมูลที่ฝังลึกและซ่อนเร้นอยู่ภายในฐานข้อมูลขนาดใหญ่ โดยใช้สถาปัตยกรรม (Client- Server) (Client/Server Architecture) ใช้เครื่องมือสมัยใหม่ที่สามารถแสดงผลแบบกราฟิก ผู้ใช้สามารถดูข้อมูลแบบเจาะลึก (Data Mining) และสามารถใช้เครื่องมือใน

การสอบถามข้อมูลได้ อย่างง่ายดายโดยไม่ตองอาศัยความชำนาญของโปรแกรมเมอร์ บ่อยครั้งเราอาจค้นพบผลลัพธ์ที่เรา ไม่คาดหวังมาก่อนการมีโปรแกรม Data Mining ใช้เป็นเครื่องมือจัดการข้อมูลจำนวนมาก จะสามารถช่วยให้เราทำงานกับข้อมูล หรือหยิบข้อมูลมาใช้งานได้ง่ายขึ้น ซึ่งนอกจากจะแสดงผลทางกราฟิกได้แล้ว ยังสามารถแก้ไข Spreadsheets ซึ่งเราใช้รวบรวมข้อมูลจำนวนมาก รวมถึงวิเคราะห์ ผ่านตัวโปรแกรมได้ด้วย

2.3.4 ขั้นตอนของการค้นหาความสัมพันธ์จาก Data Mining (Steps of KDD Process)

1. เรียนรู้และศึกษาเกี่ยวกับโปรแกรมที่จะใช้ (Learning the application domain)

2. คัดเลือกข้อมูล (Data selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาทำ Mining รวมถึงการนำข้อมูลที่ต้องการหรือไม่ต้องการออกจากฐานข้อมูล เพื่อสร้าง กลุ่มข้อมูลสำหรับพิจารณาในเบื้องต้น

3. การกรองข้อมูลและประมวลผล (Data cleaning and Processing) ข้อมูลที่เก็บรวม รวมมามีจำนวนมากจะต้องนำมากรอง เพื่อเลือกข้อมูลที่ตรงประเด็น เพราะบาง ขอมูลอาจจะไม่เป็นประโยชน์ ในขันตอนนี้เป็นขั้นตอนที่จะได้ข้อมูลที่มีคุณภาพที่จะ นำไปใช้วิเคราะห์

4. การแปลงรูปแบบข้อมูล (Data reduction and transformation) ลดรูปแบบและจัดข้อมูลให้อยู่ในรูปแบบเดียวกัน มีรูปแบบ (Format) ที่เป็นมาตรฐาน และเหมาะสมที่ จะนำไปใช้กับ Algorithm และแบบจำลองที่ใช้ทำ Data Mining

5.เลือก Algorithm เพื่อที่จะใช้ในการทำ data mining เช่น Classification , Regression , Association และ Clustering เป็นต้น

6. เลือก Algorithm ที่ใช้งานใน Data Mining เป็นเทคนิคสำหรับการ Mine ข้อมูลด้วย วิธีทางสถิติและการให้ความสัมพันธ์ระหว่างข้อมูลแบบด่างๆ

7. ทำการค้นหาความสัมพันธ์ (Pattern) ที่เราสนใจ จากความสัมพันธ์หลายๆแบบที่สามารถสังเคราะห์ได้จากการทำ Data Mining

8. ประเมินผล (Evaluate) และนำเสนอองค์ความรู้ ในขั้นตอนนี้จะเป็นการวิเคราะห์ ผลลัพธ์ที่ได้ และแปลความหมาย และประเมินผลว่าผลลัพธ์นั้นเหมาะสมหรือตรา วัตถุประสงค์หรือไม่

9. ใช้องค์ความรูที่ค้นพบ (Use of Discovered Knowledge)

2.3.5 ชนิดขององค์ความรู้ที่ค้นพบ (Types of Knowledge to be mined)

1. องค์ความรู้เกี่ยวกับคุณลักษณะของข้อมูล เช่น รู้ว่าคนที่สามารถ เรียนต่อในระดับ ปริญญาเอกได้จะพิจารณาได้จากคุณลักษณะใด

2. องค์ความรู้เกี่ยวกับการจำแนกข้อมูล

3. องค์ความรู้เกี่ยวกับความสัมพันธ์ของข้อมูล เช่น มีความสัมพันธ์ของ การซื้อสินค้าพบว่า ถ้าลูกค้าป็อบคอร์นจะต้องซื้อเป๊ปซี่

4. องค์ความรู้เกี่ยวกับการแยกประเภทข้อมูลพยากรณ์
5. องค์ความรู้เกี่ยวกับการจัดกลุ่มข้อมูล
6. องค์ความรู้เกี่ยวกับการวิเคราะห์ข้อมูลจากภายนอก
7. องค์ความรู้เกี่ยวกับข้อมูลอื่นๆ ในงานที่ค้นพบ

2.3.6 ประเภทของข้อมูลที่สามารถใช้ ข้อมูลที่มาจากฐานข้อมูลเชิงสัมพันธ์

1. ข้อมูลจากคลังข้อมูล
2. ข้อมูลจากฐานข้อมูลรายการปรับปรุง
3. ข้อมูลจากฐานข้อมูลรายการปรับปรุง
4. จากฐานขอมูลพิเศษหรือทีเก็บข่าวสารพิเศษ ซึ่งได้แก่
   - ฐานข้อมูลเชิงวัตถุ
   - ข้อมูลเกี่ยวกับเวลา
   - ฐานข้อมูลข้อความ (Text database) และฐานข้อมูลมัลติมีเดีย
   - ฐานข้อมูลแบบเก่าในอดีตหรือข้อมูลที่มาจากต่างฐานข้อมูลกัน
   - ข้อมูลจากอินเทอร์เน็ต เช่น WWW

2.3.7 Data Mining Functionalities (Data Mining Task)

งานของ Data Mining สามารถทำงานในการขุดค้นและสร้างฐานข้อมูล ได้ดังต่อไปนี้

1. การวิเคราะห์คุณสมบัติและการแยกแยะข้อมูล
2. การหาความสัมพันธ์ของข้อมูล
3. การจัดหมวดหมู่และการวิเคราะห์การ

3.1 การจัดหมวดหมู่ *(Classification)*

มีหลายเทคนิคของ Data Mining ที่ใช้ในการแก้ปัญหาแบบ Classification แต่ละเทคนิคก็จะมีหลาย Algorithm ให้เลือกและแต่ละ Algorithm จะให้ผลลัพธ์ที่ต่างกัน ซึ่งปัญหาประเภทนี้จะให้ผลลัพธ์เป็นค่าที่แน่นอน เช่น อาจจะได้คำตอบเป็น (Yes,No) หรือ (High, Medium, Low) เป็นต้น คำตอบที่เราต้องการ (Output) ถือเป็นตัวแปรตาม (Dependent variable) ซึ่งผลของตัวแปรตามจะขึ้นอยู่กับตัวแปรอิสระ (Independent variable) หรือพารามิเตอร์ชนิดต่างๆที่ต้องการใช้เป็นตัวแปรต้นเพื่อใช้ในการสังเคราะห์เพื่อสร้างแบบจำลอง เทคนิคของ Data Mining ที่ใช้ในการแก้ปัญหาแบบ Classification ได้แก่

1. Decision Tree
2. Neuro Networks
3. Naïve-Bayes

3.2 การวิเคราะห์การถดถอย (Regression)

ปัญหาแบบ Regression จะเหมือนกับแบบ Classification ต่างกันตรงที่ผลลัพธ์ ที่ได้จาก Regression เป็นค่าแน่นอน ไม่จำกัดและจะเป็นค่าอะไรก็ได้ เช่น แบบจำลองทำนายว่านาย B จะตอบรับข้อเสนอของบริษัท ถ้านาย B ได้รับผลกำไร 1,000 บาท (1,000 เป็นคำตอบเฉพาะที่แน่นอน แต่ไม่จำกัด ซึ่งตัวเลขอาจจะ เป็นค่าอื่นไปได้เรื่อยๆ ต่างจากคำตอบแบบ Yes, No)

3.3 การวิเคราะห์การรวมกลุ่ม หรือการแบ่งแยกข้อมูล (Cluster analysis or Segmentation)

เป็นการรวมกลุ่มข้อมูลที่มีลักษณะเหมือนกัน รูปแบบหรือแนวโน้มคล้ายกัน การใช้เทคนิค Clustering จะไม่มีผลลัพธ์ (Output) ไม่มีตัวแปรอิสระ (Independent variable) ไม่มีการจัดโครงร่างของวัตถุ เราจะเรียกเทคนิคของ Clustering ว่าเป็นรูปแบบการเรียนรู้ข้อมูลโดยไม่ต้องอาศัยครูสอน (Unsupervised Learning) การทำ Clustering จะทำบนพื้นฐานของข้อมูลในอดีต เช่น องค์กรต้องการทราบความเหมือนที่มีในกลุ่มของลูกข้าของตน เพื่อที่จะได้เข้าใจลักษณะเฉพาะของลูกค้ากลุ่มเป้าหมาย และสร้างกลุ่มของลูกค้าเพื่อที่องค์กรจะได้สามารถขายสินค้าได้ในอนาคต องค์กรจะทำ การแยกกลุ่มของข้อมูลลูกค้าออกเป็นกลุ่มๆ (หาส่วนที่เป็น Intersect และ Union) เทคนิคของ Data Mining เพื่อแก้ ปัญหาแบบ Clustering คือวิธี Demographic Clustering แล Neuro Clustering

4. การประเมินและการพยากรณ์ (Estimate and Prediction)

4.1 การประเมิณ (Estimate)

เป็นการประเมินที่ไม่สามารถกำหนดค่าหรือคุณสมบัติที่ชัดเจนได้ ใช้จัดการกับค่าที่มีผลแบบต่อเนื่อง เช่น ใช้ประเมินรายได้ของครอบครัว ประเมินความสูงของบุคคลในครอบครัว ประเมินจำนวนเด็กในครอบครัว

4.2 การพยากรณ์ (Prediction)

จะเหมือนกับ Classification และ Estimation ต่างกันตรงที่ Record ถูกแยก จัดลำดับในการทำนายค่าในอนาคต และนำข้อมูลในอดีตมาสร้างเป็นแบบจำลอง ใช้ทำนายสิ่งที่จะเกิดขึ้นในอนาคต เช่น การทำนายว่าลูกข้ากลุ่มใดที่องค์กรจะสูญเสียไปในอีก 6 เดือนข้างหน้า หรือ การทำนายยอดของลูกค้าจะเป็นเท่าใด ถ้าบริษัทลดราคาสินค้าลง 10%

5. การบรรยายและการแสดงภาพของข้อมูล (Description and Visualization)

5.1 การบรรยาย (Description)

เป็นการหาคำอธิบายถึงสิ่งทีจะเกิดขึ้น โดยอาศัยข้อมูลจากฐานข้อมูล เช่น กลุ่มคนที่มีการศึกษาหรือรายได้น้อย จะเลือกนักการเมืองที่มีนโยบายทุนนิยม

5.2 การแสดงภาพของข้อมูล (Visualization)

เป็นการนำเสนอข้อมูลในรูปแบบกราฟฟิค หรืออาจนำเสนอในแบบ 2 มิติ สร้างรายละเอียดในการนำเสนอให้เข้าใจมากยิ่งขึ้น เช่น องค์กรต้องการหาสถานที่ในการขยายสาขาใหม่ที่อยู่ในเขตพื้นที่ภาคเหนือของประเทศ ดังนั้นองค์กรจึงใช้แผนที่พลอต ที่ตั้งขององค์กรคู่แข่งที่มีสาขาอยู่ในเขตนั้น เพื่อพิจารณาสถานที่ตั้งที่เหมาะสมที่สุด

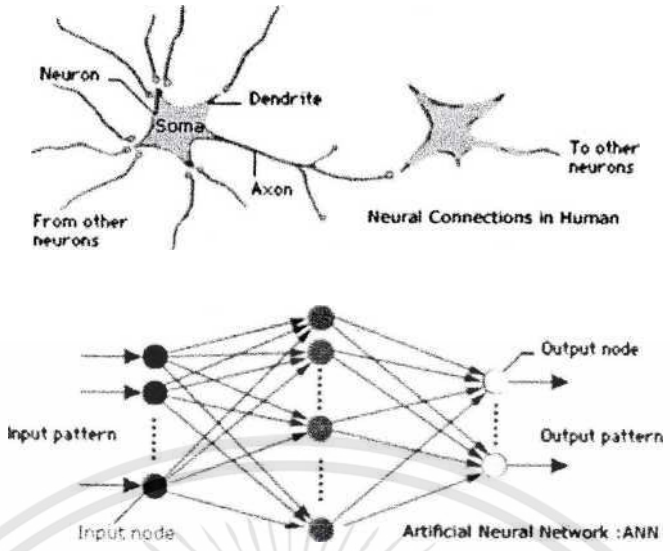
2.3.8 เครื่องมือและเทคโนโลยีของ Data Mining (Data Mining Tools and Technologies)

2.3.8.1 Neuro Network

เป็นแนวคิดให้คอมพิวเตอร์ทำงานสมองของมนุษย์ เปลี่ยนตัวเองจากการประมวลผลตามลำดับ (Sequential Processing) ให้เป็นการประมวลผลแบบคู่ขนานได้ (Parallel Processing) มีลักษณะการทำงานโดยแต่ละ Process จะรับ Input เข้าไปคำนวณและสร้าง Output ออกมาในลักษณะที่ไม่ใช่การทำงาน แบบเชิงเต้นตรง เพราะ Input แต่ละตัวจะถูกให้ลำดับความสำคัญของค่าไม่เท่ากัน ค่าของ Output ที่ได้จากการเชื่อมโยงกันนี้จะถูกนำมาเปรียบเทียบ กับ Output ที่ได้ตั้งเอาไว้ ถ้าค่าที่ออกมาเกิดความคลาดเคลื่อน ก็จะนำไปสู่การปรับค่าหรือน้ำหนัก (Weigh) ของค่าที่ใส่ไว้ให้แต่ละ Input Neuro network เป็นการสร้างแบบจำลองที่เลียนแบบการทำงานของสมองมนุษย์ มีโครงสร้างเป็นกลุ่มของ Node ที่เชื่อมโยง ถึง กันในแต่ละ Layer คือ Input layer, Hidden layer และ Output layer

2.3.8.2 Decision Tree

เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบโครงสร้างต้นไม้ (Decision Trees) ซึ่ง Decision Trees จะมีการทำงานแบบ Supervised learning (คือการเรียนรู้แบบมีครูสอน) สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างข้อมูลที่กำหนดไว้ก่อนล่วงหน้า เรียกว่า Training set ได้อัตโนมัติ และพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ ได้ด้วยรูปแบบของ Tree โครงสร้างประกอบด้วย Root Node และ Leaf Node



รูปที่ 2.6 โครงสร้าง Neuro Network (ณัฏภัทรศญา ทับทิมเทศ, 2550)

2.3.8.3 Memory Based Reasoning (MBR)

เปรียบเหมือนกับประสบการณ์การเรียนรู้ของมนุษย์ซึ่งอาศัยการสังเกตที่เกิดขึ้นแล้วสร้างรูปแบบของสิ่งนั้นขึ้นมา เราใช้ MBR เพื่อวิเคราะห์ฐานข้อมูลที่มีอยู่และกำหนดลักษณะพิเศษของข้อมูลที่อยู่ในนัน ซึงข้อมูลจะต้องมีลักษณะที่สมบูรณ์ การสังเกตจึงจะสมบูรณ์และทำนายผลได้แม่นยำยิ่งขึ้น แบบจำลองจะถูกบอกคำตอบที่ถูกต้อง มีการเก็บคำตอบสำหรับแก้ปัญหาไว้ก่อนล่วงหน้าแล้ว (Supervised learning)

2.3.8.4 Cluster Detection

คือจะแบ่งฐานข้อมูลออกเป็นส่วนๆ เรียกว่า Segment (กลุ่ม Record ที่มีลักษณะคล้ายกัน) ส่วน Record ที่ต่างกันก็จะอยู่นอก Segment, Cluster Detection ถูกใช้เพื่อค้นหากลุ่มย่อย (Sub group) ที่เหมือนๆกันในฐานข้อมูล เพื่อที่จะเพิ่มความถูกต้องในการวิเคราะห์ และสามารถมุ่งไปยังกลุ่มฟ้าหมายได้ถูกต้อง

2.3.8.5 Link Analysis

มุ่งเน้นทำงานบน Record ที่มีความสัมพันธ์กัน หรือเรียกว่า Association เทคนิคนี้จะมุ่งไปที่รูปแบบการซื้อหรือเหตุการณ์ที่เกิดขึ้นเป็นลำดับ มีอยู่ 3 เทคนิค คือ

1. Association Discovery

ใช้วิเคราะห์การซื้อขายสินล้าในรายการ เดียวกัน ศึกษาความสัมพันธ์อย่างใกล้ชิดที่ถูกซ่อนอยู่ของสินค้า ซ่ง สินล้าเหล่านั้นอาจมีแนวโน้มที่จะถูกซื้อควบคู่กันไป การวิเคราะห์แบบนี้เรียกว่า Market Basket Analysis คือรายการนั้งหมดที่ลูกล้าซื้อต่อครั้งที่ Super market การวิเคราะห์นี้สามารถนำมาใช้ประโยชน์ ในการดัดสินใจ เช่น การเตรียมสินค้าคงเหลือ การวางแผนจัดชั้นวางสินค้า การทำ Mailing list ส่าหรับ Direct mail การวางแผน

เพื่อจัด Promotion สนับสนุนการขาย ตัวอย่างของ Association เช่น 75% ของผู้ซื้อน้ำอัดลมจะซื้อข้าวโพดคั่วด้วย

2. Sequential Pattern Discovery

ถูกใช้ระบุความเกี่ยวเนื่องกันของการซื้อสินค้าของลูกค้า มีจุดหมายที่จะเข้าใจพฤติกรรมการซื้อสินค้าของลูกค้าในลักษณะ Long term เช่น ผู้ขายอาจพบว่าลูกค้าที่ ซื้อทีทัศน์มีแนวโน้มที่จะซื้อเครื่องเล่นวิดีโอในเวลาต่อมา

3. Similar Time Sequence Discovery

ค้นหาความเกี่ยวเนื่องกันระหว่างข้อมูล 2 กลุ่มซึ่งส่งผลต่อกันด้วยเวลา โดยมีรูปแบบการเคลื่อนที่เหมือนกัน ผู้ขายสินค้ามักใช้เพื่อดูแนวโน้มเพื่อเตรียม Stock เช่น เมื่อไรก็ตามที่ยอดขายสินค้าน้ำอัดลมสูงขึ้น ยอดขายมันฝรั่งจะสูงขึ้นตาม

2.3.8.6 Genetic Algorithm (GA)

เปรียบเสมือนเป็นการสร้างพันธุกรรมที่ดีสุด บนชั้นตอนของวิวัฒนาการทางชีวภาพ แนวคิดหลักคือ เมื่อเวลาผ่านไปวิวัฒนาการของเซลล์สิ่งมีชีวิตจะเลือกสายพันธุ์ที่ดีที่สุด “Fittest Species” GA มีความสามารถในการทำงานแบบรวมกลุ่มเข้าด้วยกัน เช่น มีการแบ่งกลุ่มและจัดรวมกลุ่มข้อมูลเป็น 3 ชุด ขั้นตอนการทำงานของ GA เริ่มจาก

1. จับกลุ่มข้อมูลเป็นกลุ่มๆ ด้วยการสุ่มเดา เปลียบกลุ่ม 3 กลุ่มนี้เป็นเซลล์ของสิ่งมีชีวิต GA จะมี Fittest Function ที่จะบอกว่ากลุ่มข้อมูลใดเหมาะสมกับกลุ่มใดๆ โดย Fittest Function จะเป็นตัวบ่งชี้ว่าข้อมูลเหมาะกับกลุ่มมากกว่าข้อมูลอื่นๆ

2. GA จะมี Operator ซึ่งยอมให้มีการเลียนแบบและแก้ไขลักษณะของกลุ่มข้อมูล Operator จะจำลองหน้าที่ของลักษณะที่ถูกพบในธรรมชาติ คือ มีการแพร่พันธ์ จับคู่ผสมพันธ์ และเปลี่ยนรูปร่างตามต้นแบบของพันธุกรรมเปรียบกับข้อมูลถ้ามีข้อมูบใดในกลุ่มถูกตรวจพบว่าตรงกับคุณสมบัติของ Fittest Function แล้ว มันจะคงอยู่และถูกถ่ายเข้าไปในกลุ่มนั้น แต่ถ้าไม่ตรงกับคุณสมบัติ ยังมีโอกาสที่จะถ่ายข้ามไปยังกลุ่มอื่นได้

2.3.8.7 Rule Induction

ดึงเอาชุดเกณฑ์ต่างๆมาสร้างเป็นเงื่อนไขหรือกรณีวิธีการของ Rule Induction จะสร้างชุดของกฎที่เป็นอิสระ ซึ่งไม่จำเป็นต้องอยู่ในรูปแบบของโครงสร้างต้นไม้

2.3.8.8 K-nearest neighbor (K-NN)

จะใช้วิธีในการจัดแบ่งคลาส โดยจะตัดสินใจว่าคลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ๆได้บ้าง โดยการตรวจสอบจำนวนบางจำนวนของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count up) ของจำนวนเงื่อนไขหรือกรณีต่างๆสำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกับตัวมันมากที่สุด K-NN ค่อนข้างใช้ปริมาณงานในการคำนวณสูงมาก เพราะเวลาสำหรับการคำนวณจะเพิ่มขึ้นแบบแฟคทอเรียล ตามจำนวนจุด

ทั้งหมด เทคนิคของ K-NN จะมีการคำนวณเกิดขึ้นทุกครั้งที่มีกรณีใหม่ๆเกิดขึ้น ดังนั้นถ้าจะให้เทคนิคแบบ K-NN ทำงานได้เร็วขึ้น ข้อมูลที่ใช้บ่อยควรเก็บอยู่ใน MBR

2.3.8.9 Association and Sequence Detection

1. Association ใช้หากฎความสัมพันธ์ที่เกิดขึ้นระหว่างกลุ่มของข้อมูล (Item) ต่าง ๆ ใช้ใน Market-basket analysis อาจใช้เพื่อวิเคราะห์การสั่งซื้อสินค้า

2. Sequence Detection เหมือนกับ Association แต่จะนำเหตุการณ์ที่เกิดขึ้น และเพิ่มตัวแปรด้านเวลาเข้ามาเกี่ยวข้องด้วย เพื่อใช้วิเคราะห์พฤติกรรมของ ข้อมูล

2.3.8.10 Logic Regression

เป็นการวิเคราะห์ความถดถอยแบบเค้นตรงทั่ว ๆ ไป ใช้ในการพยากรณ์ผลลัพธ์ของ 2 ตัวแปร เช่น Yes/No, 0/1 แต่เนื่องจากตัวแปรตาม (Dependent Variable) มี ค่าเพียง 2 อย่างเท่านั้น จึงไม่สามารถสร้างแบบจำลอง (Model) ได้สำหรับการ วิเคราะห์แบบ Logic Regression ดังนั้นแทนที่จะทำการพยากรณ์โดยอาศัยเพียง ค่าของตัวแปรตามที่ได้ เราจะสร้าง Model โดยอาศัย Algorithm ของความน่าจะเป็นของการเกิดเหตุการณ์ เราเรียก Algorithm นี้ว่า Log Odds

2.3.8.11 Discriminant Analysis

เป็นวิธีการทางคณิตศาสตร์ที่เก่าแก่วิธีหนึ่งซึ่งใช้ในการจำแนกและวิเคราะห์ วิธีนี้ ได้รับการเผยแพร่ครั้งแรกในปี 1936 โดย R.A Fisher เพื่อแยกดัน Iris ออกเป็น 2 พันธ์ วิธีการนี้ทำให้ค้นพบต้นไม้ประเภทอื่นๆอีกมาก ผลลัพธ์ที่ได้จาก แบบจำลองชนิดนี้ง่ายต่อการทำความเข้าใจ เพราะผู้ใช้งานทั่วๆไปก็สามารถพิจารณาได้ว่าผลลัพธ์จะอยู่ทางด้านใดของเส้นทางในแบบจำลอง การเรียนรู้สามารถทำได้ง่าย วิธีการที่ใช้มีความไวต่อรูปแบบของข้อมูล

2.3.8.12 Generalized Additive Models (GAM)

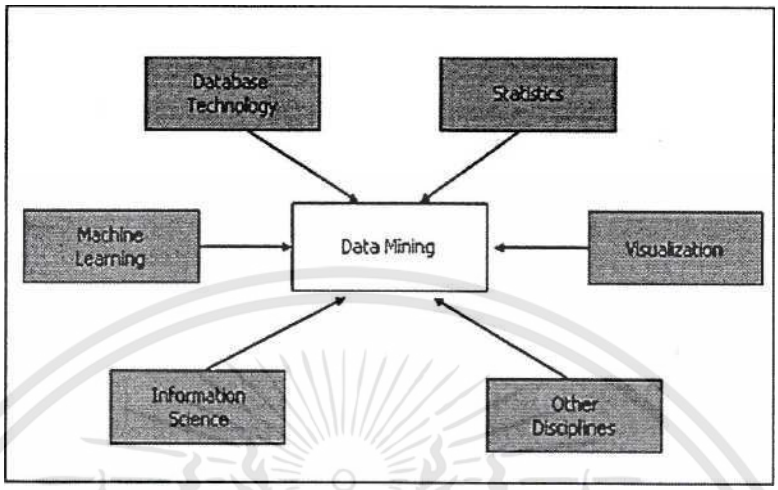
พัฒนามาจาก Linear Regression และ Logistic Regression มีการ ตั้งสมมติฐานว่า Model สามารถเขียนออกมาได้ในรูปของผลรวมของ Possibly Non-Linear Function GAM สามารถใช้ได้กับปัญหาแบบ Regression และ Classification, GAM จะใช้ความสามารถของคอมพิวเตอร์ในการด้นหารูปแบบอง Function ที่ ให้ Curve ที่เหมาะสม ทำการรวมค่าความสัมพันธ์ต่าง ๆ เขาด้วยกัน แทนที่จะ ใช้ Parameter จำนวนมาก เหมือนที่ Neural Network ใช้ แต่ GAM จะก้าวไป เหนือกว่าอีกขั้นหนึ่ง GAM จะประเมินค่าของ Output ในแต่ ละ Input 2.3.8.13

2.3.8.13 Multivariate Adaptive Regression Splits (MARS)

ถูกคิดด้นเมื่อกลางทศวรรษที่ 80 โดย Jerome H. Friedman หนึ่งในผู้คิดค้น CART MARS สามารถที่จะค้นหาและแสดงรายการตัวแปรอิสระที่มีความสำคัญ สูงสุดเช่นเดียวกับปฏิสัมพันธ์ระหว่างตัวแปรอิสระและ MARS สามารถ Plot จุด แสดงความเป็นอิสระของแต่ละตัวแปรอิสระ

ออกมาได้ ผลลัพธ์ที่ได้ก็คือ Non Linear Step-wise regression tools

2.3.9 ความสามารถที่หลากหลายของ Data Mining และการนำไปประยุกต์ใช้



รูปที่ 2.7 การนำไปประยุกต์ใช้ ของ Data Mining (ณัฏภัทรศญา ทับทิมเทศ, 2550)

สามารถนำเทคนิคของ Data Mining ไปวิเคราะห์ข้อมูลในฐานข้อมูล เพื่อนำข้อมูลที่ได้ไปใช้ประโยชน์ในงานด้านต่างๆดังนี้

1. งานด้านการตลาด (Marketing) เช่น การทำ Promotion ส่งเสริมการขาย
2. งานธนาคารด้านการเงิน (Banking / Financial Analysis) เช่น ใช้ในการวิเคราะห์การให้สินเชื่อแก่ลูกค้า การจัดทำ Package ในการกู้ยืม การทำนายอัตราการ จ่ายเงินกู้ การแบ่งกลุ่มลูกค้าเพื่อหาเปาหมายทางการตลาด (ลูกค้าชั้นดี)
3. งานด้านการขายปลีก (Retailing and sales) เป็นงานที่มีการเก็บข้อมูลจำนวนมากมาประยุกต์ใช้เพื่อหากลยุทธิ์ ทำให้เกิดการได้เปรียบคู่แข่งทางการค้าในการหาลักษณะการซื้อของลูกค้า ความสัมพันธ์ของการซื้อกับช่วงเวลา ความสัมพันธ์ระหว่างตัว สินค้า และการวิเคราะห์ประสิทธิภาพของการโฆษณา เป็นต้น ช่วยให้สามารถหา วิธีการตอบสนองความต้องการของลูกค้าได้มากที่สุด และอาจหมายถึงส่วนแบ่งทาง การตลาดที่เพิ่มขึ้นมั่นเอง
4. งานด้านการวางแผนในการผลิตสินค้า (Manufacturing and production) เช่น การพยากรณ์ยอดจำนวนการผลิตสินค้าเพื่อให้ได้กำไรมากสุด
5. งานด้านนายหน้าและความปลอดภัยด้านการค้า (Brokerage and securities trading) เช่น การพัฒนาวิธีการเพื่อสร้างความเชื่อมั่นในเรื่องความปลอดภัยของ ข้อมูลในขณะที่มีการพัฒนาวิธีการเข้าถึง และการ Mining ข้อมูลให้สะดวกต่อ การใช้งานมากขึ้น
6. งานด้านชีวการแพย์และวิเคราะห์ DNA (Biomedical an DNA) งานด้านชีวการแพย์และวิเคราะห์ DNA (Biomedical an DNA Analysis) เช่น การวิเคราะหรูปแบนการ

เรียงตัวของหน่วยพันธุกรรมเพื่อหาสาเหตุ ความผิดปกติทีทำให้เกิดโรค รวมไปถึงด้านการวินิจฉัยโรค การปองกัน และการ รักษา

นอกจากที่กล่าวมา ยังนำไปประยุกใช้กับธุรกิจทางด้านประกันภัย (Insurance), Computer hardware และ software, หน่วยงานรัฐบาลและกระทรวงกลาโหม (Government and defense), สายการบิน (Airlines), งานด้านสุขภาพ (Health care), งานด้านการข่าว (Broadcasting) และงานด้านกฎหมาย (Law enforcement) ได้ อีกด้วย

2.3.10 Intelligent Data Mining

ใช้ Intelligent Data Mining เพื่อการค้นพบข้อมูลและข่าวสารภายในคลังข้อมูล (Data warehouses) การสอบถามและการออกรายงาน (Reports) นั้นจะไม่แสดงผล ออกมา เช่น การค้นหา Patterns ความสัมพันธ์ในข้อมูลและลงความเห็นตานกฎทีเราไข้ กาหนดไว การใช Patterns และRules ในการแนะแนวทางการติดสินใจและคารทำนายผล

2.4 งานวิจัยที่เกี่ยวข้อง

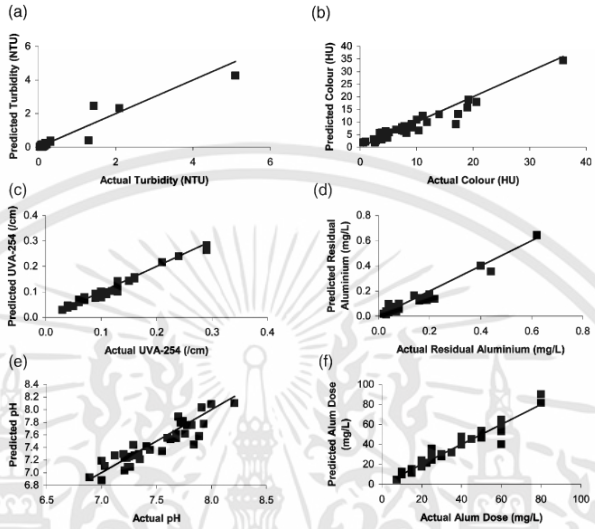
ปัจจุบันนั้น การนำเทคโนโลยีด้านการขุดเหมืองข้อมูล เข้ามาใช้เพื่อช่วยอำนวยความสะดวกในด้านต่าง ๆ เริ่มแพร่หลายมากขึ้น ไม่ว่าจะเป็นด้านสิ่งแวดล้อม เพื่อนำมาทำนายหรือพยากรณ์ผลต่าง ๆ ที่กำลังจะเกิดขึ้น หรือด้านสถิติ เพื่อวิเคราะห์ข้อมูลซับซ้อนที่ต้องใช้เวลาในการวิเคราะห์

Maire และคณะ (2004) งานวิจัยชิ้นนี้เกี่ยวกับการทำนายค่าพารามิเตอร์ต่างๆของการบำบัดน้ำเสีย ไม่ใช่เพียงแค่ทำนายปริมาณสารส้ม ข้อมูลที่ใช้ในงานวิจัยชิ้นนี้ค่อนข้างเป็นข้อมูลที่สมบูรณ์ เพราะมีทั้งการเก็บข้อมูลค่าพารามิเตอร์ของทั้งน้ำดิบ และน้ำที่ได้รับการบำบัดแล้ว มีการเก็บข้อมูล 2 ชุด ทั้งน้ำดิบ และ น้ำหลังการบำบัดซึ่งจะทำให้ประสิทธิภาพการสร้างแบบจำลองสูงขึ้นมาก แต่มีความผิดปกติบางอย่างในข้อมูลคือ ค่าพารามิเตอร์มีการแกว่งเป็นอย่างมาก ตัวอย่างเช่น ค่าความขุ่นในน้ำดิบ มีค่าต่ำที่สุดคือ 0.3 NTU ซึ่งเป็นค่าที่น้อยเกินไปสำหรับน้ำดิบ และอาจจะเป็นปัญหาสำหรับการสร้างแบบจำลอง

ตารางที่ 2.1 แสดงค่า R-Square และ MAE ของแต่ละพารามิเตอร์ (Maire และคณะ, 2004)

Model	Output	R^2	Mean absolute error
1	Turbidity of treated water	0.90	0.12 NTU
1	Colour of treated water	0.92	1.45 HU
1	UVA-254 of treated water	0.98	0.01 cm^{-1}
2	Residual aluminium of treated water	0.96	0.02 mg/l
2	pH of treated water	0.85	0.11
3	Alum dose	0.94	3.2 mg/l

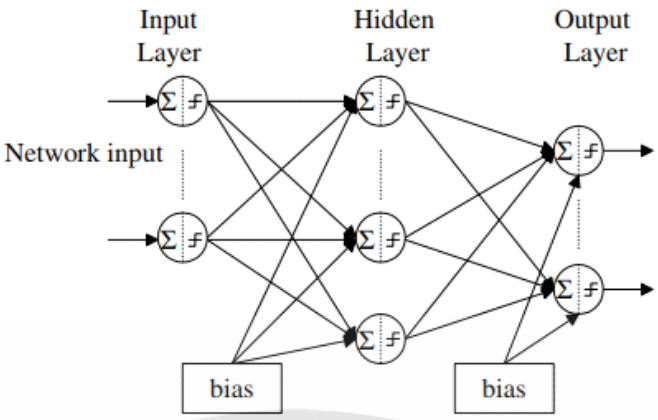
ตามตารางที่ 2.1 ค่า R-Square นั้นมีค่าสูงมาก และค่า MAE มีค่าต่ำมาก ซึ่งเป็นตัวบ่งชี้ได้ว่าแบบจำลองมีประสิทธิภาพ แต่การทำนายนั้นเป็นการทำนายตัวของมันเอง ประสิทธิภาพสูงจึงไม่ใช่เรื่องแปลก ควรมีการปรับปรุงเรื่องการแบ่งชนิดของข้อมูล เพื่อที่จะสร้างแบบจำลอง และทำนาย



รูปที่ 2.8 กราฟเปรียบเทียบระหว่าง ค่าจริง และค่าจากแบบจำลองของแต่ละพารามิเตอร์ (Maire และคณะ, 2004)

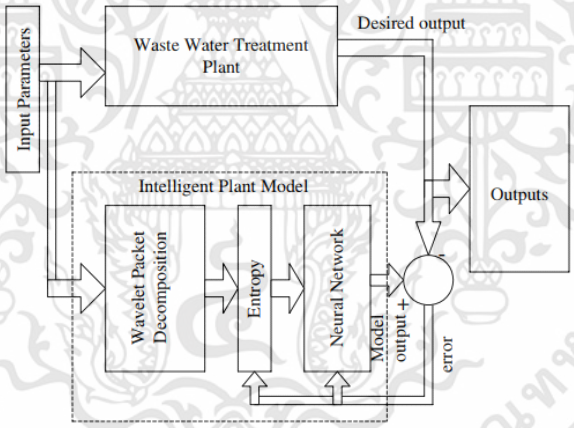
ตามรูปที่ 2.8 เป็นการเปรียบเทียบค่าจากการทำนาย และค่าจริง โดยค่าจากการทำนายนั้นจำอยู่แนวแกนตั้ง และค่าจากค่าจริงจะเป็นแนวแกนนอน โดนการกำหนดจุด แล้วลากเส้นผ่าน ข้อเสียของกราฟจากงานวิจัยชิ้นนี้คือ ขนาดของแกนนอนและแกนตั้งมีค่าไม่เท่ากัน อาจจะทำให้การอ่านกราฟด้วยตาเปล่านั้นทำยาก

Hanbay และคณะ (2008) ได้ทำงานวิจัยเกี่ยวกับการทำนายค่าประสิทธิภาพของการบำบัดน้ำในโรงประปาโดยใช้ Neural Networks โดยในแต่ละวันค่าพารามิเตอร์ต่างๆถูกเก็บโดยห้องแลป ในประเทศมาเลเซีย และ ตุรกี และจะวัดค่าประสิทธิภาพของแบบจำลองจากค่า Total Suspended Solid ส่วนค่าที่ได้มาจากแบบจำลอง เมื่อเทียบกับค่าที่ต้องการนั้นค่อนข้างเป็นที่น่าพอใจ



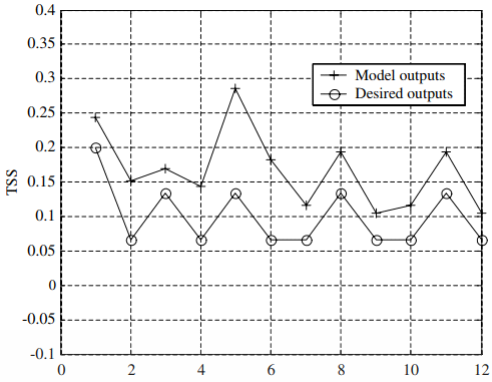
รูปที่ 2.9 โครงสร้างของ Neural Network (Hanbay และคณะ, 2008)

ตามรูปที่ 2.9 Neural Network นั้นจะใช้เวลาในการสร้างแบบจำลองนานกว่า การสร้างแบบจำลองแบบ Tree มาก แต่เรื่องประสิทธิภาพนั้นค่อนข้างดี เหมาะสำหรับใช้กับข้อมูลที่มีจำนวนมาก เพื่อประสิทธิภาพที่สูงขึ้น



