การพยากรณ์ราคาด้วยการเรียนรู้ของเครื่องจักรด้วยโปรแกรมเวก้า

ชุษณา เทียนทอง[1*] ศุภกร เจริญประสิทธิ์[2] และสุธีธิดา นภาอรุณชัย[3]
คณะพัฒนาธุรกิจและอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ[1*,2,3]
อีเมล : choosana.t@bid.kmutnb.ac.th[1*]

* **วันที่รับบทความ** 3 สิงหาคม 2563 **วันที่แก้ไขบทความ** 21 กันยายน 2563 **วันที่ตอบรับบทความ** 11 มกราคม 2564

บทคัดย่อ

งานวิจัยนี้เป็นการศึกษาเครื่องมือสำหรับการพยากรณ์ราคาด้วยการเรียนรู้ของเครื่องจักรด้วยโปรแกรมเวก้า เวอร์ชัน 3.8 มีวัตถุประสงค์เพื่อ 1) สร้างแบบจำลองการพยากรณ์ราคาโดยเทคนิคการเรียนรู้ของเครื่องจักร 2) สร้างแบบจำลองที่มีประสิทธิภาพในการพยากรณ์ราคาผักและผลไม้ ซึ่งสามารถนำไปประยุกต์ใช้ประกอบการพิจารณายอดคำสั่งซื้อและช่วยให้เกิดการจัดการข้อมูลราคาอย่างเป็นระบบ พร้อมทั้งรับมือการเปลี่ยนแปลงราคาในอนาคตที่อาจเกิดขึ้นได้ จากการศึกษาด้วยการเปรียบเทียบความคลาดเคลื่อนของข้อมูลก่อนและหลังปรับค่าพารามิเตอร์ด้วยอัลกอริทึม 3 รูปแบบ ประกอบไปด้วย การแจกแจงแบบเกาส์เซียน การถดถอยเชิงเส้น และเทคนิค SMOreg โดยโปรแกรมเวก้า เวอร์ชัน 3.8 จากผลการศึกษาพบว่า ในการทดลองนำข้อมูลราคาผักและผลไม้ย้อนหลัง 6 ปี ไปสร้างเป็นรูปแบบจำลองเพื่อพยากรณ์ราคาในอนาคตนั้น ก่อนปรับค่าพารามิเตอร์ อัลกอริทึม SMOreg ได้ให้ค่าความคลาดเคลื่อนเฉลี่ยรวมน้อยที่สุดจากทั้ง 3 รูปแบบ ด้านการวัดประสิทธิภาพด้วยค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ จากค่าจริงเปรียบเทียบกับค่าพยากรณ์ ได้เท่ากับ 2.206 และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย เท่ากับ 3.594 โดยการพยากรณ์แสดงให้เห็นว่า แบบจำลองมีแนวโน้มสามารถช่วยลดค่าใช้จ่ายลงได้ประมาณ 50,000 บาทต่อเดือน

คำสำคัญ : การเรียนรู้ของเครื่องจักร แบบจำลองการพยากรณ์ ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย โปรแกรมเวก้า

Price Forecasting Study by Machine Learning Using Weka Program

Choosana Tiantong[1*], Supakorn Charoenprasi[2] and Suteeetida Napaarunchai[3]
Department of Manufacturing and Service Industry Management, Faculty of Business and Industrial Development, King Mongkut's University of Technology North Bangkok[1*,2,3]
E-mail: choosana.t@bid.kmutnb.ac.th[1*]

* **Received:** August 3, 2020 **Revised:** September 21, 2020 **Accepted:** January 11, 2021

Abstract

This research aimed to study machine learning models for price forecasting, using the Weka program (version 3.8). The objectives were to 1) create price forecasting models utilizing machine learning techniques and 2) create models that are effective in forecasting prices of fruits and vegetables, in order to support the systemic planning of future orders. Future price changes would be calculated by comparing the (Mean Absolute Error: MAE) values before and after adjusting parameters for 3 algorithms - Gaussian Distributor, Linear Regression and SMOreg technique, using the Weka program (version 3.8). After using the data of fruits and vegetables goes back for 6 years to create models to predict future prices, the results revealed that SMOreg algorithm gave the smallest total mean absolute error, before adjusting any parameters. Measuring performance with the mean absolute error (actual value versus forecast) gave a result of 2.206 and the Root Mean Square Error was 3.594. It was found that the model could be saved costs of 50,000 Baht per month.

Keywords: machine learning, forecasting model, mean absolute error, weka program

1. บทนำ

ประเทศไทยนับเป็นจุดหมายปลายทางยอดนิยมแห่งหนึ่งของนักท่องเที่ยวต่างชาติส่วนหนึ่ง เป็นผลจากการมีแหล่งท่องเที่ยวที่ดึงดูดความสนใจของนักท่องเที่ยวติดอันดับโลกกระจายอยู่ตามภาคต่าง ๆ ทั่วประเทศ โดยเฉพาะแหล่งท่องเที่ยวทางทะเลภาคใต้ ภาคตะวันออก หรือแหล่งท่องเที่ยวเชิงนิเวศน์ เป็นต้น โรงแรมกรณีศึกษาเป็นโรงแรมที่ตั้งอยู่ใจกลางเมืองกรุงเทพมหานคร และเป็นโรงแรมที่ให้บริการหลากหลายประเภท ทั้งในส่วนของห้องพัก ห้องจัดเลี้ยง ห้องสัมมนา ห้องอาหาร สปา และกีฬา ด้วยการให้บริการที่หลากหลายนี้เอง จึงมีความจำเป็นที่จะต้องสั่งซื้อวัสดุและวัตถุดิบในแต่ละวันเป็นจำนวนมาก โดยเฉพาะผักและผลไม้ที่เป็นวัตถุดิบหลักที่ใช้ในการประกอบอาหาร มีการสั่งซื้อผลไม้เป็นจำนวนหลายร้อยกิโลกรัมต่อเดือน ส่งผลให้มีค่าใช้จ่ายต่อเดือนสูงประมาณเดือนละ 500,000 บาท