รูปที่ 2.10 แบบแผนขั้นตอนการทำนาย (Hanbay และคณะ, 2008)

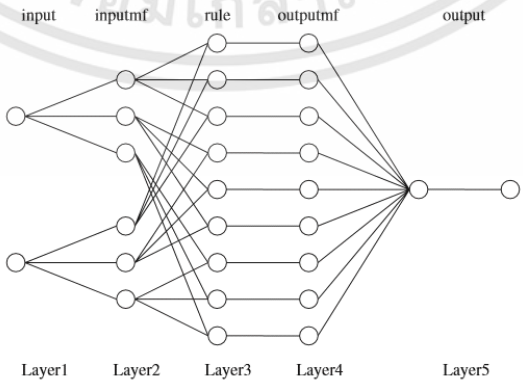
ตามรูปที่ 2.10 ขั้นแรกของการทำนายนั้น ข้อมูลพารามิเตอร์ถูกส่งไปยัง ส่วนที่คำนวณหาประสิทธิภาพน้ำที่ควรเป็น และส่วนที่ทำนาย แล้วถ้าส่วนที่ทำนายมีค่าต่างจากส่วนที่ต้องการมากเกินไป จะถูกส่งกลับไป ตรวจสอบอีกรอบในขั้นตอนการสร้างโมเดล ถ้าผลออกมาดีจึงจะถูกนำออกไป



รูปที่ 2.11 กราฟเปรียบเทียบค่า TSS จากการทำนาย และค่าที่ต้องการ (Hanbay และคณะ, 2008)

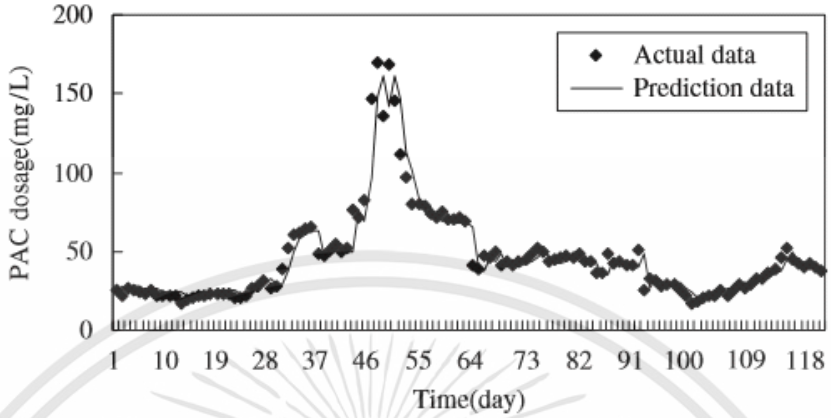
ตามรูปที่ 2.11 กราฟออกมาค่อนข้างดี เพราะว่าข้อมูลมี่ใช้เปรียบเทียบกันมีเพียง 12 ตัว ซึ่งน้อยเกินไป ดังนั้นผู้วิจัยสามารถเลือกช่วงการทำนายที่ดีที่สุดมาเป็นข้อมูลอ้างอิงเพราะส่วนอื่นของการทำนายอาจจะออกมาไม่ดีเลยก็เป็นได้ รูปทรงของกราฟทั้ง 2 ค่อนข้างใกล้เคียงกัน แต่ค่าที่ออกมานั้น อาทิเช่น ค่าที่ต้องการอยู่ที่เกือบ 0.05 แต่ค่าที่ทำนายได้อยู่ที่ 0.15 ถึงจะต่างกันแค่ 0.1 แต่เมื่อต้องการเปรียบเทียบโมเดลโดยใช้ค่า เช่น R Square หรือค่า RMSE แล้วนั้น ประสิทธิภาพของโมเดลจะต่ำมาก

Wu และคณะ (2008) ได้มีการวิจัยเกี่ยวกับ การทำนายปริมาณการเติมสาร Coagulant หรือ Poly Aluminum Chloride ใน Northern Taiwan's Surface Water แบบ Real-Time โดยการใช้เทคโนโลยี Artificial Neural Networks และ Adaptive Network-Based Fuzzy Inference System ซึ่งถือว่าเป็นเทคโนโลยีที่มีประสิทธิภาพสูงในขณะนั้น แต่ปัจจุบันเทคโนโลยีได้พัฒนาไปอย่างรวดเร็ว และมีประสิทธิภาพสูงขึ้น จึงเป็นการดีที่จะนำเอาเทคโนโลยีปัจจุบันมาใช้ในการทำนายค่า เพื่อความแม่นยำที่มากขึ้น โดยเฉพาะค่าการเติมสาร Coagulant ที่ต้องใช้การวิเคราะห์ ประมวลผลอย่างละเอียดเพื่อให้ได้ค่าจากการทำนายที่นำไปใช้จริงได้



รูปที่ 2.12 แผนภาพแสดงการทำงานของระบบ Adaptive Network-Based Fuzzy Inference System (Wu และคณะ 2008)

ตามรูปที่ 2.12 ระบบ ANFIS มีความคลายคลึงกับ Multilayer Perceptron มาก เพียงแต่ต่างกันที่ความซับซ้อนในการคำนวณที่มากกว่า

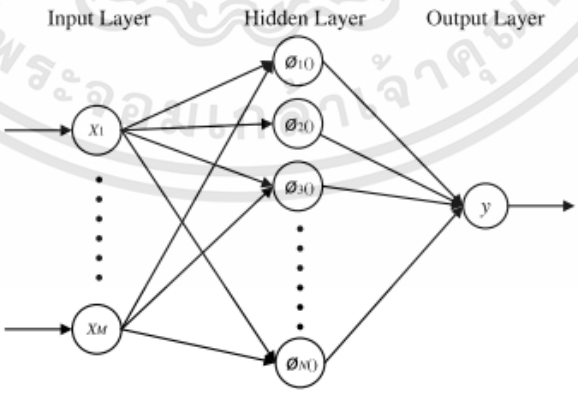


รูปที่ 2.13 กราฟระหว่างข้อมูลจากค่าจริงและค่าจากการทำนายของ ANFIS (Wu และคณะ, 2008)

ตามรูปที่ 2.13 จะเห็นได้ว่าการทำนายมีความแม่นยำสูง เพราะเป็นข้อมูลชุดเดียวกัน และจำนวนชุดข้อมูลที่นำมาทำนาย มีจำนวนน้อยมากค่าความแม่นยำเลยสูง เพราะถ้าคำนวนจากชุดข้อมูลที่ใหญ่กว่าความผิดพลาดก็ยอมมากกว่า

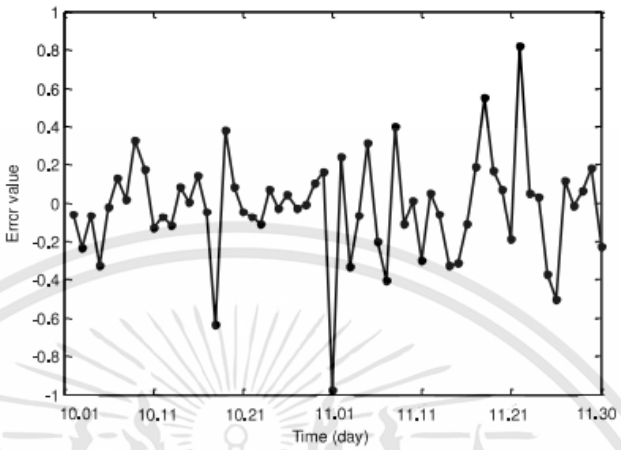
Hong และคณะ (2011) ได้ทำวิจัยเกี่ยวกับ Flexible Structure Radial Basis Function Neural Network หรือ FS-RBFNN และการใช้โมเดลจาก FS-RBFNN เพื่อทำนายคุณภาพน้ำ รวมไปถึงการเปรียบเทียบประสิทธิภาพของ FS-RBFNN กับวิธีอื่นอีกด้วย

ตามรูปที่ 2.14 โครงสร้างของ RBFNN นั้นค่อนข้างเหมือนกัน Neuro Networks ธรรมดา แต่จะต่างกันตรงที่แต่ละ Layer จะมีการนำค่าต่างๆ เข้าสมการต่อไป



รูปที่ 2.14 โครงสร้างของ RBFNN (Hong และคณะ 2011)

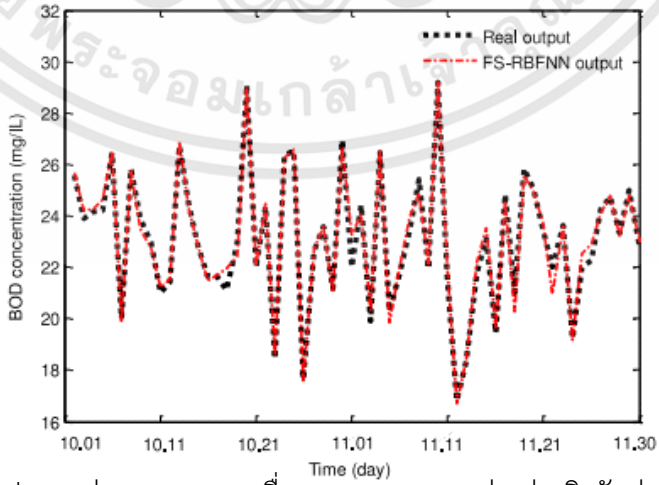
ตามรูปที่ 2.14 โครงสร้างของ RBFNN นั้นค่อนข้างเหมือนกัน Neuro Networks ธรรมดา แต่จะต่างกันตรงที่แต่ละ Layer จะมีการนำค่าต่างๆ เข้าสมการแล้วนำไปแปลผล ถึงจะไป Layer ต่อไป โดยขั้นตอนสุดท้ายมีแค่ Output เดียว



รูปที่ 2.15 กราฟเปรียบเทียบคุณภาพน้ำระหว่างค่าจริงกับค่าทำนาย (Hong และคณะ, 2011)

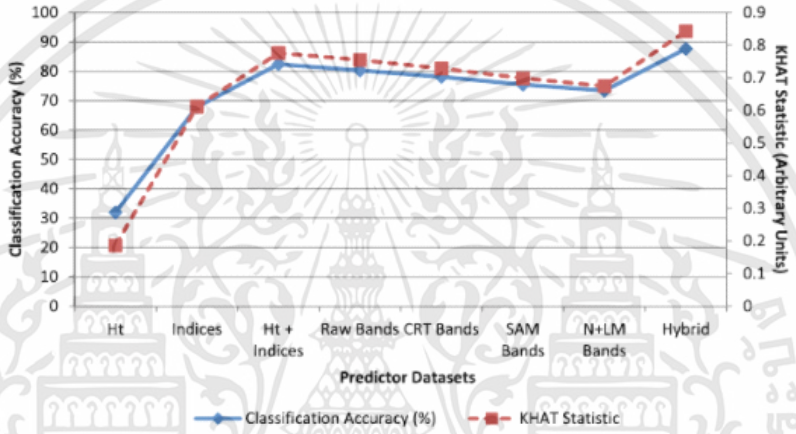
ตามรูปที่ 2.15 ค่าที่ได้จากการทำนายเมื่อเทียบกับค่าจริงแล้ว ผลออกมาถือว่าดีเยี่ยม ถึงข้อมูลจริงจะมีการแกว่งก็ตามแต่ก็สามารถทำนายได้อย่างมีประสิทธิภาพ โดยการทำนายนั้นได้ทำนายค่า BOD โดยใช้ ค่า Input เป็น COD, TSS, pH, Oil และ ซึ่งข้อมูล Input นั้นถือว่าหลากหลาย ซึ่งเป็นผลดี

ตามรูปที่ 2.16 ค่าที่เกิดจากความต่างของค่าจริงกับค่าทำนายนั้น จะเห็นได้ว่า มีความคลาดเคลื่อนน้อยมาก ในบางจุดแทบจะเป็น 0 หรือมากที่สุดก็มีความคลาดเคลื่อนไม่เกิน 1 ซึ่งแสดงให้เห็นว่าโมเดลที่สร้างจากวิธี FS-RBFNN มีประสิทธิภาพเป็นอย่างมาก และเหมาะสมที่จะทดสอบกับการใช้งานในแบบอื่น เพื่อดูว่าประสิทธิภาพยังคงสูงอยู่หรือไม่



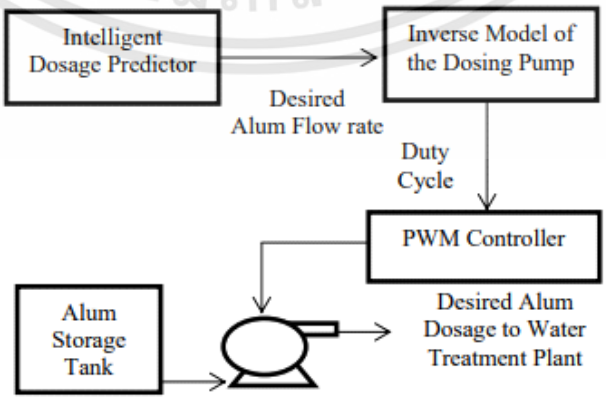
รูปที่ 2.16 กราฟแสดงค่าความคลาดเคลื่อนของ BOD ระหว่างค่าจริงกับค่าจากการทำนาย (Hong และคณะ, 2011)

Naidoo และคณะ (2012) ได้มีการใช้การขุดเหมืองข้อมูล เพื่อที่จะสร้างแบบจำลองในการทำนายค่า Savanna Tree Species ในพื้นที่ The Greater Kruger National Park, South Africa โดยใช้ ฟังก์ชัน Random Forest ในการทำนายค่าชุดข้อมูล โดยมีค่าต่างๆ เพื่อนำไปสร้างแบบจำลอง เช่น ค่าความสูงของต้นไม้ และค่า คลอโรฟิลล์บี ผลออกมาคือแบบจำลองมีความแม่นยำในการทำนายสูงถึง 87.68% ซึ่งในระดับความแม่นยำนี้ สามารถนำแบบจำลองไปประยุกต์ใช้ในการทำนายค่าในระบบใหม่ ที่คล้ายกันได้ แต่ควรจะมีการเพิ่มเติมในด้านของการเปรียบเทียบผลระหว่างฟังก์ชันที่เลือกใช้ ว่าเหตุใดจึงเลือกใช้ Random Forest แทนที่จะเป็นฟังก์ชันอื่น



รูปที่ 2.17 กราฟเปรียบเทียบความแม่นยำของค่าทำนายและค่าจริง (Naidoo และคณะ, 2012)

Kumar และคณะ (2013) ได้มีการใช้ Artificial Intelligence ในการควบคุมปริมาณการเติมสารส้มในโรงบำบัดน้ำเสีย ทั้งการคำนวณปริมาณสารส้มและการเติมอัตโนมัติ รวมอยู่ในหัวข้อเดียวกัน การทำนายค่าของสารส้มนั้นใช้ ระบบ ANN และ ANFIS เหมือนกันหัวข้อข้างบน



รูปที่ **2.18** แผนภาพการดำเนินการของระบบเติมสารส้มอัตโนมัติ (Kumar และคณะ, 2013)

ตามรูปที่ 2.18 กระบวนการเดียวที่เกิดจากการคำนวณคือกระบวนการแรก (Intelligent Dosage Predictor) นอกจากนั้นเป็นกระบวนการต่อเนื่อง จากกระบวนการแรก นั่นหมายความว่า ถ้าเกิดการคำนวณผิดพลาดในกระบวนการแรก จะเกิดการผิดพลาด

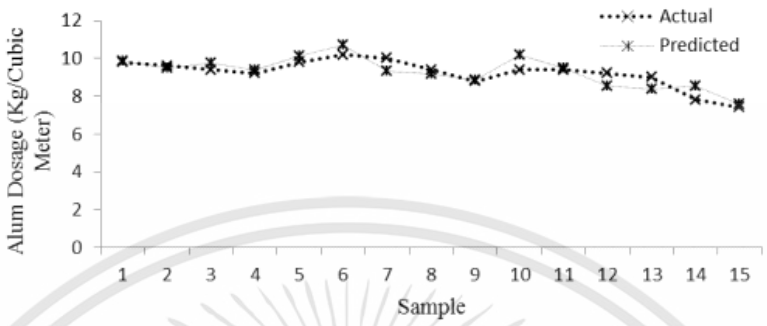
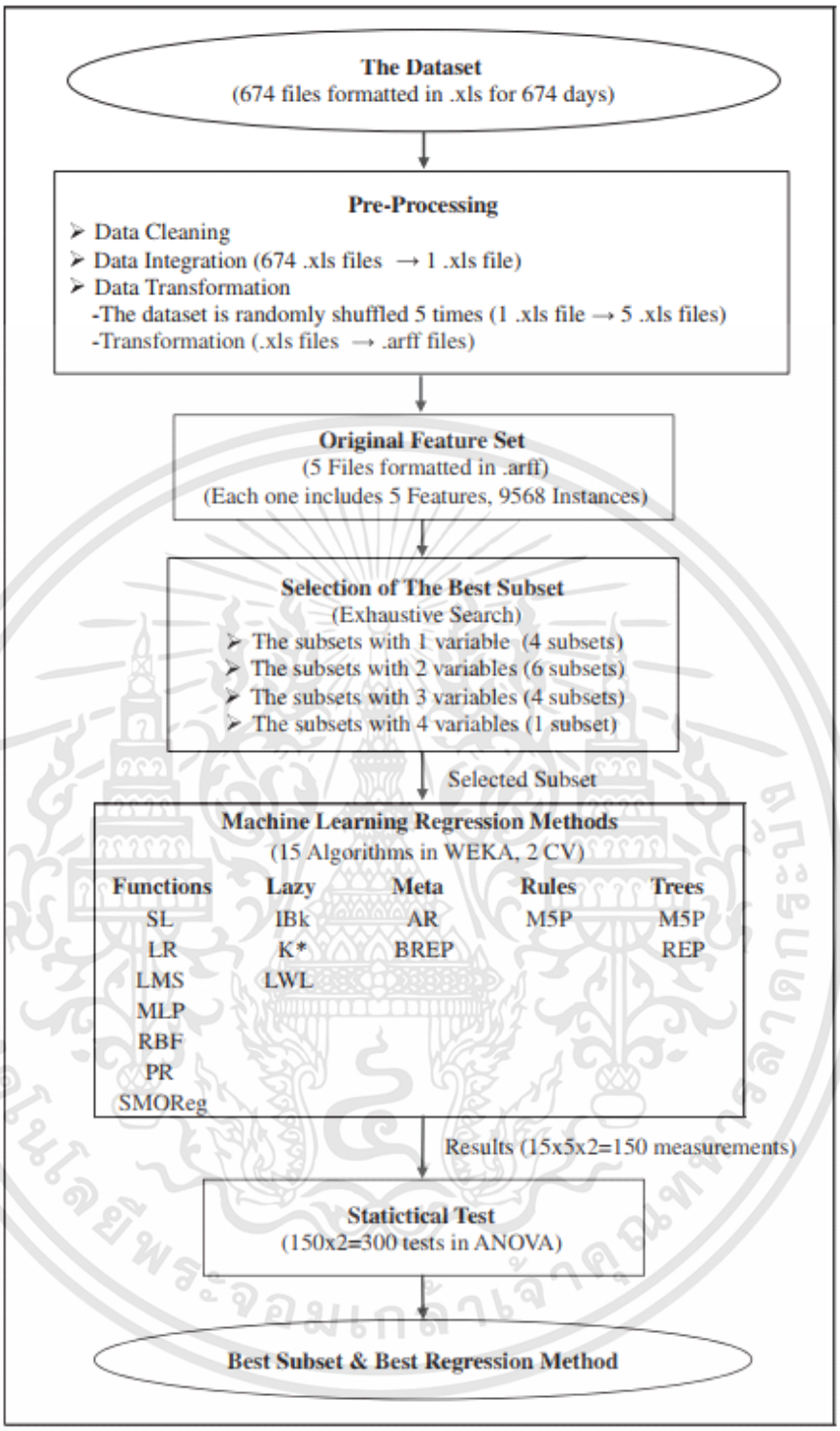


Fig.7 Actual and predicted values of tested data in ANFIS

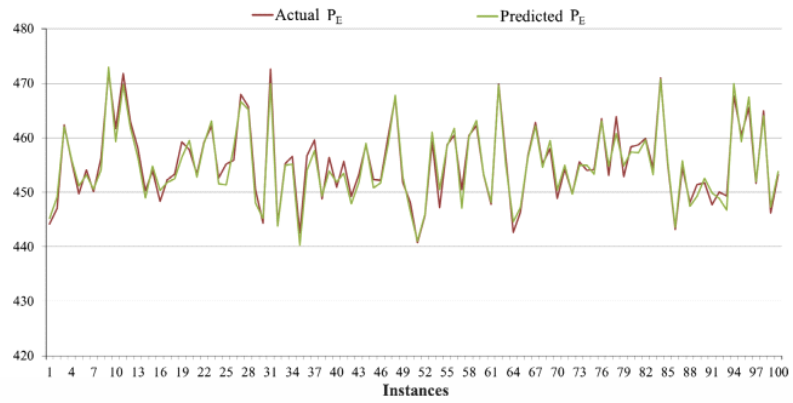
รูปที่ 2.19 กราฟข้อมูลจากค่าจริง และค่าจากการทำนายของ ANFIS (Kumar และคณะ, 2013)

ตามรูปที่ 2.19 จะเห็นได้ว่าข้อมูลมีความแม่นยำสูง เพราะค่าจริงนั้นแทบจะเป็นเส้นตรง ซึ่งในทางการทดสอบ ข้อมูลจะไม่ใช้ช่วงข้อมูลที่ง่ายต่อการทำนายจนเกินไป จุดประสงค์เพื่อที่จะทราบถึงประสิทธิภาพจริงของแบบจำลอง

Pinar (2014) ได้ทำการทำนาย ภาระสูงสุดของกำลังไฟฟ้าขาออก ของโรงไฟฟ้าหลักที่ทำงานร่วมกับโรงไฟฟ้าพลังความร้อนร่วม โดยใช้ Machine Learning เพื่อทำกำไรสูงสุด ในงานวิจัยชิ้นนี้มีการเปรียบเทียบ และยกตัวอย่าง ทฤษฎีที่ใช้ เพื่อประสิทธิภาพสูงสุดของการสร้างแบบจำลอง โดยใช้ทั้งหมด 15 ทฤษฎี และมี 4 ตัวแปรหลักในการผลิตพลังงาน ประกอบด้วย อุณหภูมิโดยรอบ, ความกดอากาศ, ความชื้นสัมพัทธ์, ความดันของท่อปล่อยไอน้ำ ซึ่งตัวแปรเหล่านี้ล้วนมีผลกระทบต่อพลังงานที่ผลิตได้ เก็บข้อข้อมูลย้อนหลังทั้งหมด 6 ปี ในการสร้างแบบจำลอง โดยใช้โปรแกรม WEKA ในการสร้างแบบจำลองของตัวแปรที่มีผลต่อการผลิตพลังงาน มีข้อมูลจาก 674 วัน นำไปผ่านการ ทำให้ข้อมูลที่มีอยู่สมบูรณ์ขึ้น ข้อมูลมีทั้งหมด 9568 ตัว หลังจากนั้นได้ทำการแบ่งชุดข้อมูลเพื่อหาชุดข้อมูลที่ดีที่สุด แล้วนำมาผ่านการสร้างแบบจำลอง ขั้นตอนนำไปทดสองทางสถิติ เพื่อที่จะได้ทราบชุดข้อมูล และแบบจำลองที่ดีที่สุด และจำใช้ในการทำนายผลในลำดับต่อไป แบบจำลองจากทฤษฎี BREP ในหมวด Meta-Learning Algorithms มีความแม่นยำที่สุด โดยมีค่า MAE = 3.220 และ RMSE = 4.239 ส่วนรองลงมานั้นเป็น M5P ในหมวด Tree-Based Learning Algorithms มีค่า MAE = 3.428 และ RMSE = 4.428 และ Reptree ในหมวด Tree-Based Learning Algorithms มีค่า MAE = 3.424 และ RMSE = 4.518 ซึ่งถือว่าผลการทดสอบแบบจำลองออกมาใกล้เคียงกันมาก ถือว่าชุดข้อมูลที่ใช้นั้นค่อนข้างมีคุณภาพ



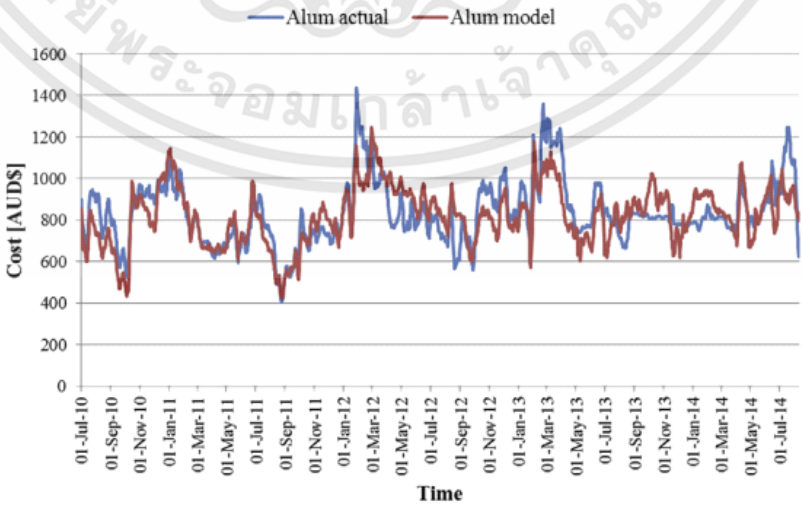
รูปที่ 2.20 แผนภาพการดำเนินการสร้างแบบจำลองและทำนายผล (Pinar, 2014)



รูปที่ 2.21 กราฟเปรียบเทียบพลังงานไฟฟ้าสูงสุดที่ผลิตได้ระหว่างค่าจริงกับค่าจากการทำนาย (Pinar, 2014)

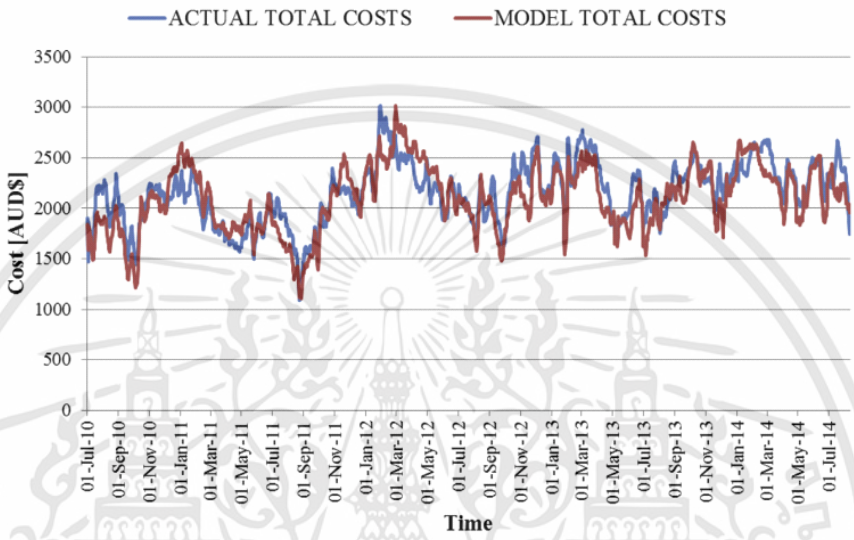
ตามรูปที่ 2.21 ค่าที่ได้จากการทำนายโดยใช้โมเดลจาก BREP นั้น ถือว่าผลออกมาดีมาก และใช้งานได้จริง เพื่อการทำกำไรสูงสุดในการผลิตไฟฟ้า แต่เป็นเพียงการทำนายข้อมูลชุดเดียวกับที่นำมาสร้างโมเดล ซึ่งความเป็นจริงนั้น ถ้านำค่าจริงมาทดสอบ ผลที่ได้อาจจะไม่สูงตาม ที่แสดงผลออกมาในกราฟ ก็เป็นได้ ส่วนที่ต้องปรับปุงคือ ควรใช้แบบจำลองในการทำนาย ข้อมูลชุดที่ไม่เกี่ยวข้อง จะทำให้ได้ค่าประสิทธิภาพจริงที่นำมาใช้งานได้

Bertone และคณะ (2016) ได้ทำวิจัยเกี่ยวกับการทำนายค่าใช้จ่ายของการบำบัดน้ำเสีย เพื่อจัดการปริมาณน้ำดิบที่เข้ามาในโรงบำบัด เนื่องจากค่าใช้จ่ายของการบำบัดน้ำนั้นมาจากค่าสารเคมีที่ต้องเติมลงไป ค่าของพลังงานที่ถูกใช้ในปั๊มน้ำ และค่าอื่นๆ ดังนั้นถ้าสามารถทำนายค่าใช้จ่ายที่เกิดขึ้น ก็จะสามารถรู้ปริมาณน้ำดิบที่ต้องรับเข้ามาได้ และบริหารได้อย่างมีประสิทธิภาพ การทำนายค่าใช้จ่ายในงานวิจัยชิ้นนี้ ได้รวมไปถึงการทำนายปริมาณสารเคมีที่ต้องเติมลงไปในน้ำด้วย เพราะสารเคมีที่เติมลงไปถือว่าเป็นค่าใช้จ่ายที่เพิ่มขึ้น



รูปที่ 2.22 กราฟระหว่างค่าสารส้มจริง กับค่าจากแบบจำลอง (Bertone และคณะ, 2016)

ตามรูปที่ 2.22 เป็นเพียงการทดลองแบบจำลอง ไม่ใช่การทำนายที่เกิดขึ้นจริง แต่ถือได้ว่าค่าจากแบบจำลอง เมื่อเทียบกับค่าจริง ผลที่ออกมาค่อนข้างดี แต่เมื่อนำไปทำนายจริงแล้ว ค่าจะเพี้ยนต่างจากนี้มาก และทางผู้วิจัยไม่ได้มีการระบุ วิธีการสร้างแบบจำลองดังกล่าวจึงไม่สามารถนำไปใช้งานได้ รวมถึงค่าที่จะนำไปทำนายค่าใช้จ่ายนั้น ไม่ได้มีเพียวแค่สารส้ม แต่มี ค่าต่างๆจำนวนมาก



รูปที่ 2.23 กราฟระหว่างค่าใช้จ่ายจริง และค่าใช้จ่ายจากแบบจำลอง (Bertone และคณะ, 2016)

ตามรูปที่ 2.23 ผลเทียบจากค่าจริงและค่าจากแบบจำลองมีประสิทธิภาพค่อนข้างสูงเนื่องจาก งานวิจัยชิ้นนี้ได้ทำการใส่ข้อมูลค่าใช้จ่ายทั้งเข้ามาในระบบเลย แทนที่จะเป็นการรวมค่าสารเคมีและค่าพลังงานจากแบบจำลองเข้าด้วยกัน แล้วนำมาเปรียบเทียบ ดังนั้นถึงแบบจำลองนี้จะมีประสิทธิภาพสูง แต่ก็ไม่สามารถใช้ได้จริง

Kiyoki และคณะ (2016) ได้มีการใช้ Data mining ในการวิเคราะห์ข้อมูลทางด้านต่างๆของสิ่งแวดล้อมไม่ว่าจะเป็น ความหลากหลายของธรรมชาติและสัตว์ โดยใช้ อุณหภูมิ ระดับของคาร์บอนไดออกไซด์ ระดับน้ำทะเล เป็นต้น มาใช้เป็นตัวแปรต้น และรวมไปถึงการวิเคราะห์ความสัมพันธ์ของตัวแปรในน้ำอีกด้วย ได้มีการใช้ Semantic computing เข้ามาวิเคราะห์ข้อมูลที่เก็บรวบรวมจำนวนมากผ่านทางด้านของ โปรแกรม 5D world map ซึ่งเป็น Data base ขนาดใหญ่ เพื่อการวิเคราะห์ข้อมูลที่ดีขึ้นไปอีกระดับ

ตารางที่ 2.2 ตารางแสดงค่า RMSE ของปริการเติมสารส้มในแบบจำลอง (Kiyoki และคณะ, 2016)

Result	Natural Networks	SVM	Decision tree	Decision tree forest
RMSE	4.09	3.66	3.78	2.37

ตามตารางที่ 2.2 ได้มีการทำนายค่าการเติมปริมาณสารส้มในแหล่งน้ำแหล่งหนึ่งโดยการนำข้อมูลมาใช้งาน จากฐานข้อมูล 5D world map ซึ่งค่าที่ออกมานั้น Natural Networks = 4.09, SVM = 3.66, Decision tree = 3.78, Decision tree forest = 2.37 จะเห็นได้ว่าทฤษฎีที่มีค่า RMSE ต่ำที่สุดนั้นคือ Decision tree forest ซึ่งในกรณีนี้ใช้ข้อมูลทั้งหมดจำนวน 2014 ชุด ในการเรียนรู้ ถือว่าจำนวนมากพอสมควร แต่ในการเปรียบเทียบกับแหล่งน้ำอื่นนั้น จำเป็นต้องเพิ่มการวัดการคำนวณค่าอื่นๆเข้าไปด้วย เพื่อเป็นการเปรียบเทียบแบบจำลองได้อย่างมีประสิทธิภาพ

Almasi และคณะ (2017) ได้มีการใช้ Decision Tree ในการทำนายราคาน้ำมัน มีการเก็บข้อมูลของราคาน้ำมัน 24 ปีและใช้ 3 ทฤษฎีในการสร้างแบบจำลอง ได้แก่ Decision stump, Random forest, Random tree, M5P และ Reptree และทำการวิเคราะห์ประสิทธิภาพของแบบจำลอง โดยใช้ค่า RMSE (Root Mean Square Error), MAE (Mean Absolute Error), CC (Correlation Coefficient), RRSE (Root Relative Squared Error), RAE (Root Absolute Error) ในการวัดผล ตัวแปรมีทั้งหมด 8 ตัว ได้แก่ ปี, ฤดูกาล(แบ่งเป็นหน้าร้อนกับหน้าหนาว), ราคาเฉลี่ยของสัปดาห์ที่ผ่านมา, สัปดาห์ที่เท่าไหร่ของข้อมูลทั้งหมด, สัปดาห์ที่เท่าไหร่ของปี, ค่าสัมประสิทธิ์การเปลี่ยนแปลงราคาของน้ำมันโลก, ความต้องการของตลาดโลก, ราคาการซื้อขายน้ำมัน

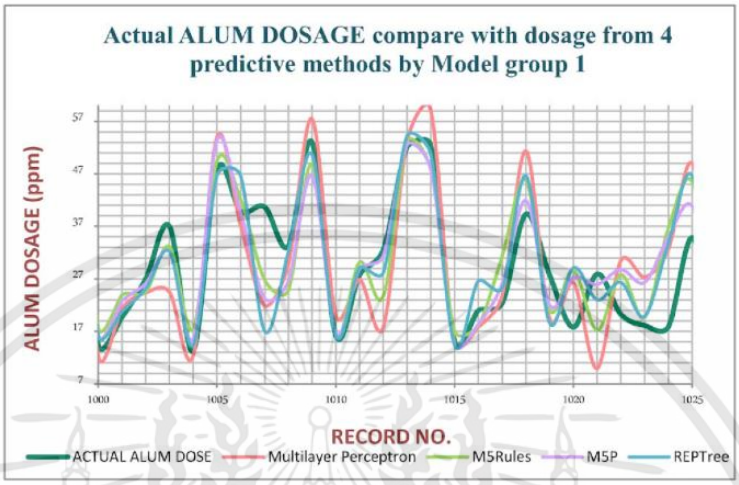
ตารางที่ 2.3 การเปรียบเทียบค่าจากการทดสอบแบบจำลอง (Almasi และคณะ, 2017)

	Decision Stump	M5P	Random Forest	Random Tree	REPTree
CC	0.8709	0.9999	0.9994	0.9982	0.9982
MAE	7.5655	0.1135	0.4275	0.8136	0.6344
RMSE	11.3595	0.275	0.8057	1.403	1.3688
RAE	44.3986%	0.6662%	2.509%	4.7749%	3.7227%
RRSE	49.1033%	1.1888%	3.4826%	6.0648%	5.9167%

ตามตารางที่ 2.3 ทฤษฎีที่มีประสิทธิภาพสูงที่สุดคือ M5P ซึ่งมีค่า RMSE เพียง 0.1135 และมีค่า MAE 0.275 แม่นยำรองลงมาคือ Random Forest มีค่า RMSE = 0.4275 และมีค่า MAE 0.8057 ถ้าสังเกตจากค่าดังกล่าว M5P นั้นมีความแม่นยำสูงมาก และมากกว่า ทฤษฎีอื่นเป็นเท่าตัวเลยทีเดียว ซึ่งสามารถนำไปใช้งานได้จริง ค่าตัวแปรที่ใช้มีความสมเหตุสมผล มีการใช้ข้อมูลชุดเก่ามาวิเคราะห์ และใช้ข้อมูลจำนวนมากถึง 24 ปี แบบจำลอง M5P จากงานวิจัยชิ้นนี้ถือว่ามีประสิทธิภาพเป็นอย่างมาก

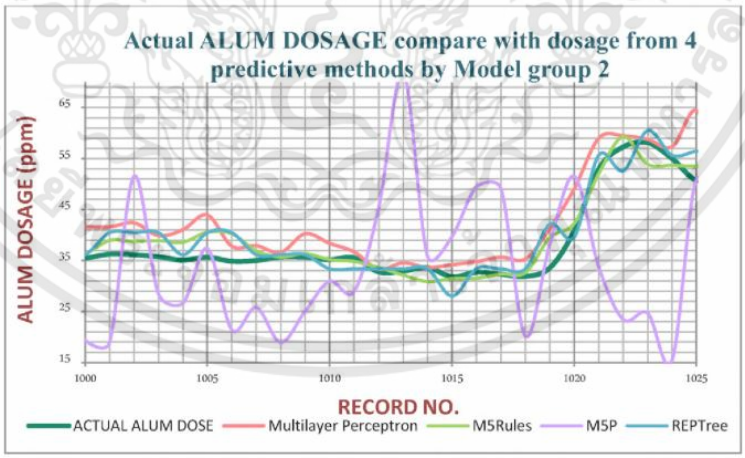
Chawakitchareon และคณะ (2017) งานวิจัยชิ้นนี้ได้มีการทำนายปริมาณการเติมสารส้ม โดยใช้ข้อมูลตัวแปรต้นทั้งหมด 5 ตัว ความขุ่น (Turbidity), ค่าความเป็นด่าง (Alkalinity), การนำไฟฟ้า (Conductivity), ค่าพีเอช (pH) และ สี (Color) ส่วนตัวแปรตาม

คือ ปริมาณการเติมสารส้ม และได้แบ่งข้อมูลออกเป็น 2 กลุ่ม แบ่งตามการคลีนข้อมูล คือ กลุ่มที่ 1 คลีนข้อมูลที่มีช่องว่างหรือค่ามากกว่าความเป็นจริงโดยการตัดหรือทดแทนข้อมูลเข้าไปด้วยค่าเฉลี่ย ส่วน กลุ่มที่ 2 คลีนข้อมูลมีช่องว่างหรือค่ามากกว่าความเป็นจริง โดยการลบข้อมูลนั้นทิ้งไป



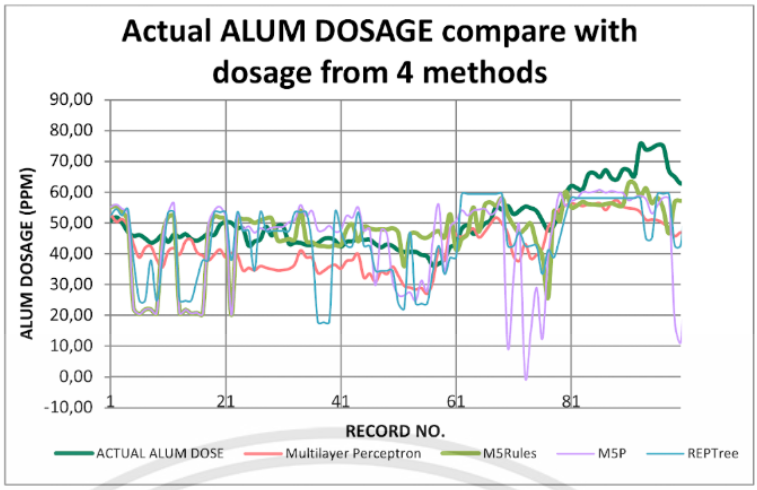
รูปที่ 2.24 กราฟเปรียบเทียบแบบจำลองโดยใช้ข้อมูล กลุ่มที่ 1 (Chawakitchareon และคณะ, 2017)

ตามรูปที่ 2.24 ค่าการทำนายของแต่ละทฤษฎีนั้น ออกมาค่อนข้างดีเพราะ ได้มีการทดสอบกับข้อมูลเพียง 25 ตัว และทุกทฤษฎีผลออกมาใกล้เคียงกัน



รูปที่ 2.25 กราฟเปรียบเทียบแบบจำลองโดยใช้ข้อมูล กลุ่มที่ 2 (Chawakitchareon และคณะ, 2017)

ตามรูปที่ 2.25 ค่าการทำนายของแต่ละทฤษฎีนั้น ไม่ค่อยสอดคล้องกับค่าจริง ซึ่งทดสอบกับข้อมูลชุดเดียวกับกลุ่มที่ 1 แต่ไม่ทราบว่าข้อมูลตัวเดียวกันหรือไม่



รูปที่ 2.26 กราฟเปรียบเทียบค่าจากการทำนายจริงโดยใช้แบบจำลองที่สร้างจากข้อมูลกลุ่มที่ 1 (Chawakitchareon และคณะ, 2017)

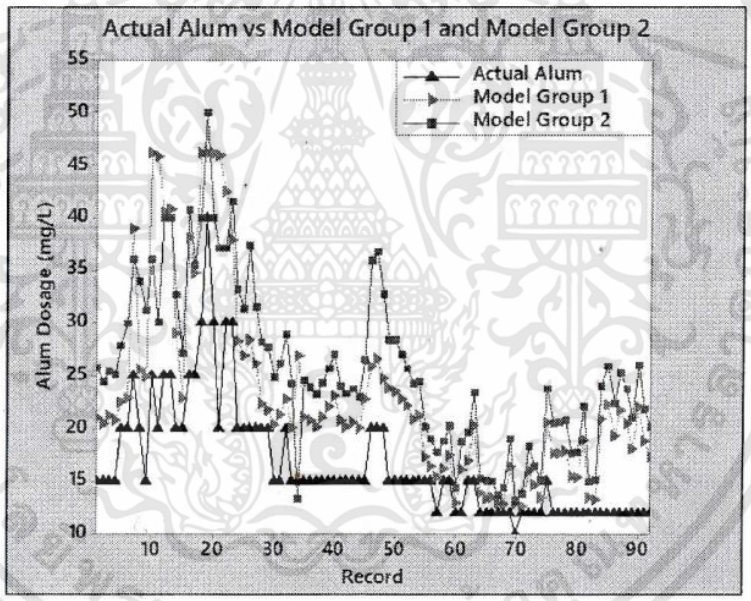
ตามรูปที่ 2.26 จะเห็นได้ว่าเมื่อเพิ่มตัวทดสอบเข้าไปจาก 100 ตัว ผลที่ออกมาค่อนข้างคลาดเคลื่อน ต่างจาก กราฟก่อนหน้านั้นที่ใช้ ตัวทดสอบเพียง 25 ตัว เพราะการใช้ตัวแปรจำนวนน้อยมาทำการสร้างแบบจำลอง แล้วต้องทำนายตัวทดสอบจำนวนมาก ค่าก็จะเพี้ยนไปมากกว่าปกติ โดยทางผู้วิจัยได้นำความความคลาดเคลื่อนจากงานวิจัยชิ้นนี้มาพิจารณาการคัดแยกข้อมูล และการตัดตัวแปรที่ไม่จำเป็นออก

Ladsavong และคณะ (2017) ได้ทำการทำนายปริมาณการเติมสารส้มของ Dongmarkkaiy Water Treatment Plant (DWTP) และได้เก็บข้อมูลตัวแปรของน้ำดิบทั้งสิ้น 2,891 ชุด โดยใช้ตัวแปรต้นเป็น ค่าความขุ่น (Turbidity), ค่าความเป็นด่าง (Alkalinity) และ ค่าพีเอช (pH) ส่วนตัวแปรตามเป็น ปริมาณการเติมสารส้ม (Alum Dosage) ซึ่งผู้วิจัยได้นำบทความชิ้นนี้มาพิจารณา และได้เปลี่ยนจากตัวแปร ค่าความเป็นด่าง (Alkalinity) เป็น ปริมาณของแข็งแขวนลอย (Suspended solids) เพราะว่าแบบจำลองจะสร้างความสัมพันธ์ของตัวแปรได้ง่ายกว่า และได้แบ่งข้อมูลออกเป็น 2 กลุ่ม กลุ่มแรกคือฤดูแล้ง กลุ่มที่ 2 คือฤดูฝนโดยในบทความนี้ได้ใช้วิธีการ Cross validation ในการจัดการกับข้อมูล ซึ่งทางผู้วิจัยได้เปลี่ยนมาเป็นแบบ Split แทน เพราะว่าสามารถกำหนดของเขตของการสร้างแบบจำลองได้มีประสิทธิภาพมากกว่า

ตารางที่ 2.4 ตารางแสดงค่า จากการทำนายข้อมูลโดยใช้แบบจำลองจากข้อมูล กลุ่มที่ 2 (Ladsavong และคณะ, 2017)

	MLP	M5Rules	M5P	REPTree
RMSE	2.29	3.47	6.26	6.47
MAE	1.37	2.84	2.92	2.77

ตามตารางที่ 2.4 จะเห็นได้ว่าค่า RMSE และ MAE นั้นมีค่าน้อยมากซึ่งอาจเป็นผลมาจากแหล่งน้ำดิบของ Dongmarkkaiy Water Treatment Plant (DWTP) มีความสะอาดมากกว่าที่ไทย จึงได้มีการเติมปริมาณสารส้มน้อย ค่า RMSE จึงน้อยตามไปด้วย ดังนั้นจึงไม่สามารถบอกถึงประสิทธิภาพของแบบจำลองได้ทั้งหมด ผู้วิจัยได้เพิ่มค่าชี้วัดในงานวิจัยชิ้นนี้เพิ่มมา 2 ค่า คือ Correlation coefficient และ R-square เพื่อที่จะสามารถชี้วัดค่าประสิทธิภาพที่แท้จริงของแบบจำลองได้

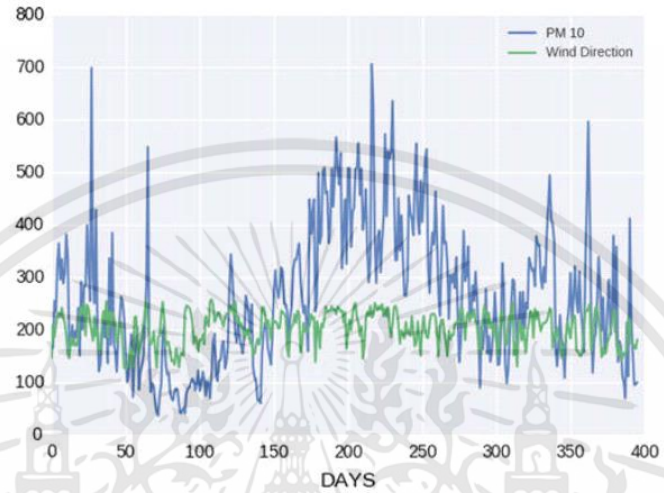


รูปที่ 2.27 กราฟแสดงผลการทำนาย ระหว่างค่าจริง และค่าจากแบบจำลอง ทั้ง 2 กลุ่ม (Ladsavong และคณะ, 2017)

ตามรูปที่ 2.27 จะเห็นได้ว่าถึงแม้ค่า RMSE จะน้อย แต่ผลการทำนายโดยใช้แบบทดสอบ 100 ตัวนั้นผลที่ออกมา ไม่ค่อยมีความสัมพันธ์กับค่าจริง ดังนั้นในการวิเคราะห์และสังเคราะห์ข้อมูลควรมีค่าที่ชี้วัด แบบจำลองที่มากกว่านี้

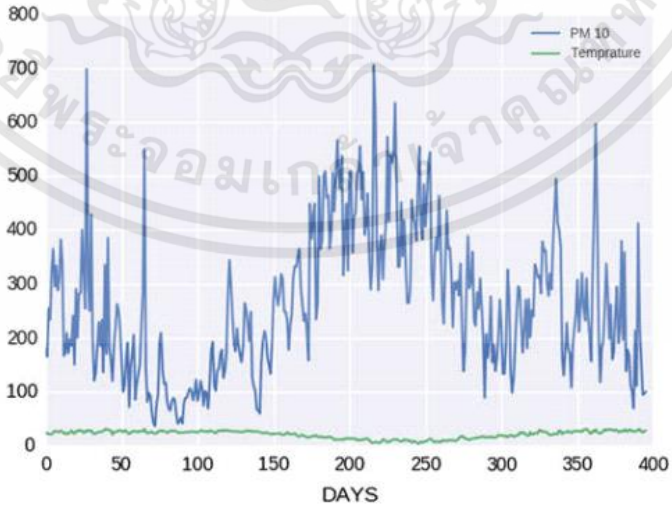
Akhtar และคณะ (2018) งานวิจัยชิ้นนี้ได้ทำการทำนาย และวิเคราะห์ระดับของมลพิษในเมือง Delhi ประเทศ India โดยใช้ Multilayer Perceptron มีการเก็บข้อมูลตัวแปรทั้งหมด 396 วัน ตัวแปรทั้งหมด 9 ตัวแปร ได้แก่ อุณหภูมิเฉลี่ย อุณหภูมิสูงสุด อุณหภูมิ

ต่ำสุด ความชื้น ความดัน ความเร็วลมสูงสุด ทัศนวิสัย ทิศทางของลม และฝุ่นควัน (PM10) ได้แบ่งข้อมูลสำหรับการสร้างแบบจำลอง 316 ตัวอย่าง และ ข้อมูลที่ใช้ทดสอบแบบจำลอง 80 ตัวอย่าง ในการสร้างแบบจำลองนั้นได้ใช้ทฤษฎีทั้งหมด 3 ทฤษฎี คือ Multilayer Perceptron, Support Vector Machines และ Naïve Bayes



รูปที่ 2.28 กราฟความสัมพันธ์สูงที่สุดระหว่างฝุ่นควันกับทิศทางของลม (Akhtar และคณะ, 2018)

ตามรูปที่ 2.28 คำบรรยายในงานวิจัยเขียนว่า กราฟดังกล่างมีความสัมพันธ์กันของตัวแปรสูงที่สุด ซึ่งการขึ้นลงของกราฟมีความเกี่ยวเนื่องกันพอสมควร แต่สามารถรับรู้ได้จากการสังเกตเท่านั้น ไม่สามารถวัดได้จากค่าทางสถิติ เพราะ ค่าตัวแปรทั้ง 2 มีหน่วยวัดที่ต่างกัน



รูปที่ 2.29 กราฟความสัมพันธ์ต่ำที่สุดระหว่างฝุ่นควันกับอุณหภูมิ (Akhtar และคณะ, 2018)

ตามรูปที่ 2.29 คำบรรยายในงานวิจัยเขียนว่า กราฟดังกล่างมีความสัมพันธ์กันของตัวแปรต่ำที่สุด ซึ่งการขึ้นลงของกราฟจากการสังเกตนั้น แทบไม่มีความเกี่ยวข้องกันเลยตามที่งานวิจัยชิ้นนี้กล่าว และไม่สามารถวัดได้จากค่าทางสถิติ

ตารางที่ 2.5 การเปรียบเทียบค่าจากการทดสอบแบบจำลองของ 3 ทฤษฎี (Akhtar และคณะ, 2018)

Technique	Accuracy (%)	Precision	Recall	F-measure
MLP	98.10	0.98	0.95	0.97
SVM	92.50	0.92	0.90	0.91
Naïve Bayes	91.25	0.90	0.87	0.89

ตามตารางที่ 2.5 ผลที่ได้ออกมาจากการทดสอบแบบจำลองนั้น Multilayer Perceptron มีความแม่นยำมากที่สุดคือ 98.1% รองลงมาคือ Support Vector Machines 92.5% และ Naïve Bayes 91.25% ซึ่งทั้ง 3 ทฤษฎีนั้นมีค่าความแม่นยำอยู่ในเกณฑ์ที่สูงมาก ทั้งๆที่ความเกี่ยวข้องของตัวแปรบางตัวนั้นแทบไม่มีความเกี่ยวข้องกันเลย ถ้ามีกราฟเปรียบเทียบค่าจากการทำนายจะเป็นประโยชน์กับผู้ที่อ่านงานวิจัยชิ้นนี้มาก และค่าตัวแปรตัวอื่นที่ไม่ระบุความสัมพันธ์ก็ไม่มีการชี้แจง

2.4.1 สรุปผลงานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่ได้มีมาก่อนแล้วนั้น ทางผู้วิจัยได้ความรู้และเทคนิคเพื่อเพิ่มประสิทธิภาพการทำนายค่าโดยผ่านการวิเคราะห์โดยการทำเหมือนข้อมูลเป็นอย่างมาก

แรงจูงใจที่คิดจะทำวิจัยในหัวข้อนี้คือ มีความคิดว่าการทำการทำลองในห้องแลปนั้นใช้เวลาค่อนข้างนาน และในขอบเขตการใช้งานของ Data Mining จากงานวิจัยที่กล่าวมาข้างต้น ครอบคลุมการนำมาทำนายปริมาณสารส้มในน้ำประปา และสามารถทำนายได้อย่างมีประสิทธิภาพอีกด้วย

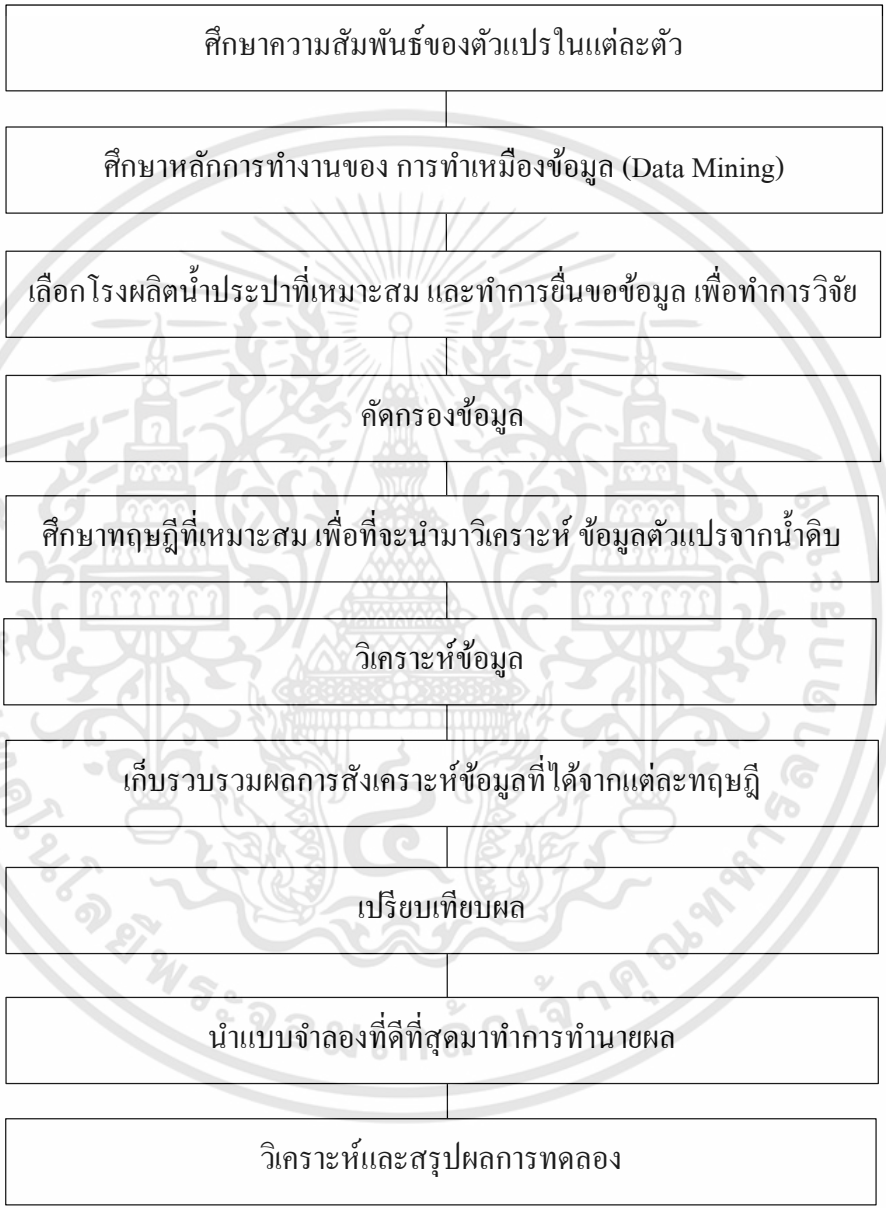
สิ่งที่ได้จากงานวิจัยที่ได้นำมาทบทวน คือ

1. การเลือกใช้ตัววัดค่าประสิทธิภาพแบบจำลองแบบเหมาะสม
2. การวิเคราะห์แบบจำลองอย่างมีประสิทธิภาพก่อนนำไปทำนายจริง
3. การใช้ทฤษฎีที่หลากหลายเพื่อทำให้มีทางเลือกมากขึ้นในการทำแบบจำลอง
4. ผลการทำลองไม่สามารถเปรียบเทียบกันได้เสมอไป โดยเฉพาะค่า RMSE และ MAE
5. การทำนายนั้น ค่าวัดประสิทธิภาพอาจจะออกมามีค่าสูง แต่พอนำไปใช้จริงอาจจะไม่สอดคล้องในบางกรณี ให้ดูค่าอื่นประกอบด้วย
6. จำนวนข้อมูลในการเรียนรู้ ควรที่จะมากกว่าจำนวนข้อมูลที่ใช้ทดสอบ

บทที่ 3

วิธีดำเนินการ

3.1 แผนผังการดำเนินงานวิจัย



รูปที่ 3.1 แผนผังการดำเนินงานวิจัย

3.2 การเลือกโรงประปาที่เหมาะสม

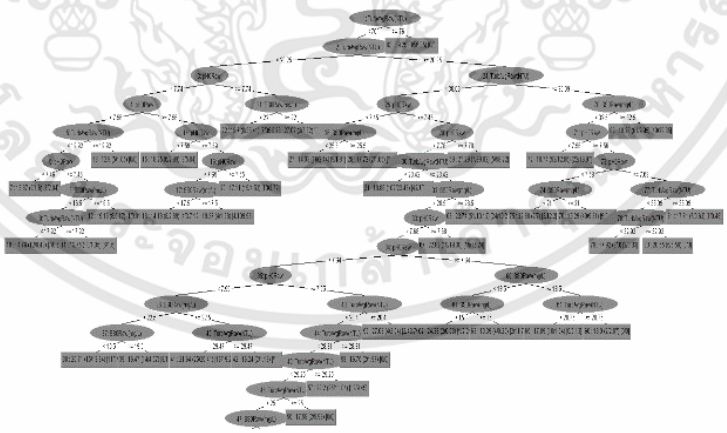
ในการเลือกสถานที่ที่เหมาะสมในการเก็บข้อมูลตัวแปรต่างๆในน้ำดิบนั้น จำเป็นจะต้องเลือกโรงผลิตน้ำประปาที่มีการเก็บข้อมูลทุกวันอย่างต่อเนื่องมาเป็นเวลานาน และอุปกรณ์การเก็บต้องเป็นมาตรฐานสากล ทางผู้วิจัยได้เลือกโรงผลิตน้ำประปาบางเขนเพื่อทำการเก็บข้อมูล และขอข้อมูลย้อนหลัง เพื่อที่จะนำข้อมูลมาวิเคราะห์ และสังเคราะห์ในลำดับต่อไป

3.3 แนวคิดในการคัดเลือกทฤษฎีที่นำมาใช้สังเคราะห์ข้อมูล

ในการเลือกทฤษฎีที่จะนำมาสังเคราะห์ข้อมูลนั้น ทางผู้วิจัยได้ใช้ทฤษฎีที่มีการสังเคราะห์เกี่ยวกับตัวเลข และการให้เหตุผล เข้ามาร่วมในการสังเคราะห์เพื่อความแม่นยำในการวิเคราะห์ผล และสร้างแบบจำลอง จึงได้เลือกทั้งหมด 4 ทฤษฎี

3.3.1 REPTree

REPTree ใช้ the regression tree logic และ creates multiple trees ในการทำซ้ำที่ต่างกัน หลังจากนั้นเลือกหนึ่งตัวที่ดีที่สุดจากต้นไม้ที่สร้างมาทั้งหมด เอามาเป็นตัวแทน ในการตัดต้นไม้ออก ในการทำนายข้อมูลนั้นจะใช้ค่า RMSE เข้ามาช่วยในการลดค่าความผิดพลาด REPTree นั้นย่อมาจาก Reduced Error Pruning Tree นั่นคือการเรียนรู้จากการตัดสินใจที่รวดเร็วและสร้าง Decision tree จากฐานข้อมูลโดยลดหรือเพิ่มความแปรปรวนเข้าไป วิธีนี้จะรับข้อมูลเฉพาะตัวเลข และข้อมูลที่ผิดพลาด จะถูกจัดการโดยใช้ C4.5’s Method

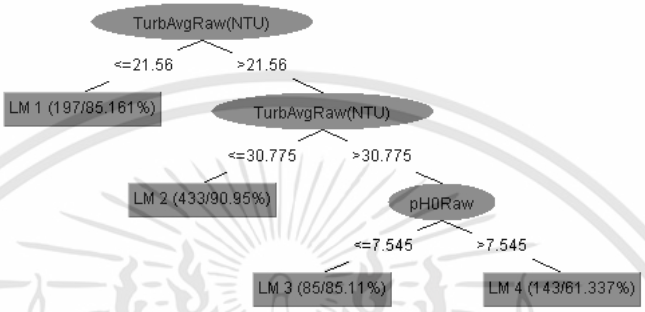


รูปที่ 3.2 แบบจำลอง REPTree

3.3.2 M5P

M5P เป็นตัวจำแนกประเภทที่นิยมที่สุดในการตัดสินใจ แบบต้นไม้ ในด้านของโครงสร้างนั้น แบบจำลองของการตัดสินใจแบบต้นไม้ ร่วมกับ ฟังก์ชันการถดถอยเชิงเส้น นำมาแทนที่ในส่วน

สุดท้ายของใบ แบบจำลอง M5 คือวิธีการทำนายค่าทางตัวเลขวิธีหนึ่ง และโหนดของต้นไม้จะมีน้ำหนักมากกว่าตัวแปร นั่นจะทำให้ค่าความผิดพลาดลดลงเป็นอย่างมาก ถ้าเทียบกับฟังก์ชันมาตรฐาน แบบจำลอง M5P ถูกค้นพบโดย Quinlan ในปี 1992 และทฤษฎีของเขาถูกแก้ไขโดย Wang ในปี 1977 แบบจำลองประเภทต้นไม้มีประโยชน์หลากหลายใช้ได้ในหลายกรณี ซึ่งทำให้มันเหมาะสมที่จะใช้เพื่อเพิ่มประสิทธิภาพของการสังเคราะห์ข้อมูลเชิงถดถอย



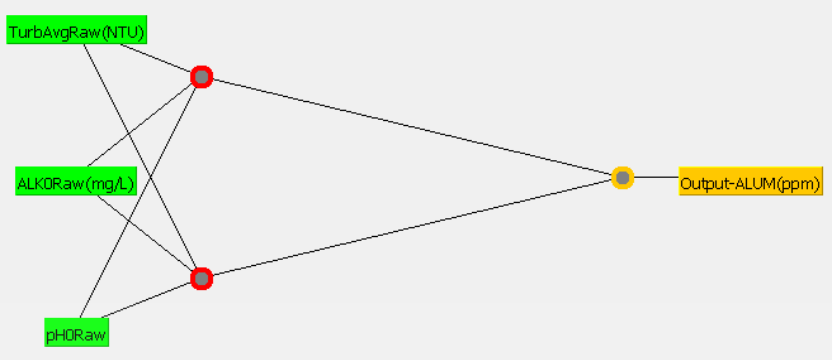
รูปที่ 3.3 แบบจำลอง M5P

3.3.3 M5Rules

วิธี M5rules สร้างประพจน์เชิงถดถอยในรูปแบบ IF-THEN rule โดยใช้งานในการสร้างรายการการตัดสินใจจากแบบจำลอง M5 เป็นหลัก วิธีนี้สามารถใช้งานได้กับ Continuous variables และ Nominal variables ได้ทั้งคู่ เหมาะแก่การใช้งานในรูปแบบการตัดสินใจที่ไม่เกี่ยวกับตัวเลข

3.3.4 Multilayer Perceptron

MLP เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายชั้น ใช้สำหรับงานที่มีความซับซ้อน โดยมีกระบวนการฝึกฝนเป็นแบบมีผู้สอน (Supervise) และใช้ขั้นตอนการส่งค่าย้อนกลับ (Backpropagation) สำหรับการฝึกฝนกระบวนการส่งค่าย้อนกลับ ประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า (Forward Pass) การส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อมูลเข้า และจะส่งผ่าน จากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (Error-Correction) คือผลต่างของผลตอบที่แท้จริง (Actual Response) กับผลตอบเป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาด (Error Signal) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ และค่าน้ำหนักของการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย



รูปที่ 3.4 แบบจำลอง Multilayer Perceptron

3.4 สังเคราะห์ข้อมูลจากทฤษฎีที่เลือกมาโดยใช้ Weka Software

Weka ย่อมาจาก Waikato Environment for Knowledge Analysis ซึ่งเป็นซอฟต์แวร์สำเร็จรูปที่สามารถดาวน์โหลดได้จากเว็บไซต์ซึ่งอยู่ภายใต้การควบคุมของ GPL License ซึ่งโปรแกรม Weka ได้ถูกพัฒนามาจากภาษาจาวาทั้งหมด ซึ่งเขียนมาโดยเน้นกับงานทางด้านการเรียนรู้ด้วยเครื่อง (Machine Learning) และการทำเหมืองข้อมูล (Data Mining) โปรแกรมจะประกอบไปด้วยโมดูลย่อย ๆ สำหรับใช้ในการจัดการข้อมูล และเป็นโปรแกรมที่สามารถที่ใช้ Graphic User Interface (GUI) และใช้คำสั่งในการให้ซอฟต์แวร์ประมวลผล

รูปที่ 3.5 Weka data mining Software

3.4.1 ข้อได้เปรียบของ Weka

1. สามารถดาวน์โหลดได้อย่างเสรีภายใต้สัญญาอนุญาติแบบสาธารณะทั่วไป
2. รวบรวมข้อมูลที่ครอบคลุม Preprocessing และเทคนิคการสร้างโมเดล

3.4.2 ส่วนประกอบหลักของการทำงาน

1. Weka มีสิ่งอำนวยความสะดวกสำหรับการนำเข้าข้อมูลด้วยวิธีการ Preprocessing ฐานข้อมูลเป็นไฟล์ ชนิด .CSV และข้อมูลนี้ใช้เพื่อเรียกการกรอง มีอัลกอริธึม Preprocessing เป็นฟิลเตอร์สามารถนำมาใช้เพื่อแก้ไขข้อมูลได้
2. มี Algorithms Classifiers หลากหลายชนิดที่จะช่วยให้ผู้ใช้เลือกใช้หมวดหมู่ Classify ข้อมูลโดยการสร้างโมเดลแบบแบ่งบานและแบบถดถอย เพื่อประเมิณผลข้อมูลสามารถสร้างโมเดลการพยากรณ์และแบบจำลอง
3. ช่วยให้สามารถเข้าถึงความสัมพันธ์ที่ถูกจำกัดไว้ สามารถระบุถึงความสำคัญระหว่างข้อมูลในลักษณะ Interrelationships

3.4.3 การจัดเตรียมไฟล์เพื่อนำเข้าโปรแกรม

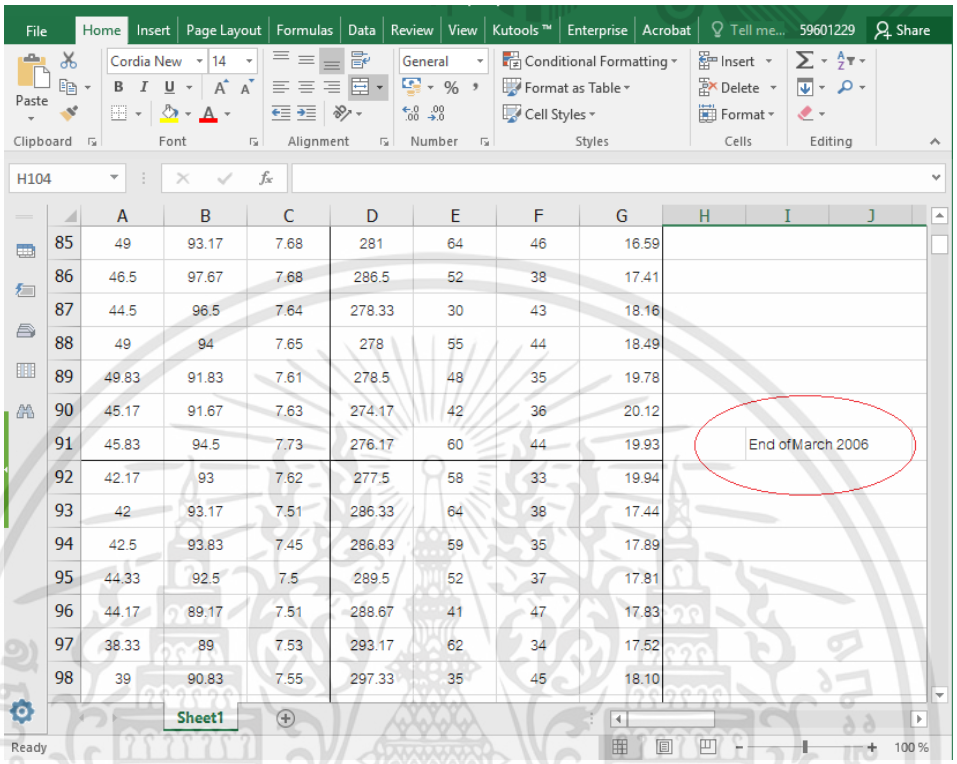
1. วิธีการวิเคราะห์และจัดการฐานข้อมูล

ขั้นตอนในการวิเคราะห์และจัดการฐานข้อมูลเป็นเทคนิคส่วนบุคคลขึ้นอยู่กับดุลพินิจของผู้วิเคราะห์ข้อมูลว่าจะปรับแต่งข้อมูลส่วนใด, ตัดข้อมูลส่วนใด หรือแทนค่าข้อมูลส่วนใด เนื่องจากการจัดการๆในขั้นตอนนี้ล้วนแล้วแต่ส่งผลต่อการสังเคราะห์แบบจำลองทั้งสิ้น ทางผู้วิจัยได้แบ่งการจัดการฐานข้อมูลสำหรับนำเข้า โปรแกรมเป็น 2 ประเภท

- การจัดกลุ่มแบบที่ 1: เป็นข้อมูลที่จัดการโดยการแทนค่าข้อมูลชุดที่มีค่าตัวแปรขาดหายไป ผิดเพี้ยน อันเนื่องจากการใช้สัญลักษณ์พิเศษ – , . = / [ ] & # $! _ @ ^ * หรือเว้นว่างใส่ลงไปในฐานข้อมูล เรียกว่ามี missing values ใน input ทั้ง 3 ตัว ตัวใดตัวหนึ่งหรือหลายตัวขาดหายไป โดยการนำค่าเฉลี่ยของพารามิเตอร์ชนิดนั้น จากข้อมูลในแต่ละเดือนมาแทนค่าส่วนที่หายไป
- การจัดกลุ่มแบบที่ 2: เป็นข้อมูลที่จัดการโดยการตัดข้อมูลชุดที่มีค่าตัวแปร Input ตัวใดตัวหนึ่งหรือหลายตัวผิดเพี้ยนหรือขาดหายไป รวมทั้งชุดข้อมูลที่ไม่มีค่าการเติมปริมาณสารส้ม ทิ้งทั้งบันทึก เพื่อลดค่า bias ที่จะเกิดขึ้นระหว่างการสร้างแบบจำลอง

2. วิธีการข้อมูลเพื่อใช้ในการจัดกลุ่ม เพื่อสังเคราะห์แบบจำลอง และเพื่อปรับแต่งแบบจำลอง

- เตรียม Database file: เมื่อไฟล์ผ่านการตรวจสอบความถูกต้อง จากนั้นลบข้อมูลลำดับหรือวันที่ ที่ไม่เกี่ยวข้องกับตัวแปร แล้วบันทึกไฟลล์เป็น .CSV

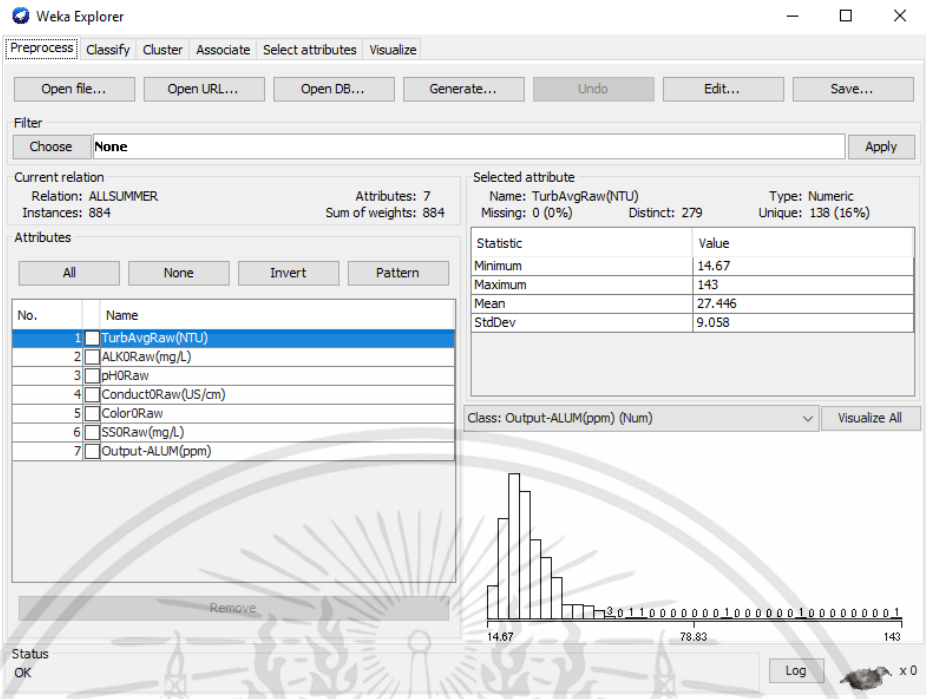


รูปที่ 3.6 ข้อมูลที่ไม่เกี่ยวข้องกับตัวแปร เช่นวันที่ในวงกลมสีแดง

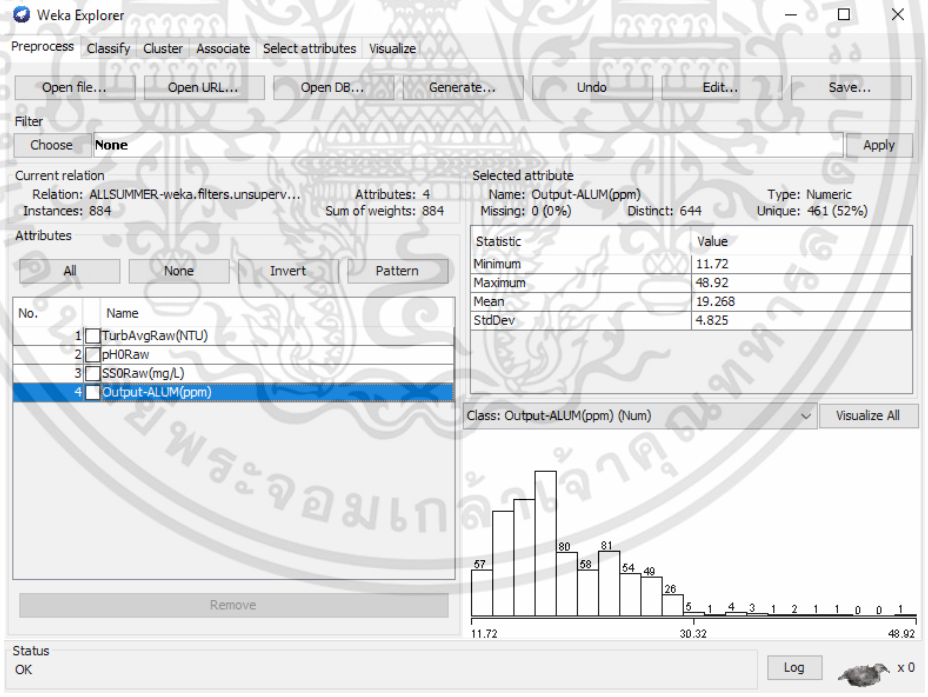
- การสังเคราะห์แบบจำลอง: เปิดโปรแกรม Weka เลือกไปที่แท็บ Preprocess เลือก Open file เปิดไฟล์ .CSV ที่บันทึกไว้ หากไฟล์ถูกจัดเตรียมมาถูกวิธีและไม่มีข้อมูลผิดเพี้ยน หรือมีสัญลักษณ์รวมอยู่ในตัวข้อมูล จะสามารถเปิดขึ้นตัวอย่างดังรูป 3.6 แต่ถ้าไม่สามารถเปิดได้ ให้กลับไปตรวจสอบฐานข้อมูลตามหัวข้อวิธีการวิเคราะห์และจัดการฐานข้อมูล ว่ามีสัญลักษณ์ – , . = / [ ] & # $! _ @ ^ * หรือช่องเว้นว่าง ในฐานข้อมูลหรือไม่ ถ้าตรวจพบให้ตรวจสอบและแก้ไขให้เรียบร้อย แล้วนำไฟล์ .CSV มาเปิดในโปรแกรม Weka อีกรอบ