ในปัจจุบัน การสั่งซื้อเป็นจำนวนมากเช่นนี้ อาจจะมีปัญหาตามมาได้ เช่น ราคาสินค้าบางประเภทปรับสูงขึ้นตามสภาวะเศรษฐกิจอย่างกะทันหัน หรือสินค้าเกษตรบางประเภทขาดตลาดเนื่องจากภัยพิบัติทางธรรมชาติ ประกอบกับทางครัวและผู้ที่เกี่ยวข้องในการจัดเตรียมวัตถุดิบขาดการเตรียมความพร้อมในการวางแผนใช้วัตถุดิบภายใต้สถานการณ์ดังกล่าว

ด้วยเหตุนี้เอง จึงเกิดแนวความคิดในการประยุกต์ใช้การเรียนรู้ของเครื่องจักร (machine learning) ด้วยโปรแกรมเวก้า (WEKA) ซึ่งเป็นโปรแกรมสำหรับการสร้างเหมืองข้อมูลในการพยากรณ์ราคาผักและผลไม้ที่มียอดคำสั่งซื้อสูงเป็น 5 ลำดับแรก โดยแบ่งเป็นผลไม้ 5 ชนิด และผัก 5 ชนิด ประกอบด้วยชุดข้อมูลฮันนี่ดิว แคนตาลูป มะม่วงน้ำดอกไม้ ส้มสายน้ำผึ้ง แก้วมังกร เห็ดฟาง คื่นช่าย ซูชินี มะเขือเทศราชินี และกรีนโอ๊ค เพื่อคาดการณ์แนวโน้มราคาที่ผันผวนและการวางแผนการบริหารต้นทุนสำหรับอนาคต

2. วัตถุประสงค์

2.1 เพื่อศึกษาการสร้างแบบจำลองการพยากรณ์ราคาโดยเทคนิคการเรียนรู้ของเครื่องจักร

2.2 เพื่อสร้างแบบจำลองที่มีประสิทธิภาพในการพยากรณ์ราคาผักและผลไม้

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การทำเหมืองข้อมูลเป็นกระบวนการที่คัดเลือกข้อมูลที่มีประโยชน์จากฐานข้อมูลหรือแหล่งข้อมูลที่มีขนาดใหญ่เพื่อให้ได้ข้อมูลที่มีเหตุผลและช่วยในการตัดสินใจอีกทั้งยังช่วยในการทำนายสิ่งที่อาจเกิดขึ้นในอนาคตโดยลักษณะการทำเหมืองข้อมูลคือ นำเอาข้อมูลที่เก็บในอดีตมาสร้างแบบจำลอง (model) โดยการสร้างแบบจำลองประกอบด้วย 4 ขั้นตอน คือ

3.1 การทำความเข้าใจข้อมูล

โดยการเก็บรวบรวมข้อมูลทำความคุ้นเคยกับข้อมูล ระบุปัญหาคุณภาพของข้อมูลแบบเชิงลึก ตั้งสมมติฐานเกี่ยวกับข้อมูลในมุมมองต่าง ๆ

3.2 การเตรียมข้อมูล

โดยการสร้างชุดข้อมูลการคัดเลือกตัวแปรอิสระที่พบว่าบางตัวแปรอาจจะไม่มีความสำคัญ และไม่มีความสัมพันธ์กับตัวแปรตาม จึงต้องทำการคัดเลือกตัวแปรอิสระเพื่อให้ได้ตัวแบบที่เหมาะสมที่ประกอบด้วยตัวแปรอิสระที่สำคัญและมีความสัมพันธ์กับตัวแปรตามเท่านั้นมาทำการเตรียมข้อมูล

3.3 การสร้างแบบจำลอง

การเตรียมข้อมูลมาทำการประเมินประสิทธิภาพของแบบจำลองแต่ละรูปแบบด้วยเทคนิคการจำแนกข้อมูล แล้วเลือกรูปแบบที่ให้ประสิทธิภาพสูงสุดในการสร้างแบบจำลอง

3.4 การประเมินผล

ประเมินผลลัพธ์ของแบบจำลองที่ได้จากการสร้างโดยดูจากค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE) ที่ดีที่สุดในการพยากรณ์แล้วนำไปสร้างตัวแบบการพยากรณ์ โดยการวิเคราะห์เพื่อหาข้อมูลชุดการเรียนรู้ (training data set) แล้วทำการเปรียบเทียบค่าพยากรณ์ที่ได้กับค่าจริงที่เกิดขึ้นด้วยการวัดประสิทธิภาพด้วยเปอร์เซ็นต์ค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์