3.4.4 การใช้โปรแกรมหลังจากที่เปิดไฟล์ข้อมูลสำเร็จ

เมื่อเปิดไฟล์แล้วให้เลือกตัดตัวแปรที่ไม่ต้องการออก โดยการเลือกเครื่องหมายถูกต้องที่ช่องสี่เหลี่ยมข้างหน้าตัวแปรที่ไม่ต้องการ แล้วกด Remove



รูปที่ 3.7 หน้าโปรแกรมที่สามารถเปิดไฟล์ฐานข้อมูลได้



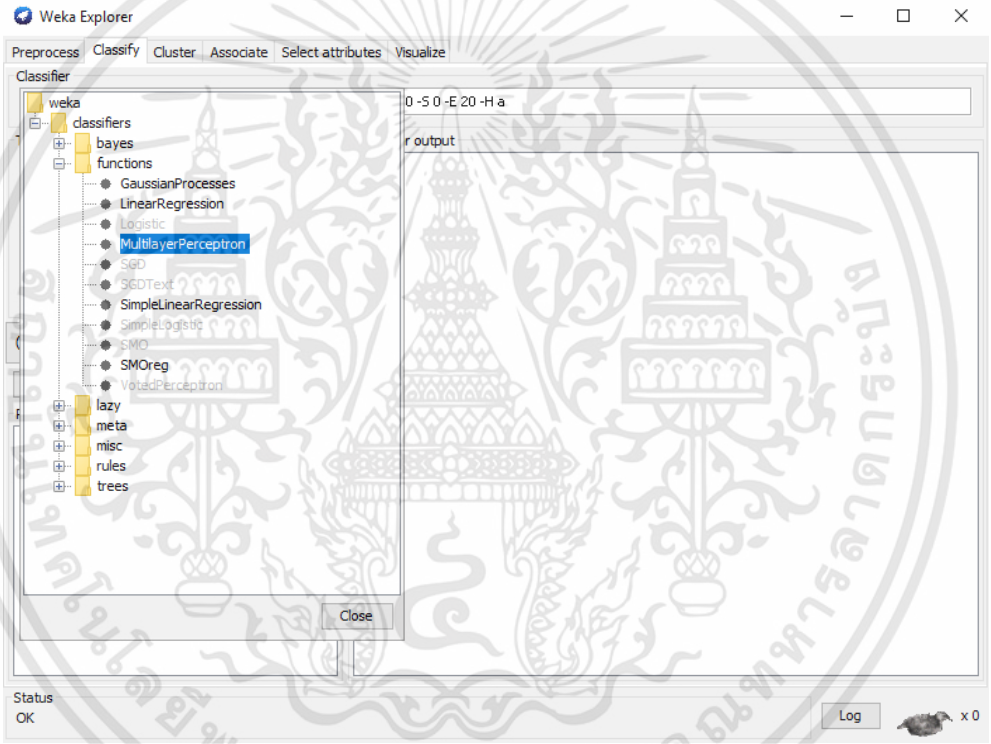
รูปที่ 3.8 หน้าโปรแกรมหลังจากตัวแปรที่ไม่ต้องการถูก Remove ออกไปแล้ว

เมื่อลบตัวแปรที่ไม่ต้องการออกไปแล้วให้เลือกไปที่แถว Classify เลือก Test option เป็น Percentage split ที่ 90%, 75% และ 50% ตามต้องการ ต่อไปเลือกที่ More option... แล้วเลือก

เครื่องหมายถูกที่หน้า Output model หลังจากนั้น ที่หัวข้อ Classifier ให้เปลี่ยนเป็นแบบจำลองที่เราต้องการที่จะเปลี่ยนทั้ง 4 ชนิด ได้แก่

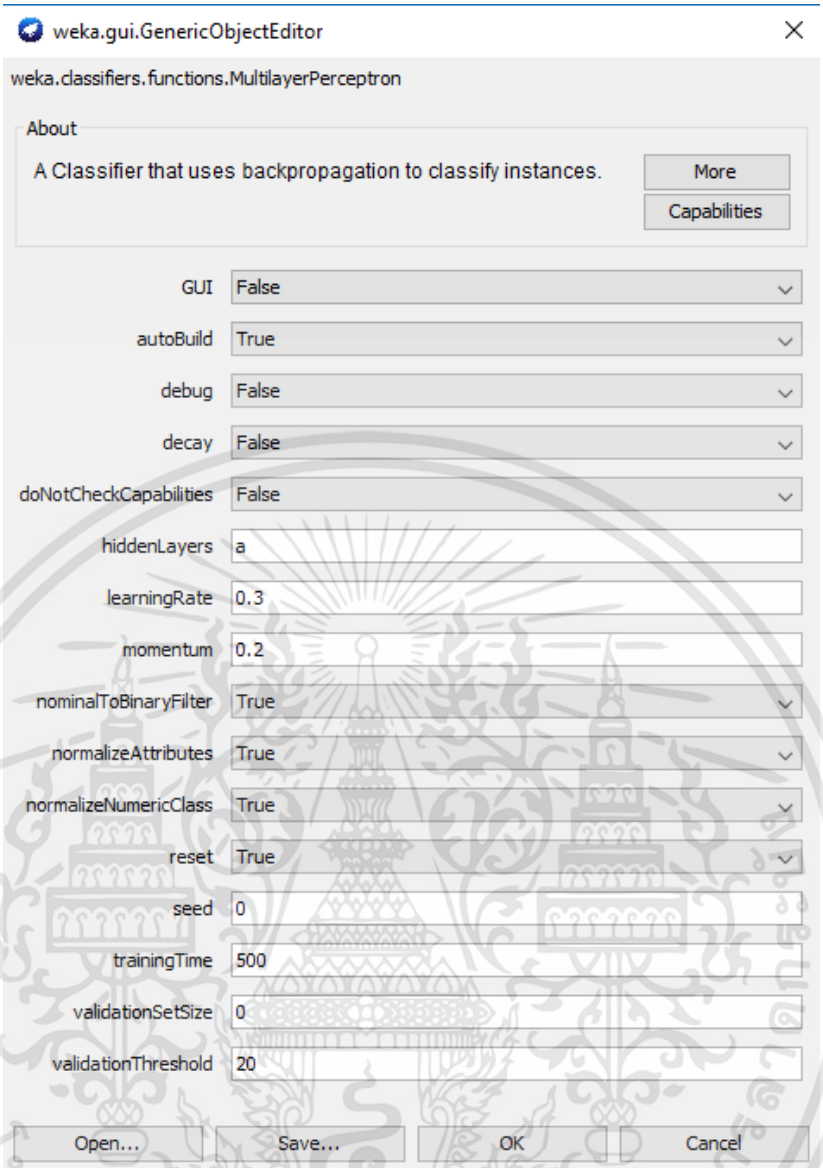
1. REPTree
2. M5P
3. M5Rules
4. Multilayer Perceptron

หลังจากเลือกแล้ว ให้กดที่ปุ่ม Start รอให้รูปนกด้านขวาล่างขยับไปเรื่อยๆจนหยุด แล้ว Status ด้านซ้ายล่าง ขึ้นคำว่า OK ก็ถือว่าการสังเคราะห์โมเดลเสร็จสิ้นเรียบร้อย



รูปที่ 3.9 การเลือก Classifier ที่ต้องการ

ในการปรับแต่งแบบจำลอง แต่ละ Algorithms ในแท็บ Classifier สามารถสังเคราะห์แบบจำลองด้วยวิธีการทางสถิติและการให้ความสัมพันธ์หลายรูปแบบ แต่ละรูปแบบจะผลิตแบบจำลองที่ให้ค่าทางคณิตศาสตร์ที่แตกต่างกัน โดยแต่ละค่าจะบอกความแม่นยำ และประสิทธิภาพของแบบจำลองว่ามีค่าสูงหรือต่ำ โดยทางผู้วิจัยนั้นสามารถปรับแต่งแบบจำลองได้หลายรูปแบบโดยการกดเลือกที่ชื่อ Algorithms ที่เลือกใช้ในหัวข้อ Classifier และทดลองปรับค่าตัวแปรในการสร้างให้ผลที่ออกมามีค่าความแม่นยำมากที่สุด และบันทึกค่าที่ปรับแต่งเก็บไว้เปรียบเทียบผล

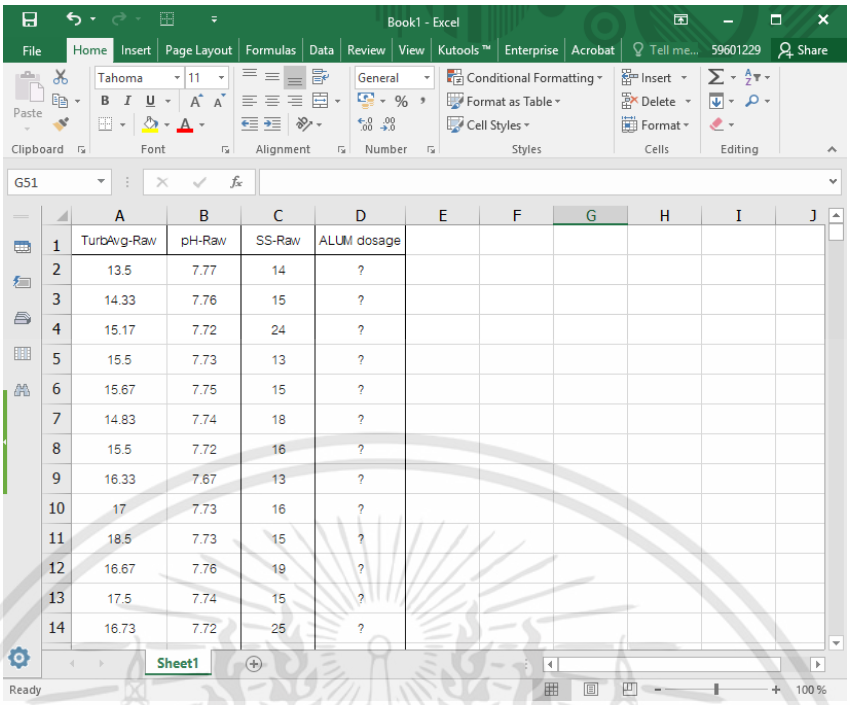


รูปที่ 3.10 การปรับแต่งแบบจำลองของวิธี Multilayer Perceptron

3.4.5 การใช้แบบจำลองทำนายข้อมูล

1. การเตรียมไฟล์ข้อมูลสำหรับการทำนาย

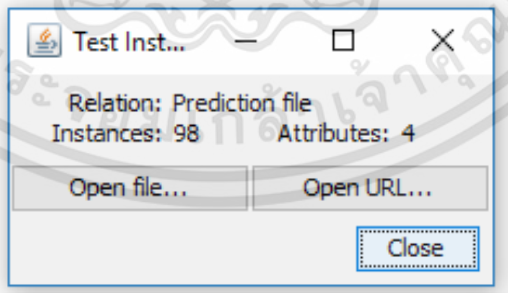
ไฟล์ข้อมูลที่ใช้ในการทำนายมีรูปแบบเดียวกับฐานข้อมูลที่นำมาสร้างแบบจำลอง โดยหัวตารางข้อมูลที่จะทำนาย ต้องเป็นชื่อตัวแปร และลำดับของตัวแปรชนิดเดียวกับไฟล์ฐานข้อมูล เช่น ตัวแปรลำดับที่ 2 ของไฟล์ฐานข้อมูลชื่อ pH-raw ย่อมากจาก ในไฟล์ที่ต้องการทำนายในตัวแปรลำดับที่ 2 ก็ต้องชื่อ pH-raw เช่นกัน และในช่อง Alum dosage ให้แทนที่ค่าตัวเลขด้วยเครื่องหมาย ? แทน แล้วบันทึก (รูปที่ 3.10)



รูปที่ 3.11 การปรับแต่งแบบจำลองของวิธี Multilayer Perceptron

2. การนำไฟล์ที่ต้องการทำนายเข้าสู่โปรแกรม

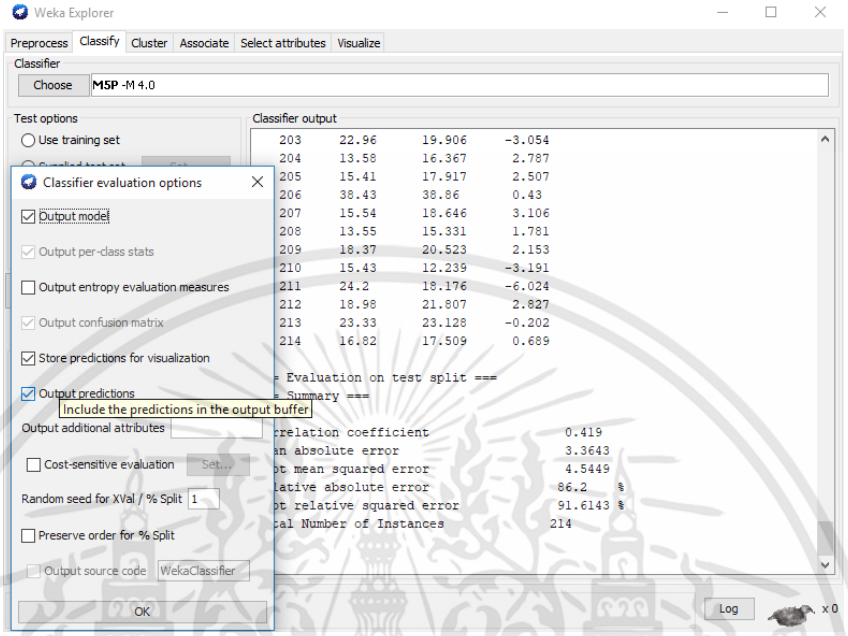
ในส่วนของ Test options ให้เลือกหัวข้อ Supplied test set แล้วกดที่ Set... เลือก Open file... เปิดไฟล์ที่เตรียมไว้ ถ้าไฟล์ที่เตรียมนั้นถูกต้อง โปรแกรมจะนับจำนวนชุดข้อมูล จำนวนตัวแปรทั้งหมด ทั้ง Input และ Output และแสดงตัวเลข (รูป 3.11)



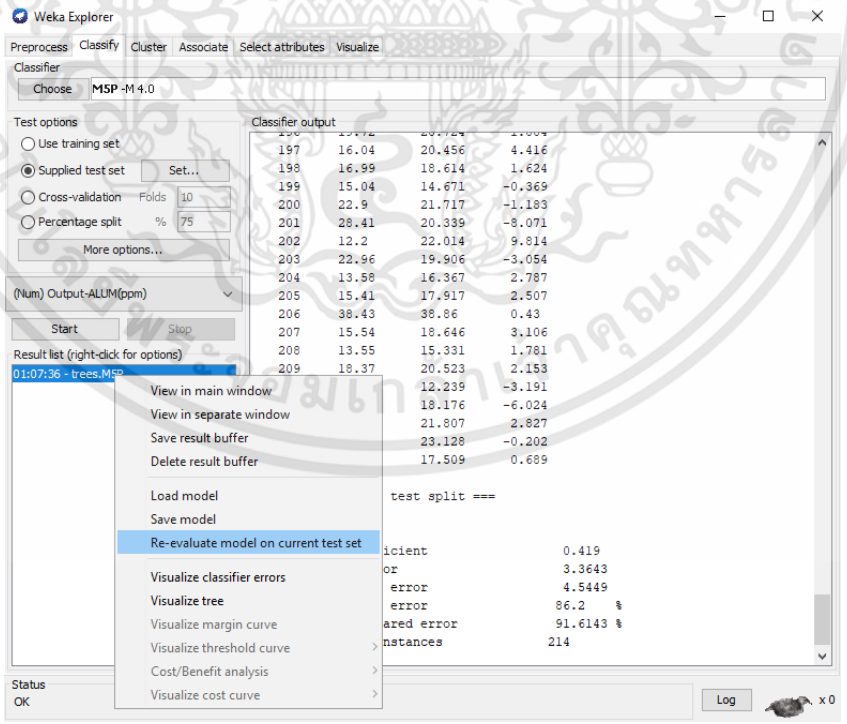
รูปที่ 3.12 การเปิดไฟล์ข้อมูลที่ใช้ทำนาย

เมื่อเปิดไฟลล์ข้อมูลที่จะทำนายเสร็จแล้ว ในหัวข้อ Test options ตั้งค่าการแสดงผลข้อมูล ให้กดไปที่ More options... เลือกเครื่องหมายถูกที่ช่อง Output Prediction

(รูป 3.12) หลังจากนั้นคลิกขวาที่โมเดลไปที่แบบจำลองที่สร้างไว้ เลือกไปที่ Re-evaluate model on current test set (รูป 3.13)

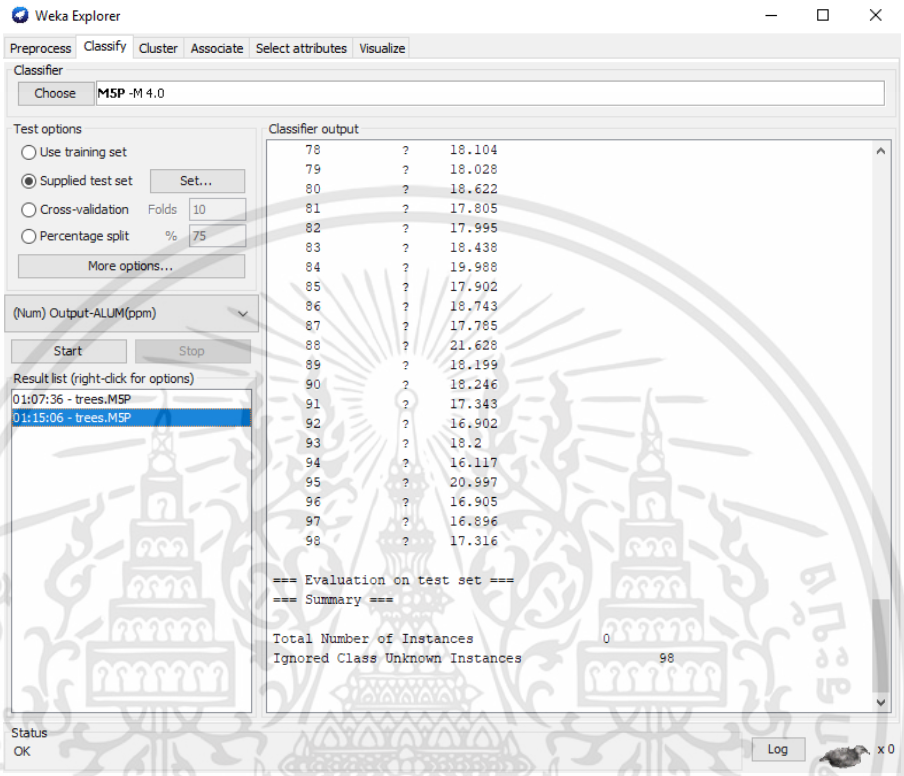


รูปที่ 3.13 เลือก Output prediction



รูปที่ 3.14 เลือก Re-evaluate model on current test set

เมื่อโปรแกรมทำนายข้อมูลโดยใช้แบบจำลองที่เลือกไว้สำเร็จจะแสดงผลการทำนายตั้งแต่ลำดับแรกจนถึงลำดับสุดท้าย จาก Classifier Output (รูป 3.13) แล้วนำข้อมูลที่ได้ไปเหรียบเทียบค่าจริง เพื่อหาค่าตัวชี้วัดทางคณิตศาสตร์ที่ต้องการ



รูปที่ 3.15 แสดงผลการทำนายค่า Alum Dosage

3.5 การทดสอบความแม่นยำของแบบจำลองที่ได้

3.5.1 Root Mean Square Error (RMSE)

การประเมินผลด้วย Root Mean Square Error (RMSE) คือการวัดค่าความแตกต่างระหว่างค่าจริงและค่าประมาณได้จากแบบจำลอง หากค่า RMSE มีค่าน้อยแสดงว่าแบบจำลองสามารถประมาณค่าได้ใกล้เคียงกับค่าจริง ดังนั้นหากค่านี้มีค่าเท่ากับศูนย์ นั่นหมายความว่าแบบจำลองนั้นมีความแม่นยำสูงสุด ค่า RMSE สามารถคำนวณได้จาก

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x)^2}{n}}$$

x_i = ปริมาณการเติมสารส้ม (mg/L) ที่ใช้จริง

x = ปริมาณการเติมสารส้ม (mg/L) จากการทำนาย

n = จำนวนชุดของข้อมูล

3.5.2 Correlation Coefficient (CC)

Correlation Coefficient คือ การวัดเชิงตัวเลขของความสัมพันธ์ทางสถิติระหว่างสองตัวแปร ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์นี้จะมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 ซึ่งหากมีค่าใกล้ -1.0 นั้นหมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันอย่างมากในเชิงตรงกันข้าม หากมีค่าใกล้ +1.0 นั้นหมายความว่า ตัวแปรทั้งสองมีความสัมพันธ์กันโดยตรงอย่างมาก และหากมีค่าเป็น 0 นั้นหมายความว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์ต่อกัน มีสมการดังต่อไปนี้

$$CC = \frac{\sum_{i=1}^{n}(x_i - \bar{x_i})\,(x - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x_i})^2 \sum_{i=1}^{n}(x - \bar{x})^2}}$$

x_i = ปริมาณการเติมสารส้ม (mg/L) ที่ใช้จริง

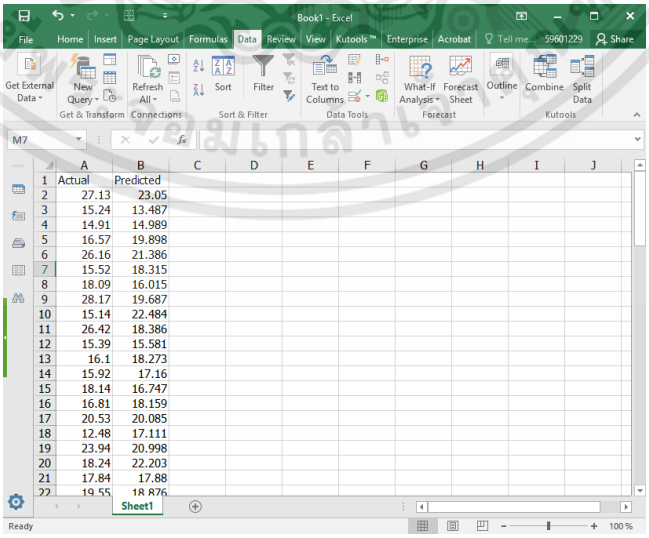
x = ปริมาณการเติมสารส้ม (mg/L) จากการทำนาย

n = จำนวนชุดของข้อมูล

3.5.3 R-Square

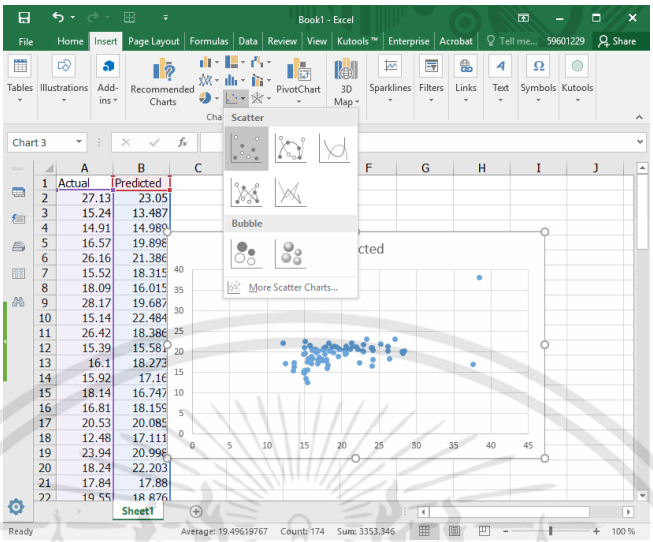
R-square คือสัดส่วนของความแปรปรวนในตัวแปรตามที่สามารถคาดการณ์ได้จากตัวแปรต้น และเป็นสถิติที่ใช้ในบริบทของแบบจำลองทางสถิติซึ่งมีวัตถุประสงค์หลักคือการคาดการณ์ผลลัพธ์ในอนาคตหรือการทดสอบสมมติฐาน บนพื้นฐานของข้อมูลที่เกี่ยวข้องกัน ที่ซึ่ง 1.00 คือค่าที่สูงที่สุดในการหาค่า R-Square ในการหาค่า R-Square โดยปกติจะหาจากสมการ ซึ่งข้อมูลในส่วนของปริมาณการเติมสารส้ม ข้อมูลนั้นมีค่าค่อนข้างกระจายจึงไม่สามารถคำนวณจากสมการได้อย่างแม่นยำ เพราะต้องลากเส้น Trendline ขึ้นมา ทางผู้วิจัยจึงได้หาค่า R-Square จากฟังก์ชันของ Excel มีวิธีดังต่อไปนี้

1. คัดลอกข้อมูลจริง และข้อมูลจากการทำนายมาไว้ในหน้าต่างเดียวกัน (รูป 3.16)



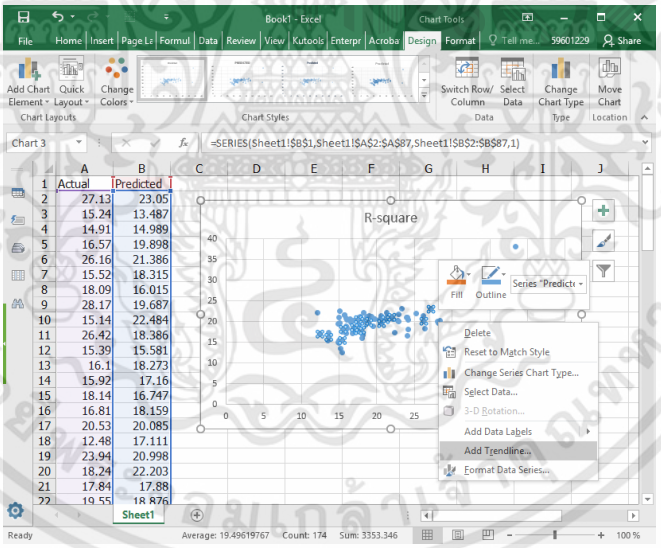
รูปที่ 3.16 การจัดข้อมูลเพื่อหา R-square

1. ครอบดำข้อมูลทั้ง 2 ชุด แล้วเข้าไปที่ Insert แล้วเลือกสร้างกราฟแบบ Scatter (รูป 3.17)



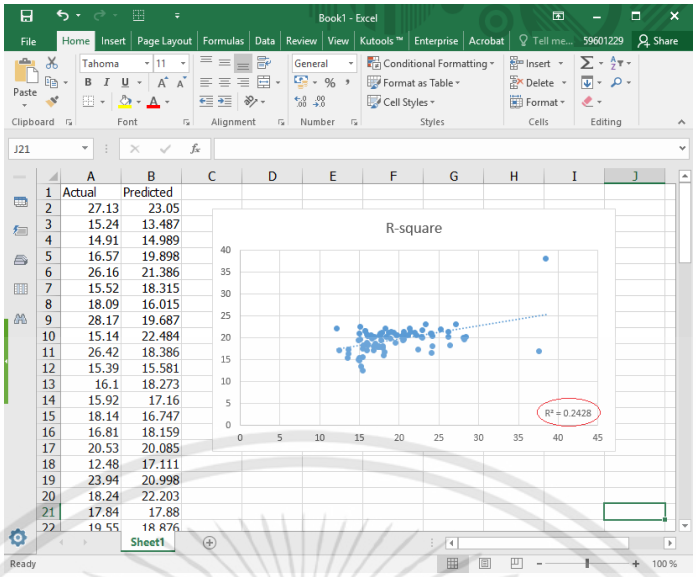
รูปที่ 3.17 การสร้างกราฟแบบ Scatter

2. เลือกคลิกขวาที่จุดในกราฟจุดใดจุดหนึ่งแล้วจากนั้น เลือกที่ Add Trendline...



รูปที่ 3.18 การสร้าง Trendline

4. พอมี Trendline ขึ้นมาจะมีแถบด้านขวาขึ้นตามมาด้วย ให้กดเลือกที่ช่อง Display R-square value on chart ก็จะได้ค่า R-square ออกมา (รูป 3.18) เพื่อนำไปเปรียบเทียบกับแบบจำลองตัวอื่น



	A	B
1	Actual	Predicted
2	27.13	23.05
3	15.24	13.487
4	14.91	14.989
5	16.57	19.898
6	26.16	21.386
7	15.52	18.315
8	18.09	16.015
9	28.17	19.687
10	15.14	22.484
11	26.42	18.386
12	15.39	15.581
13	16.1	18.273
14	15.92	17.16
15	18.14	16.747
16	16.81	18.159
17	20.53	20.085
18	12.48	17.111
19	23.94	20.998
20	18.24	22.203
21	17.84	17.88
22	19.55	18.876

รูปที่ 3.19 กราฟแสดงค่า R-square

บทที่ 4
ผลการทดลอง

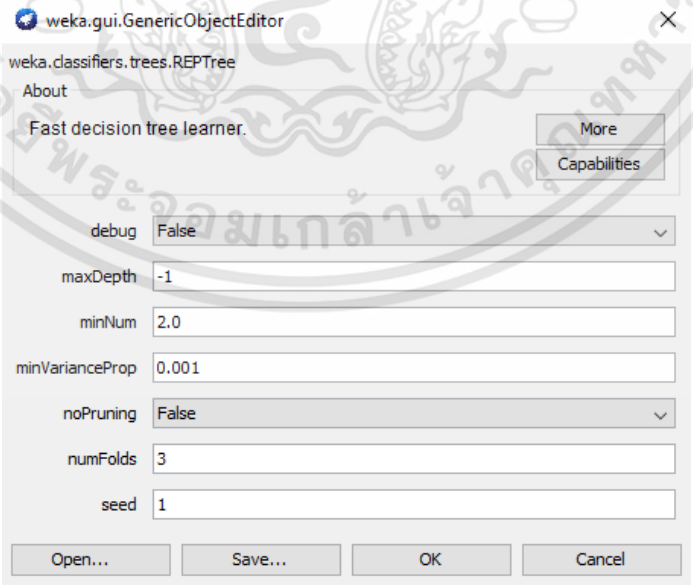
4.1 กล่าวนำ

งานวิจัยฉบับนี้เสนอถึงวิธีการทำนายปริมาณสารส้มที่ใช้ในกระบวนการผลิตน้ำประปา โดยใช้โปรแกรม Weka เพื่อสร้างแบบจำลองทางคณิตศาสตร์ใช้ทำนายค่าปริมาณการเติมสารส้ม ในงานวิจัยนี้เปรียบเทียบผลลัพธ์การทำนาย RMSE, CC และ R-square จากโมเดลที่สร้างด้วยวิธี REPTree, M5P, M5Rule, MLP เพื่อหาว่าแบบจำลองใดสามารถทำนายค่าได้แม่นยำที่สุด ตัวแปรต้นที่ใช้ในการศึกษามีทั้งหมด 3 ตัว ได้แก่ ความขุ่น (Turbidity), ค่าพีเอช (pH), ปริมาณของแข็งแขวนลอย (Suspended solids) ตัวแปรตามคือปริมาณการเติมสารส้ม (Alum dosage) ข้อมูลที่ใช้ในการสร้างแบบจำลองคือข้อมูลน้ำดิบขาเข้า รวบรวมจากโรงผลิตน้ำประปาบางเขน จำนวน 3,500 ชุด (คัดกรองแล้ว) เพื่อใช้ในการสร้างแบบจำลอง

4.2 การปรับแต่งตัวแปรของทฤษฎีที่ใช้สร้างแบบจำลอง

เพื่อการสร้างแบบจำลองของแต่ละทฤษฎีอย่างมีประสิทธิภาพ ทางผู้วิจัยได้ทำการปรับแต่งค่าคงที่ของแต่ละทฤษฎีโดยใช้ข้อมูลที่กรองแล้วทั้งหมดและสลับข้อมูลจำนวน 1,220 ชุด เพื่อให้เข้ากับชุดข้อมูลมากที่สุด

4.2.1 REPTree



รูปที่ 4.1 หน้าต่างแสดงตัวแปรของทฤษฎี REPTree

debug คือ ถ้าเลือกไปที่ True เพิ่มการควบคุมข้อมูลให้มากขึ้น เช่น การคัดกรองข้อมูล

maxDepth คือ ความลึกของต้นไม้ ถ้าใส่ -1 คือไม่จำกัดความลึก

minNum คือ ค่าที่น้อยที่สุดของผลรวมน้ำหนักในใบไม้ แต่ละใบ

minVarianceProp คือ สัดส่วนที่น้อยที่สุดของการเปลี่ยนแปลงในข้อมูลทั้งหมด

noPruning คือ ถ้าเลือกไปที่ True จะแสดงข้อมูลที่ถูกตัดแต่งข้อมูล

numFolds คือ การกำหนดจำนวนการแต่งข้อมูล

seed คือ จำนวนการสุ่มข้อมูล

ตารางที่ 4.1 แสดงผลการปรับแต่งตัวแปรเพื่อสร้างแบบจำลองของทฤษฎี REPTree

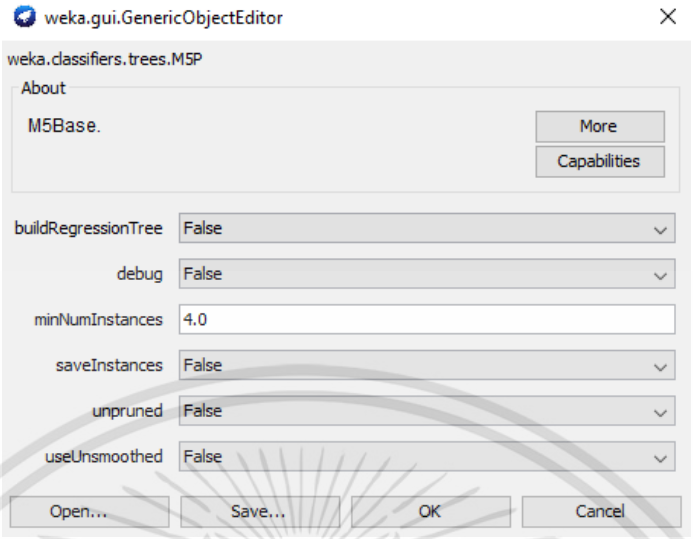
No.	max depth	min num	Variance prop	noPruning	Folds	seed	CC	RMSE
1	-1	1000	0.0010	FALSE	3	1	-0.0382	14.0919
2	-1	100	0.0010	TRUE	3	1	-0.0382	14.0919
3	-1	10	0.0001	TRUE	4	2	0.6512	10.7537
4	-1	2	0.1000	FALSE	4	3	0.5826	11.5681
5	-1	2	0.0010	FALSE	3	1	0.5780	11.6370
6	-1	3	1.0000	TRUE	5	4	0.3581	13.1556
7	-1	4	0.0010	FALSE	5	4	0.5055	12.2065
8	-1	5	0.0010	TRUE	3	3	0.6580	10.7126
9	10	2	0.0010	TRUE	4	2	0.6465	11.095
10	100	2	0.0010	FALSE	5	1	0.6018	11.3976
11	1000	2	0.0010	FALSE	3	1	0.5780	11.6370

ตามตารางที่ 4.1 สรุปได้ว่า การปรับค่าตัวแปรต่างๆใน REPTree มีผลทำให้ค่า CC และ RMSE เปลี่ยน จากข้อมูลในตาราง ครั้งที่มีค่า CC มากที่สุด และ RMSE ที่น้อยที่สุด คือ ครั้งที่ 8 และจะเห็นได้ว่ามีการสังเคราะห์แบบจำลอง 2 ครั้งที่ผล เท่ากัน โดย 2 ครั้งนั้นมีค่า minnum, Variance prop, noPruning, Folds และ seed ที่เหมือนกัน แต่ตัวแปรที่ต่างกันคือ max depth ซึ่งไม่มีผลทำให้ค่า CC และ RMSE เปลี่ยน จากการสร้างแบบจำลองด้วยข้อมูลค่าพารามิเตอร์จากน้ำดิบ

ตารางที่ 4.2 ผลจากการปรับแต่งตัวแปรจากทฤษฎี REPTree ที่ดีที่สุด

No.	max depth	min num	Variance prop	noPruning	Folds	seed	CC	RMSE
8	-1	5	0.0010	TRUE	3	3	0.6580	10.7126

4.2.2 M5P



รูปที่ 4.2 หน้าต่างแสดงตัวแปรของทฤษฎี M5P

buildRegressionTree คือ ถ้าเลือกไปที่ True/ False สร้างแบบจำลองการถดถอย แบบต้นไม้/กฎ
debug คือ ถ้าเลือกไปที่ True เพิ่มการควบคุมข้อมูลให้มากขึ้น เช่น การคัดกรองข้อมูล
minNumInstances คือ จำนวนตัวแปรที่อณุญาตให้ทำงานในโหนดของใบไม้
saveInstances คือ ถ้าเลือกไปที่ True จะทำการรักษาตัวแปรในแต่ละ Node เพื่อนำไปแสดงผล
unpruned คือ ถ้าเลือกไปที่ True ต้นไม้/กฎ จะไม่ถูกตัดแต่ง
unsmoothed คือ ถ้าเลือกไปที่ True จะได้ผลการทำนายที่ไม่ถูกตัดแต่งหรือ ทำให้เรียบ

ตารางที่ 4.3 แสดงผลการปรับแต่งตัวแปรเพื่อสร้างแบบจำลองของทฤษฎี M5P

No.	Build Regression Tree	Min num Instances	save Instances	unpruned	Use unsmoothed	CC	RMSE
1	FALSE	4	FALSE	FALSE	FALSE	0.2324	8.1787
2	FALSE	4	TRUE	FALSE	FALSE	0.2324	8.1787
3	FALSE	5	FALSE	TRUE	FALSE	0.2332	7.8431
4	FALSE	5	TRUE	TRUE	FALSE	0.2332	7.8431
5	FALSE	5	FALSE	FALSE	FALSE	0.2322	8.1811
6	TRUE	4	TRUE	FALSE	TRUE	0.7221	7.0580
7	TRUE	4	FALSE	TRUE	TRUE	0.7051	7.0714
8	TRUE	4	TRUE	TRUE	TRUE	0.7051	7.0714
9	TRUE	2	FALSE	FALSE	TRUE	0.7221	7.0580
10	TRUE	2	TRUE	FALSE	TRUE	0.7221	7.0580

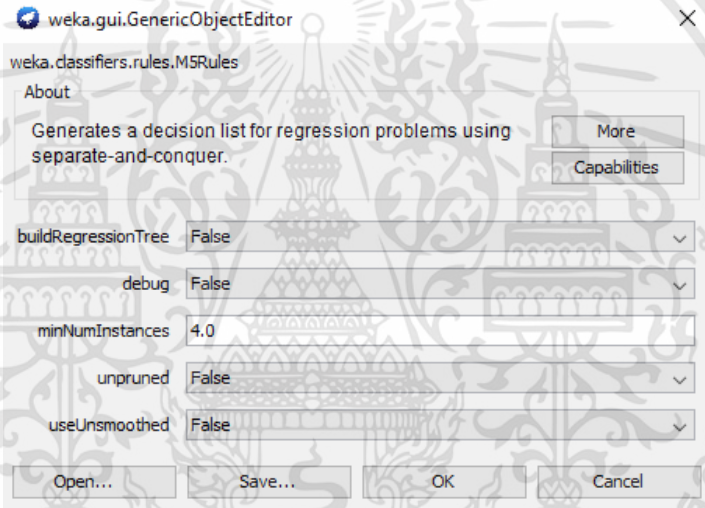
ตามตารางที่ 4.3 สรุปได้ว่า การปรับค่าตัวแปรต่างๆใน M5P มีผลทำให้ค่า CC และ RMSE เปลี่ยน จากข้อมูลในตาราง ครั้งที่มีค่า CC มากที่สุด และ RMSE ที่น้อยที่สุด คือ ครั้งที่ 6, ครั้งที่ 9 และ ครั้6ที่ 10 จะเห็นได้ว่าทั้ง 3 ครั้งนั้น มีค่า Build Regression Tree, unpruned, และ Use

unsmoothed ที่เหมือนกัน แต่ตัวแปรที่ต่างกันคือ Min num Instances และ save Instances ซึ่งไม่มีผลทำให้ค่า CC และ RMSE เปลี่ยนแปลง จากการสร้างแบบจำลองด้วยข้อมูลค่าพารามิเตอร์จากน้ำดิบ

ตารางที่ 4.4 ผลจากการปรับแต่งตัวแปรจากทฤษฎี M5P ที่ดีที่สุด

No.	Build Regression Tree	Min num Instances	save Instances	unpruned	Use unsmoothed	CC	RMSE
6	TRUE	4	TRUE	FALSE	TRUE	0.7221	7.0580
9	TRUE	2	FALSE	FALSE	TRUE	0.7221	7.0580
10	TRUE	2	TRUE	FALSE	TRUE	0.7221	7.0580

4.2.3 M5Rules



รูปที่ 4.3 หน้าต่างแสดงตัวแปรของทฤษฎี M5Rules

buildRegressionTree คือ ถ้าเลือกไปที่ True/ False สร้างแบบจำลองการถดถอย แบบต้นไม้/กฎ
debug คือ ถ้าเลือกไปที่ True เพิ่มการควบคุมข้อมูลให้มากขึ้น เช่น การคัดกรองข้อมูล
minNumInstances คือ จำนวนตัวแปรที่อณุญาตให้ทำงานในโหนดของใบไม้
unpruned คือ ถ้าเลือกไปที่ True ต้นไม้/กฎ จะไม่ถูกตัดแต่ง
unsmoothed คือ ถ้าเลือกไปที่ True จะได้ผลการทำนายที่ไม่ถูกตัดแต่งหรือ ทำให้เรียบ

ตารางที่ 4.5 แสดงผลการปรับแต่งตัวแปรเพื่อสร้างแบบจำลองของทฤษฎี M5P

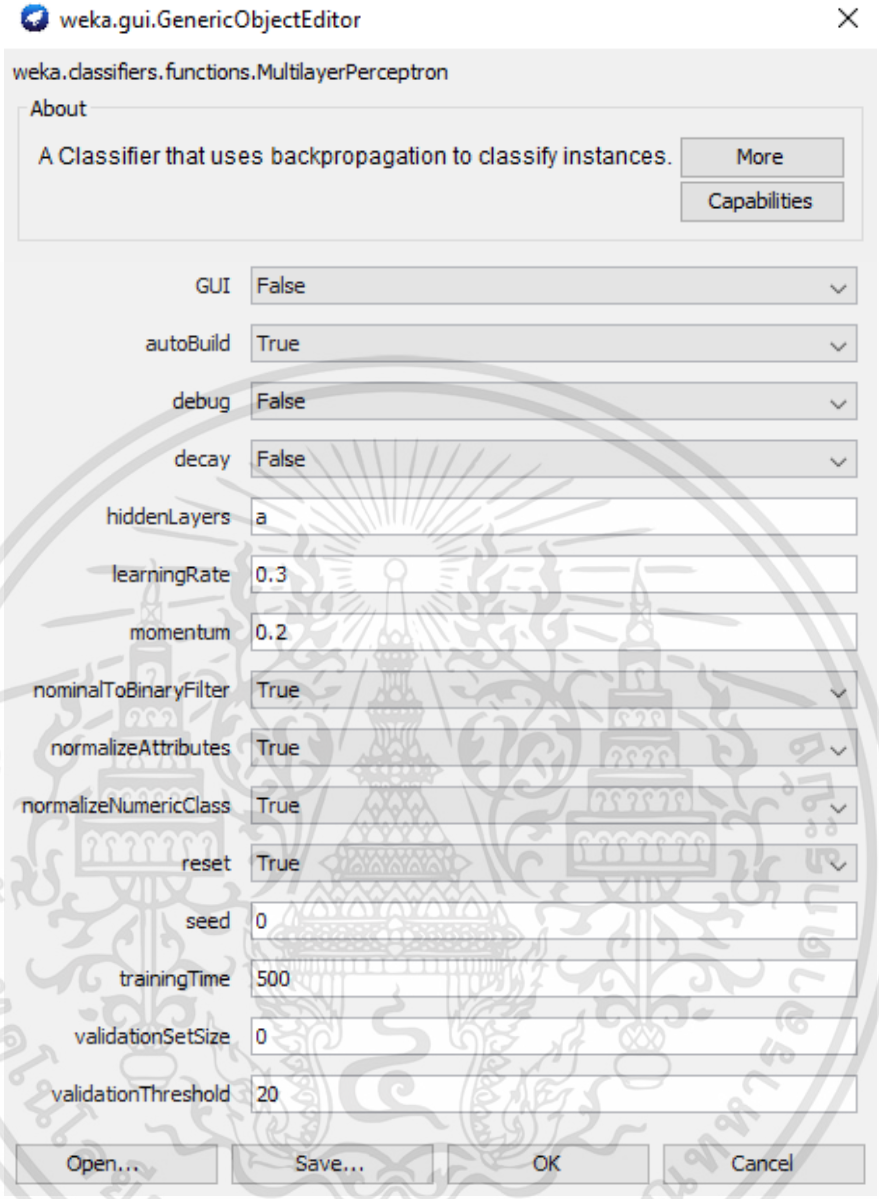
No.	Build Regression Tree	Min num Instances	unpruned	Use unsmoothed	CC	RMSE
1	FALSE	4	FALSE	FALSE	0.1813	53.3215
2	FALSE	5	FALSE	FALSE	0.4622	17.9604
3	FALSE	4	TRUE	FALSE	0.2437	31.7418
4	FALSE	5	TRUE	FALSE	0.2692	28.0142
5	FALSE	2	FALSE	FALSE	0.1813	52.3215
6	TRUE	3	FALSE	TRUE	0.7446	9.4140
7	TRUE	2	TRUE	TRUE	0.6739	10.9211
8	TRUE	3	TRUE	TRUE	0.6739	10.9211
9	TRUE	2	FALSE	TRUE	0.7446	9.4140
10	TRUE	3	FALSE	TRUE	0.7446	9.4140

ตามตารางที่ 4.5 สรุปได้ว่า การปรับค่าตัวแปรต่างๆใน M5Rules มีผลทำให้ค่า CC และ RMSE เปลี่ยน จากข้อมูลในตาราง ครั้งที่มีค่า CC มากที่สุด และ RMSE ที่น้อยที่สุด คือ ครั้งที่ 6, ครั้งที่ 9 และ ครั้งที่ 10 เช่นเดียวกับ M5P จะเห็นได้ว่าทั้ง 3 ครั้งนั้น มีค่า Build Regression Tree, unpruned, และ Use unsmoothed ที่เหมือนกัน แต่ตัวแปรที่ต่างกันคือ Min num Instances และ save Instances ซึ่งไม่มีผลทำให้ค่า CC และ RMSE เปลี่ยนแปลง จากการสร้างแบบจำลองด้วยข้อมูลค่าพารามิเตอร์จากน้ำดิบ

ตารางที่ 4.6 ผลจากการปรับแต่งตัวแปรจากทฤษฎี M5Rules ที่ดีที่สุด

No.	Build Regression Tree	Min num Instances	unpruned	Use unsmoothed	CC	RMSE
6	TRUE	3	FALSE	TRUE	0.7446	9.4140
9	TRUE	2	FALSE	TRUE	0.7446	9.4140
10	TRUE	3	FALSE	TRUE	0.7446	9.4140

4.2.4 Multilayer Perceptron(MLP)



รูปที่ 4.4 หน้าต่างแสดงตัวแปรของทฤษฎี MLP

GUI คือ ถ้าเลือกไปที่ True สามารถทำการหยุดได้ระหว่าง Training

autoBuild คือ ถ้าเลือกไปที่ True จะทำการเพิ่มและเชื่อมต่อ Hidden Layer ด้วยตัวโปรแกรมเอง

debug คือ ถ้าเลือกไปที่ True เพิ่มการควบคุมข้อมูลให้มากขึ้น เช่น การคัดกรองข้อมูล

decay คือ ถ้าเลือกไปที่ True จะทำกำหนด LearningRate ด้วยตัวโปรแกรมเอง

hiddenLayers คือ การกำหนดจำนวน hiddenLayers

learningRate คือ การกำหนดน้ำหนัก หรือ อัตราในการเรียนรู้

momentum คือ การให้น้ำหนักระหว่างการเรียนรู้

nominalToBinaryFilter คือ ถ้าเลือกไปที่ True จะทำการกรองข้อมูลตัวแปร ก่อนที่จะเริ่มเรียนรู้

normalizeAttributes คือ ถ้าเลือกไปที่ True จะทำให้ตัวแปรไม่แกว่ง

normalizeNumericClass คือ ถ้าเลือกไปที่ True จะทำให้ตัวแปรไม่แกว่ง ในกรณีตัวแปรเป็นตัวเลข

reset คือ ถ้าเลือกไปที่ True จะมีการเริ่มต้นใหม่หากผลที่ออกมามีค่าต่ำ

seed คือ จำนวนการสุ่มของเริ่มต้นในการสร้าง

trainingTime คือ จำนวนรอบของการ เรียนรู้

validationSetSize คือ ขนาดของการสุ่มข้อมูลโดยการ validation

validationThreshold คือ จำนวนครั้งที่ต้องการให้สุ่มข้อมูลแบบ validation

ตารางที่ 4.7 แสดงผลการปรับแต่งตัวแปรเพื่อสร้างแบบจำลองของทฤษฎี MLP

No.	Hidden layers	Learning rate	Momentum	Seed	Training time	CC	RMSE
1	1	0.3	0.2	1	500	0.2873	15.6411
2	2	0.5	0.2	1	1,000	0.2880	16.9851
3	3	0.3	0.2	1	2,000	0.4318	13.9670
4	4	0.5	0.2	2	500	0.2455	16.8189
5	5	0.4	0.2	2	500	0.3172	16.1401
6	6	0.6	0.2	2	1,000	0.1410	20.6592
7	7	0.4	0.2	3	2,000	0.3810	15.8261
8	8	0.6	0.2	3	500	0.1692	16.8564
9	9	0.7	0.2	3	1,000	0.1671	15.3527
10	10	0.9	0.2	4	2,000	0.0295	16.4612
11	1	0.7	0.3	4	3,000	0.0191	15.3413
12	2	0.9	0.3	4	4,000	0.0303	16.8655
13	3	0.7	0.3	5	500	0.0905	22.2357
14	4	0.9	0.3	5	500	0.0005	21.575
15	5	0.7	0.3	5	1,000	0.0307	19.9869
16	6	0.9	0.3	1	1,000	-0.0363	17.7619
17	7	0.7	0.3	2	2,000	0.0490	22.4253
18	8	0.9	0.3	3	2,000	-0.0013	16.1013
19	9	0.7	0.3	4	3,000	0.1564	14.889
20	10	0.9	0.3	5	3,000	0.0003	21.5731

ตามตารางที่ 4.7 สรุปได้ว่า การปรับค่าตัวแปรต่างๆใน MLP มีผลทำให้ค่า CC และ RMSE เปลี่ยน จากข้อมูลในตาราง ครั้งที่มีค่า CC มากที่สุด และ RMSE ที่น้อยที่สุด คือ ครั้งที่ ครั้งที่ 9 และครั้งที่ 10 มีค่า และครั้งที่ 7 ตามลำดับ จะเห็นได้ว่าค่า CC กับ RMSE ของการสังเคราะห์แบบจำลองแต่ละครั้งมีค่อค่อนข้างแกว่ง เนื่องจากตัวแปรแต่ละตัวมีผลค่อนข้างมากต่อแบบจำลอง และในการสังเคราะห์แบบจำลองโดยทฤษฎี MLP นั้นมีความแต่งต่างกับอีก 3 ทฤษฎีที่ผ่านมา คือ ค่า CC จะไม่แปรผันตามค่า RMSE

ตารางที่ 4.8 ผลจากการปรับแต่งตัวแปรจากทฤษฎี MLP ที่ดีที่สุด

No.	Hidden layers	Learning rate	Momentum	Seed	Training time	CC	RMSE
3	3	0.3	0.2	1	2,000	0.4318	13.9670

ตารางที่ 4.9 ผลการสังเคราะห์ที่ดีที่สุดของแต่ละทฤษฎี

Algorithms	CC	RMSE
REPTree	0.6580	10.7126
M5P	0.7221	7.0580
M5Rules	0.7446	9.4140
MLP	0.4318	13.9670

จากตารางที่ 4.9 จะเห็นได้ว่า ทฤษฎี M5P จะมีค่า CC ที่มากที่สุด และค่า RMSE ที่ต่ำที่สุด โดยมีค่า CC อยู่ที่ 0.7221 และค่า RMSE อยู่ที่ 9.4140 ซึ่งเป็นค่าที่ทำให้แบบจำลองถือว่ามีความน่าเชื่อถือมากเลยทีเดียว ส่วนทฤษฎีที่มีค่า CC น้อยที่สุด และมีค่า RMSE สูงที่สุดคือ MLP มีค่า CC อยู่ที่ 0.4318 และค่า RMSE อยู่ที่ 13.9670 ซึ่งยังถือว่าอยู่ในเกณฑ์ที่ใช้งานได้ ในส่วนต่อไปนั้น ก็จะนำแบบจำลองของแต่ละทฤษฎีที่ถูกดัดแปลงปรับแต่งค่าตัวแปร ไปทดสอบและทำนายผล ของข้อมูลจริง

4.3 การทดสอบแบบจำลองกับข้อมูลจริง

ในการสังเคราะห์ข้อมูล ข้อมูลจะถูกแบ่งออกตามฤดูกาลเพื่อเพิ่มประสิทธิภาพในการสังเคราะห์ เนื่องจากในแต่ละฤดูมีค่าตัวแปรของน้ำดิบที่ต่างกันค่อนข้างมาก แบ่งออกเป็น 3 ชุดด้วยกัน

1.ชุดฤดูฝน (กลางเดือนพฤษภาคม – กลางเดือนตุลาคม) จำนวน 1,433 ชุด

2.ชุดฤดูหนาว (กลางเดือนตุลาคม – กลางเดือนกุมภาพันธ์) จำนวน 852 ชุด

3.ชุดฤดูร้อน (กลางเดือนกุมภาพันธ์ – กลางเดือนพฤษภาคม) จำนวน 1,215 ชุด

และในข้อมูลแต่ละชุดได้มีการจัดข้อมูลเป็น 3 รูปแบบ เพื่อหาแบบจำลองที่มีประสิทธิภาพเหมาะสมกับการนำไปใช้จริง

1.Percentage Split 90%

2.Percentage Split 75%

3.Percentage Split 50%

โดยที่ Split 90% หมายถึง ถ้ามีการแบ่งข้อมูลออกเป็น 100 ส่วน จะใช้ข้อมูลที่รับเข้ามา เป็น Training Data 90 ส่วน และจะเป็น Testing Data 10 ส่วน เกณฑ์ในการคัดเลือก Percent Split ที่นำมาใช้นั้น การเลือกนั้นควรมากกว่า 50% ขึ้นไป เพื่อให้ข้อมูลในการเรียนรู้นั้นมากกว่าข้อมูลที่ใช้

ทดสอบ และ ช่วงเปอร์เซ็นต์ที่สมบูรณ์ที่สุดอยู่ที่ 75% - 80% (Bringmann และคณะ, 2010) ส่วน 90% นั้นที่เลือกมาเพราะ จะนำไปใช้เพื่อทดสอบประสิทธิภาพของแบบจำลองที่ 75%

4.3.1 REPTree

4.3.1.1 ชุดฤดูฝน

ตารางที่ 4.10 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูฝนจากแบบวิธี REPTree

	CC	R-Square	RMSE
Percentage Split 90%	0.7651	0.5855	9.3566
Percentage Split 75%	0.8036	0.6457	8.1006
Percentage Split 50%	0.7937	0.6300	8.1657

ตามตารางที่ 4.10 ในข้อมูลชุดฤดูฝนจากวิธี REPTree การจัดข้อมูลที่ให้ผลดีที่สุด คือ Percentage Split 75% ให้ค่า CC = 0.8036, R-square = 0.6457, RMSE = 8.1006 ซึ่งมีค่าที่ถือว่าอยู่ในเกณฑ์ที่ดีมาก ส่วนค่าที่เปอร์เซ็นต์อื่นๆ ผลที่ออกมาก็ถือว่าอยู่ในเกณฑ์ที่สูงมากเช่นกัน

4.3.1.2 ชุดฤดูหนาว

ตารางที่ 4.11 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูหนาวจากวิธี REPTree

	CC	R-Square	RMSE
Percentage Split 90%	0.6826	0.4660	9.4669
Percentage Split 75%	0.7051	0.4972	10.2128
Percentage Split 50%	0.5894	0.3474	11.4094

ตามตารางที่ 4.11 ในข้อมูลชุดฤดูหนาวจากวิธี REPTree การจัดข้อมูลที่ให้ผลดีที่สุดคือ Percentage Split 75% ให้ค่า CC = 0.7051, R-square = 0.4972, RMSE = 10.2128 ซึ่งมีค่าที่ถือว่าอยู่ในเกณฑ์ที่ดีสามารถนำไปใช้ได้ ส่วนค่าที่เปอร์เซ็นต์อื่นๆ ผลที่ออกมาถือว่าใช้ได้

4.3.1.3 ชุดฤดูร้อน

ตารางที่ 4.12 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูร้อนจากวิธี REPTree

	CC	R-Square	RMSE
Percentage Split 90%	0.5458	0.2979	4.0290
Percentage Split 75%	0.1961	0.0385	4.9566
Percentage Split 50%	0.3771	0.1422	4.9553

ตามตารางที่ 4.12 ในข้อมูลชุดฤดูร้อนจากวิธี REPTree การจัดข้อมูลที่ให้ผลดีที่สุด คือ Percentage Split 90% ให้ค่า CC = 0.5458, R-square = 0.2979, RMSE = 4.0290 ถ้าเจาะจงดูไปที่ค่า RMSE เพียงอย่างเดียว ค่าที่ออกมานั้นกลับมีค่าน้อยมาก ซึ่งเป็นผลดี

ตารางที่ 4.13 ตารางเปรียบเทียบระหว่างแต่ละฤดูกาลของทฤษฎี REPTree

	CC	R-Square	RMSE
ฤดูฝน, Percentage Split 75%	0.8036	0.6457	8.1006
ฤดูหนาว, Percentage Split 75%	0.7051	0.4972	10.2128
ฤดูร้อน, Percentage Split 90%	0.5458	0.2979	4.0290

ตามตารางที่ 4.13 ในทฤษฎี REPTree ชุดข้อมูล และการจัดข้อมูลที่ให้ผลออกมาดีที่สุดคือ ฤดูฝน ที่ Percentage Split 75% ให้ค่า CC = 0.8036, R-square = 0.6457, RMSE = 8.1006 แต่ถ้ามองเฉพาะค่า RMSE จะเป็นทางด้านของ ชุดข้อมูล ฤดูร้อนที่ Percentage Split 90% RMSE = 4.0290

4.3.2 M5P

4.3.2.1 ชุดฤดูฝน

ตารางที่ 4.14 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูฝนจากแบบวิธี M5P

	CC	R-Square	RMSE
Percentage Split 90%	0.7929	0.6287	8.7450
Percentage Split 75%	0.7823	0.6120	8.4748
Percentage Split 50%	0.7686	0.5908	8.5953

ตามตารางที่ 4.14 ในข้อมูลชุดฤดูฝนจากวิธี M5P การจัดข้อมูลที่ให้ผลดีที่สุดคือ Percentage Split 90% ให้ค่า CC = 0.7929, R-square = 0.6287, RMSE = 8.7450 ซึ่งมีค่าที่ถือว่าอยู่ในเกณฑ์ที่ดีมาก ส่วนค่าที่เปอร์เซ็นต์อื่นๆ ผลที่ออกมาก็ถือว่าอยู่ในเกณฑ์ที่สูงมากเช่นกัน

4.3.2.2 ชุดฤดูหนาว

ตารางที่ 4.15 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูหนาวจากวิธี M5P

	CC	R-Square	RMSE
Percentage Split 90%	0.7033	0.4946	9.4015
Percentage Split 75%	0.8028	0.6446	8.7679
Percentage Split 50%	0.6544	0.4283	10.5477

ตามตารางที่ 4.15 ในข้อมูลชุดฤดูหนาวจากวิธี M5P การจัดข้อมูลที่ให้ผลดีที่สุดคือ Percentage Split 75% ให้ค่า CC = 0.8028, R-square = 0.6446, RMSE = 8.7679 ซึ่งมีค่าที่ถือว่าอยู่ในเกณฑ์ที่ดีมาก ส่วนค่าที่เปอร์เซ็นต์อื่นๆ ผลที่ออกมาก็ถือว่าอยู่ในเกณฑ์ที่ดี

4.3.2.3 ชุดฤดูร้อน

ตารางที่ 4.16 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูร้อนจากวิธี M5P

	CC	R-Square	RMSE
Percentage Split 90%	0.4928	0.2428	4.2163
Percentage Split 75%	0.4190	0.1756	4.5449
Percentage Split 50%	0.4038	0.1631	4.4766

ตามตารางที่ 4.16 ในข้อมูลชุดฤดูร้อนจากวิธี M5P การจัดข้อมูลที่ให้ผลดีที่สุดคือ Percentage Split 90% ให้ค่า CC = 0.4928, R-square = 0.2428, RMSE = 4.2163 ถ้าเจาะจงดูไปที่ค่า RMSE เพียงอย่างเดียว ค่าที่ออกมานั้นกลับมีค่าน้อยมาก ซึ่งเป็นผลดี

ตารางที่ 4.17 ตารางเปรียบเทียบระหว่างแต่ละฤดูกาลของทฤษฎี M5P

	CC	R-Square	RMSE
ฤดูฝน, Percentage Split 90%	0.7929	0.6287	8.7450
ฤดูหนาว, Percentage Split 75%	0.8028	0.6446	8.7679
ฤดูร้อน, Percentage Split 90%	0.4928	0.2428	4.2163

ตามตารางที่ 4.17 ในทฤษฎี M5P ชุดข้อมูล และการจัดข้อมูลที่ให้ผลออกมาดีที่สุด คือ ฤดูหนาว ที่ Percentage Split 75% ให้ค่า CC = 0.8028, R-square = 0.6446, RMSE = 8.7679 แต่ถ้ามองเฉพาะค่า RMSE จะเป็นทางด้านของ ชุดข้อมูล ฤดูร้อนที่ Percentage Split 90% RMSE = 4.2163

4.3.3 M5Rules

4.3.3.1 ชุดฤดูฝน

ตารางที่ 4.18 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูฝนจากแบบวิธี M5Rules

	CC	R-Square	RMSE
Percentage Split 90%	0.7448	0.5547	9.5745
Percentage Split 75%	0.7851	0.6164	8.4262
Percentage Split 50%	0.7832	0.6134	8.3423

ตามตารางที่ 4.18 ในข้อมูลชุดฤดูฝนจากวิธี M5Rules การจัดข้อมูลที่ให้ผลดีที่สุด คือ Percentage Split 75% ให้ค่า CC = 0.7851, R-square = 0.6164, RMSE = 8.4263 และมีค่าใกล้เคียงกับการจัดข้อมูลแบบ Percentage Split 50% เป็นอย่างมาก

4.3.3.2 ชุดฤดูหนาว

ตารางที่ 4.19 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูหนาวจากวิธี M5Rules

	CC	R-Square	RMSE
Percentage Split 90%	0.7574	0.5736	8.5215
Percentage Split 75%	0.7994	0.6390	8.6624
Percentage Split 50%	0.7436	0.5529	9.3346

ตามตารางที่ 4.19 ในข้อมูลชุดฤดูหนาวจากวิธี M5Rules การจัดข้อมูลที่ให้ผลดีที่สุดคือ Percentage Split 75% ให้ค่า CC = 0.7994, R-square = 0.6390, RMSE = 8.6624 ซึ่งมีค่าที่ถือว่าอยู่ในเกณฑ์ที่ดีมาก ส่วนค่าที่เปอร์เซ็นต์อื่นๆ ผลที่ออกมาก็ถือว่าอยู่ในเกณฑ์ที่ดีเช่นกัน

4.3.3.3 ชุดฤดูร้อน

ตารางที่ 4.20 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูร้อนจากวิธี M5Rules

	CC	R-Square	RMSE
Percentage Split 90%	0.5029	0.2529	4.1965
Percentage Split 75%	0.3729	0.1391	4.7709
Percentage Split 50%	0.3907	0.1526	4.5727

ตามตารางที่ 4.20 ในข้อมูลชุดฤดูร้อนจากวิธี M5Rules การจัดข้อมูลที่ให้ผลดีที่สุด คือ Percentage Split 90% ให้ค่า CC = 0.5029, R-square = 0.2529, RMSE = 4.1965 ซึ่งมีค่าที่ถือว่าอยู่ในเกณฑ์ที่ดีมาก ส่วนค่าที่เปอร์เซ็นต์อื่นๆ ผลที่ออกมาก็ถือว่าอยู่ในเกณฑ์ที่ดีเช่นกัน ถ้าเจาะจงดูไปที่ค่า RMSE เพียงอย่างเดียว ค่าที่ออกมานั้นกลับมีค่าน้อยมาก ซึ่งเป็นผลดี

ตารางที่ 4.21 ตารางเปรียบเทียบระหว่างแต่ละฤดูกาลของทฤษฎี M5Rules

	CC	R-Square	RMSE
ฤดูฝน, Percentage Split 75%	0.7851	0.6164	8.4262
ฤดูหนาว, Percentage Split 75%	0.7994	0.6390	8.6624
ฤดูร้อน, Percentage Split 90%	0.5029	0.2529	4.1965

ตามตารางที่ 4.21 ในทฤษฎี M5Rules ชุดข้อมูล และการจัดข้อมูลที่ให้ผลออกมาดีที่สุดคือ ฤดูหนาว ที่ Percentage Split 75% ให้ค่า CC = 0.7994, R-square = 0.6390,

RMSE = 8.6624 ซึ่งชุดข้อมูล ฤดูฝน ที่ Percentage Split 75% ก็อยู่ในเกณฑ์ที่ดีมาก เช่นกันแตกต่างกันเพียงเล็กน้อย แต่ถ้ามองเฉพาะค่า RMSE จะเป็นทางด้านของ ชุดข้อมูล ฤดูร้อนที่ Percentage Split 90% RMSE = 4.2163

4.3.4 Multilayer Perceptron

4.3.4.1 ชุดฤดูฝน

ตารางที่ 4.22 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูฝนจากแบบวิธี MLP

	CC	R-Square	RMSE
Percentage Split 90%	0.7456	0.5559	10.0718
Percentage Split 75%	0.7758	0.6018	8.6990
Percentage Split 50%	0.7703	0.5935	8.5508

ตามตารางที่ 4.21 ในข้อมูลชุดฤดูฝนจากวิธี MLP การจัดข้อมูลที่ให้ผลดีที่สุดคือ Percentage Split 75% ให้ค่า CC = 0.7758, R-square = 0.6018, RMSE = 8.6690 และมีค่าใกล้เคียงกับการจัดข้อมูลแบบ Percentage Split 50% เป็นอย่างมาก

4.3.4.2 ชุดฤดูหนาว

ตารางที่ 4.23 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูหนาวจากวิธี MLP

	CC	R-Square	RMSE
Percentage Split 90%	0.4020	0.1616	12.5865
Percentage Split 75%	0.3500	0.1225	14.8013
Percentage Split 50%	0.3679	0.1456	14.5238

ตามตารางที่ 4.23 ในข้อมูลชุดฤดูหนาวจากวิธี MLP การจัดข้อมูลที่ให้ผลดีที่สุดคือ Percentage Split 90% ให้ค่า CC = 0.4020, R-square = 0.1616, RMSE = 12.5865 ถือว่าอยู่ในเกณฑ์ที่ต่ำ ส่วนค่าที่เปอร์เซ็นต์อื่นๆ ผลที่ออกมาก็ถือว่าอยู่ในเกณฑ์ต่ำเช่นเดียวกัน

4.3.4.3 ชุดฤดูร้อน

ตารางที่ 4.24 ค่าชี้วัดประสิทธิภาพของแบบจำลองชุดข้อมูลฤดูร้อนจากวิธี MLP

	CC	R-Square	RMSE
Percentage Split 90%	0.3488	0.1217	4.9945
Percentage Split 75%	0.1623	0.0263	6.1986
Percentage Split 50%	0.2046	0.0418	5.3986

ตามตารางที่ 4.24 ในข้อมูลชุดฤดูร้อนจากวิธี MLP การจัดข้อมูลที่ให้ผลดีที่สุดคือ Percentage Split 90% ให้ค่า CC = 0.3488, R-square = 0.1217, RMSE = 4.9945 ซึ่งมีค่าที่ถือว่าอยู่ในเกณฑ์ที่ต่ำ ไม่สามารถนำไปใช้งานได้ ส่วนค่าที่เปอร์เซ็นต์อื่นๆ ผลที่ออกมาก็ถือว่าอยู่ในเกณฑ์ต่ำเช่นเดียวกัน เพียงแต่ค่า RMSE ที่ออกนั้นมีค่าน้อยกว่า ซึ่งถือว่าดี

ตารางที่ 4.25 ตารางเปรียบเทียบระหว่างแต่ละฤดูกาลของทฤษฎี MLP

	CC	R-Square	RMSE
ฤดูฝน, Percentage Split 75%	0.7758	0.6018	8.6990
ฤดูหนาว, Percentage Split 90%	0.4020	0.1616	12.5865
ฤดูร้อน, Percentage Split 90%	0.3488	0.1217	4.9945

ตามตารางที่ 4.25 ในทฤษฎี MLP ชุดข้อมูล และการจัดข้อมูลที่ให้ผลออกมาดีที่สุด คือ ฤดูฝน ที่ Percentage Split 75% ให้ค่า CC = 0.7994, R-square = 0.6390, RMSE = 8.6624 ซึ่งชุดข้อมูล ที่เหลือนั้น ค่าที่ออกมา ไม่สามารถนำไปใช้ได้ แต่ถ้ามองเฉพาะค่า RMSE จะเป็นทางด้านของ ชุดข้อมูล ฤดูร้อนที่ Percentage Split 90% RMSE = 4.9945

4.4 เปรียบเทียบแบบจำลองที่ได้จากการวิเคราะห์ของแต่ละทฤษฎี

ตารางที่ 4.26 ตารางเปรียบเทียบแบบจำลองที่ดีที่สุดของแต่ละทฤษฎี

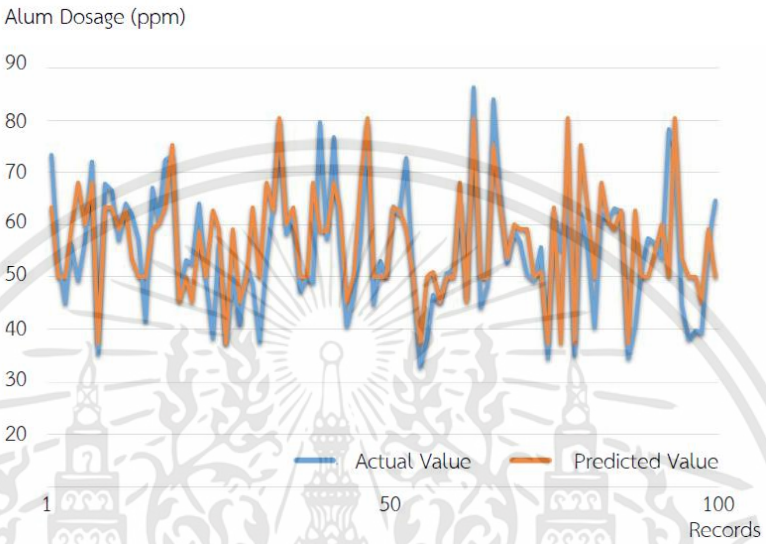
	CC	R-Square	RMSE
REPTree, ฤดูฝน, Percentage Split 75%	0.8036	0.6457	8.1006
M5P, ฤดูหนาว, Percentage Split 75%	0.8028	0.6446	8.7679
M5Rules, ฤดูหนาว, Percentage Split 75%	0.7994	0.6390	8.6624
MLP, ฤดูฝน, Percentage Split 75%	0.7758	0.6018	8.6990

ตามตาราง 4.26 จะเห็นได้ว่าค่า CC, R-square และ RMSE ของแบบจำลองแต่ละตัวนั้นมีค่าค่อนข้างใกล้เคียงกันมาก ซึ่งการจัดชุดข้อมูลที่ดีที่สุดคือ ที่ Percentage Split 75% เหมือนกันทุกแบบจำลอง ส่วนหน้าฝนกับหน้าหนาวนั้น ให้ค่า CC กับ R-Square สูงที่สุด หมายความว่าพารามิเตอร์ในน้ำดิบที่รับเข้ามานั้น มีค่าค่อนข้างคงที่ และ แบบจำลองที่ดีที่สุด เป็นทางด้านของ REPTree ชุดข้อมูลฤดูฝนที่ Percentage Split 75% มีค่า CC = 0.8036, R-Square = 0.6457 และ RMSE = 8.1006

4.5 นำแบบจำลองที่ดีที่สุดของแต่ละทฤษฎี ไปทำนายค่าการเติมปริมาณสารส้ม

4.5.1 การนำแบบจำลองทฤษฎี REPTree มาทำนายค่าสารส้มจริง

ในการทำนายนั้นได้ใช้ชุดข้อมูลในการทำนาย 100 ชุด ซึ่งแบบจำลองในการทำนายนั้นนำมาจาก ชุดข้อมูลฤดูฝนที่ Percentage Split 75% CC = 0.8036 และ R-square = 0.6457, RMSE = 8.1006 มาสร้างแบบจำลอง

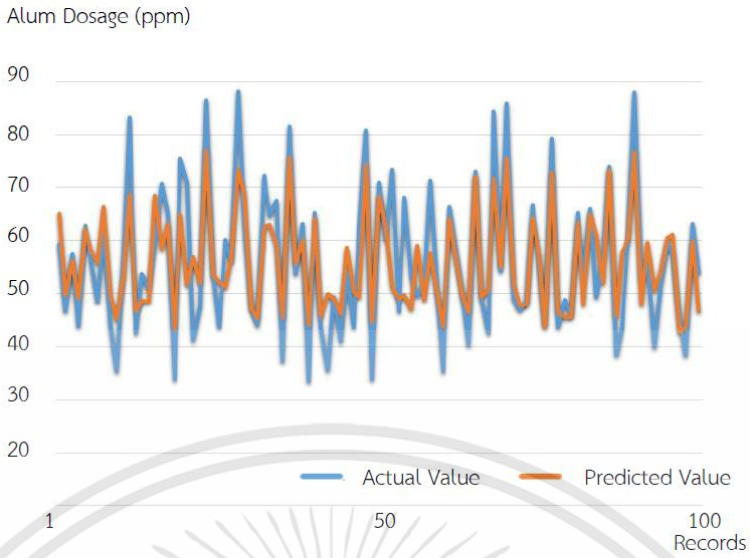


รูปที่ 4.5 กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี REPTree

ตามรูปที่ 4.5 กราฟระหว่างค่าจริงกับค่าจากการทำนายมีค่าค่อนข้างใกล้เคียงกันอยู่ในเกณฑ์ที่สามารถนำมาใช้จริงได้ ถึงแม้ว่ากราฟอาจจะไม่ลงลอยกันสนิท แต่การขึ้นลงของกราฟนั้น ขึ้นลงตามกัน และแบบจำลองนี้เป็นเพียงการเรียนรู้จากข้อมูลเพียง 1,433 ชุดเท่านั้น

4.5.2 การนำแบบจำลองทฤษฎี M5P มาทำนายค่าสารส้มจริง

ในการทำนายนั้นได้ใช้ชุดข้อมูลในการทำนาย 100 ชุด ซึ่งแบบจำลองในการทำนายนั้นนำมาจาก ชุดข้อมูลฤดูหนาวที่ Percentage Split 75% ค่า CC = 0.8028, R-square = 0.6446, RMSE = 8.7679 มาสร้างแบบจำลอง