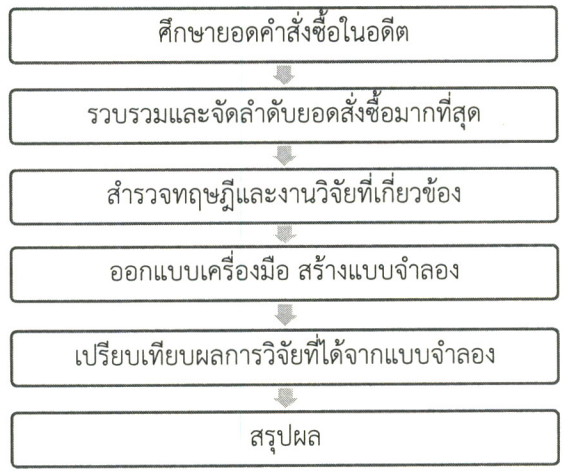
จากการศึกษางานวิจัยที่ใกล้เคียง พบว่า ภูริท [1] ได้สร้างอัลกอริทึมแบบรวมสำหรับการคัดเลือกคุณสมบัติของข้อมูลว่าวิธีการคัดเลือกแบบใด ได้สร้างรูปแบบการทำนายผลได้ดีที่สุด พบว่าการปรับค่าพารามิเตอร์เป็นปัจจัยหนึ่งมีผลทำให้ประสิทธิภาพในการจัดลำดับคุณสมบัติของข้อมูลที่ผู้วิจัยนำมาทดสอบดีขึ้น ชัชชญา [2] การพยากรณ์ความต้องการใช้ปูนซีเมนต์ของประเทศไทย ทำการเปรียบเทียบตัวแบบพหุคูณและตัวแบบอนุกรมเวลาตัวแบบพหุคูณจากการรวมผลพยากรณ์เป็นตัวแบบที่เหมาะสมที่สุดเนื่องจากให้ค่า RMSE ต่ำที่สุดที่ร้อยละ 3.58 สุภาพร [3] พยากรณ์การใช้ลูกกอล์ฟในสนามฝึกกอล์ฟโดยใช้ตัวแปรอนุกรมเวลา ผลการวิจัยสามารถสรุปได้ว่าในรอบ 7 วันสามารถสร้างแบบจำลองในการพยากรณ์การจัดเตรียมจำนวนลูกกอล์ฟที่ใช้ในวันถัดไปได้มีประสิทธิภาพที่สุดด้วยเทคนิคการถดถอยเชิงเส้น มีค่า MAE เท่ากับ 34.320 และพบว่าเทคนิคที่สามารถสร้างแบบจำลองการพยากรณ์การจัดเตรียมจำนวนลูกกอล์ฟที่ใช้ต่อแพ็กเกจได้มีประสิทธิภาพที่สุด คือ เทคนิคการถดถอยเชิงเส้น ที่มีค่า MAE เท่ากับ 8.135, 1.507 5.417 และ 6.026 ของคูปองชุดในรอบ 7 วัน คูปองเล่มในรอบ 21 วัน ลูกชุดในรอบ 21 วันและลูกถาดในรอบ 7 วัน ตามลำดับ วิทิดา [4] การวิจัยเรื่องการพยากรณ์กระแสไฟฟ้าขัดข้องโดยใช้เทคนิคการทำเหมืองข้อมูลรวบรวมข้อมูลจากเหตุการณ์ของการไฟฟ้าหาลักษณะเด่นของแต่ละกลุ่มตัวอย่าง โดยทำการประมวลผลผ่านโปรแกรม WEKA พบว่าสามารถช่วยพยากรณ์การเกิดกระแสไฟฟ้าขัดข้องได้อย่างแม่นยำ เสกสรร และคณะ [5] สร้างตัวแบบพยากรณ์ผลการเรียนของนักเรียนโรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์วิทยาเขตกำแพงแสนศูนย์วิจัยและพัฒนาการศึกษาโดยใช้ข้อมูลนักเรียนระดับมัธยมศึกษาระหว่างปีการศึกษา 2548 - 2556 เพื่อพัฒนาคลังข้อมูลโดยใช้โครงสร้างแบบสโนว์เฟลกสกีมามาสร้างตัวแบบพยากรณ์ผลการเรียนโดยใช้ชุดข้อมูล 2 แบบ คือ ข้อมูลแบบไม่จัดกลุ่มและข้อมูลแบบจัดกลุ่ม จากนั้นนำไปผ่านกระบวนการคัดเลือกคุณลักษณะซึ่งใช้วิธีเทคนิคในการเลือกคุณสมบัติของ คุณลักษณะโดยใช้การพิจารณาบนพื้นฐานความสัมพันธ์ (Correlation-based Feature Selection: CFS) และวิธีเทคนิคการหาผลต่าง

ของสถานะข้อมูล (Information Gain: IG) แล้วใช้เทคนิคเหมืองข้อมูลแบบ การแจกแจงแบบเกาส์เซียน วราพร [6] การพยากรณ์ อนุกรมเวลาโดยการเปรียบเทียบวิธีแบบฉบับ และวิธีบอกซ์-เจนกินส์ กรณี ศึกษาจำนวนการเกิดอุบัติเหตุในประเทศไทย พบว่าค่า RMSE ที่ได้จากการพยากรณ์โดยวิธีแบบฉบับ 520,707.69 ครั้ง ส่วนค่า RMSE ที่ได้จากการพยากรณ์โดยวิธีบอกซ์–เจนกินส์ 2,662.514 ครั้ง โดยค่า ประมาณที่ดีที่สุดของพารามิเตอร์ ซึ่งค่าประมาณที่ดีที่สุดหมายถึงค่า ที่ทำให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย มีค่าน้อยที่สุด ซึ่งพบว่าการพยากรณ์โดยวิธีบอกซ์–เจนกินส์จะให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำกว่าการพยากรณ์โดยวิธีแบบฉบับ เนื่องจากข้อมูลอนุกรมเวลาชุดนี้เป็นข้อมูลที่มีฤดูกาล ซึ่งการพยากรณ์โดยวิธีแบบฉบับจะสมมติว่าค่าแนวโน้มเป็นเส้นตรง ซึ่งแท้ที่จริงแล้วอาจมีสมการเป็นอย่างอื่น จึงทำให้มีค่า RMSE สูง กว่าการพยากรณ์โดยวิธีบอกซ์–เจนกินส์ พันทิพา [7] ได้ศึกษาการพยากรณ์การใช้พลังงานรวมทั้งหมดของประเทศไทยโดยวิธีของบอกซ์เจนกินส์และโครงข่ายประสาทเทียมโดยใช้ข้อมูลจากสำนักงานนโยบายและแผนพลังงานกระทรวงพลังงานจำแนกเป็นรายเดือน พบว่าวิธีที่เหมาะสม คือ โครงข่ายประสาทเทียมตัวแบบ 4:3:1 โดยมีตัวชี้วัดทางเศรษฐกิจที่ใช้พยากรณ์ คือ ผลิตภัณฑ์มวลรวมภายในประเทศมูลค่าการส่งออกสินค้ามูลค่าการนำเข้าสินค้าจำนวนประชากรซึ่งมีความแม่นยำในการพยากรณ์สูงกว่าวิธีของบอกซ์–เจนกินส์ นิสานันท์ [8] พัฒนาแบบจำลองสำหรับการขายผลิตภัณฑ์ประกันชีวิตใช้เทคนิคการทำเหมืองข้อมูล (data mining) ตามกรอบ (Cross Industry Standard Process for Data-Mining: CRISP-DM) โดยการสร้างแบบจำลองการแบ่งกลุ่ม (clustering) ด้วยวิธี Simple K-Means เพื่อใช้ในการจัดกลุ่มลูกค้าที่ซื้อกรมธรรม์ประกันชีวิตนำไปหาความสัมพันธ์ (association rule) โดยใช้อัลกอริทึม apriori ซึ่งผลที่ได้จากการทำเหมืองข้อมูลสามารถนำไปออกแบบกิจกรรมส่งเสริมการตลาดเพื่อกระตุ้นยอดขายให้แก่บริษัทได้ นิฉา [9] การพยากรณ์ปริมาณการใช้ไฟฟ้าของประเทศไทย โดยใช้ตัวแบบ SARIMA และตัวแบบการถดถอยที่มีความคลาดเคลื่อนเป็นตัวแบบ ARMA สรุปได้ว่าตัวแบบการถดถอยที่มีความคลาดเคลื่อนเป็นตัวแบบ ARMA เป็นวิธีที่เหมาะสมที่สุดโดยมีค่าเฉลี่ยของร้อยละความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Percentage Error: MAPE) ต่ำที่สุดเท่ากับ 1.7898 % และมีค่าสัมประสิทธิ์การตัดสินใจมากที่สุดเท่ากับ 98.35 % สุธีรา ต้องใจ และอรวรรณ [10] การใช้เทคนิคเหมืองข้อมูลในการประเมินความรู้ และหาความถนัด เพื่อพัฒนาศักยภาพของนักศึกษา วิจัยนี้เป็นประโยชน์สำหรับอาจารย์ ในการดูแลนักศึกษาช่วยลดความเสี่ยงของนักศึกษาที่จะมีผลการเรียนที่ไม่ผ่านเกณฑ์ และได้แบบทดสอบวัดความถนัดเฉพาะสาขาทางวิทยาการคอมพิวเตอร์ เพื่อหาความถนัดในด้านสายอาชีพของนักศึกษา