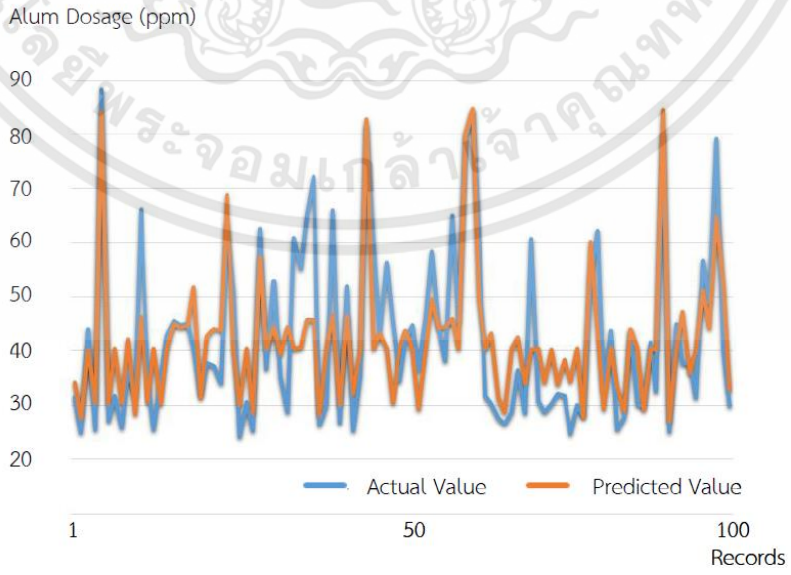


รูปที่ 4.6 กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี M5P

ตามรูปที่ 4.6 กราฟระหว่างค่าจริงกับค่าจากการทำนายมีค่าค่อนข้างใกล้เคียงกันอยู่ในเกณฑ์ที่สามารถนำมาใช้จริงได้ ถึงแม้ว่ากราฟอาจจะไม่ลงลอยกันสนิท แต่การขึ้นลงของกราฟนั้น ขึ้นลงตามกัน และแบบจำลองนี้เป็นเพียงการเรียนรู้จากข้อมูลเพียง 852 ชุดเท่านั้น

4.5.3 การนำแบบจำลองทฤษฎี M5Rules มาทำนายค่าสารส้มจริง

ในการทำนายนั้นได้ใช้ชุดข้อมูลในการทำนาย 100 ชุด ซึ่งแบบจำลองในการทำนายนั้นนำมาจาก ชุดข้อมูลฤดูฝนที่ Percentage Split 75% CC = 0.7994, R-square = 0.6390 และ RMSE = 8.6624 มาสร้างแบบจำลอง

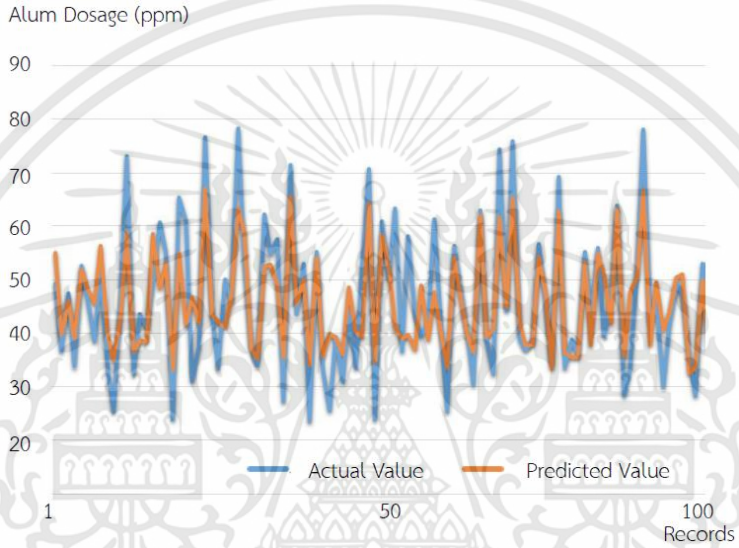


รูปที่ 4.7 กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี M5Rules

ตามรูปที่ 4.7 กราฟระหว่างค่าจริงกับค่าจากการทำนายมีค่าค่อนข้างใกล้เคียงกันอยู่ในเกณฑ์ที่สามารถนำมาใช้จริงได้ มีการขึ้นลงของกราฟที่ลักษณะ ขึ้นลงตามกัน และแบบจำลองนี้เป็นเพียงการเรียนรู้จากข้อมูลเพียง 852 ชุดเท่านั้น

4.5.4 การนำแบบจำลองทฤษฎี MLP มาทำนายค่าสารส้มจริง

ในการทำนายนั้นได้ใช้ชุดข้อมูลในการทำนาย 100 ชุด ซึ่งแบบจำลองในการทำนายนั้นนำมาจาก ชุดข้อมูลฤดูฝนที่ Percentage Split 75% CC = 0.7994, R-square = 0.6390 และ RMSE = 8.6624 มาสร้างแบบจำลอง



รูปที่ 4.8 กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี MLP

ตามรูปที่ 4.8 กราฟระหว่างค่าจริงกับค่าจากการทำนายมีค่าค่อนข้างใกล้เคียงกันอยู่ในเกณฑ์ที่สามารถนำมาใช้จริงได้ ถึงแม้ว่ากราฟอาจจะไม่ลงลอยกันสนิท แต่การขึ้นลงของกราฟนั้น ขึ้นลงตามกัน และแบบจำลองนี้เป็นเพียงการเรียนรู้จากข้อมูลเพียง 1,433 ชุดเท่านั้น

4.6 นำแบบจำลองที่ดีที่สุด ไปทำนายค่าการเติมปริมาณสารส้ม ในพื้นที่อื่น

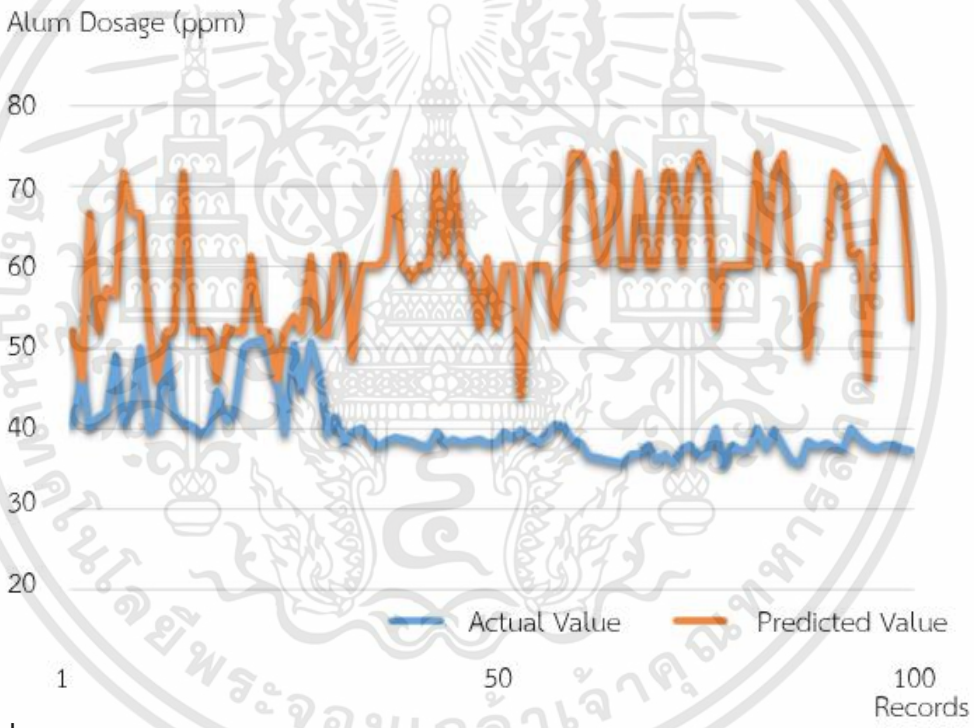
แบบจำลองที่มีประสิทธิภาพที่สุด 2 แบบจำลอง คือ แบบจำลองของ Reptree ชุดข้อมูลฤดูฝน สังเคราะห์แบบจำลองโดย Percentage Split 75% มีค่า CC = 0.8028, R-square = 0.6446, RMSE = 8.7679 และ M5P ชุดข้อมูลฤดูหนาว สังเคราะห์แบบจำลองโดย Percentage Split 75% มีค่า CC = 0.8028, R-square = 0.6446 และ RMSE = 8.7679 โดยจะนำแบบจำลองดังกล่าว มาทำนายค่าปริมาณการเติมสารส้ม ของโรงผลิตน้ำประปาธนบุรี ซึ่งจะใช้ข้อมูลในการทำนาย 100 ชุด จากข้อมูลทั้งหมด 4,114 ตัว เก็บข้อมูลตั้งแต่วันที่ 1 มกราคม พ.ศ.2549 ถึง 31 ธันวาคม พ.ศ. 2559 ตัวแปรต้นใช้ 3 ตัว คือ ค่าความขุ่น (Turbidity) ค่าปริมาณของแข็งแขวนลอย (Suspended solids)

และ ค่าพีเอช (pH) และตัวแปรตามคือ ปริมาณการเติมสารส้ม (Alum dosage) เช่นเดียวกันกับแบบจำลอง

ตารางที่ 4.27 ตารางเปรียบเทียบค่าจากการทำนายปริมาณการเติมสารส้ม ของโรงผลิตน้ำประปาในพื้นที่อื่น

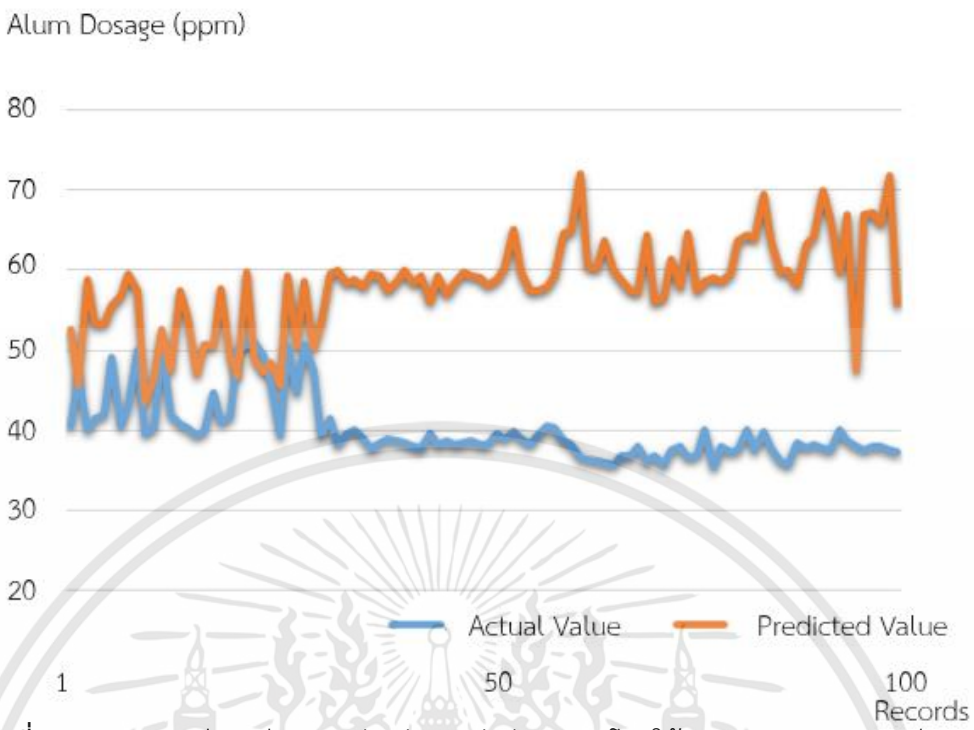
	CC	R-Square	RMSE
REPTree, ฤดูฝน, Percentage Split 75%	-0.3785	0.1443	23.0491
M5P, ฤดูหนาว, Percentage Split 75%	-0.4622	0.2137	20.0064

ตามตาราง 4.27 จะเห็นได้ว่า M5P ชุดข้อมูลฤดูหนาว ที่ Percentage Split 75% มีผลที่ดีกว่า ที่ CC = -0.4622, R-Square = 0.2137 และ RMSE = 20.0064 ถึงแม้ผลแบบจำลองของ M5P จะดีกว่า REPTree แต่ค่าที่ออกมาถือว่าต่ำ ไม่สามารถใช้งานได้จริง



รูปที่ 4.9 กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี REPTree ชุดข้อมูลฤดูฝน ทดสอบกับข้อมูลจาก โรงผลิตน้ำประปาธนบุรี

ตามรูปที่ 4.9 กราฟระหว่างค่าจริงกับค่าจากการทำนายมีค่าค่อนข้างต่างกันมาก อยู่ในเกณฑ์ที่ไม่สามารถนำมาใช้จริงได้ การขึ้นลงของกราฟ ไม่มีความสัมพันธ์กัน และมีความคลาดเคลื่อนสูง ที่ CC = -0.3785, R-Square = 0.1443 และ RMSE = 23.0491



รูปที่ 4.10 กราฟเปรียบเทียบระหว่างค่าจริงกับค่าทำนายโดยใช้แบบจำลองจากทฤษฎี M5P ชุดข้อมูลฤดูหนาว ทดสอบกับข้อมูลจาก โรงผลิตน้ำประปาธนบุรี

ตามรูปที่ 4.10 กราฟระหว่างค่าจริงกับค่าจากการทำนายมีค่าค่อนข้างต่างกันมาก อยู่ในเกณฑ์ที่ไม่สามารถนำมาใช้จริงได้ การขึ้นลงของกราฟ ไม่มีความสัมพันธ์กัน และมีความคลาดเคลื่อนสูง ที่ CC = -0.4622, R-Square = 0.2137 และ RMSE = 20.0064

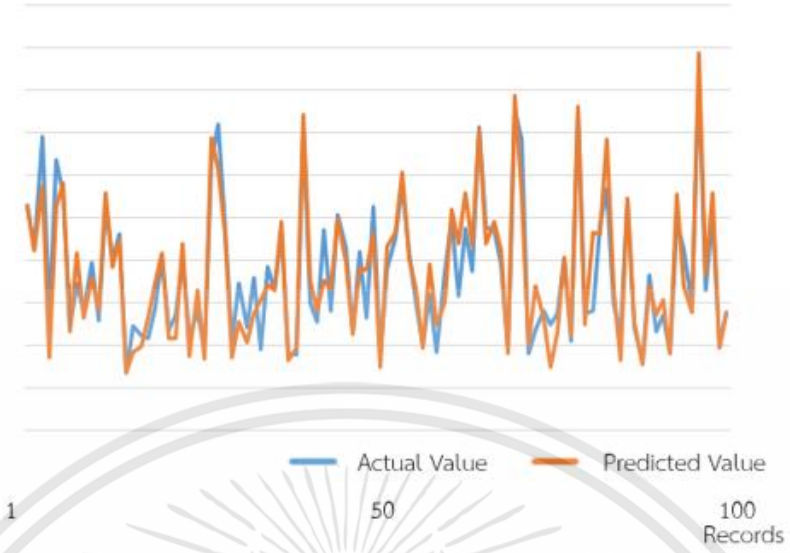
4.7 การเพิ่มประสิทธิภาพ โดยใช้โมเดลจากซอฟแวร์ตัวอื่น

การเพิ่มประสิทธิภาพนั้นได้ทำโดยการนำซอฟแวร์ตัวอื่น มาทำการทำนายผลเปรียบเทียบกัน เพื่อวิเคราะห์ความแตกต่าง และประสิทธิภาพระหว่างซอฟแวร์ในแต่ละตัว โดยจะใช้ ซอฟแวร์ที่นำมาเพื่อใช้ร่วมกัน อีก 2 ตัว

1. Python

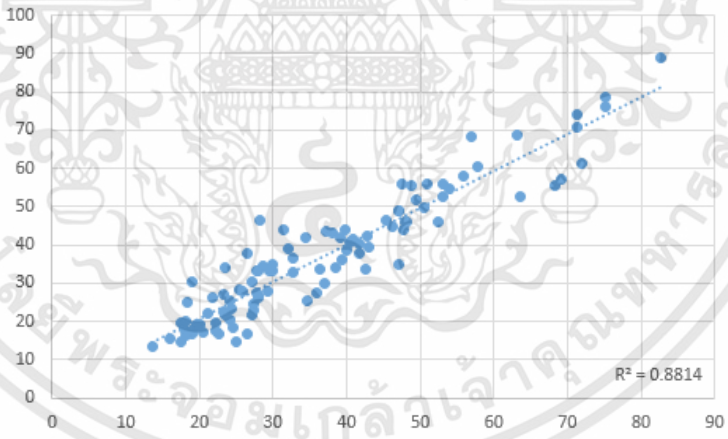
ได้ทำนายการเติมสารส้มโดยใช้แบบจำลองจาก ทฤษฎี ที่มีอัลกอริทึม คล้ายกับ REPTree เพื่อเปรียบเทียบกัน มีชื่อว่า Extra Tree และนำมาทำนายค่าปริมาณการเติมสารส้ม จำนวน 100 ชุด โดยเรียนรู้จากข้อมูลจำนวน 3,400 ชุด

Alum Dosage



รูปที่ 4.11 การทำนายค่าปริมาณการเติมสารส้มโดย Extra Tree

ตามรูปที่ 4.11 ค่าการทำนายที่ออกมานั้นมีค่าใกล้เคียงกับค่าจริงมาก มีการขึ้นลงของกราฟตามค่าจริง RMSE = 5.272 ซึ่งมีค่าน้อยมาก แต่ความละเอียดในการลู่ตามหยักที่ขึ้นลงของกราฟนั้น Weka ยังถือว่าทำได้ดีกว่า

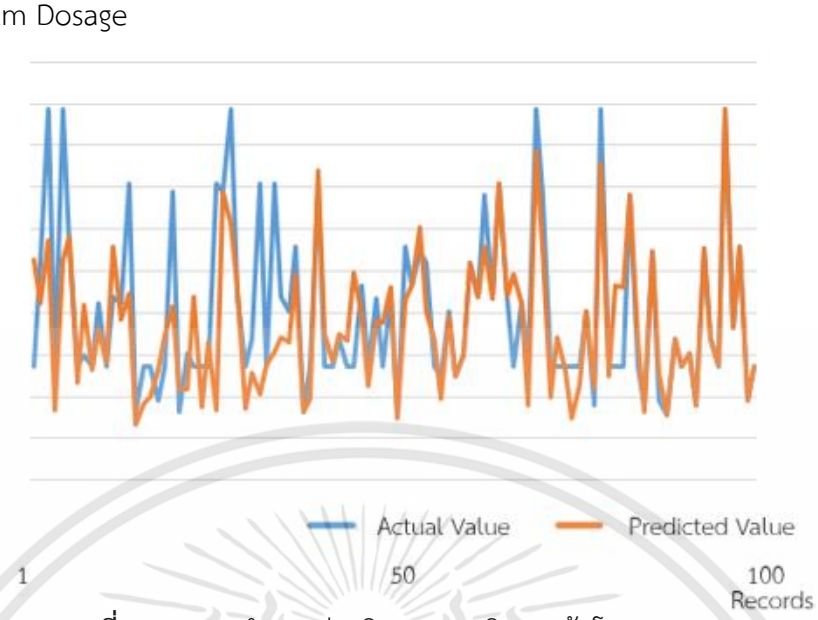


รูปที่ 4.12 กราฟเปรียบเทียบค่าจริงกับค่าจากการทำนายโดย Extra Tree เพื่อหาค่า R-square

ตามรูป 4.12 จะเห็นได้ว่าค่า R-square นั้นมีค่าสูงมาก สูงถึง 0.8814 ซึ่งแสดงถึงความสัมพันธ์กันเป็นอย่างมากของตัวแปรในแบบจำลอง

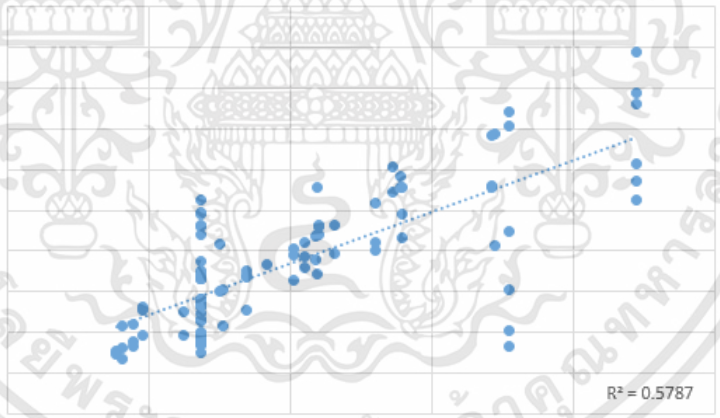
2. Rapid Miner

ได้ทำนายการเติมสารส้มโดยใช้แบบจำลองจาก ทฤษฎี REPTree ที่คล้ายกับ REPTree จาก Weka พอสมควรแต่แตกต่างกันที่วิธีนำเข้าของข้อมูล นำมาทำนายค่าปริมาณการเติมสารส้ม จำนวน 100 ชุด โดยเรียนรู้จากข้อมูลจำนวน 3,400 ชุด



รูปที่ 4.13 การทำนายค่าปริมาณการเติมสารส้มโดย REPTree

ตามรูปที่ 4.13 ค่าการทำนายที่ออกมานั้นมีค่าค่อนข้างดี มีการขึ้นลงของกราฟตามค่าจริง RMSE = 13.1942 ซึ่งมีอยู่ในเกณฑ์ที่ใช้ได้ แต่ความละเอียดในการหยักขึ้นลง Weka ทำได้ดีกว่า



รูปที่ 4.14 กราฟเปรียบเทียบค่าจริงกับค่าจากการทำนายโดย REPTree เพื่อหาค่า R-square

ตามรูป 4.14 จะเห็นได้ว่าค่า R-square = 0.5787 นั้นมีค่าปานกลาง ถือว่าแบบจำลองนำไปใช้ได้ในบางกรณี

บทที่ 5
วิเคราะห์ผลงานวิจัย

จากการดำเนินงานวิจัยเพื่อให้บรรลุตามวัตถุประสงค์ ผู้วิจัยได้รวบรวมข้อมูลและวิเคราะห์ผลตามหัวข้อต่อไปนี้

1. วิเคราะห์ข้อดีและข้อเสียของแต่ละทฤษฎี
2. วิเคราะห์ความแม่นยำจากการทำนายว่าสามารถนำมาใช้จริงได้หรือไม่
3. วิเคราะห์การทำนายโดชใช้ ข้อมูลทดสอบ (Testing set) ที่มาจากโรงผลิตน้ำประปาต่างแห่ง

5.1 วิเคราะห์ข้อดีและข้อเสียของแต่ละทฤษฎี

ในแต่ละทฤษฎีนั้น มีผลการสังเคราะห์ข้อมูลที่ต่างกัน และค่าต่างๆที่ได้มาค่อนข้างต่างกัน ซึ่งมีผลกับการนำไปประยุกต์ใช้จริง โดยความเหมาะสมของแบบจำลองสามารถดูได้จากหัวข้อดังต่อไปนี้

5.2.1 REPTree

ตารางที่ 5.1 ผลการสังเคราะห์แบบจำลองจาก ทฤษฎี REPTree

	CC	R-Square	RMSE
ฤดูฝน, Percentage Split 90%	0.7651	0.5855	9.3566
ฤดูฝน, Percentage Split 75%	0.8036	0.6457	8.1006
ฤดูฝน, Percentage Split 50%	0.7937	0.6300	8.1657
ฤดูหนาว, Percentage Split 90%	0.6826	0.4660	9.4669
ฤดูหนาว, Percentage Split 75%	0.7051	0.4972	10.2128
ฤดูหนาว, Percentage Split 50%	0.5894	0.3474	11.4094
ฤดูร้อน, Percentage Split 90%	0.5458	0.2979	4.0290
ฤดูร้อน, Percentage Split 75%	0.1961	0.0385	4.9566
ฤดูร้อน, Percentage Split 50%	0.3771	0.1422	4.9553

ตามตารางที่ 5.1 แสดงให้เห็นว่า การสร้างแบบจำลองจากชุดข้อมูล ฤดูฝน เป็นชุดข้อมูลที่มีความเกี่ยวข้องกันของแต่ละตัวแปรมากที่สุด และผลที่ดีที่สุดอยู่ที่ Percentage Split 75% แต่ไม่ว่าจะใช้ Percentage Split เท่าใด ผลก็ออกมาดีเสมอ ลำดับต่อมาคือผลการสังเคราะห์ จากข้อมูลชุดฤดูหนาว ผลที่ออกมานั้นถือว่าอยู่ในเกณฑ์ที่สามารถนำไปใช้ได้ โดยค่าที่ดีที่สุดอยู่ที่ Percentage Split 75% เช่นเดียวกันกับฤดูหนาว และชุดข้อมูลที่มีผลการสังเคราะห์แย่ที่สุดคือ ฤดูร้อน ซึ่งข้อมูลไม่สามารถนำไปใช้ได้

5.2.2 M5P

ตารางที่ 5.2 ผลการสังเคราะห์แบบจำลองจาก ทฤษฎี M5P

	CC	R-Square	RMSE
ฤดูฝน, Percentage Split 90%	0.7929	0.6287	8.7450
ฤดูฝน, Percentage Split 75%	0.7823	0.6120	8.4748
ฤดูฝน, Percentage Split 50%	0.7686	0.5908	8.5953
ฤดูหนาว, Percentage Split 90%	0.7033	0.4946	9.4015
ฤดูหนาว, Percentage Split 75%	0.8028	0.6446	8.7679
ฤดูหนาว, Percentage Split 50%	0.6544	0.4283	10.5477
ฤดูร้อน, Percentage Split 90%	0.4928	0.2428	4.2163
ฤดูร้อน, Percentage Split 75%	0.4190	0.1756	4.5449
ฤดูร้อน, Percentage Split 50%	0.4038	0.1631	4.4766

ตามตารางที่ 5.2 แสดงให้เห็นว่า การสร้างแบบจำลองจากชุดข้อมูล ฤดูฝน เป็นชุดข้อมูลที่มีความเกี่ยวข้องกันของแต่ละตัวแปรมากที่สุด และผลที่ดีที่สุดอยู่ที่ Percentage Split 90% แต่ไม่ว่าจะใช้ Percentage Split เท่าใด ผลก็ออกมาดีเสมอ ลำดับต่อมาคือผลการสังเคราะห์ จากข้อมูลชุดฤดูหนาว ผลที่ออกมานั้นค่อนข้างดีกว่าฤดูหนาวของ M5P โดยค่าที่สูงที่สุดอยู่ที่ Percentage Split 75% มีค่าสูงถึง CC = 0.8028, R-square = 0.6446 และ RMSE = 8.7679 และชุดข้อมูลที่มีผลการสังเคราะห์แย่ที่สุดคือ ฤดูร้อน ซึ่งข้อมูลไม่สามารถนำไปใช้ได้ แต่ข้อมูลที่สังเคราะห์ได้นั้น ดีกว่าแบบ REPTree

5.2.3 M5Rules

ตารางที่ 5.3 ผลการสังเคราะห์แบบจำลองจาก ทฤษฎี M5Rules

	CC	R-Square	RMSE
ฤดูฝน, Percentage Split 90%	0.7448	0.5547	9.5745
ฤดูฝน, Percentage Split 75%	0.7851	0.6164	8.4262
ฤดูฝน, Percentage Split 50%	0.7832	0.6134	8.3423
ฤดูหนาว, Percentage Split 90%	0.7574	0.5736	8.5215
ฤดูหนาว, Percentage Split 75%	0.7994	0.6390	8.6624
ฤดูหนาว, Percentage Split 50%	0.7436	0.5529	9.3346
ฤดูร้อน, Percentage Split 90%	0.5029	0.2529	4.1965
ฤดูร้อน, Percentage Split 75%	0.3729	0.1391	4.7709

ฤดูร้อน, Percentage Split 50%	0.3907	0.1526	4.5727

ตามตารางที่ 5.3 แสดงให้เห็นว่า การสร้างแบบจำลองจากชุดข้อมูล ฤดูฝน เป็นชุดข้อมูลที่มีความเกี่ยวข้องกันของแต่ละตัวแปรมากที่สุด และผลที่ดีที่สุดอยู่ที่ Percentage Split 75% และ Percentage Split 50% ผลก็ออกมาใกล้เคียงกันมาก แต่ไม่ว่าจะใช้ Percentage Split เท่าใด ผลก็ออกมาดีเสมอ ลำดับต่อมาคือผลการสังเคราะห์ จากข้อมูลชุดฤดูหนาว ผลที่ออกมานั้นมีค่าดีที่สุดในทุกทฤษฎีในชุดข้อมูลฤดูหนาว โดยค่าที่สูงที่สุดอยู่ที่ Percentage Split 75% มีค่าสูงถึง CC = 0.7994, R-square = 0.6390 และ RMSE = 8.6624 และไม่ว่าจะใช้ Percentage Split เท่าใด ผลก็ออกมาดีเสมอ ชุดข้อมูลที่มีผลการสังเคราะห์แย่ที่สุดคือ ฤดูร้อน เหมือนทฤษฎีต่างๆที่ผ่านมา ซึ่งข้อมูลไม่สามารถนำไปใช้ได้

5.2.4 Multilayer Perceptron

ตารางที่ 5.4 ผลการสังเคราะห์แบบจำลองจาก ทฤษฎี Multilayer Perceptron

	CC	R-Square	RMSE
Percentage Split 90%	0.7456	0.5559	10.0718
Percentage Split 75%	0.7758	0.6018	8.6990
Percentage Split 50%	0.7703	0.5935	8.5508
ฤดูหนาว, Percentage Split 90%	0.4020	0.1616	12.5865
ฤดูหนาว, Percentage Split 75%	0.3500	0.1225	14.8013
ฤดูหนาว, Percentage Split 50%	0.3679	0.1456	14.5238
ฤดูร้อน, Percentage Split 90%	0.3488	0.1217	4.9945
ฤดูร้อน, Percentage Split 75%	0.1623	0.0263	6.1986
ฤดูร้อน, Percentage Split 50%	0.2046	0.0418	5.3986

ตามตารางที่ 5.4 แสดงให้เห็นว่า การสร้างแบบจำลองจากชุดข้อมูล ฤดูฝน เป็นชุดข้อมูลที่มีความเกี่ยวข้องกันของแต่ละตัวแปรมากที่สุด และผลที่ดีที่สุดอยู่ที่ Percentage Split 75% แต่ไม่ว่าจะใช้ Percentage Split เท่าใด ผลก็ออกมาดีเสมอ ถ้าเทียบกับ 3 ทฤษฎีที่กล่าวมาก่อนหน้านี้ MLP ถือว่ามีผลที่ออกมาประสิทธิภาพต่ำที่สุด เพราะอาจจะไม่เหมาะกับการสังเคราะห์ข้อมูลที่เป็นตัวเลข โดยค่า R-Square ของแบบจำลองจาก MLP จะเฉลี่ยนอกจากฤดูฝน มีค่าต่ำมาก ซึ่งไม่สามารถนำมาใช้งานจริงได้

5.2 วิเคราะห์ความแม่นยำจากการทำนายว่าสามารถนำมาใช้จริงได้หรือไม่

ในการนำแบบจำลองมาใช้งานในการทำนายค่าจริงนั้นควรใช้แบบจำลองและชุดข้อมูลที่นำมาสร้างแบบจำลองให้ตรงกับ ข้อมูลที่ต้องการทำนาย เช่น ต้องการทำนายข้อมูลในฤดูหนาว ก็ควร

ใช้แบบจำลองที่สร้างมาจากฤดูหนาว หรือ ต้องการทำนายข้อมูลในฤดูฝน ก็ควรใช้แบบจำลองที่สร้างมาจากฤดูฝน เพราะค่าตัวแปรต่างๆ ที่นำมาสร้างแบบจำลอง ค่อนข้างมีค่าที่ต่างกัน เนื่องจากปริมาณน้ำฝน และ อุณหภูมิเป็นต้น

5.3 วิเคราะห์การทำนายโดยใช้ ข้อมูลทดสอบ (Testing set) ที่มาจากโรงผลิตน้ำประปาธนบุรี

ในการทำนายค่าปริมาณการเติมสารส้มในโรงประปาต่างแห่งนั้น ในกรณีนี้ใช้ข้อมูลจากโรงผลิตน้ำประปาธนบุรี ซึ่งมีผลดังต่อไปนี้

ตารางที่ 4.26 ตารางเปรียบเทียบค่าจากการทำนายปริมาณการเติมสารส้ม ของโรงผลิตน้ำประปาในพื้นที่อื่น

	CC	R-Square	RMSE
REPTree, ฤดูฝน, Percentage Split 75%	-0.3785	0.1443	23.0491
M5P, ฤดูหนาว, Percentage Split 75%	-0.4622	0.2137	20.0064

ตามตาราง 4.26 ได้เห็นได้ว่า ค่า CC มีการติดลบ นั่นหมายความว่า มีความสัมพันธ์กันในเชิงตรงกันข้าม ซึ่งจะถือว่าแบบจำลองไม่สามารถเข้ากันได้ เพราะการรูปแบบการบำบัดน้ำเสีย คือต้องมีการแปรผันตรงของตัวแปรต้นและตัวแปรตาม แต่ในกรณีนี้ค่าที่ออกมากลับเป็นการแปรผกผัน ทำให้สามารถระบุได้จากผลการวิเคราะห์ว่า ไม่สามารถใช้แบบจำลองร่วมกันได้ จากผลของการสังเคราะห์ที่ออกมา

5.4 วิเคราะห์ผลการเพิ่มประสิทธิภาพ โดยใช้โมเดลจากซอฟแวร์ตัวอื่น

ในการทดสอบแบบจำลองกับซอฟแวร์ตัวอื่นนั้นได้แก่ Python และ Rapid Miner ผลออกมาค่อนข้างดี โดยที่ Python มีค่า R-square ที่ต่ำมาก และ Rapid Miner ฏ้ถือว่าอยู่เกณฑ์ที่ใช้ได้

ตารางที่ 5.5 ค่าแสดงประสิทธิภาพของซอฟท์แวร์ตัวเสริม

	R-Square	RMSE
Extra tree (Python)	0.8814	5.2720
REPTree (Rapid Miner)	0.5787	13.1942
REPTree (Weka)	0.8028	8.7679

ตามตารางที่ 5.5 จะเห็นได้ว่า Extra tree นั้นให้ค่า CC ที่มีค่าสูงมากถึง 0.8817 ลำดับรองลงมาคือ Weka = 0.8028 ส่วนน้อยที่สุดนั้นคือ Rapid Miner CC = 0.5787 ซึ่งที่มีค่าน้อยนั้นอาจจะเป็นเพราะการนำเข้าของข้อมูลที่ยาก และซอฟแวร์ไม่สามารถแยกชนิดของข้อมูลได้ เลยทำให้ ผลออกมาไม่ดีเท่าซอฟแวร์ตัวอื่น ส่วนผลของ Python นั้น สามารถกำหนดค่าได้ว่าจะรับข้อมูลเป็นข้อมูลชนิดใด การประมวลผลก็เลยง่ายตามไปด้วย ทำให้ผลที่ออกมา ได้ประสิทธิภาพสูงสุดของแบบจำลอง

บรรณานุกรม

ณัฏภัทรศญา ทับทิมเทศ. **“ระบบสนับสนุนการตัดสินใจ.”** [Online]. Available : http://www.no-poor.com/dssandos/Chapter5-dss.html 2550.

มั่นสิน ตัณฑุลเวศม์. **วิศวกรรมประปา.** กรุงเทพฯ : สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย. 2537.

พรศักดิ์ สมรไกรสรกิจ. **“กระบวนการโคแอกกูเลชั่น (Coagulation) และ ฟล็อคคูเลชั่น (Flocculation).”** กรุงเทพมหานคร. : การประปานครหลวง. 2550.

Akhtar, A., Masood, S., Gupta, C., and Masood, A. 2018. "Prediction and Analysis of Pollution Levels in Delhi Using Multilayer Perceptron." **Data Engineering and Intelligent Computing. Springer, Singapore.** 563-572.

Amirtharajah, A., and Kirk M.M. 1982. "Rapid-mix design for mechanisms of alum coagulation." **Journal American Water Works Association.** 210-216.

Almasi, S.N., Bagherpour, R., Mikaeil, R., Ozcelik, Y., and Kalhori, H. 2017. "Predicting the Building Stone Cutting Rate Based on Rock Properties and Device Pullback Amperage in Quarries Using M5P Model Tree." **Geotechnical and Geological Engineering.** 35(4): 1311-1326.

Bertone, E., Stewart, R.A., Zhang, H., and O'Halloran, K. 2016. "Hybrid water treatment cost prediction model for raw water intake optimization." **Environmental Modelling & Software.** 75:230-242.

Chawakitchareon, P., Nattanan B., and Pakorn C. 2017. "Prediction of Alum Dosage in Water Supply by WEKA Data Mining Software." **Information Modelling and Knowledge Bases XXVIII**. 292: 83.

Driscoll, C.T., and William D.S. 1990. "The chemistry of aluminum in the environment." **Environmental Geochemistry and Health.** 12(1): 28-49.

Bringmann, B., Nijssen, S., and Zimmermann, A. 2010. "From Local Patterns to Classification Models. " **Inductive Databases and Constraint-Based Data Mining**. 127-154.

Hanbay, D., Ibrahim T., and Yakup D. 2008. "Prediction of wastewater treatment plant performance based on wavelet packet decomposition and neural networks." **Expert Systems with Applications.** 34(2): 1038-1043.

Hong-Gui, Qi-li C., and Jun-Fei, Q. 2011. "An efficient self-organizing RBF neural network for water quality prediction." **Neural Networks.** 24(7): 717-725. Kumar, J., Satheesh, P.P. and P. Balakumaran. 2013. "Artificial Intelligence Based Alum Dosage Control in Water Treatment Plant." **Int. J. Eng. Technol.** 5: 3344-3350.

Kiyoki, Y., Chen, x., Heimberger, A., Chawakitchareon, P., and Sornlertlamvanich, V. 2016. "Cross - cultural and Environmental Data Analysis in Data Mining Processes for a Global Resilient Society." **Information Modelling and Knowledge Bases XXVII**. 281-298.

Kumar, J., Satheesh, P.P. and P. Balakumaran. 2013. "Artificial Intelligence Based Alum Dosage Control in Water Treatment Plant." **Int. J. Eng. Technol.** 5: 3344-3350.

Ladsavong, k., Chawakitchareon, P., and Kiyoki, Y. 2017. "Application of Data Mining Software to Predict the Alum Dosage in Coagulation Process: A Case Study of Vientaine, Lao PDR**." Information Modelling and Knowledge Bases XXIX.** 110-124.

Maier, H.R., Nicolas, M., and Christopher, W.K.C. 2004. "Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters." **Environmental Modelling & Software.** 19(5): 485-494.

Naidoo, L., Cho, M.A., Mathieu, R., and Asner, G. 2012. "Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment." **ISPRS journal of Photogrammetry and Remote Sensing.** 69: 167-179.

Pınar, T. 2014. "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods." **International Journal of Electrical Power & Energy Systems.** 60: 126-140.Han,

Wu, Guan-De., and Shang-Lien L. 2008. "Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system." **Engineering Applications of Artificial Intelligence.** 21(8): 1189-1195.

ภาคผนวก

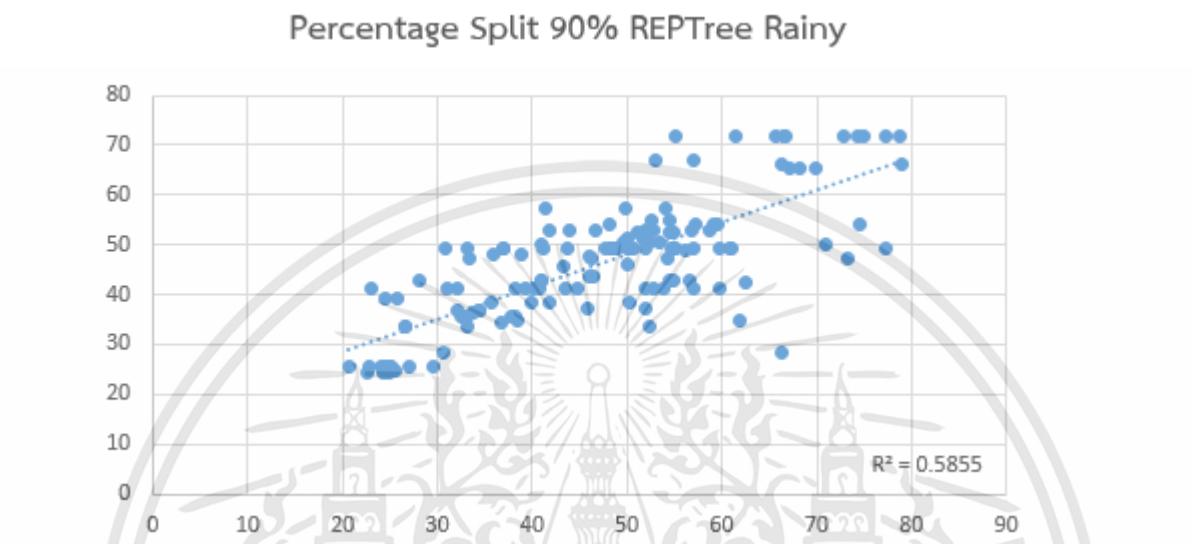
ภาคผนวก ก.

ข้อมูลผลการทดลอง

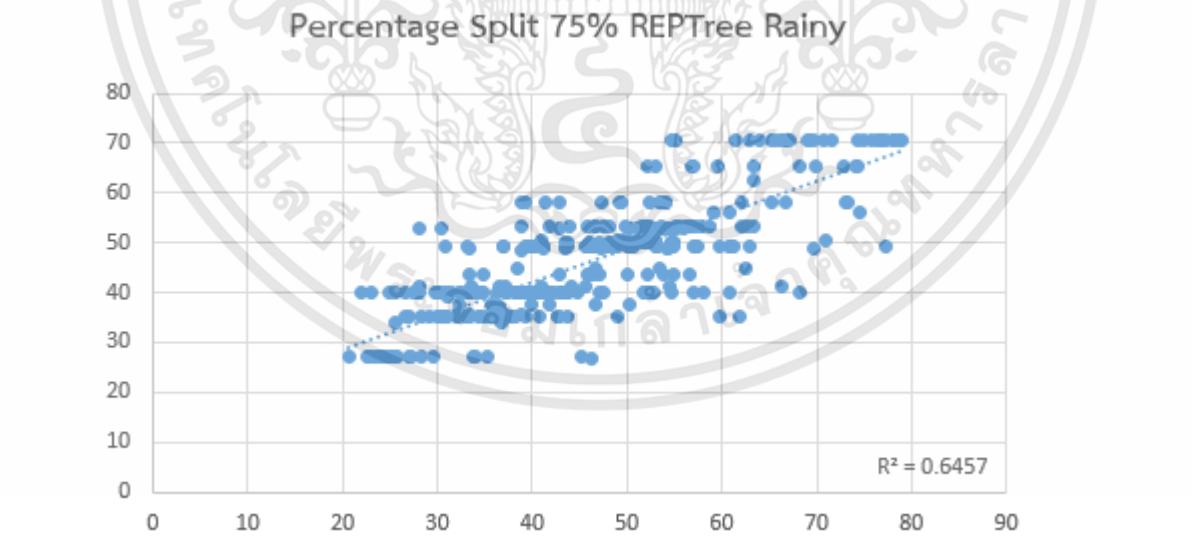
1.ผลการหาค่า R-square จากแบบจำลอง

1.1 REPTree

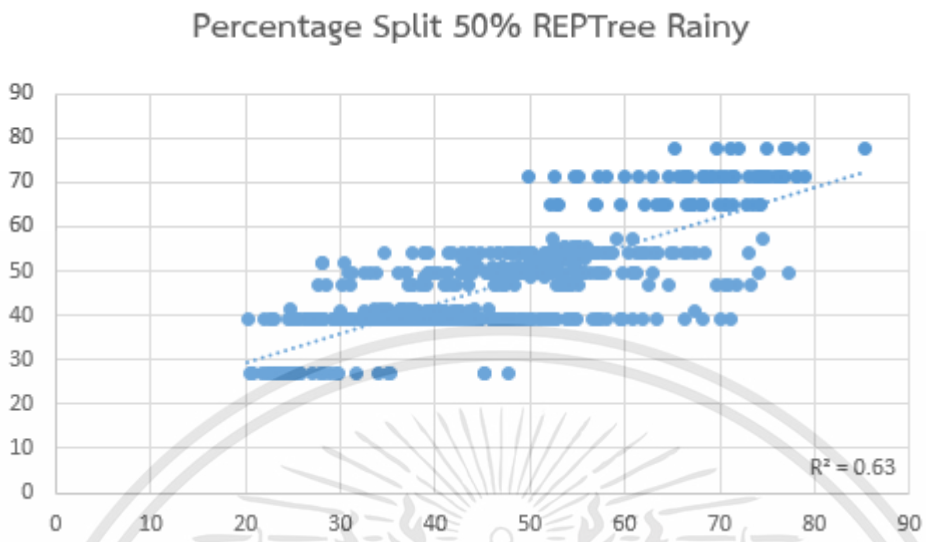
รูปที่ ผก. 1 แบบจำลองจากทฤษฎี REPTree ชุดข้อมูล ฤดูฝน ที่ Percentage split 90%



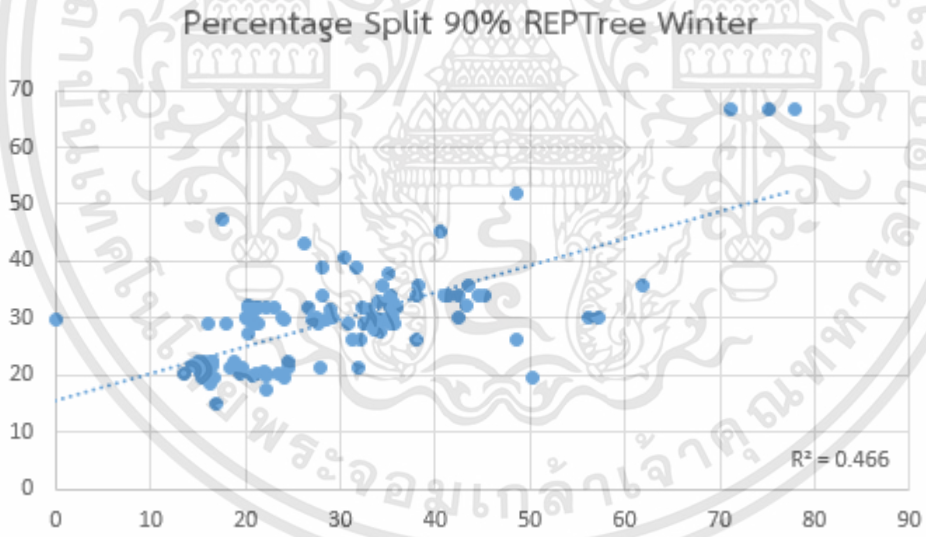
รูปที่ ผก. 4 แบบจำลองจากทฤษฎี REPTree ชุดข้อมูล ฤดูฝน ที่ Percentage split 75%



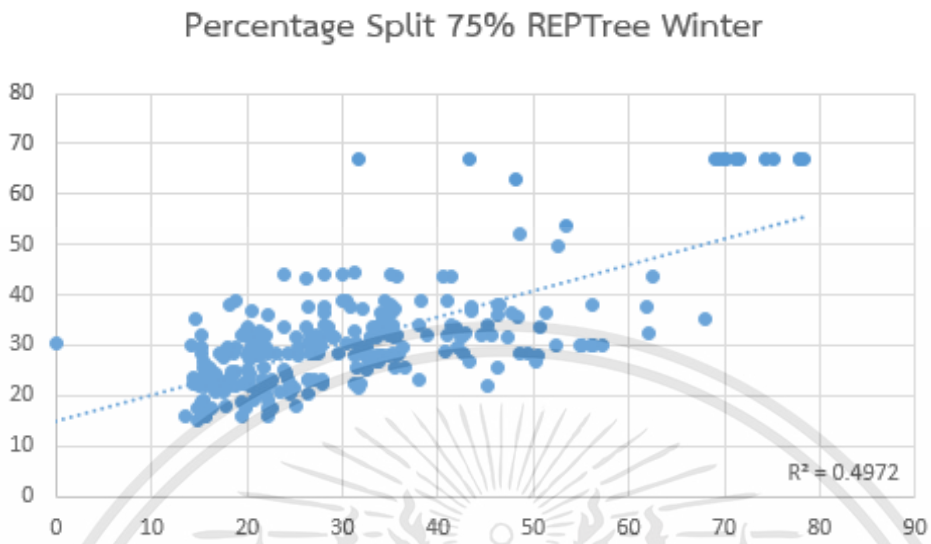
รูปที่ ผก. 3 แบบจำลองจากทฤษฎี REPTree ชุดข้อมูล ฤดูฝน ที่ Percentage split 50%



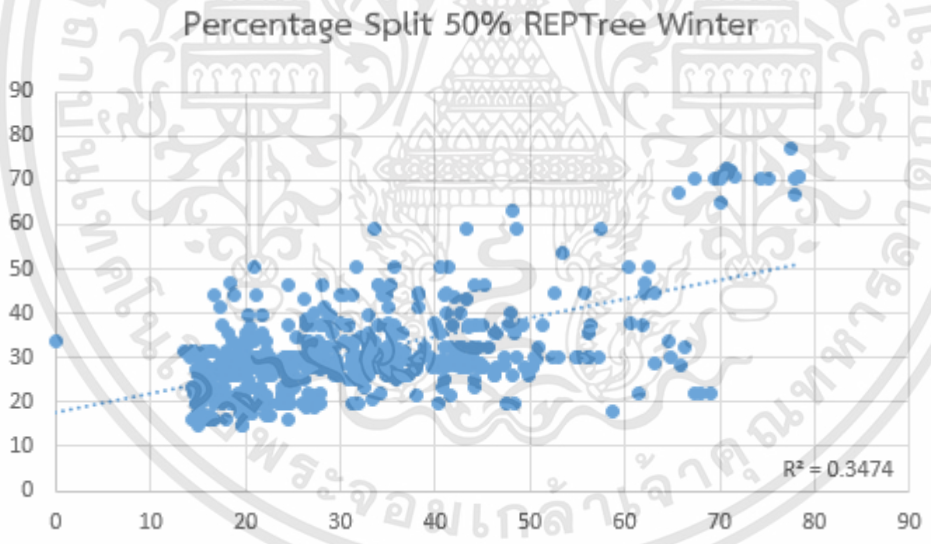
รูปที่ ผก. 4 แบบจำลองจากทฤษฎี REPTree ชุดข้อมูล ฤดูหนาว ที่ Percentage split 90%



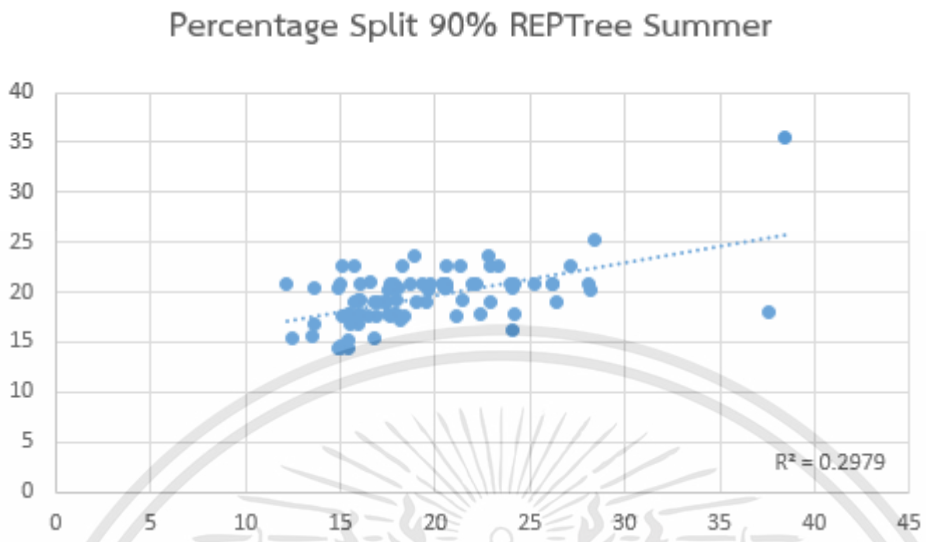
รูปที่ ผก. 5 แบบจำลองจากทฤษฎี REPTree ชุดข้อมูล ฤดูหนาว ที่ Percentage split 75%



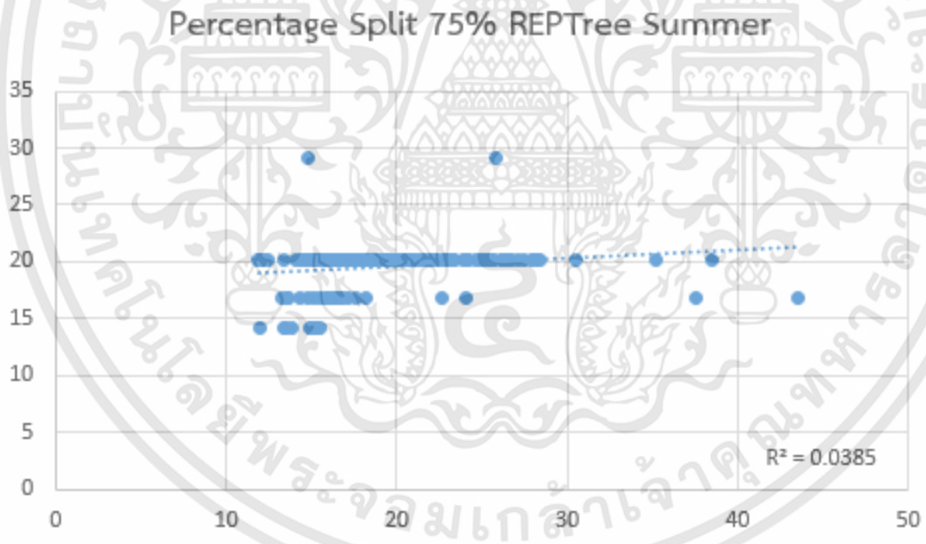
รูปที่ ผก. 6 แบบจำลองจากทฤษฎี REPTree ชุดข้อมูล ฤดูหนาว ที่ Percentage split 50%



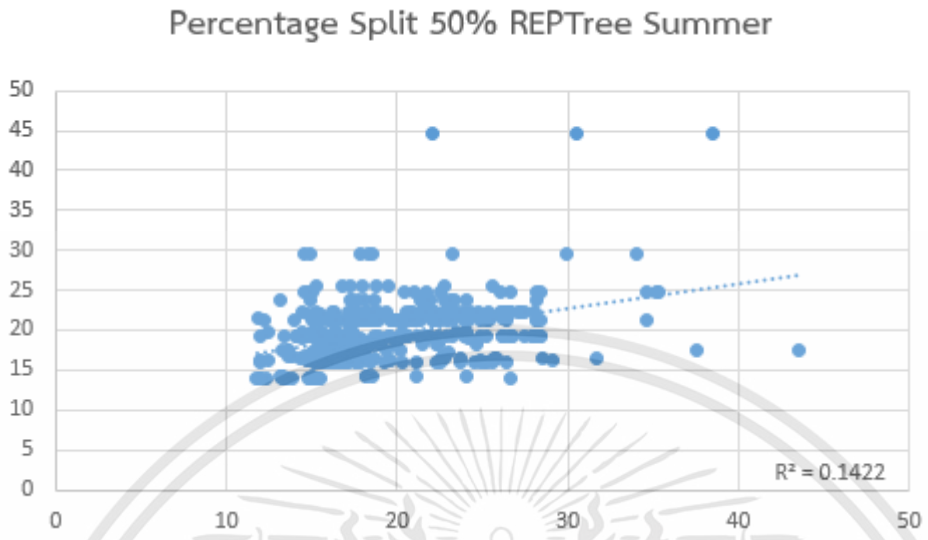
รูปที่ ผก. 7 แบบจำลองจากทฤษฎี REPTree ชุดข้อมูล ฤดูร้อน ที่ Percentage split 90%



รูปที่ ผก. 8 แบบจำลองจากทฤษฎี REPTree ชุดข้อมูล ฤดูร้อน ที่ Percentage split 75%

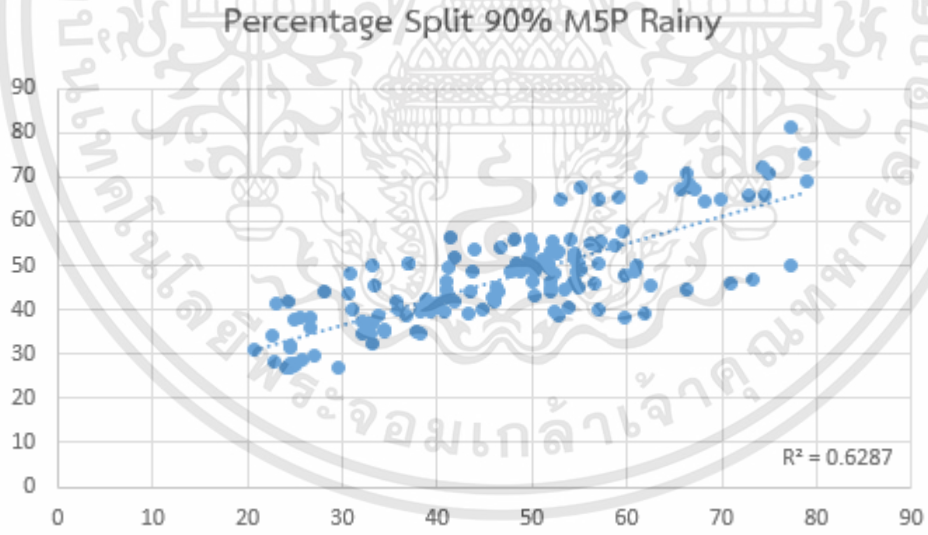


รูปที่ ผก. 9 แบบจำลองจากทฤษฎี REPTree ชุดข้อมูล ฤดูร้อน ที่ Percentage split 50%

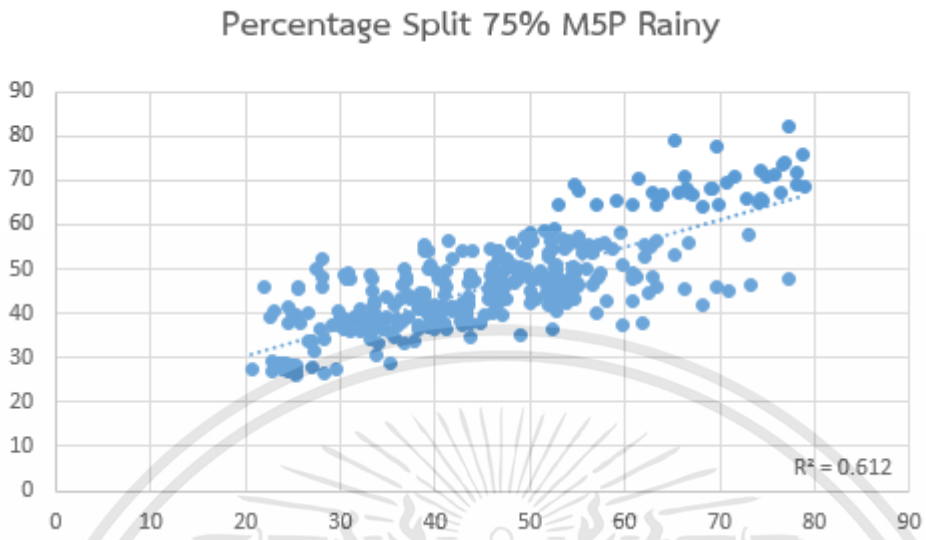


1.2 M5P

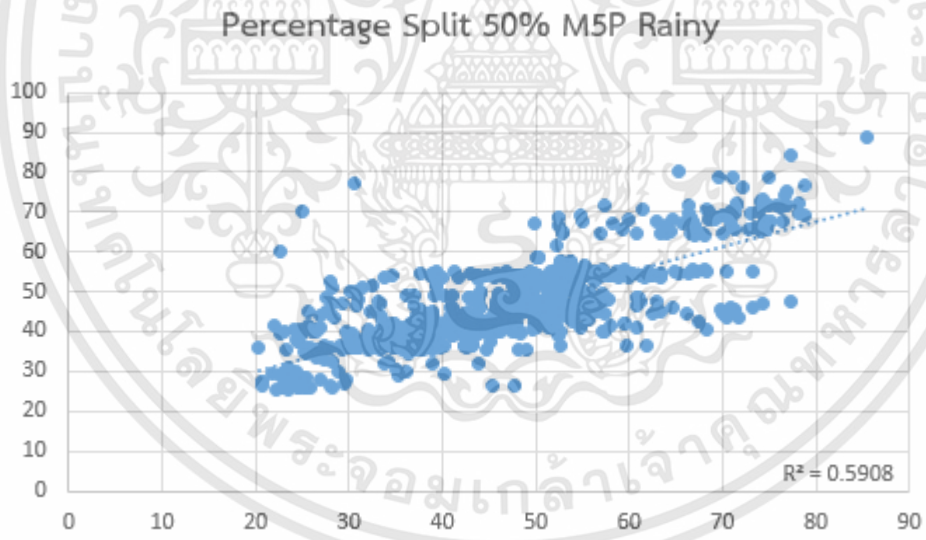
รูปที่ ผก. 10 แบบจำลองจากทฤษฎี M5P ชุดข้อมูล ฤดูฝน ที่ Percentage split 90%



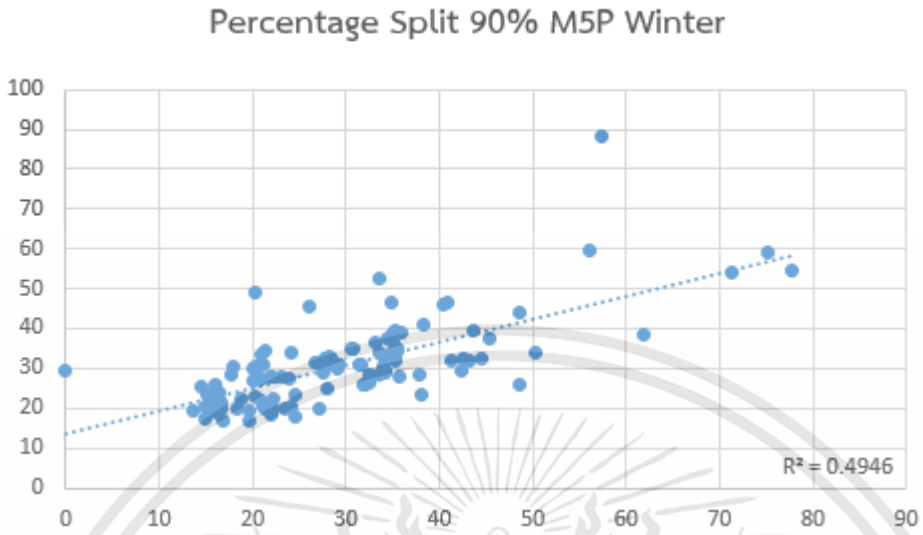
รูปที่ ผก. 11 แบบจำลองจากทฤษฎี M5P ชุดข้อมูล ฤดูฝน ที่ Percentage split 75%



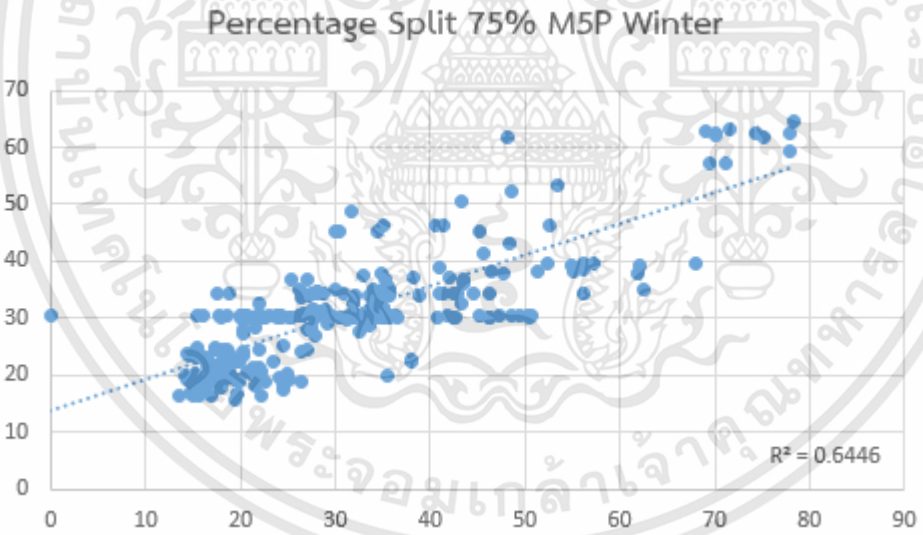
รูปที่ ผก. 12 แบบจำลองจากทฤษฎี M5P ชุดข้อมูล ฤดูฝน ที่ Percentage split 50%



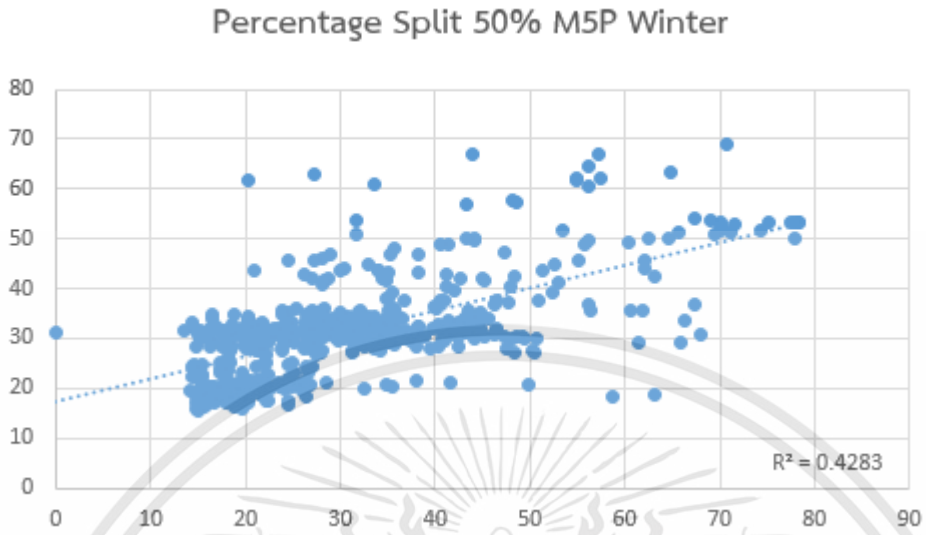
รูปที่ ผก. 13 แบบจำลองจากทฤษฎี M5P ชุดข้อมูล ฤดูหนาว ที่ Percentage split 90%



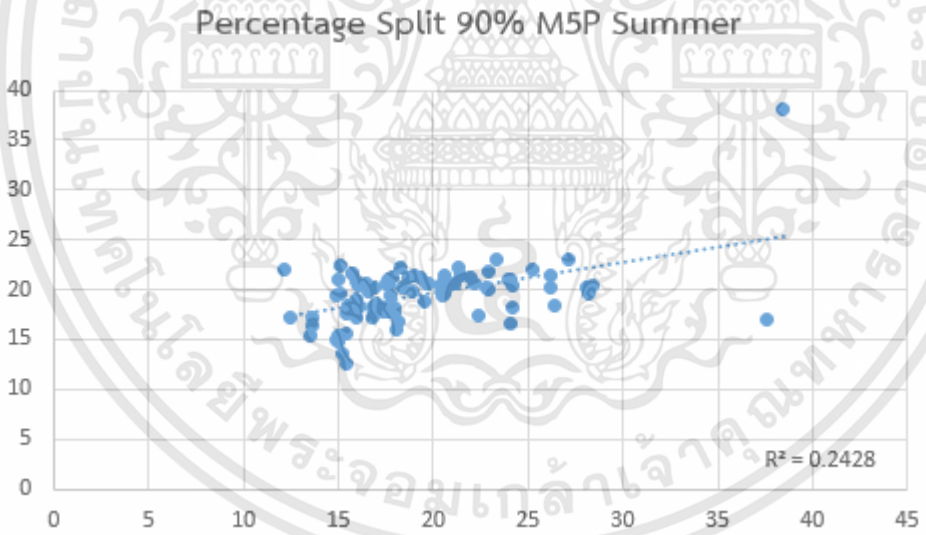
รูปที่ ผก. 14 แบบจำลองจากทฤษฎี M5P ชุดข้อมูล ฤดูหนาว ที่ Percentage split 75%



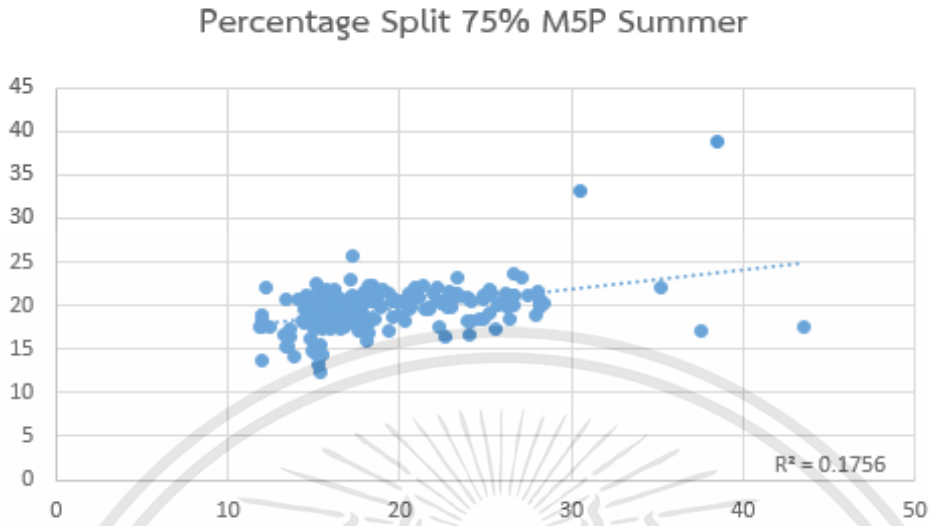
รูปที่ ผก. 15 แบบจำลองจากทฤษฎี M5P ชุดข้อมูล ฤดูหนาว ที่ Percentage split 50%



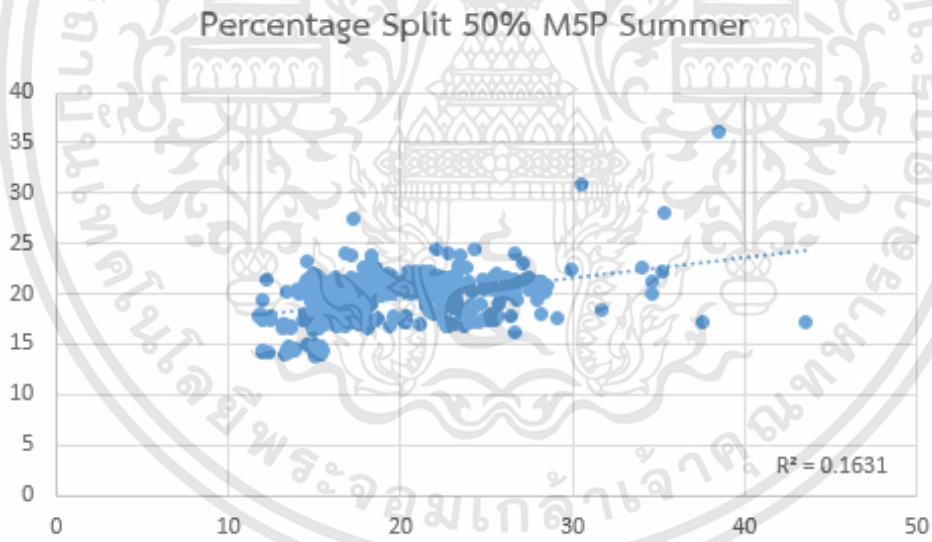
รูปที่ ผก. 16 แบบจำลองจากทฤษฎี M5P ชุดข้อมูล ฤดูร้อน ที่ Percentage split 90%



รูปที่ ผก. 17 แบบจำลองจากทฤษฎี M5P ชุดข้อมูล ฤดูร้อน ที่ Percentage split 75%

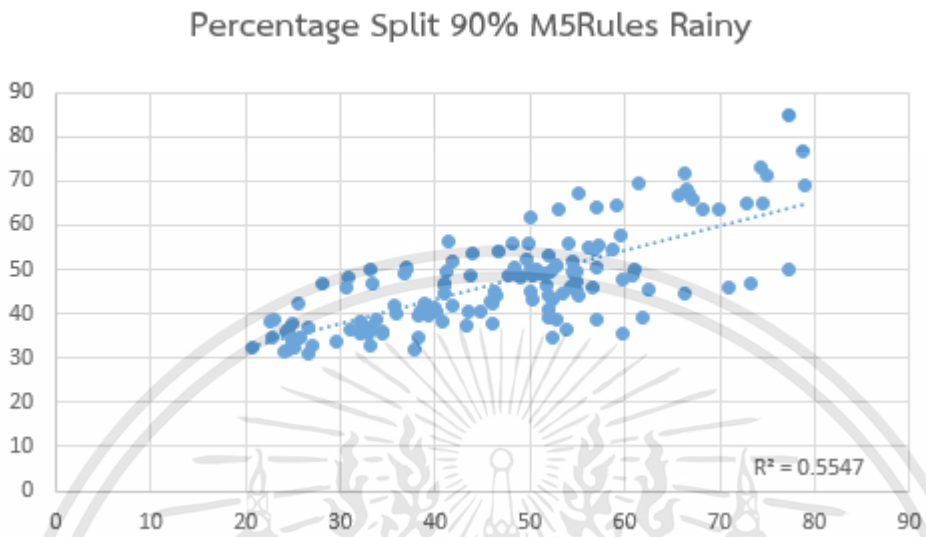


รูปที่ ผก. 18 แบบจำลองจากทฤษฎี M5P ชุดข้อมูล ฤดูร้อน ที่ Percentage split 50%

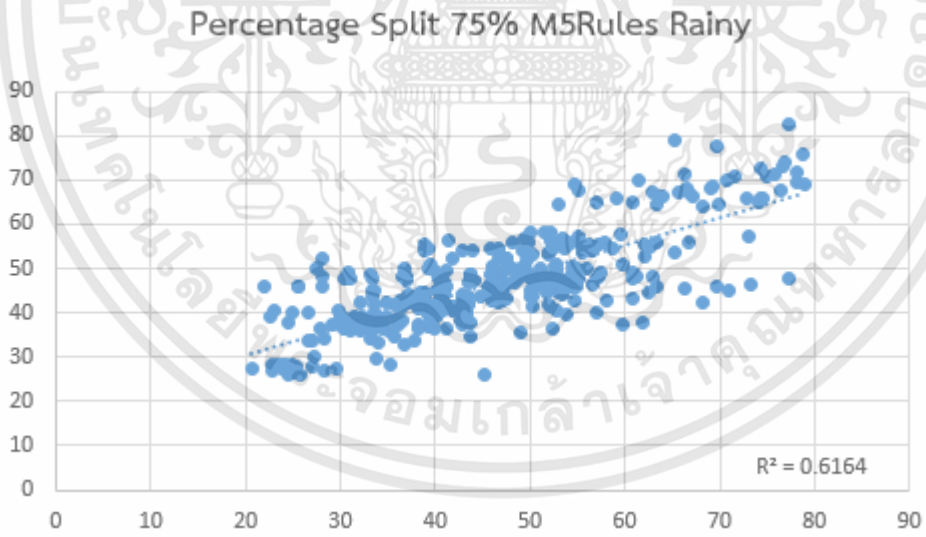


1.5 M5Rules

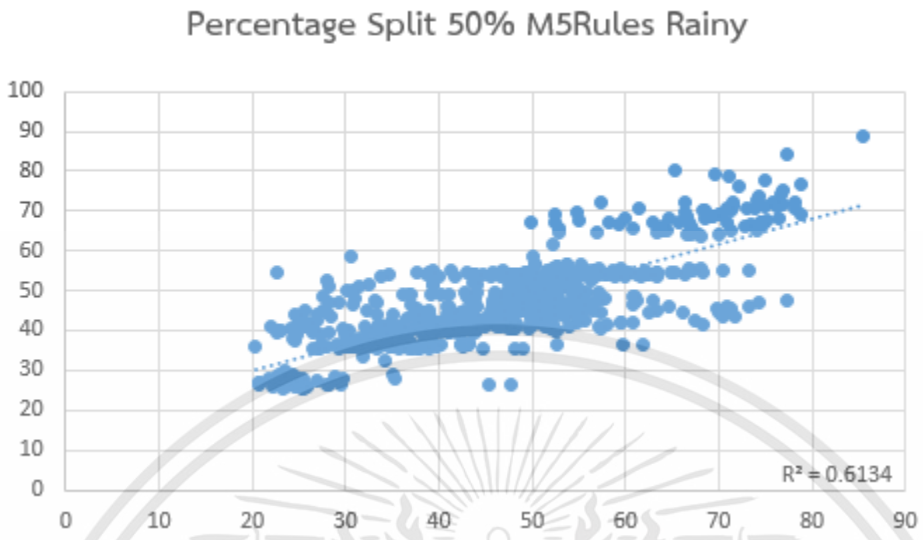
รูปที่ ผก. 19 แบบจำลองจากทฤษฎี M5Rules ชุดข้อมูล ฤดูฝน ที่ Percentage split 90%