เมื่อศึกษางานวิจัยทั้งหมดแล้ว จึงได้แนวโน้มข้อสรุปว่า เทคนิคเหมืองข้อมูลสามารถนำไปพยากรณ์ได้ดี โดยสามารถดูได้จากค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย

4. วิธีการวิจัย

ภายหลังจากการศึกษางานวิจัยที่เกี่ยวข้อง และรวบรวมข้อมูลพื้นฐานต่าง ๆ จึงได้นำมาออกแบบกระบวนการวิจัย ดังแสดงในรูปที่ 1 เพื่อดำเนินการวางแผนแต่ละส่วนตั้งแต่การศึกษา เก็บรวบรวมข้อมูลที่เกี่ยวข้อง ตลอดจนการสร้างแบบจำลองเปรียบเทียบผลลัพธ์ที่ได้



รูปที่ 1 กระบวนการวิจัย

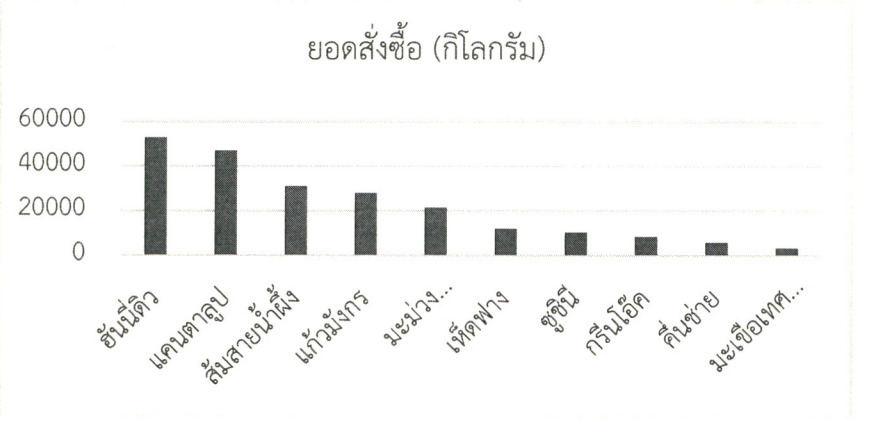
จากรูปที่ 1 สามารถอธิบายกระบวนการได้ดังนี้ โดย (1) ศึกษายอดคำสั่งซื้อในอดีต ซึ่งได้ทำการวิเคราะห์ข้อมูลจากต้นทุนการสั่งซื้อสินค้า 6 ปีย้อนหลัง โดยศึกษาชุดข้อมูลที่สนใจ ได้แก่ ร้านค้า ยอดขั้นต่ำการสั่งซื้อ ยอดขั้นต่ำในการจัดส่ง เดือน อุณหภูมิ ยอดสั่งซื้อ ปริมาณน้ำฝนเฉลี่ย ราคาในอดีต (2) การรวบและจัดลำดับยอดสั่งซื้อมากที่สุด โดยทำการจัดลำดับข้อมูลยอดคำสั่งซื้อที่มากที่สุด 10 ลำดับแรก แบ่งออกเป็น ผลไม้ 5 ชนิด และผัก 5 ชนิด (3) สำรวจทฤษฎีและงานวิจัยที่เกี่ยวข้อง เพื่อหาข้อมูลอ้างอิงเพื่อสรุปผลหาเทคนิคที่เหมาะสมในการพยากรณ์ (4) ออกแบบเครื่องมือ สร้างแบบจำลอง ด้วยโปรแกรมเวก้า เวอร์ชัน 3.8 ซึ่งเป็นโปรแกรมด้านการทำเหมืองข้อมูลและเป็นเป็นซอฟต์แวร์เปิด (open source) จากที่ทำการศึกษาชุดข้อมูลเรียบร้อยแล้ว ได้ทำการเลือกชุดข้อมูลเพื่อมาทำการสร้างเหมืองข้อมูล ได้แก่ เดือน อุณหภูมิ ยอดสั่งซื้อ ปริมาณน้ำฝนเฉลี่ย ราคาในอดีต จำนวน 360 ข้อมูล จากนั้นนำมาสร้างตัวแบบพยากรณ์ (5) เปรียบเทียบผลการวิจัยที่ได้จากแบบจำลอง ด้วยเทคนิคเหมืองข้อมูลทั้ง 3 เทคนิค ประกอบไปด้วย การแจกแจงแบบเกาส์เซียน (gaussian distribution) การถดถอยเชิงเส้น (linear regression) และวิธี SMOreg ซึ่งได้เลือกเปรียบเทียบผลจากเปอร์เซ็นต์ค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (6) สรุปผล

5. ผลการวิจัย

5.1 การวิเคราะห์ข้อมูล

5.1.1 ต้นทุนในการสั่งซื้อสินค้าประเภทอาหารของโรงแรมกรณีศึกษา มียอดการสั่งซื้อเฉลี่ยอยู่ที่ 8 ล้านบาทต่อปี โดยการสั่งซื้อผักและผลไม้มีสัดส่วนในการสั่งซื้อมากที่สุดเมื่อเทียบกับสินค้าประเภทอื่น นับเป็นสัดส่วนสำคัญที่ส่งผลต่อต้นทุนโรงแรม อีกทั้งวัตถุดิบดังกล่าวเป็นของสด มีการเน่าเสียได้หากเก็บไว้นาน โดยข้อมูลที่เก็บเป็นยอดคำสั่งซื้อสินค้าในแต่ละเดือนระหว่างปี พ.ศ. 2556 - 2561 ด้วยไมโครซอฟต์ เอกซ์เซล (microsoft excel) ของแผนก

ดังแสดงในรูปที่ 2 เมื่อพิจารณาผักและผลไม้ทั้งหมด พบว่าประเภทของผักและผลไม้ที่มียอดการสั่งซื้อมากที่สุด 10 ลำดับแรก ประกอบไปด้วย ฮันนี่ดิว แคนตาลูป ส้มสายน้ำผึ้ง แก้วมังกร มะม่วงน้ำดอกไม้ เห็ดฟาง ชูชินี กรีนโอ๊ค คื่นช่าย และมะเขือเทศราชินี โดยนำข้อมูลดังกล่าวมาใช้ในการสร้างเหมืองข้อมูลและสร้างแบบทดสอบการเรียนรู้ของเครื่องจักร



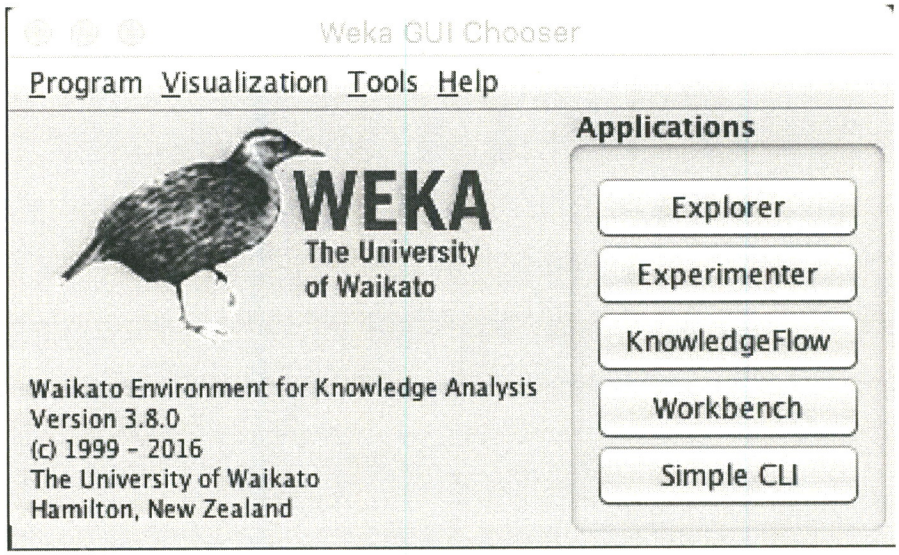
รูปที่ 2 กราฟแสดงจำนวนการสั่งซื้อผัก ผลไม้ ระหว่างปี พ.ศ. 2556-2561

5.1.2 ด้านชุดข้อมูลสำหรับการสร้างเหมืองข้อมูลเพื่อการเรียนรู้ของเครื่องจักร ได้ใช้การเก็บรวบรวมข้อมูลย้อนหลังในอดีตจากฐานข้อมูลในระบบและปัจจัยที่ส่งผลต่อราคา ประกอบไปด้วย ร้านค้า ยอดขั้นต่ำการสั่งซื้อ ยอดขั้นต่ำในการจัดส่ง เดือน อุณหภูมิ ยอดสั่งซื้อ ปริมาณน้ำฝนเฉลี่ย และราคาในอดีต จากนั้นนำไปทำการคัดเลือกคุณสมบัติของข้อมูลเพื่อหาชุดข้อมูลที่มีความสัมพันธ์ในการเรียนรู้ของเครื่องจักร ด้วยเทคนิคที่ใช้การหากลุ่มคุณลักษณะที่ได้รับการประเมินค่าจากความสามารถในการคาดการณ์ (correlation based feature selection) โดยคัดเลือกคุณสมบัติที่มีความสัมพันธ์กับราคามากกว่าร้อยละ 80 ดังแสดงในตารางที่ 1

ตารางที่ 1 ผลการคัดเลือกคุณลักษณะข้อมูล (correlation feature selection)

การคัดเลือกคุณลักษณะข้อมูล			
ข้อมูล	Correlation	**ข้อมูล**	Correlation
เดือน	0.8745	ยอดขั้นต่ำในการจัดส่ง	0.3175
ยอดสั่งซื้อ	0.7998	ยอดขั้นต่ำการสั่งซื้อ	0.3023
ปริมาณน้ำฝนเฉลี่ย	0.7654	ร้านค้า	0.2024
อุณหภูมิ	0.7589		

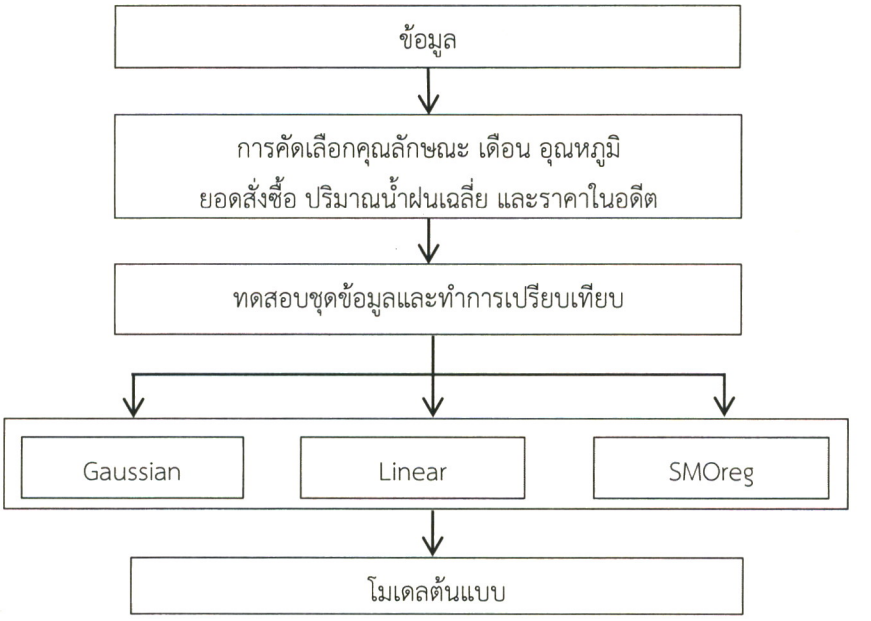
5.1.3 การศึกษาการเรียนรู้ของเครื่องจักรเพื่อพยากรณ์ราคาในอนาคตนั้น ได้เลือกใช้โปรแกรมเวก้า เวอร์ชัน 3.8 ในการศึกษาเพื่อสนับสนุนเกี่ยวกับการทำเหมืองข้อมูล โดยโปรแกรมมีรูปแบบการเรียนรู้ของการทำเหมืองข้อมูลให้เลือกใช้อย่างครบถ้วน ทั้งยังสนับสนุนการทำงานทางด้านการเรียนรู้ด้วยเครื่อง โดยโปรแกรมจะประกอบไปด้วยโมดูลย่อย ๆ สำหรับใช้ในการจัดการข้อมูล สามารถทำงานได้ทุกระบบปฏิบัติการ สามารถเขียนฟังก์ชันเพิ่มเข้าไปในโปรแกรมเองได้ และเป็นเป็นซอฟต์แวร์เปิด ทั้งนี้ ได้ดำเนินการตั้งสมมติฐานไว้โดยให้ผลทำนายที่มีค่าความแม่นยำไม่น้อยกว่าร้อยละ 90



รูปที่ 3 โปรแกรม WEKA

5.2 การทำเหมืองข้อมูล

5.2.1 ด้านการจัดเตรียมข้อมูลนั้น ข้อมูลที่ใช้ในการสร้างเหมืองข้อมูลเพื่อการเรียนรู้ของเครื่องจักรประกอบไปด้วย ร้านค้า ยอดขั้นต่ำการสั่งซื้อ ยอดขั้นต่ำในการจัดส่ง เดือน อุณหภูมิ ยอดสั่งซื้อ ปริมาณน้ำฝนเฉลี่ย และราคาในอดีต โดยนำข้อมูลดังกล่าวไปทำการคัดเลือกคุณสมบัติของข้อมูลเพื่อหาชุดข้อมูลที่มีความสัมพันธ์ในการเรียนรู้ของเครื่องจักร และนำมาทำการกำหนดคุณลักษณะของข้อมูล เพื่อใช้ในการสร้างตัวแบบพยากรณ์ ดังรูปที่ 4 ข้อมูลที่ได้ทำการคัดคุณสมบัติด้วยกระบวนการคัดเลือกคุณลักษณะซึ่งใช้วิธีการหาค่าสหสัมพันธ์



รูปที่ 4 กระบวนการสร้างตัวแบบพยากรณ์

5.2.2 ด้านการสร้างตัวแบบพยากรณ์ ได้มีการนำข้อมูลหลังจากผ่านกระบวนการคัดเลือกคุณลักษณะไปทำการแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ ชุดข้อมูลการเรียนรู้ เป็นชุดข้อมูลสำหรับการสร้างแบบจำลอง และชุดข้อมูลทดสอบ (testing data set) เพื่อเปรียบเทียบเทคนิคเหมืองข้อมูลที่เหมาะสมเพื่อสร้างตัวแบบพยากรณ์ ด้วยเทคนิคเหมืองข้อมูลที่สามารถให้ค่าพยากรณ์แบบตัวเลข (numeric) ค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ และค่ารากที่สองของความคลาดเคลื่อน เพื่อตรวจสอบหาความคลาดเคลื่อนหรือผิดพลาดของการพยากรณ์ ว่ารูปแบบใดมีความคลาดเคลื่อนมากน้อยกว่ากันคิดเป็นกี่เปอร์เซ็นต์ โดยหากค่าทั้งสองเข้าใกล้ศูนย์มากแสดงว่าประสิทธิภาพของแบบจำลองก็ยิ่งมากขึ้น ส่งผลถึงความสอดคล้องของข้อมูลกับรูปแบบการเรียนรู้ในการพยากรณ์ จากนั้นนำไปสร้างตัวแบบการพยากรณ์โดยการวิเคราะห์เพื่อหาข้อมูลชุดเรียนรู้ที่มีความสามารถในการสร้างแบบจำลองจากเทคนิคเหมืองข้อมูลทั้ง 3 เทคนิค ประกอบไปด้วย การแจกแจงแบบเกาส์เซียน การถดถอยเชิงเส้น และวิธี SMOreg

5.2.3 เปรียบเทียบประสิทธิภาพแบบจำลอง การสร้างเหมืองข้อมูลเพื่อทำการทดสอบแบบจำลองนั้น เทคนิคแต่ละเทคนิคจะแสดงผลลัพธ์ ประสิทธิภาพการเรียนรู้ ด้วยค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์และค่ารากที่สองของค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง โดยงานวิจัยนี้ ได้สร้างเหมืองข้อมูลเพื่อทำการทดสอบด้วยการสร้างชุดข้อมูลการเรียนรู้ของตัวอย่างจำนวนทั้งหมด 10 ชุด ประกอบไปด้วย ฮันนี่ดิว แคนตาลูป ส้มสายน้ำผึ้ง แก้วมังกร มะม่วงน้ำดอกไม้ เห็ดฟาง ชูชินี กรีนโอ๊ค คื่นช่าย และมะเขือเทศราชินี จากนั้นทำการเปรียบเทียบพารามิเตอร์ก่อนและหลังจากปรับค่า ซึ่งจะส่งผลต่อความแม่นยำในการประมวลผลและระยะเวลาในการประมวลผลที่แตกต่างกัน เพื่อให้

ได้พารามิเตอร์ที่ให้ค่าความคลาดเคลื่อนน้อยที่สุด และนำไปใช้เป็นโมเดลต้นแบบสำหรับการพยากรณ์ราคาผักและผลไม้ในอนาคต ซึ่งการปรับค่าพารามิเตอร์แต่ละเทคนิคได้กำหนดค่าก่อนและหลัง ดังตารางที่ 2

ตารางที่ 2 ค่าพารามิเตอร์ก่อนและหลังปรับของ 3 เทคนิค

อัลกอริทึม	ดูค่าจาก	ก่อนปรับค่าพารามิเตอร์	หลังปรับค่าพารามิเตอร์
Gaussian Processes	Nois	1.0	0.1
Linear Regression	Ridge	1.0	0.5
SMOreg	The Complexity Parameter	1.0	1.5

โดยเมื่อสร้างชุดข้อมูลต้นแบบแล้วทำการประมวลผลจะได้ผลลัพธ์ของแต่ละเทคนิค ดังตารางที่ 3

ตารางที่ 3 ค่าพารามิเตอร์ที่มีค่าความคลาดเคลื่อนน้อยที่สุดในแต่ละชุดข้อมูลเรียนรู้

ชุดข้อมูล	เทคนิควิธีเหมืองข้อมูล					
	gaussian processes		linear regression		SMOreg	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
ชุดข้อมูลฮันนี่ดิว	0.486	0.764	0.436	0.752	0.460	0.790
ชุดข้อมูลแคนตาลูป	2.937	3.517	3.052	3.641	2.767	3.554
ชุดข้อมูลส้มสายน้ำผึ้ง	0.913	1.286	0.952	1.332	0.909	1.390
ชุดข้อมูลแก้วมังกร	4.322	5.485	4.662	5.701	4.361	5.616
ชุดข้อมูลมะม่วงน้ำดอกไม้	3.798	6.065	2.793	5.833	1.287	5.371
ชุดข้อมูลเห็ดฟาง	1.817	3.112	1.318	3.066	1.118	2.906
ชุดข้อมูลชูชินี	4.835	6.195	4.957	6.174	4.880	6.249
ชุดข้อมูลกรีนโอ๊ค	2.742	4.178	2.784	4.456	1.672	4.234
ชุดข้อมูลคื่นช่าย	0.743	0.980	0.458	0.685	0.420	0.658
ชุดข้อมูลมะเขือเทศราชินี	3.832	4.890	4.171	5.159	4.186	5.167

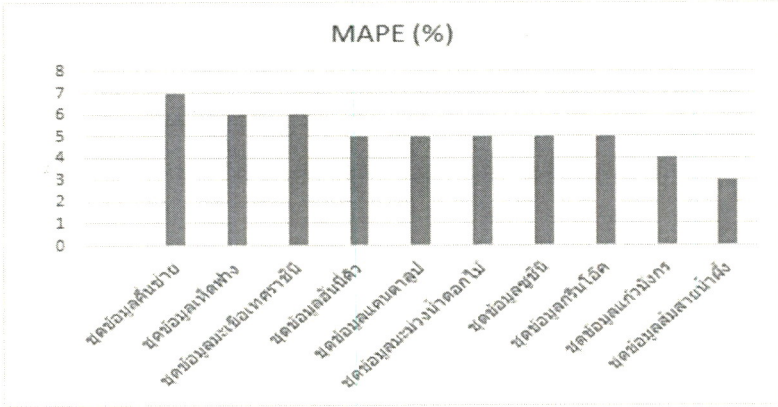
ตารางที่ 4 เปรียบเทียบค่าความคลาดเคลื่อนพารามิเตอร์เฉลี่ยรวมก่อนปรับและหลังปรับ

พารามิเตอร์	ค่าความคลาดเคลื่อนเฉลี่ย			
	รูปแบบเริ่มต้น		หลังจากปรับค่า	
	MAE	RMSE	MAE	RMSE
Gaussian Processes	4.602	5.204	2.643	3.647
Linear Regression	2.565	3.698	2.558	3.680
SMOreg	2.214	3.607	2.206	3.594

จากตารางที่ 4 แสดงผลการเปรียบเทียบค่าความคลาดเคลื่อนพารามิเตอร์ก่อน และหลังปรับด้วยค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ และค่ารากที่สองของค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง จากผลแสดงให้เห็นว่าพารามิเตอร์ด้วยรูปแบบเริ่มต้นของการแจกแจงแบบเกาส์เซียนให้ค่าความคลาดเคลื่อนมากกว่าพารามิเตอร์หลังจากปรับค่าของ การแจกแจงแบบเกาส์เซียน อยู่ที่ 1.959 และ 1.557 พารามิเตอร์ด้วยรูปแบบเริ่มต้นของการถดถอยเชิงเส้น ให้ค่าความคลาดเคลื่อนมากกว่าพารามิเตอร์หลังจากปรับค่าของการถดถอยเชิงเส้นอยู่ที่ 0.007 และ 0.018 พารามิเตอร์ด้วยรูปแบบเริ่มต้นของ SMOreg ให้ค่าความคลาดเคลื่อนน้อยกว่าพารามิเตอร์หลังจากปรับค่าของ SMOreg อยู่ที่ 0.008 และ 0.013 จากค่าเฉลี่ยรวมของค่าความคลาดเคลื่อนแสดงให้เห็นว่าอัลกอริทึม SMOreg ก่อนปรับมีค่าความคลาดเคลื่อนเฉลี่ยรวมน้อยที่สุด

5.3 การเปรียบเทียบประสิทธิภาพการทำงานของเหมืองข้อมูล

5.3.1 การทดสอบแบบจำลองกับชุดข้อมูลทดสอบในการวิจัยครั้งนี้ ได้ทดสอบการพยากรณ์ด้วยการเปรียบเทียบประสิทธิภาพชุดข้อมูล ทดสอบด้วยค่าพารามิเตอร์รูปแบบเริ่มต้นและหลังจากปรับค่าดังตารางที่ 2 โดยนำข้อมูลในชุดทดสอบประกอบไปด้วยผักและผลไม้ทั้งหมด 10 ตัวอย่าง เดือน อุณหภูมิ ยอดสั่งซื้อ และปริมาณน้ำฝนเฉลี่ยเป็นจำนวนทั้งสิ้น 10 ชุดข้อมูล มาทดสอบการเรียนรู้ ทำการเปรียบเทียบและทำนายผลข้อมูล เพื่อให้ได้ราคาสำหรับการพยากรณ์ราคาในปีถัดไปในที่นี้ใช้ค่าความคลาดเคลื่อนสัมบูรณ์เป็นตัวชี้วัดประสิทธิภาพการพยากรณ์ในแต่ละเดือน โดยใช้หาความคลาดเคลื่อนหรือความผิดพลาดของการพยากรณ์คิดออกมาเป็นร้อยละ ซึ่งการวัดความคลาดเคลื่อนระหว่างค่าจริงและค่าพยากรณ์ค่าเฉลี่ยของร้อยละความคลาดเคลื่อนสัมบูรณ์จะเป็นบวกเสมอ ดังแสดงในรูปที่ 5



รูปที่ 5 การเปรียบเทียบค่าเฉลี่ยของร้อยละความคลาดเคลื่อนสัมบูรณ์ชุดข้อมูลผักและผลไม้ 10 ชนิด

5.4 ผลการทดสอบสมมติฐาน

จากสมมติฐานที่ตั้งไว้ ค่าความแม่นยำของการทำนายผลข้อมูลรูปแบบที่ได้จะต้องมีความแม่นยำไม่น้อยกว่า 90 เปอร์เซ็นต์ จากผลการวิเคราะห์แบบจำลองด้วยชุดข้อมูล ทดสอบจากการศึกษาพบว่า SMOreg ได้ให้ค่าความแม่นยำที่มากกว่า 90 เปอร์เซ็นต์โดยมีค่าความคลาดเคลื่อนของข้อมูลประกอบไปด้วย คื่นช่ายคลาดเคลื่อนร้อยละ 7 เห็ดฟาง และมะเขือเทศราชินีคลาดเคลื่อนร้อยละ 6 ฮันนี่ดิว แคนตาลูป มะม่วงน้ำดอกไม้ ซูชินี และกรีนโอ๊คคลาดเคลื่อนร้อยละ 5 แก้วมังกรคลาดเคลื่อนร้อยละ 4 ส้มสายน้ำผึ้งคลาดเคลื่อนร้อยละ 3

6. อภิปรายและสรุปผลการวิจัย

จากการศึกษา ผลลัพธ์ที่ได้ เป็นไปตามวัตถุประสงค์ที่ตั้งไว้คือ ได้แบบจำลองที่มีประสิทธิภาพในการพยากรณ์ราคาผักและผลไม้ ที่ช่วยในการวางแผนต้นทุนที่จะเกิดขึ้นในอนาคต โดยสามารถพยากรณ์ราคาด้วยเทคนิคการเรียนรู้ของเครื่องจักร ร่วมกับโปรแกรมเวก้า ส่งผลให้เกิดการวางแผนบริหารจัดการให้สอดคล้องกับราคาผักและผลไม้ ซึ่งรูปแบบการเรียนรู้ของเครื่องจักรต้นแบบมีความแม่นยำในการทำนายผลข้อมูลในอนาคต เป็นไปตามสมมติฐานที่ตั้งไว้ โดยให้ผลทำนายที่มีค่าความแม่นยำไม่น้อยกว่าร้อยละ 90 จากการทดลองนำมาพยากรณ์เพื่อปรับแผนการใช้วัตถุดิบส่งผลให้เป็นการลดต้นทุนด้านของโรงแรมกรณีศึกษา สามารถพยากรณ์ยอดการสั่งซื้อของโรงแรมกรณีศึกษาและลดต้นทุนได้ประมาณ 50,000 บาทต่อเดือน

7. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการสนับสนุนจาก คณะพัฒนาธุรกิจและอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ประจำปี พ.ศ. 2562 และบริษัทกรณีศึกษา ที่ให้การสนับสนุนด้านข้อมูลในงานวิจัยจนกระทั่งสำเร็จลุล่วงด้วยดี

8. เอกสารอ้างอิง

[1] Thongkam P. Ensemble Algorithm for Feature Selection [Thesis]. Pathum thani; Thammasat University; 2016. (in Thai)

[2] Sermpongpan C. Forcasting of cement consumption in thailand [Thesis]. Pathum thani; Thammasat University; 2017. (in Thai)

[3] Bundasak S. Forecasting golf ball consumption with time series analysis [Thesis]. Chonburi; Kasetsart University; 2016. (in Thai)

[4] Kapol W. Forecasting of power outages cause using data mining techniques a case study of provincial electricity [Thesis]. Pathum thani; Thammasat University; 2015. (in Thai)

[5] Vilailuck S, Jaroenpuntaruk V, Wichadakul D. Utilizing Data Mining Techniques to Forecast Student Academic Achievement of Kasetsart University Laboratory School Kamphaeng Saen Campus Educational Research and Development Center. Veridian E-Journal, Science and Technology Silpakorn University. 2015; 2(2): 1-17.

[6] Ngamsuk W. Time series forecasting by the comparision of classical and box-jenkinsmethods case study the number of accidents in the country. [Thesis]. Chonburi: Burapha University; 2012. (in Thai)

[7] Konchalad P. The total energy consumption forecasting in Thailand by Box-Jenkins method and artificial neural network [Thesis]. Mahasarakham: Mahasarakham University; 2011. (in Thai)

[8] Polasa N. To sell product modeling and sales forecasting of life insurance using data mining techniques Case study: Life insurance company [Thesis]. Pathum thani; Thammasat University; 2015. (in Thai)

[9] Kaewhawong N. Forecasting Electricity Consumption of Thailand by Using SARIMA and Regression Models with ARMA Errors. Thai Journal of Science and Technology. 2015; 4(1): 24-36.

[10] Vonganansup S, Yampaka T, Moosika O. Applied Data Mining Techniques To Analize Pre-Test And Skill-Test For Increased Student Potential. Rajamangala University of Technology Tawan-ok Research Journal. 2016; 5(1): 12-9.