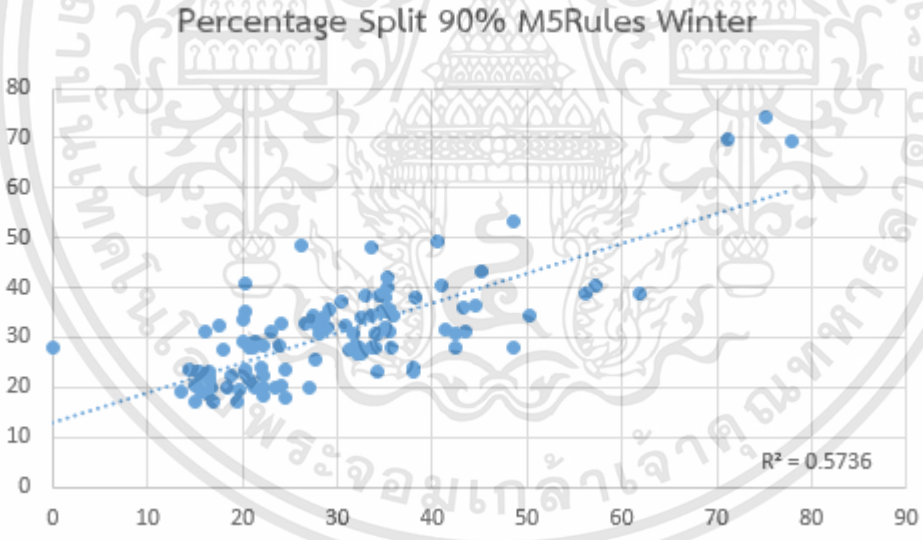
รูปที่ ผก. 20 แบบจำลองจากทฤษฎี M5Rules ชุดข้อมูล ฤดูฝน ที่ Percentage split 75%



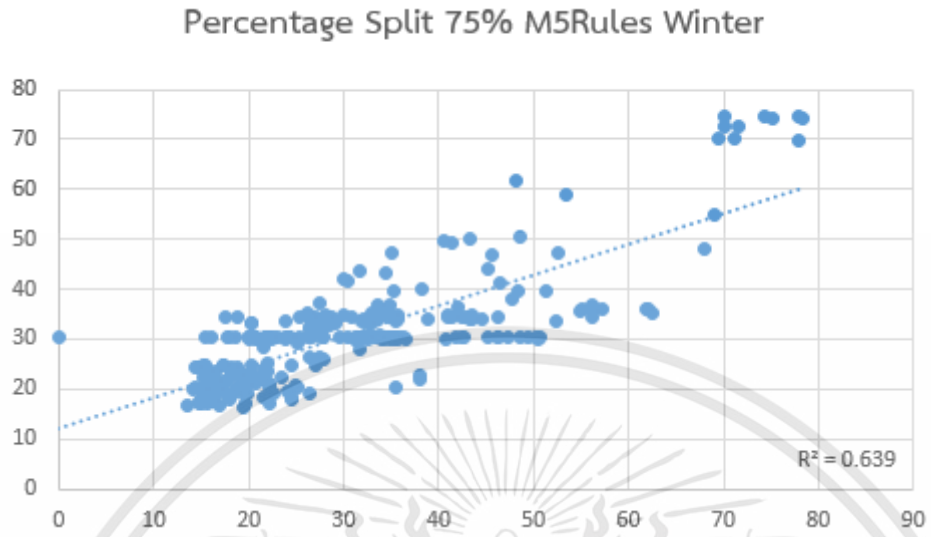
รูปที่ ผก. 21 แบบจำลองจากทฤษฎี M5Rules ชุดข้อมูล ฤดูฝน ที่ Percentage split 50%



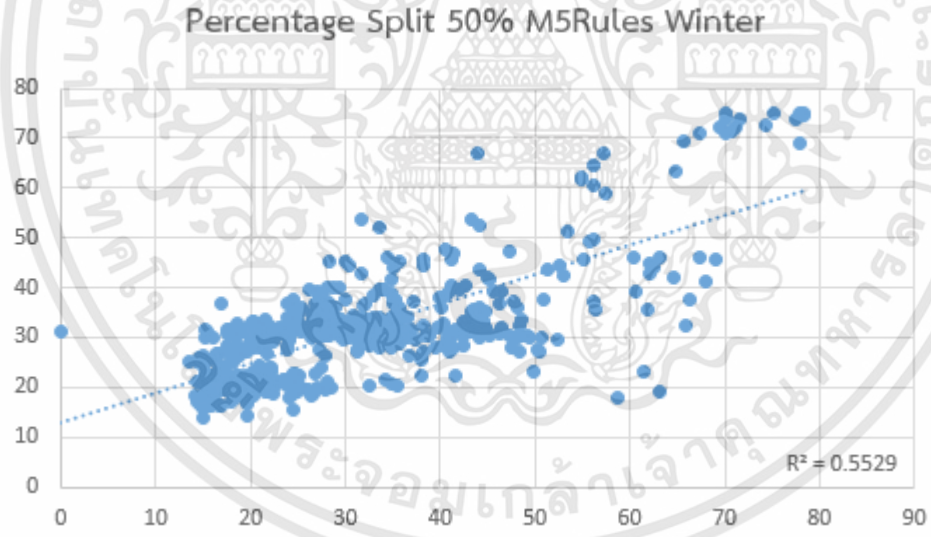
รูปที่ ผก. 22 แบบจำลองจากทฤษฎี M5Rules ชุดข้อมูล ฤดูหนาว ที่ Percentage split 90%



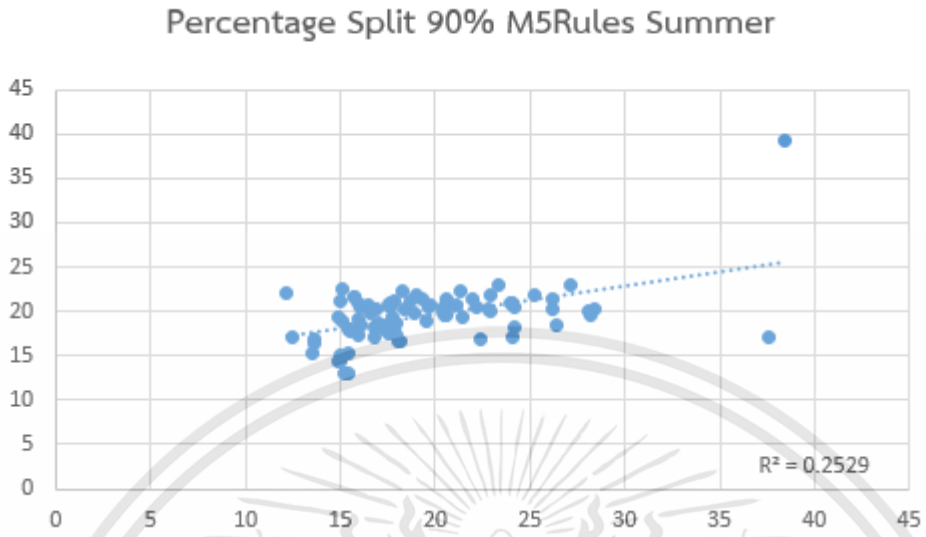
รูปที่ ผก. 23 แบบจำลองจากทฤษฎี M5Rules ชุดข้อมูล ฤดูหนาว ที่ Percentage split 75%



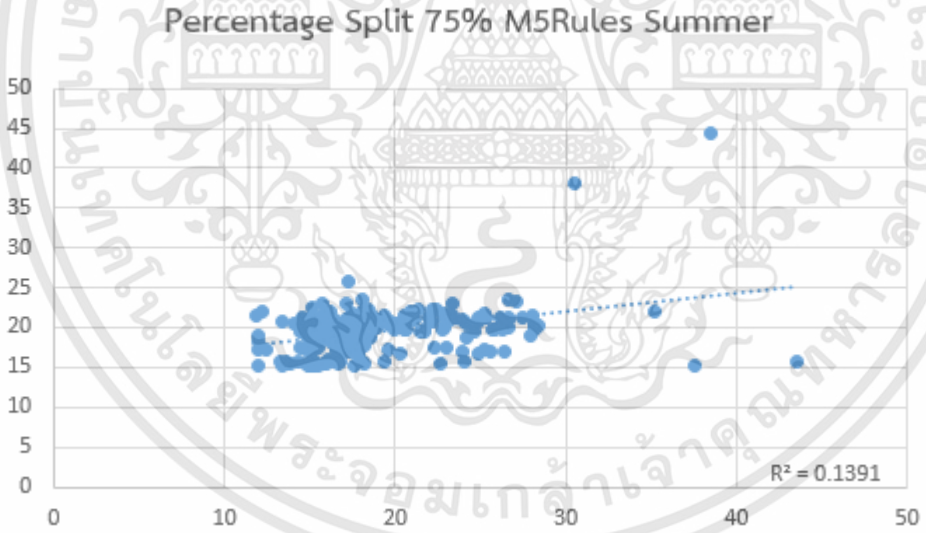
รูปที่ ผก. 24 แบบจำลองจากทฤษฎี M5Rules ชุดข้อมูล ฤดูหนาว ที่ Percentage split 50%



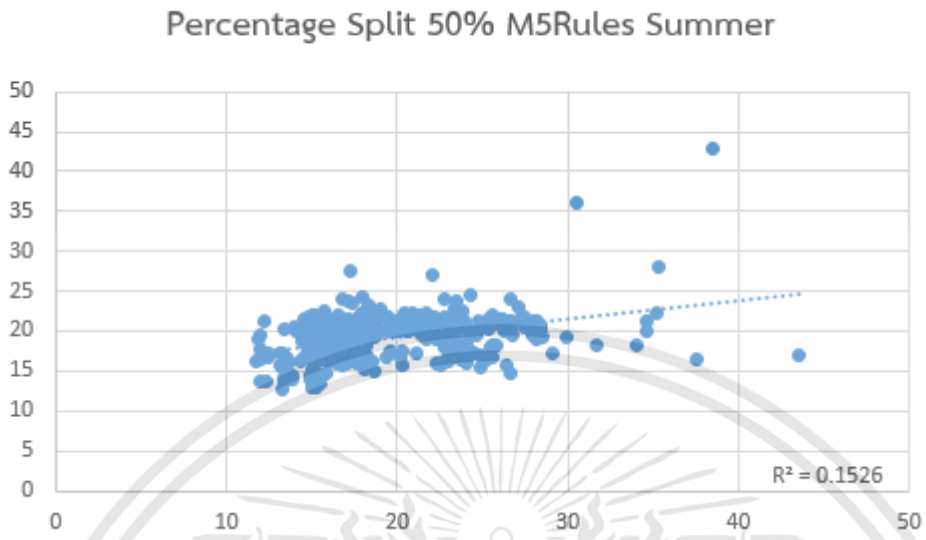
รูปที่ ผก. 25 แบบจำลองจากทฤษฎี M5Rules ชุดข้อมูล ฤดูร้อน ที่ Percentage split 90%



รูปที่ ผก. 26 แบบจำลองจากทฤษฎี M5Rules ชุดข้อมูล ฤดูร้อน ที่ Percentage split 75%

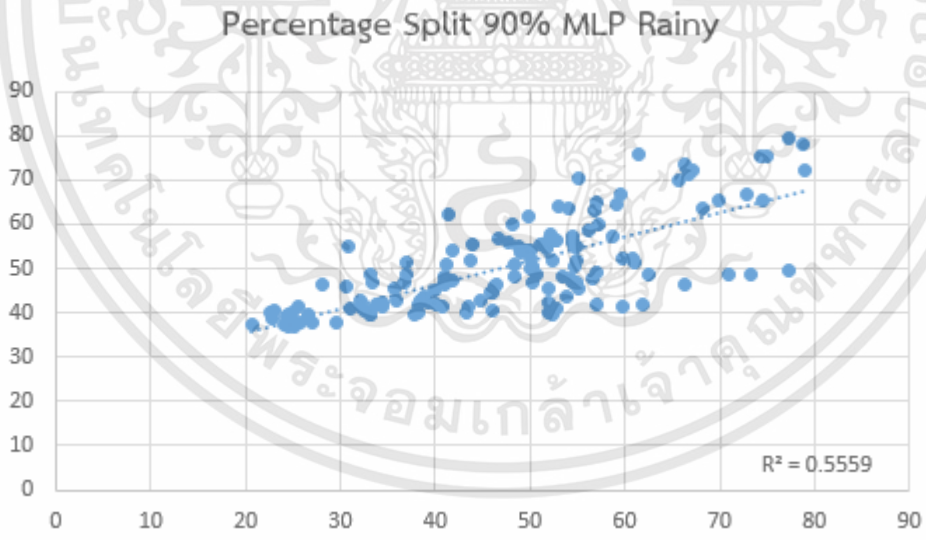


รูปที่ ผก. 27 แบบจำลองจากทฤษฎี M5Rules ชุดข้อมูล ฤดูร้อน ที่ Percentage split 50%

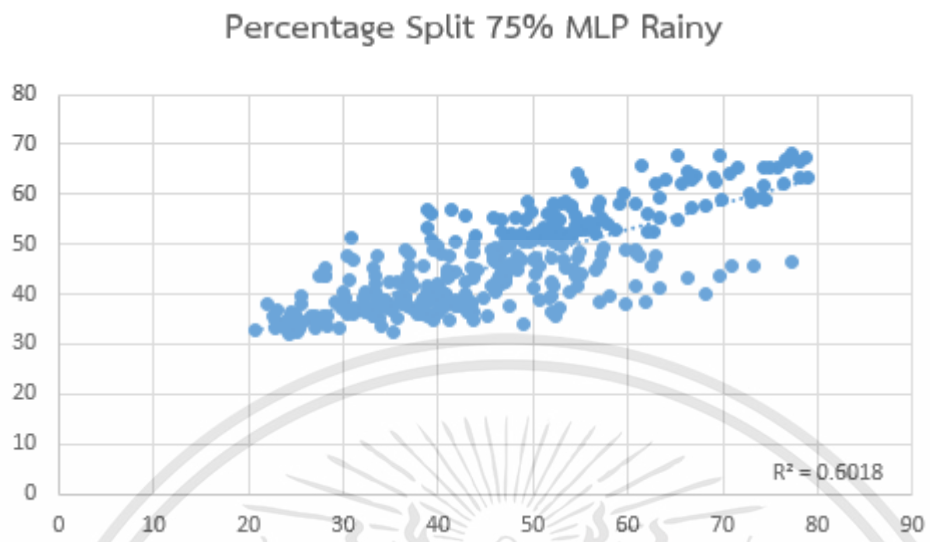


1.4 Multilayer Perceptron (MLP)

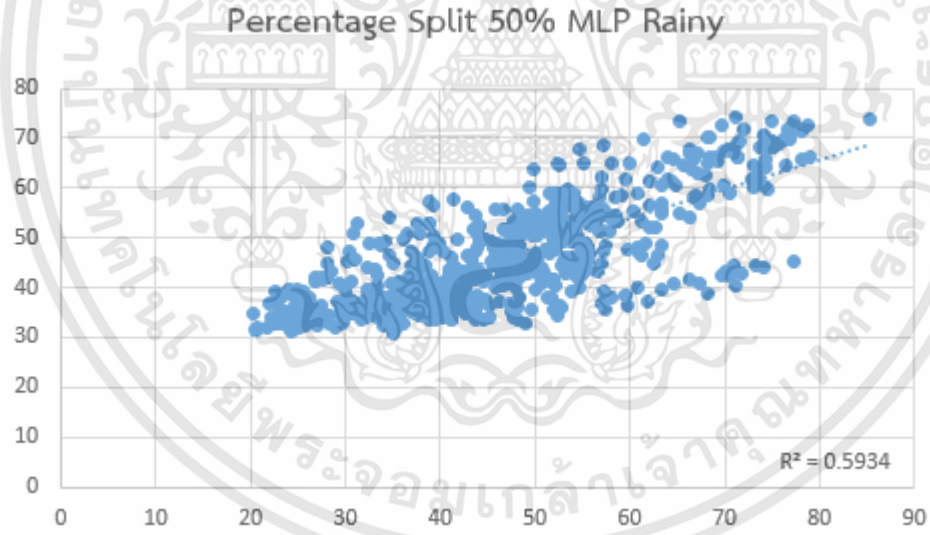
รูปที่ ผก. 28 แบบจำลองจากทฤษฎี MLP ชุดข้อมูล ฤดูฝน ที่ Percentage split 90%



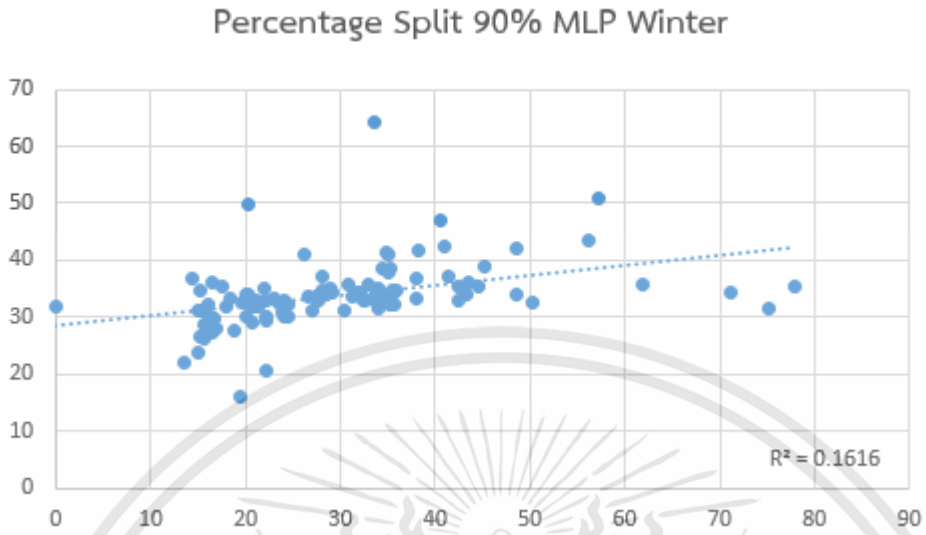
รูปที่ ผก. 29 แบบจำลองจากทฤษฎี MLP ชุดข้อมูล ฤดูฝน ที่ Percentage split 75%



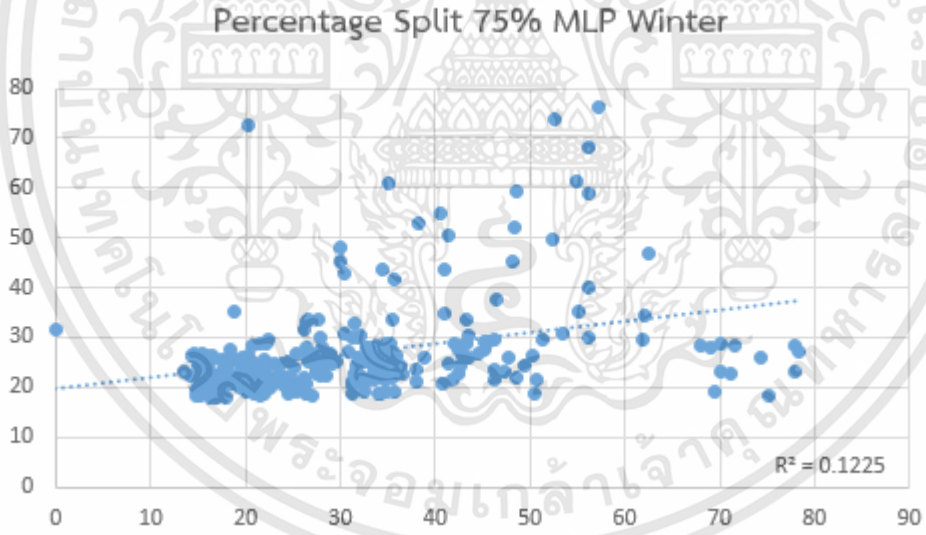
รูปที่ ผก. 30 แบบจำลองจากทฤษฎี MLP ชุดข้อมูล ฤดูฝน ที่ Percentage split 50%



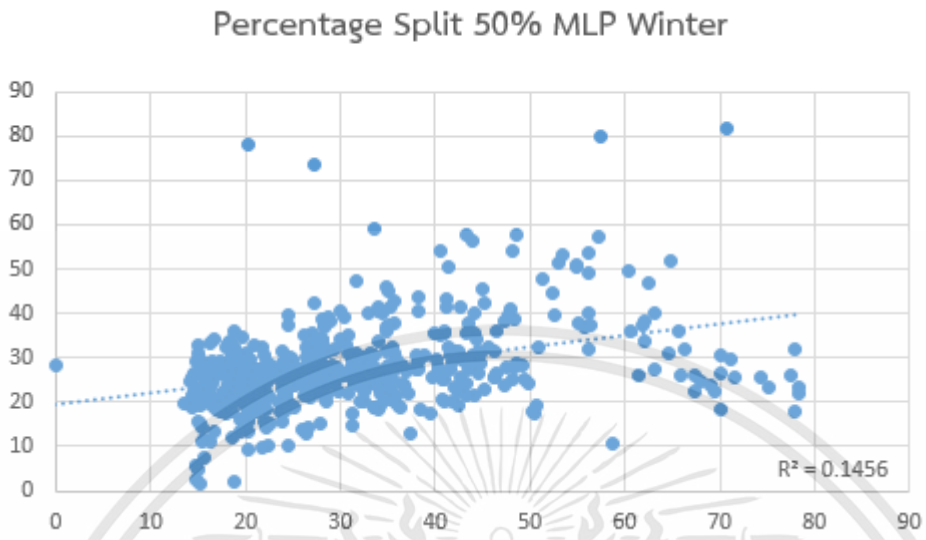
รูปที่ ผก. 31 แบบจำลองจากทฤษฎี MLP ชุดข้อมูล ฤดูหนาว ที่ Percentage split 90%



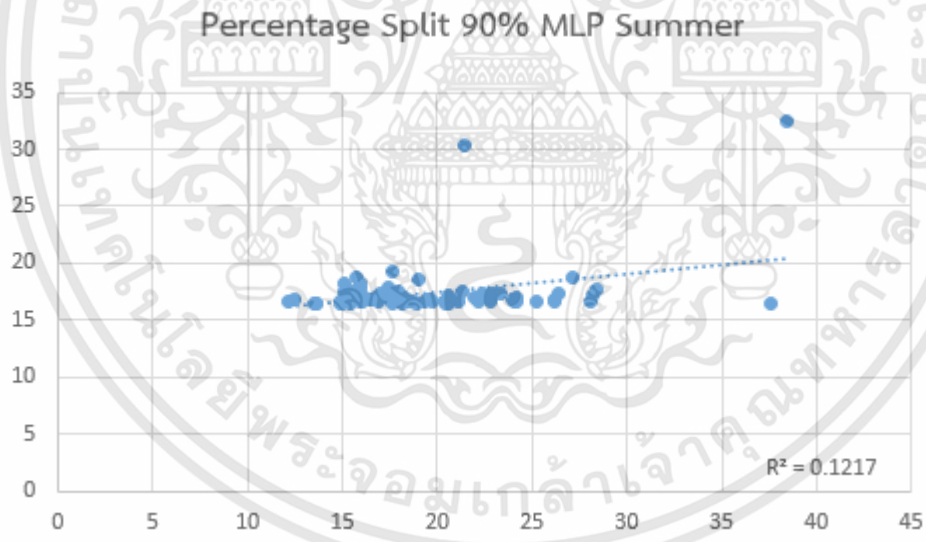
รูปที่ ผก. 32 แบบจำลองจากทฤษฎี MLP ชุดข้อมูล ฤดูหนาว ที่ Percentage split 75%



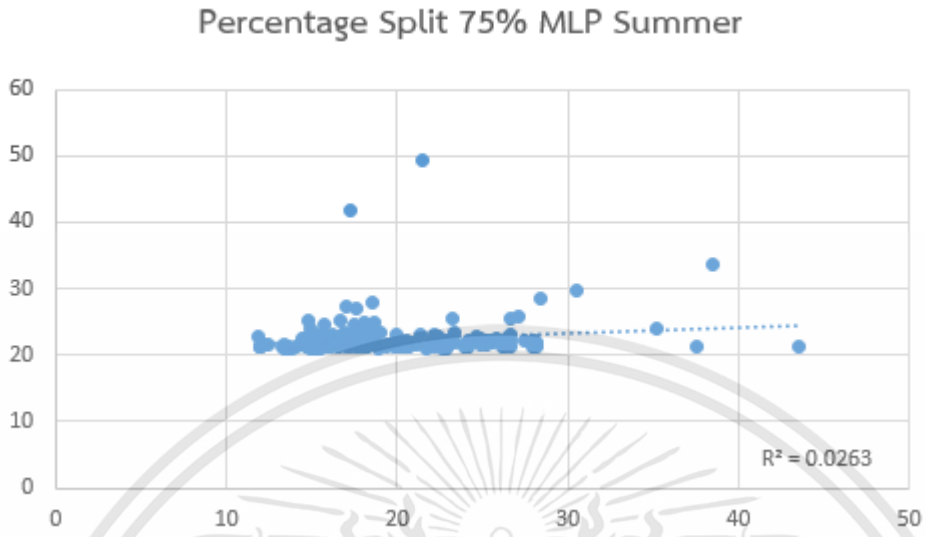
รูปที่ ผก. 33 แบบจำลองจากทฤษฎี MLP ชุดข้อมูล ฤดูหนาว ที่ Percentage split 50%



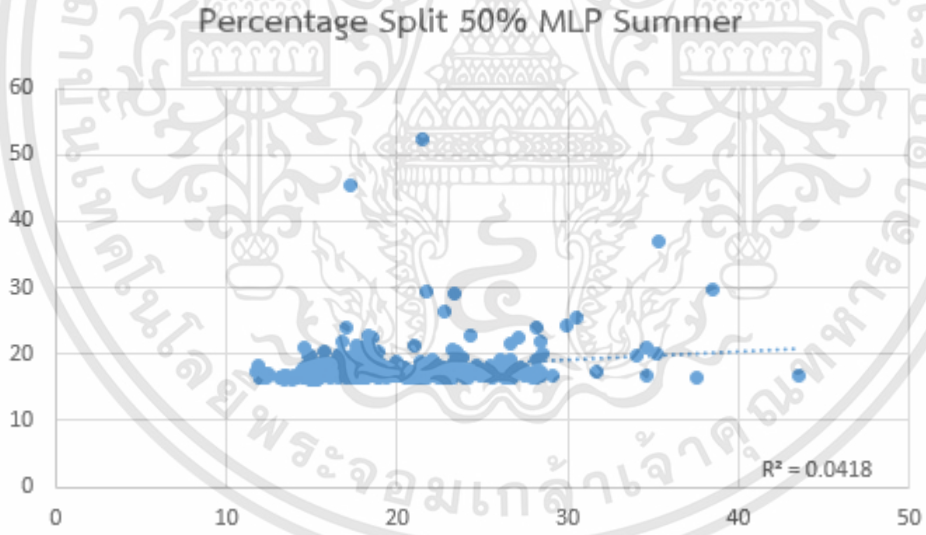
รูปที่ ผก. 34 แบบจำลองจากทฤษฎี MLP ชุดข้อมูล ฤดูร้อน ที่ Percentage split 90%



รูปที่ ผก. 35 แบบจำลองจากทฤษฎี MLP ชุดข้อมูล ฤดูร้อน ที่ Percentage split 75%



รูปที่ ผก. 36 แบบจำลองจากทฤษฎี MLP ชุดข้อมูล ฤดูร้อน ที่ Percentage split 75%



2. ค่าประสิทธิภาพของแบบจำลองที่มีผลดีที่สุดในแต่ละทฤษฎี

รูปที่ ผก. 37 การสังเคราะห์โดยวิธี REPTree ชุดข้อมูล ฤดูฝน ที่ Percentage Split 75%

Classifier output

```
|   |   |   |   |   |   pH0Raw < 7.47
|   |   |   |   |   |   |   pH0Raw < 7.44 : 71.11 (10/13.03) [1/0.03]
|   |   |   |   |   |   |   pH0Raw >= 7.44 : 68.07 (4/5.16) [1/0.95]
|   |   |   |   |   |   pH0Raw >= 7.47 : 74.08 (9/11.63) [10/16.91]
|   |   |   |   |   SS0Raw(mg/L) >= 135 : 76.53 (4/2.1) [1/7.84]
|   |   |   |   pH0Raw >= 7.53 : 69.66 (12/35.04) [7/66.57]
|   |   |   TurbAvgRaw(NTU) >= 245.75 : 80.99 (3/11.18) [0/0]

Size of the tree : 117

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.8036
Mean absolute error                      5.9679
Root mean squared error                  8.1006
Relative absolute error                 54.0895 %
Root relative squared error             59.6314 %
Total Number of Instances              359
```

รูปที่ ผก. 38 การสังเคราะห์โดยวิธี M5P ชุดข้อมูล ฤดูหนาว ที่ Percentage Split 75%

Classifier output

```
      + 4.3975 * SS0Raw(mg/L)=48.0,44.0,86.0,56.0,6.0,65.0,59.0,76.0,52.
      - 0.1766 * SS0Raw(mg/L)=44.0,86.0,56.0,6.0,65.0,59.0,76.0,52.0,9.0
      + 0.334 * SS0Raw(mg/L)=52.0,9.0,41.0,103.0,70.0,45.0,60.0,69.0,55.
      + 5.0257 * SS0Raw(mg/L)=60.0,69.0,55.0,67.0,68.0,53.0,84.0,78.0,58
      + 7.8538 * SS0Raw(mg/L)=66.0,142.0,83.0,50.0,4.0,91.0,127.0,88.0
      + 0.4646 * SS0Raw(mg/L)=83.0,50.0,4.0,91.0,127.0,88.0
      + 135.0768

Number of Rules : 14

Time taken to build model: 0.86 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.8028
Mean absolute error                      6.7557
Root mean squared error                  8.7679
Relative absolute error                 60.3033 %
Root relative squared error             60.8597 %
Total Number of Instances              274
```

รูปที่ ผก. 39 การสังเคราะห์โดยวิธี M5Rules ชุดข้อมูล ฤดูหนาว ที่ Percentage Split 75%

Classifier output

```
        + 44.222 [32/61.383%]

Rule: 8

Output-ALUM(ppm) =
        -55.8214 * pH0Raw
        + 469.2821 [9/76.993%]



Time taken to build model: 1.46 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.7994
Mean absolute error                      6.6157
Root mean squared error                  8.6624
Relative absolute error                 59.0538 %
Root relative squared error             60.1277 %
Total Number of Instances              274
```

รูปที่ ผก. 40 การสังเคราะห์โดยวิธี M5P ชุดข้อมูล ฤดูฝน ที่ Percentage Split 75%

Classifier output

```
    Inputs    Weights
    Threshold    0.9693530247235941
    Attrib TurbAvgRaw(NTU)    -5.833645072246802
    Attrib pH0Raw    14.460903355893484
    Attrib SS0Raw(mg/L)    5.869337537807114
Class
    Input
    Node 0


Time taken to build model: 0.33 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.7758
Mean absolute error                      6.7699
Root mean squared error                  8.699
Relative absolute error                 61.359  %
Root relative squared error             64.0361 %
Total Number of Instances              359
```

ภาคผนวก ข.

บทความงานวิจัยที่ได้รับตีพิมพ์เผยแพร่

Environmental Engineering Association of Thailand

18R6-02

Enhance Alum Dosage Prediction in Coagulation Process by Data Mining Software: A Case Study Bangkhen Water Supply Plant, Bangkok, Thailand

Aphisit Thipnang[1], Petchporn Chawakitchareon[2*] Paskorn Khanthongtip[3] and Yasushi Kiyoki[4]

Introduction

The purpose of this study is to find a correlation between parameters in a raw water and alum dosage, and use the correlation to predict alum dosage by data mining software. The data was collected from Bangkhen water supply plant, and the data would be linked by WEKA. Alum dosage has correlation with parameters of raw water, such as Turbidity, pH and Alkalinity. Suspended Solid and Turbidity removal increased substantially as the alum dosage is increased [1]. Alum work well in acidic waters and Alkalinity is defined as the acid absorbing property of water [2]. It shows that alum is related to Turbidity, Suspended Solid, pH and Alkalinity. In this relationship, we have the concept of modeling to make alum dosage predictions from other parameters by using data mining, It turns a large collection of data into knowledge for easily finding the relationship of the data.

Methodology

1. Data Collection

The data had been collected from Bangkhen water supply plant, Bangkok, Thailand from1 January 2006 to 31 July 2015. The Total number of collected data is 3,500 records. The data was collected from Jar-Test in each day. The histogram of attribute distribution in the dataset as shown in Figure 1.

Fig. 1 Histogram of attribute distribution in the data set.

[1]Graduate Student, [2]Assistant Professor, Department of Civil Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand.
E-mail: 59601229@kmitl.com
[2*]Associate Professor, Environmental Engineering Department, Faculty of Engineering,, Chulalongkorn University, Bangkok 10330, Thailand.
Phone : +662-2186674, Fax : 662-218-6666,
E-mail: petchporn.c@chula.ac.th
[3]Professor, Graduate School of Media and Governance, Keio University, Japan.

2. WEKA Data Mining Software

WEKA is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. it contains a collection of visualization tools and algorithms for data analysis and predictive modeling[4]. WEKA has many classification processes for finding the correlation. The usage depends on an amount of data of a type of data. In correlation between raw water and alum dosage, we used 4 classifier processes, Multilayer Perceptron (MLP), REPTree, M5P and M5Rules for building a model of correlation.

3. Multilayer Perceptron

The multilayer perceptron consists of a system of simple interconnected neurons, or nodes, as illustrated, which is a model representing a nonlinear mapping between an input vector and an output vector as shown in Figure 2.

Fig. 2 Multilayer Perceptron Process

4. M5P

M5P is the most commonly used classifier of decisions trees family. Structurally, a model tree takes the form of a decision tree with linear regression functions instead of terminal class values at its leaves. The M5 model tree is a numerical prediction algorithm and the nodes of the tree are chosen over the attribute.

5. M5Rules

M5rules algorithm produces propositional regression rules in IF-THEN rule format using routines for generating a decision list from M5 Model trees [3].

6. REPTree

REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning.

7. Enhancement the model

The data was prepared to 3 groups for finding the best correlation because the values of the parameters in each group are quite different. Group 1 is summer

season (858 records), Group 2 is winter season (1210 records) and Group 3 is rainy season (1432 records). We choose the three variables (Turbidity, pH and SS) that are directly related to the correlation, rather than using all parameters (Turbidity, pH, SS, Alkalinity, Color and conductivity) same as a previous research [4].

Result and Discussion

1. **Model building to 3 Groups**
   a. The first model group was built by using the first data group (summer season)
   b. The second model group was built by using the second data group (winter season)
   c. The Third model group was built by using the third data group (rainy season)

2. **Comparing Model Group**

For the model comparison, the best method of each model was chosen to compare the model precision and credibility of the model by RMSE and MAE. The results of this model as shown in Table 1.

Table 1. The RMSE and MAE value of the best method of 3 models.

Classifier	RMSE	MAE
Summer season: M5Rules	3.8567	2.7622
Winter season: REPTree	7.8961	5.4059
Rainy season: M5Rules	14.2928	10.7433

we found that the M5Rules method of model group a (summer season) gave the less RMSE and MAE value than another model.

3. **Comparing the model precision**

For the model credibility comparison, we compared the models of this research with previous research and we obtained the RMSE and MAE value to decide the model accuracy. We compared with

a. 2 groups of models (Dry season and Rainy season) from the previous research [5].

b. 1 group from the previous research [4].

A comparison between this research and 2 previous researches which use the same 4 methods in WEKA data mining software. The results indicated that the M5Rules method of group a of this research [5] gives the less RMSE and MAE than other method of all group of this research. And M5P method of rainy season of previous research [7] gives less RMSE and MAE than other method of all group. But if compare with the same place of plant and the same data, this research gives less RMSE and MAE (RMSE = 3.8567, MAE = 2.762 in M5Rules Method) than previous research. [4]

Applications

In the prediction by Weka, we found that the prediction of alum dosage from M5Rules of model group 1 is nearby the actual alum dosage than another method. The results are shown in Figure 3.

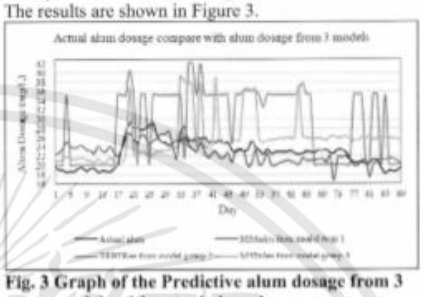


Fig. 3 Graph of the Predictive alum dosage from 3 models with actual alum dosage

In the prediction by Weka, we found that the prediction of alum dosage from M5Rules is nearby the actual alum dosage than another method.

Conclusion

This research found that each parameters of raw water Significantly related and it's relationship can be utilized such as building the model for prediction of alum dosage. This research has tried to predict and the result is satisfactory. The precision of the model depends on the accuracy of the data being collected each day. If the data is wrong or swing It may make the model distorted.

Acknowledgement

This research is supported by Chulalongkorn University and Keio University, Japan.

References

[1] Al-Mutairi, N. Z., Hamoda, M. F., and Al-Ghusain, I. (2004). Coagulant selection and sludge conditioning in a slaughterhouse wastewater treatment plant. *Bioresource technology*, *95*(2), 115-119.

[2] McDonald, J. (2006). Alkalinity and pH relationship. *CSTN. May*.

[3] Holmes, G., Hall, M., and Prank, E. (1999, December). Generating rule sets from model trees. In *Australasian Joint Conference on Artificial Intelligence* (pp. 1-12). Springer Berlin Heidelberg.

[4] Chawakitcharoen, P., Boonmee, N., and Charutragulchai, P. (2017). Prediction of Alum Dosage in Water Supply by WEKA Data Mining Software. *Information Modelling and Knowledge Bases XXVIII*, *292*, 83.

[5] Ladsavong, K., Chawakitcharoen, P. and Kiyoki, Y. 2017. Prediction of Alum Dosage in Coagulation Process using Weka program: A Case Study at Chinaimo Water Treatment Plant (CWTP) in Vientiane Capital, Lao PDR. The 9th AUN/SEED-Net Regional Conference on Environmental Engineering (pp.522-532). The Zign Hotel, Chonburi, Thailand, January 23-24, 2017. PID0059.

CERTIFICATE

This is to certify that

Mr. Aphisit Thipnang

Presented the paper entitled

"Enhance Alum Dosage Prediction in Coagulation Process by Data Mining Software: A Case Study Bangkhen Water Supply Plant, Bangkok, Thailand"

at

The 6th International Conference on Environmental Engineering, Science and Management

May 17-18, 2017 at The Twin Towers Hotel Bangkok, Thailand

May 18, 2017

Prasert Tapaneeyangkul, PhD

President of Environmental Engineering Association of Thailand

ประวัติผู้เขียน

ชื่อ-นามสกุล	นายอภิสิทธิ์ ทิพย์นาง
วัน เดือน ปีเกิด	24 ตุลาคม 2536
ที่อยู่	1/5 ถนนนครสวรรค์ ตำบลตลาด อำเภอเมือง จังหวัดมหาสารคาม 44000

ประวัติการศึกษา

พ.ศ. 2559	ศึกษาต่อระดับปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมสิ่งแวดล้อมและพลังงานเพื่อความยั่งยืน สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2555 - 2559	สำเร็จการศึกษาวิศวกรรมศาสตร์บัณฑิต สาขาวิศวกรรมพลังงานไฟฟ้า สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง