KKU Res. J. 2014; 19(1) : 31-43
http://resjournal.kku.ac.th

การศึกษาเปรียบเทียบเทคนิคเหมืองข้อมูลใช้ในการทำนายผลผลิตพืชไร่ทางการเกษตร: กรณีศึกษาข้าวไทย

A Comparative Study of Data Mining Techniques to Predict Agricultural Production: A Case Study in Thai Rice

ขนิษฐา กุลนาวิน[1], ปานจิตร์ หลงประดิษฐ์[1], ชมัยพร เจริญพร[2], ภัทรสินี ภัทรโกศล[3]*
Kanitta Kulnawin[1], Panchit Longpradit[1], Chamaiporn Charoenporn[2], Pattarasinee Bhattarakosol[3]*

[1]สาขาวิชาเทคโนโลยีสารสนเทศคุณภาพ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏเพชรบุรี
[2]โปรแกรมวิชาเกษตรศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา
[3]ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
**Correspondent author: nonone007@yahoo.com*

บทคัดย่อ

ข้าวเป็นธัญพืชที่สำคัญของประเทศไทยเนื่องจากเป็นทั้งอาหารเพื่อบริโภคในประเทศและเพื่อการส่งออก ไทยเป็นผู้ส่งออกข้าวอันดับหนึ่งของโลกหากแต่ทว่าในปี 2011 เวียดนามได้มีการปรับปรุงการผลิตและการส่งออกข้าวจนเกือบได้กลายเป็นผู้ส่งออกที่ใหญ่ที่สุดในโลก ในการนี้เทคโนโลยีสารสนเทศสามารถช่วยส่งเสริมการเพิ่มผลผลิตให้ดีขึ้นได้ ดังนั้นงานวิจัยนี้จึงมีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบเทคนิคเหมืองเพื่อใช้ในการทำนายผลผลิตข้าวไทยที่มีความแม่นยำสูง โดยเริ่มจากทำการศึกษาเปรียบเทียบเพื่อหาปัจจัยที่เหมาะสมในการทำนายผลผลิตข้าว แล้วค่อยศึกษาเทคนิคของเหมืองข้อมูลเพื่อหาเทคนิคที่ทำนายผลแม่นยำสูง ผลของการศึกษานี้พบว่าปัจจัยที่เหมาะสมในการทำนายได้แก่ ข้อมูลผลผลิตทางการเกษตร ปริมาณน้ำฝน อุณหภูมิสูงสุด อุณหภูมิต่ำสุด จำนวนวันที่ฝนตก และความชื้นสัมพัทธ์ พื้นที่เพาะปลูก พื้นที่เก็บเกี่ยวโดยเป็นข้อมูลทุกจังหวัดตามชนิดของพันธ์ข้าวและฤดูการปลูกข้าวย้อนหลังเป็นเวลา 10 ปีตั้งแต่ปี 2543-2552 ข้อมูลได้มาจากสำนักงานเศรษฐกิจการเกษตรและกรมอุตุนิยมวิทยา ในส่วนของการทำการเปรียบเทียบเทคนิคเหมืองข้อมูลได้พิจารณา 4 อัลกอริทึม ดังนี้ โครงข่ายประสาทเทียม การถดถอยเชิงเส้น ซัพพอร์ตเวกเตอร์แมช ซีน และต้นไม้การตัดสินใจ ผลของการศึกษานี้พบว่าเทคนิคการถดถอยเชิงเส้นเป็นวิธีที่มีประสิทธิภาพเพื่อใช้ทำนายผลผลิตข้าวซึ่งมีค่าเฉลี่ยสัมประสิทธิ์สหสัมพันธ์โดยรวมเท่ากับ 0.9886 และค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์โดยรวมเท่ากับ 11.1937 ผลที่ได้จากการทำนายสามารถนำมาใช้เพื่อกำหนดนโยบายการวางแผนการเพาะปลูกและการพัฒนาของการเพาะปลูกข้าวที่เหมาะสมเพื่อเพิ่มศักยภาพการทำเกษตรกรรม ซึ่งงานวิจัยต่อไปคือ การนำเทคนิคเหมืองข้อมูลไปบูรณาการร่วมกับโปรแกรมตัวแทนอัจฉริยะ เพื่อให้ระบบทำนายผลผลิตพืชไร่ที่มีประสิทธิภาพมากยิ่งขึ้นด้วยการปรับปรุงข้อมูลแบบอัตโนมัติแทนแบบคงที่

Abstract

In Thailand, rice crop is a major cereal crop that it be the main food and also for exports. Thailand was raked top exporters of high quality rice crop in this ASEAN. Until 2011, Vietnam has improved the production and the

exporting so that Vietnam's rice has become the biggest rice exporter in the world. According to, information technology can help enhancing for the better rice crop productivity. Therefore, this research aimed to compare mining techniques to predict Thai rice production with high accuracy. This research starts from the identifying the appropriate factor to predict rice production. Then determine the high accuracy of data mining techniques to predict rice production. The results showed that there are eight appropriate factors to predict the rice production that are average rainfall, max-temperature, min-temperature, number of days of rain, relative humidity, average rice area, average rice harvest and average rice area. The data was collected by using every province according to the type of rice and period of farming 10 years throughout 2000-2009, which collecting from Office of Agricultural Economics and Thai Meteorology Department. The other results of a comparative data mining techniques by using four algorithms to predict rice crop production: neural networks, linear regression, support vector machines and decision trees, showed that the most coefficient of correlation is linear regression. So, the average of efficiency relation is 0.9895; and so on average of relative absolute error is 11.1937. As a result of prediction rice production can be used in policy formulation, planning, planting, and development of rice crop that suitable for increasing agriculture potential. The future work is to integrate the data mining techniques to do automatic data updating with an intelligent agent in order to predict more effectively by means of automatic update instead of static approach.

คำสำคัญ: เหมืองข้อมูล การถดถอยเชิงเส้น ผลผลิตข้าว
Keywords: Data mining, Linear regression, Rice production

1. บทนำ

ข้าวเป็นพืชเศรษฐกิจของไทยที่ทำรายได้หลักเข้าประเทศและเป็นอาหารที่สำคัญที่สุดของคนไทย โดยประเทศไทยมีการเพาะปลูกข้าวในทุกภูมิภาค หากแต่ว่าในปี 2554 ที่ผ่านมาไทยมีพื้นที่เพาะปลูกข้าว 68.75 ล้านไร่ แต่ผลิตข้าวได้เพียง 20.46 ล้านตัน และมีผลผลิตเฉลี่ย 390 กิโลกรัมต่อไร่โดยประมาณ (1) ซึ่งผลผลิตเฉลี่ยต่ำกว่าประเทศเพื่อนบ้านที่เป็นประเทศผู้ส่งออกข้าวรายใหญ่เช่น เวียดนามที่มีพื้นที่เพาะปลูกเพียง 48.375 ล้านไร่ แต่ผลิตข้าวได้ 27.15 ล้านตัน และมีผลผลิตเฉลี่ย 561.24 กิโลกรัมต่อไร่โดยประมาณ (1) การที่ไทยสามารถผลิตและส่งออกข้าวเป็นอันดับหนึ่งของโลกนั้นเนื่องมาจากการได้เปรียบด้านพื้นที่เพาะปลูกและพันธุ์ข้าวหอมมะลิที่มีคุณภาพซึ่งมีราคาสูงในตลาดโลก แต่ประสิทธิภาพในการผลิตข้าวของประเทศไทยไม่สามารถแข่งขันกับประเทศเพื่อนบ้าน เช่น เวียดนามได้และเป็นอุปสรรคที่สำคัญ คือ ความสามารถในการแข่งขันการส่งออกข้าว เวียดนามใช้พื้นที่เพาะปลูกน้อยกว่าแต่ให้ผลผลิตสูงกว่า เนื่องจากได้นำเอาเทคโนโลยีทางเกษตรมาช่วยในการผลิตข้าวคือหลักการ ลด 3 เพิ่ม 3 คือ ลดปริมาณเมล็ดพันธุ์ให้เหมาะสม ลดการใช้ปุ๋ยเคมี ลดการใช้ยาปราบศัตรูพืช เพิ่มผลผลิต เพิ่มคุณภาพและเพิ่มกำไร (2) นอกจากนี้ต้นทุนค่าแรงที่ต่ำกว่าไทยทำให้มีผลต่อราคาการส่งออกข้าวที่ถูกกว่า ดังนั้น เพื่อให้ข้าวไทยยังคงเป็นที่ต้องการของตลาดโลกจึงจำเป็นต้องรักษาคุณภาพการส่งออกข้าวและมีการวางแผนกำหนดปริมาณการผลิตที่เหมาะสมกับฤดูกาลที่เปลี่ยนแปลงตลอดเวลา ความต้องการเหล่านี้สามารถบรรลุได้หากมีเทคโนโลยีสารสนเทศที่สามารถช่วยสนับสนุนข้อมูลในการตัดสินใจเพื่อการวางแผนปลูกข้าว อันจะช่วยลดความเสี่ยงและความผิดพลาดของการปลูกข้าวได้ โดยท้ายที่สุดแล้วจะสามารถส่งผลให้ไทยเพิ่มปริมาณผลผลิตข้าวต่อพื้นที่เพาะปลูกสูงขึ้นได้ตามที่ต้องการ งานวิจัยนี้จึงได้มีใช้การนำเทคโนโลยีเหมืองข้อมูลมาประยุกต์เพื่อใช้ในการทำนายผลผลิตข้าว โดยเทคโนโลยีเหมืองข้อมูลที่นำมาประยุกต์นี้สามารถสกัดข้อมูลที่เป็นประโยชน์มาช่วยในการทำนายผลผลิต

ข้าวที่เหมาะสม และนำไปสู่การตัดสินใจวางแผนการปลูกข้าวที่ถูกต้องตามฤดูกาล งานวิจัยนี้ได้ศึกษาเปรียบเทียบเทคนิคเหมืองข้อมูลที่นำมาประยุกต์ใช้ในการทำนายผลผลิตข้าว เพื่อศึกษาหาปัจจัยและเทคนิคที่มีประสิทธิภาพที่สุดในการทำนายผลผลิต ซึ่งการทำเหมืองข้อมูลเป็นการค้นหาองค์ความรู้ที่มีประโยชน์ที่ซ่อนอยู่ในข้อมูลขนาดใหญ่ (knowledge discovery in database) ผลลัพธ์ของการทำเหมืองข้อมูลนำมาช่วยในการวิเคราะห์แนวโน้มผลผลิตข้าวที่จะเกิดขึ้นของเกษตรกร สามารถนำแบบจำลองของการพยากรณ์ที่ได้นำไปพัฒนาระบบเทคโนโลยีสารสนเทศที่ช่วยในสนับสนุนการตัดสินใจในการเลือกพันธ์ข้าวที่เหมาะสมเพื่อช่วยในการวางแผนปลูกข้าวส่งผลให้ผลผลิตข้าวต่อไร่เพิ่มขึ้น

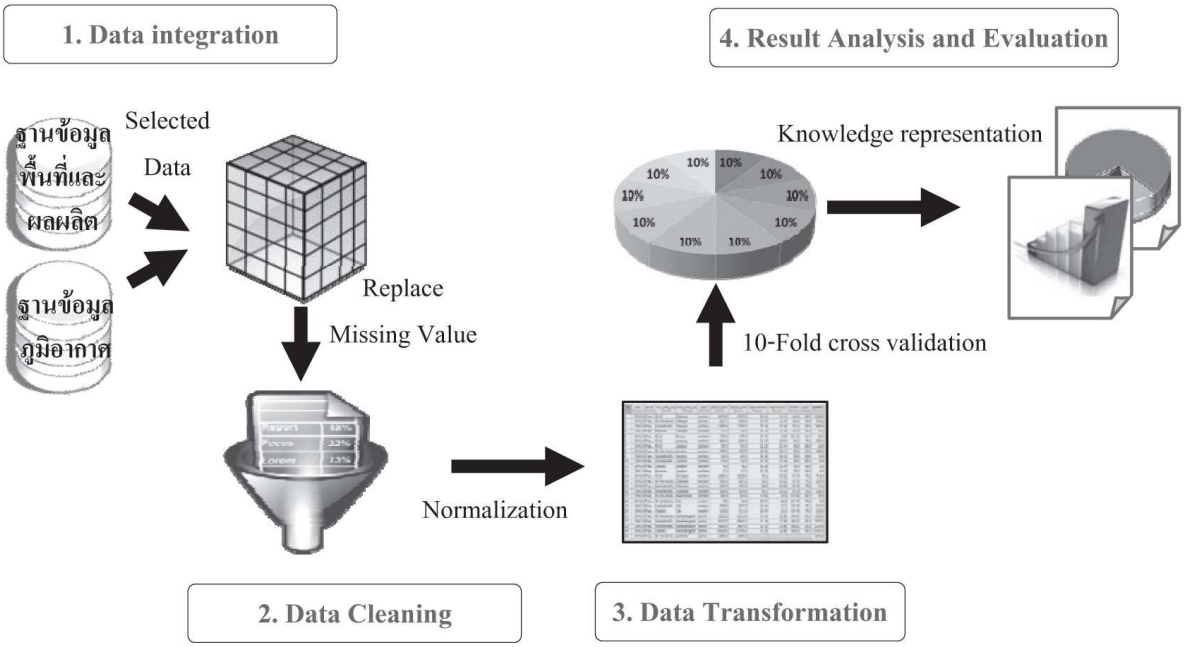
2. วิธีวิจัย

การดำเนินการวิจัยนี้ได้ใช้เหมืองข้อมูลเพื่อใช้ในการทำนายผลผลิตข้าว ผลของการทำนายเป็นสิ่งที่คาดการณ์แนวโน้มที่จะเกิดขึ้นในอนาคตสามารถนำมาใช้เพื่อลดความเสี่ยงและช่วยสนับสนุนการตัดสินใจ จากการศึกษาพบว่าในปัจจุบันการทำเหมืองข้อมูลที่ดีจะต้องสามารถจำแนกข้อมูลที่ถูกต้อง แม่นยำและรวดเร็ว โดยนักวิจัยได้พยายามพัฒนาและนำเสนอวิธีการต่างๆ ให้แบบจำลองมีประสิทธิภาพสูงขึ้น อาทิเช่น อัมรินทร์ ก้อนแพง (3) ได้เสนอการออกแบบและพัฒนาโมเดลการพยากรณ์ราคาข้าวเปลือกโดยใช้อัลกอริทึมแบบจำลองโครงข่ายประสาทเทียม (artificial neural networks: ANNs) พบว่าค่าความผิดพลาดของการทำนายน้อยกว่าแบบจำลองการปรับให้เรียบแบบเอ็กซ์โปแนนเชียลด้วย โดยทำการทำนายราคาข้าวเปลือก 4 ชนิดดังนี้ ข้าวหอมมะลิ ข้าวเจ้านาปี 5% ข้าวเจ้านาปีและนาปรัง 14-15% และข้าวเหนียว ส่วนโรเนล และ คัลคาร์นิ (4) ได้ใช้แบบจำลองโครงข่ายประสาทเทียมเพื่อทำนายผลผลิตโดยอาศัยข้อมูลในอดีตเพื่อลดความเสี่ยงของการทำเกษตร ผลผลิตทางการเกษตรจะต้องอาศัยปัจจัยทางภูมิอากาศเป็นส่วนใหญ่ ทำนายศัตรูพืชที่เกิดขึ้นในช่วงฤดูฝน พืชที่ใช้ทำการศึกษาได้แก่ ข้าว ถั่วลิสง ถั่วเหลือง และอ้อย เทคนิคข้อมูลเหมือง แต่ผลที่ได้เหมาะกับเฉพาะในแปลงทดลองเท่านั้น และดิลก ภิญโญศรี และคณะ (5) สร้างแบบจำลองเพื่อการประมาณการผลผลิตอ้อย ประกอบด้วย 2 ขั้นตอนคือ การสร้างต้นไม้การตัดสินใจ และ การประมาณค่าด้วยแบบจำลองการถดถอยเชิงเส้นโครงข่ายประสาทเทียม (General Regression Neural Network : GRNN) เริ่มจากการคัดเลือกข้อมูลที่เป็นปัจจัยที่มีผลต่อการเจริญเติบโตของอ้อยและใช้ตัวอย่างแปลงทดลองอย่างจำนวน 1,446 แปลง โดยแบ่งกลุ่มข้อมูลออกได้เป็น 5 กลุ่ม จากนั้นก็ทำการประมาณค่าด้วยแบบจำลองการถดถอยเชิงเส้นโครงข่ายประสาทเทียม แต่เนื่องจากขาดปัจจัยการจัดการแปลงที่ดีทำให้มีความแตกต่างของข้อมูลผลของการทำนายจึงไม่แตกต่างกันมากนัก การประมาณค่าที่ได้อาจส่งผลการทำนายก่อให้เกิดความผิดพลาดสูงขึ้นได้

จากการศึกษาพบว่าในประเทศไทยยังไม่มีการนำเหมืองข้อมูลมาใช้ในการทำนายผลผลิตข้าวไทย และอัลกอริทึมที่เป็นที่นิยมใช้ในการทำนายผลผลิตทางการเกษตรคือแบบจำลองโครงข่ายประสาทเทียม ดังนั้นในการวิจัยนี้จึงได้ทำการศึกษาเพื่อหาปัจจัยที่เหมาะสมและศึกษาเปรียบเทียบเทคนิคเหมืองข้อมูลสำหรับการทำนายผลผลิตข้าวไทยที่มีความแม่นยำสูงโดยเลือกใช้แบบจำลองโครงข่ายประสาทเทียมเพื่อศึกษาในการทำนาย โดยแบ่งขั้นตอนการศึกษาได้ดังรูปที่ 1

การทำเหมืองข้อมูลที่ดีควรจะเริ่มตั้งแต่การเตรียมข้อมูล (data preparation) เพื่อให้สามารถนำไปใช้งานได้ จากรูปที่ 1 แสดงขั้นการทำเหมืองข้อมูล ซึ่งมีขั้นตอนของการทำเหมืองข้อมูลดังนี้

ขั้นตอนที่ 1 เป็นการรวมฐานข้อมูลจากหลายแห่งมารวมกัน (integration data) โดยศึกษาจากลักษณะและคุณสมบัติของข้อมูล ข้อมูลที่ใช้มาจาก 2 แหล่งข้อมูล แหล่งแรกเป็นข้อมูลผลผลิตในแต่ละชนิดพันธุ์ข้าว พื้นที่เพาะปลูก พื้นที่เก็บเกี่ยว จากสำนักงานเศรษฐกิจการเกษตร และอีกแหล่งข้อมูลมาจากกรมอุตุนิยมวิทยา คือ วันที่ฝนตก ปริมาณน้ำฝน ความชื้นสัมพัทธ์ อุณหภูมิสูงสุดและต่ำสุดของทุกจังหวัด ข้อมูลทั้งหมดที่ได้มาเป็นข้อมูลย้อนหลัง 10 ปี (พ.ศ. 2543-2552) มีการจัดกลุ่มของข้อมูลตามฤดูการเพาะปลูก(นาปี, นาปรัง) โดยข้อมูลที่ได้มีทั้งหมด 8648 เรคคอร์ดดังแสดงตัวอย่างของมูลที่ได้ทำการรวบรวมจัดเก็บลงในฐานข้อมูลในตารางที่ 1



รูปที่ 1. การทำเหมืองข้อมูลเพื่อศึกษาเปรียบเทียบเทคนิคเหมืองข้อมูลใช้ในการทำนายผลผลิตข้าว

ตารางที่ 1. แสดงตัวอย่างของข้อมูลที่จัดเก็บลงในฐานข้อมูล

ลำดับ	ปี	ฤดูการปลูก	พันธ์ข้าว	พื้นที่การปลูก	...	ปริมาณน้ำฝน	จำนวนวันฝนตก	ความชื้น	ผลผลิตที่ได้
1	2543	นาปรัง	RD10	38400	...	48.23	74	33	24608
2	2543	นาปรัง	RD-Non-photosensitive	6237	...	48.23	74	33	3412
3	2543	นาปรัง	SuphanBuri60_90	11876	...	48.23	74	33	9504
4	2543	นาปรัง	khawhom	332	...	48.23	74	33	195
5	2543	นาปรัง	RD10	1770	...	56.87	73	30	966
6	2543	นาปรัง	RD10	6569	...	39.75	69	28	3503
7	2543	นาปรัง	RD10	767	...	32.68	69	23	629
8	2543	นาปรัง	RD-Non-photosensitive	1787	...	32.68	69	23	1386
9	2543	นาปรัง	ChaiNat1	42	...	32.68	69	23	32
10	2543	นาปรัง	khawhom	63	...	32.68	69	23	44
11	2543	นาปรัง	RD10	12856	...	31.72	64	27	7110
12	2543	นาปรัง	ChaiNat1	3142		46.42	65	33	2320
...	...	...	...	...		...	...	...	...
8648	2552	นาปี	ChaiNat1	12576		157.82	90	15	5571

ขั้นตอนที่ 2 เป็นการทำความสะอาดข้อมูล (data cleaning) ซึ่งเป็นกระบวนการให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนำมาใช้ แสดงถึงความน่าเชื่อถือของข้อมูล เนื่องจากข้อมูลได้มาจากคนละแหล่งข้อมูลก่อเกิดการสูญหายของข้อมูล (missing values) ส่งผลให้เกิดค่าที่ผิดปกติได้ (outlier) ทำให้ข้อมูลไม่พร้อมที่จะนำไปใช้งาน หรือไม่อยู่ในรูปแบบที่จะสามารถนำไปวิเคราะห์ด้วยเทคนิคทางเหมืองข้อมูลได้ ดังนั้นจากการศึกษาการแทนที่ค่าที่สูญหายเป็นวิธีที่ง่ายและเป็นที่นิยมมี 3 วิธี (6) ดังนี้

1. แทนที่ค่าที่สูญหายด้วยค่าคงที่

2. การแทนค่าสูญหายด้วยค่ากลาง (mean) หรือค่าฐานนิยม (mode)

3. การแทนค่าสูญหายด้วยการสุ่มค่า

เนื่องจากมีการรวมข้อมูลจาก 2 แหล่งข้อมูลเมื่อนำมาผนวกกันก่อให้เกิดการสูญหายของข้อมูล อาทิเช่น วันที่ฝนตก ปริมาณน้ำฝน ความชื้นสัมพัทธ์ อุณหภูมิสูงสุดและต่ำสุด ดังนั้นจาก 3 วิธี ข้างต้นในการวิจัยนี้ได้ทำการแทนค่าข้อมูลสูญหายด้วยค่ากลางหรือค่าฐานนิยมเป็นซึ่งเป็นวิธีการที่นิยมมากที่สุด (7) นอกจากนี้ในการวิจัยนี้การจัดกลุ่มข้อมูลที่ได้ทำการศึกษาครั้งนี้แบ่งออกเป็น 20 กรณีศึกษาโดยทำการจัดกลุ่มตามพันธ์ข้าว ฤดูการปลูกข้าว (นาปี, นาปรัง) ภูมิภาคของการปลูกข้าว (ภาคเหนือ, ภาคกลาง, ภาคตะวันออก, ภาคตะวันตอก, ภาคตะวันออกเฉียงเหนือ และ ภาคใต้) เพื่อลดค่าที่ผิดปกติของข้อมูล (outlier) ดังตารางที่ 2 แสดงกลุ่มของข้อมูลและจำนวนแอททริบิวท์ในแต่ละกลุ่มของข้อมูล

ตารางที่ 2. แสดงการจัดกลุ่มข้อมูลตามพันธ์ข้าว ฤดูการปลูกข้าว และภูมิภาคของการปลูกข้าว

กรณีศึกษาที่	กลุ่มข้อมูลตามพันธ์ข้าว ฤดูการปลูกข้าว และภูมิภาค	จำนวนแอททริบิวต์
1	ข้าวพันธุ์ปทุมธานี1 ภาคกลาง ข้าวนาปรัง	332
2	ข้าวพันธุ์หอม ภาคกลาง ข้าวนาปรัง	130
3	สุพรรณบุรี1 ภาคกลาง ข้าวนาปรัง	366
4	ข้าวพันธุ์ปทุมธานี1 ภาคกลาง ข้าวนาปี	352
5	ข้าวพันธุ์สุพรรณบุรี 60 90 ภาคกลาง ข้าวนาปรัง	388
6	ข้าวพันธุ์กข6 ภาคเหนือ ข้าวนาปี	198
7	ข้าวพันธุ์ชัยนาท1 ภาคกลาง ข้าวนาปี	166
8	ข้าวพันธุ์ชัยนาท1 ภาคใต้ ข้าวนาปรัง	96
9	ข้าวพันธุ์ชัยนาท1 ภาคตะวันออกเฉียงเหนือ ข้าวนาปรัง	180
10	ข้าวพันธุ์ปทุมธานี1 ภาคตะวันออก ข้าวนาปี	88
11	ข้าวพันธุ์กข15 ภาคเหนือ ข้าวนาปี	114
12	ข้าวพันธุ์กข6 ภาคกลาง ข้าวนาปี	118
13	ข้าวพันธุ์กข15 ภาคตะวันออกเฉียงเหนือ ข้าวนาปี	396
14	ข้าวพันธุ์ปทุมธานี1 ภาคใต้ ข้าวนาปี	112
15	ข้าวพันธุ์กข6 ภาคตะวันออกเฉียงเหนือ ข้าวนาปี	418
16	ข้าวพันธุ์ชัยนาท1 ภาคใต้ ข้าวนาปรัง	328
17	ข้าวพันธุ์ปทุมธานี1 ภาคตะวันออกเฉียงเหนือ ข้าวนาปรัง	142
18	สุพรรณบุรี1 ภาคกลาง ข้าวนาปี	380
19	ข้าวพันธุ์สุพรรณบุรี 60 90 ภาคกลาง ข้าวนาปี	394
20	ข้าวพันธุ์สุพรรณบุรี 60 90 ภาคตะวันออกเฉียงเหนือ ข้าวนาปรัง	152
รวม		**4,850**

ขั้นตอนที่ 3 เป็นการแปลงรูปแบบข้อมูล (data transformation) ภายใต้การทำงานของเหมืองข้อมูลเพื่อค้นหารูปแบบของแบบจำลองในการทำนายตามหลักของอัลกอริทึม (algorithm)

ขั้นตอนที่ 4 เป็นขั้นตอนสุดท้ายของกระบวนการ คือ การวิเคราะห์และประเมินผลลัพธ์ที่ของแบบของแบบจำลอง (result analysis and evaluation) เป็นขั้นตอนการแปลความหมาย และการประเมินแบบของแบบจำลองที่ได้ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการ โดยเริ่มจากนำเข้าข้อมูลชุดที่ใช้ในการเรียนรู้ (training data set) เพื่อสอนให้เครื่องคอมพิวเตอร์เรียนรู้ปัญหาโดยสร้างแบบจำลองออกมาในรูปแบบจัดหมวดหมู่ข้อมูลใดควรอยู่ในประเภทเดียวกันบ้าง จากงานวิจัยของ รัชภูมิ ใจกล้า (8) ได้ทำนายผลผลิตข้าวจากการเรียนรู้ของวิธีการซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) ในการทำนายน้ำหนักของไนโตรเจนในดิน ทำนายน้ำหนักของต้นข้าว และทำนายน้ำหนักของเมล็ดข้าว เมื่อนำไปเปรียบเทียบกับการถดถอยเชิงเส้นและไม่เป็นเชิงเส้นพบว่าซัพพอร์ตเวคเตอร์แมชชีนให้ผลรวมค่าเฉลี่ยผิดพลาดน้อยที่สุด แต่วิธีการของซัพพอร์ตเวกเตอร์แมชชีนจะมีปัญหาในการที่ต้องหาค่าพารามิเตอร์ที่เหมาะสม ซึ่งค่าพารามิเตอร์ที่แตกต่างกันก็จะให้ผลลัพธ์ที่แตกต่างกัน หากข้อมูลมีปัจจัยที่แตกต่างกันไปขึ้นอยู่กับการนำเข้าของข้อมูลอาจทำให้อัลกอริทึมนี้ผลลัพธ์มีค่าไม่แน่นอนได้ ถึงแม้ว่าเทคนิคที่กล่าวมาจะให้ผลลัพธ์ของการทำนายที่ดี แต่จากการศึกษาแต่ละเทคนิคของอัลกอริทึมมีลักษณะของการนำไปใช้เพื่อสร้างแบบจำลองได้มีประสิทธิภาพต่างกันขึ้นอยู่กับคุณสมบัติของข้อมูลที่เป็นปัจจัยนำเข้า ดังนั้นวิจัยนี้ได้เลือกอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน และการถดถอยเชิงเส้นซึ่งเป็นอีกอัลกอริทึมที่นิยมใช้ในการทดสอบเพื่อสร้างแบบจำลองกับข้อมูลเชิงตัวเลข (numeric) โดยทำการเปรียบเทียบกับอัลกอริทึมโครงข่ายประสาทเทียมจากข้อมูลกรณีศึกษาทั้ง 20 กรณีศึกษาเพื่อศึกษาหาอัลกอริทึมที่เหมาะสมกับข้อมูลที่จะนำมาใช้ทำนายผลผลิตข้าวและยังได้ศึกษาอัลกอริทึมเพิ่มเติมคือต้นไม้การตัดสินใจ (decision trees) โดยศึกษาจากคุณสมบัติแต่ละอัลกอริทึมที่มีประสิทธิภาพที่สุดนำมาใช้ในการพยากรณ์ครั้งนี้ดังตารางที่ 3 (9-11)

ผลลัพธ์ของของการประมวลผลแต่ละอัลกอริทึมจะได้แบบจำลองออกมาเพื่อนำไปทดสอบความแม่นยำกับข้อมูลนำเข้าอีกชุดหนึ่งที่เรียกว่า ข้อมูลชุดทดสอบ (testing data set) เป็นการตรวจสอบความถูกต้องของแบบจำลองว่าผลของการทำนายออกมาถูกต้องหรือไม่ เพื่อลดผลกระทบของความแตกต่างของข้อมูลและทำความเข้าใจลักษณะของแบบจำลอง สุดท้ายจะได้แบบจำลองข้อมูลที่แม่นยำ (12) ในงานวิจัยนี้ได้มีการใช้วิธีการประเมินประสิทธิภาพ (evaluation) ที่เรียกว่า “เคโฟลดคลอสวาลิเดชั่น” (K-Fold cross validation) เป็นวิธีการประเมินประสิทธิภาพ ทำโดยการสุ่มตัวอย่าง (resampling) ข้อมูลขึ้นมาทดสอบ โดยแบ่งข้อมูลออกเป็นส่วน ๆ ทั้งหมด K ส่วน เท่า ๆ กัน โดยจะใช้ K-1 ส่วนเป็นข้อมูลที่ใช้ในการสอนระบบ อีก 1 ส่วนใช้เป็นข้อมูลทดสอบ และทำการวนทดสอบใหม่โดยสลับส่วนข้อมูลทดสอบกับส่วนข้อมูลสอนระบบ เพื่อให้ทุกส่วนนั้นได้มีโอกาสเป็นข้อมูลทดสอบ งานวิจัยนี้ใช้การแบ่งข้อมูลสำหรับการทดสอบเป็น 10 ชุดย่อย (fold) คือ กำหนดให้ K =10 ดังนั้นจึงมีการเปลี่ยนข้อมูลทดสอบทั้งหมด 10 รอบ ต่อมาหาค่าความผิดพลาดที่ต่ำที่สุดที่เกิดขึ้นจากการสอนทั้ง 10 ครั้ง แล้วก็จะได้เป็นค่าความผิดพลาดของแบบจำลองที่เกิดขึ้น สุดท้ายก็จะได้กลุ่มของแบบจำลองที่ดีที่สุดสำหรับจะใช้ในทำนาย (13)(14)

ในส่วนของการประเมินผลการทำนายข้อมูลจากการทำเทนโฟลดคลอสวาลิเดชั่น (10-Fold cross validation) เพื่อศึกษาเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมวัดด้วยค่าสัมประสิทธิ์สหสัมพันธ์ (coefficient of correlation) เพื่อวัดค่าสัมประสิทธิ์สหสัมพันธ์เป็นการวัดทิศทางและระดับของความสัมพันธ์ระหว่างตัวแปรที่ได้ทำการวิจัย และค่าความคลาดเคลื่อนสัมพัทธ์ (relative absolute error) เป็นการแสดงค่าเปอร์เซ็นต์ความคลาดเคลื่อนได้จากค่าความคลาดเคลื่อนได้จากค่าความคลาดเคลื่อนสัมบูรณ์เปรียบเทียบกับค่าจริง ทำการวัดจากความคลาดเคลื่อนของค่าจริงและค่าพยากรณ์ที่ได้ควรพิจารณาจากค่าสัมประสิทธิ์ต่าง ๆ หรือจำนวนข้อมูลต่างๆ จะพิจารณาจากการที่ค่าจริงใกล้เคียงค่าพยากรณ์ที่สุด หรือทำให้เกิดความคลาดเคลื่อนน้อยที่สุด ย่อมเป็นค่าที่เหมาะสมกับการใช้พยากรณ์ให้ได้ผลลัพธ์ที่แม่นยำ (13) ได้ใช้วิธีวัด

ตารางที่ 3. ข้อดีและข้อด้อยของอัลกอริทึมที่ใช้ในการพยากรณ์แบบต่าง ๆ เมื่อเปรียบเทียบกับโครงข่ายประสาทเทียม

วิธีการพยากรณ์	ลักษณะของการพยากรณ์	ข้อดี	ข้อด้อย	ตัวอย่างการนำไปใช้งาน
1. มัลติเลเยอร์เพอร์เซ็ป ตรอน (multilayer perceptron)	มีการเรียนรู้ของแบบจำลองโครงข่ายประสาทเทียมทำได้โดยการส่งข้อมูลเข้ามายังส่วนที่เรียกว่าเพอร์เซ็ปตรอน (perceptron) สามารถเทียบได้กับสมองของมนุษย์ โดยที่เพอร์เซ็ปตรอนทำการรับข้อมูลที่อยู่ในรูปของตัวเลขเพื่อเข้ามาคำนวณ	1. สามารถในการเรียนรู้จำลองปัญหาได้โดยไม่จำเป็นต้องทราบถึงรูปแบบของข้อมูล 2. เหมาะกับความสัมพันธ์เป็นแบบไม่เชิงเส้นได้	1. ไม่สามารถอธิบายและให้เหตุผลของความสัมพันธ์ของตัวแปรและผลลัพธ์ที่ได้ 2. ไม่มีหลักการที่ชัดเจนในการกำหนดโครงสร้าง	1. ปริมาณน้ำฝน 2. โหลดไฟฟ้า 3. ความเร็วลม 4. ดัชนีตลาดหลักทรัพย์ 5. อัตราแลกเปลี่ยน
2. การถดถอยเชิงเส้น (linear regression)	ใช้หลักทางสถิติเพื่อสร้างแบบจำลองทางคณิตศาสตร์ ด้วยการวิเคราะห์ทางสถิติของสมการเชิงเส้นถดถอยเชิง ซึ่งทุกข้อมูลจะมีคุณสมบัติเป็นตัวเลข ซึ่งเป็นค่าที่ใช้ในการทำนายผลของข้อ มูลที่มุ่งเน้นในการทำนายค่า	1.ง่ายต่อการอธิบายความสัมพันธ์ระหว่างตัวแปร 2. ผลการพยากรณ์ที่ดีถ้าความ สัมพันธ์ของตัวแปรเป็นแบบเชิงเส้น	1. ไม่เหมาะกับการพยากรณ์ที่ความ สัมพันธ์ของตัวแปรเป็นแบบไม่เชิงเส้น 2. ให้ผลการพยากรณ์เป็นตัวเลขเท่านั้น	1. ผลผลิตการเกษตร 2. ยอดขายสินค้า 3. ความต้องการพลังงาน 4. อุณหภูมิ
3. เอสเอ็มโอาร์อีจี (SMOreg)	มีการเรียนรู้ตามหลักการซัพพอร์ตเวคเตอร์แมชชีนที่หาระนาบการตัดสินใจในการแบ่งข้อมูลตามสมการเส้นตรงเพื่อแบ่งเขตข้อมูลเป็นสองกลุ่มออกจากกัน โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด	1. สามารถคัดแยกข้อมูลในลักษณะความสัมพันธ์เป็นแบบไม่เชิงเส้นได้ดี 2. มีการเรียนรู้จำลองปัญหาเพื่อลดโครงสร้างความเสี่ยงให้น้อยที่สุด	นำเสนอผลการคัดแยกในแนวระนาบตามความสัมพันธ์ระหว่างกลุ่ม ที่ไม่สามารถแปรความสัมพันธ์ที่คัดแยกออกมาได้	1. การคัดแยกปะเภท 2. ผลผลิตทางการเกษตร 3. ตรวจจับใบหน้าหรือลายมือ
4. เอ็มไฟว์พี (M5P)	มีการเรียนรู้ในลักษณะของโครงสร้างต้นไม้ (decision trees) เริ่มด้วยการนำข้อมูลมาจัดหมวดหมู่แบ่งออกในลักษณะของโครงสร้างต้นไม้ โดยสร้างกฎต่าง ๆ เกิดขึ้นตามเป้าหมายของการใช้งานในแต่ละงาน	เป็นเทคนิคที่นิยมเนื่องจากผู้ใช้สามารถทำความเข้าใจผลลัพธ์ได้ง่าย	จะต้องมีการแบ่งให้เป็นข้อมูลแบบไม่ต่อเนื่อง (Discrete Data) จึงไม่เหมาะกับข้อมูลที่เป็นตัวเลข	1. การจำแนกประเภทนักเรียน 2. จำแนกอาการป่วยหรือเป็นโรค 3. การจำแนกประเภทลูกค้า

ค่าสัมประสิทธิ์สหสัมพันธ์ (coefficient of correlation) และค่าความคลาดเคลื่อนสัมพัทธ์ (relative absolute error) โดยวิธีวัดค่าสัมประสิทธิ์สหสัมพันธ์ใช้เพื่อวัดทิศทางและระดับของความสัมพันธ์ระหว่างตัวแปร X และ Y (15-16)

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

r คือสัมประสิทธ์สหสัมพันธ์ระหว่างตัวแปร X และ Y

$\sum X$ คือผลรวมของข้อมูลที่วัดได้จากตัวแปรที่1 (X)

คือผลรวมของข้อมูลที่วัดได้จากตัวแปรที่2 (Y)

คือผลรวมระหว่างผลคูณของข้อมูลระหว่างตัวแปรที่1 และตัวแปรที่ 2

คือผลรวมกำลังสองของข้อมูลที่วัดได้จากตัวแปรที่1 (X)

$\sum Y^2$ คือผลรวมกำลังสองของข้อมูลที่วัดได้จากตัวแปรที่2 (Y)

ส่วนการวัดค่าความคลาดเคลื่อนสัมพัทธ์ เป็นการวัดความคลาดเคลื่อนของค่าจริงและค่าที่พยากรณ์ได้โดยใช้ค่าสัมประสิทธิ์ต่าง ๆ หรือจำนวนข้อมูลต่างๆ จะพิจารณาจากการที่ค่าจริงใกล้เคียงค่าพยากรณ์ที่สุด หรือทำให้เกิดความคลาดเคลื่อนน้อยที่สุด ย่อมเป็นค่าที่เหมาะสมกับการใช้พยากรณ์ให้ได้ผลลัพธ์ที่แม่นยำซึ่งวัดด้วยค่าความคลาดเคลื่อนสัมพัทธ์ คือ ค่าเปอร์เซ็นต์ความคลาดเคลื่อนได้จากค่าความคลาดเคลื่อนได้จากค่าความคลาดเคลื่อนสัมบูรณ์เปรียบเทียบกับค่าจริง (13)

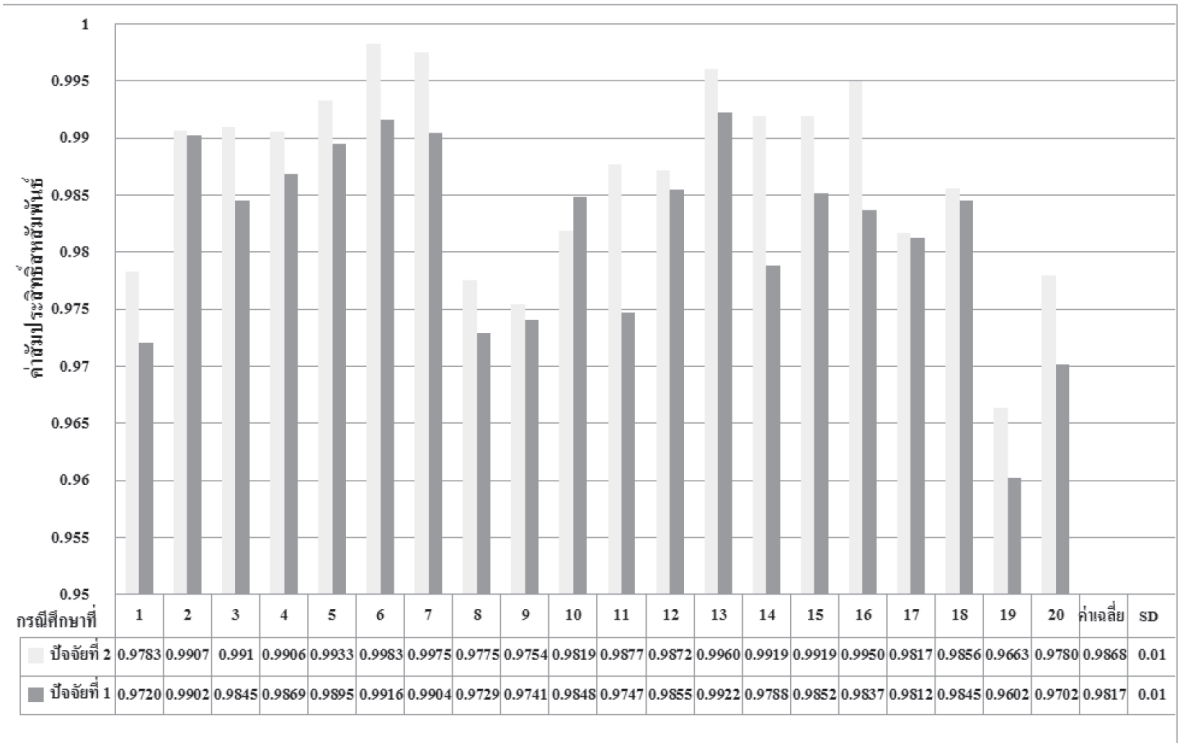
$$\%\text{Error} = \left[\frac{X_t - X_m}{X_t}\right] \times 100\,\%$$

X_t คือ ค่าจริง (True Value)

X_m คือ ค่าที่วัดได้ (Measured Value)

3. ผลการวิจัยและอภิปราย

ข้าวเป็นพืชไร่ที่ต้องอาศัยปัจจัยสิ่งแวดล้อมจากธรรมชาติส่วนใหญ่ ฤดูที่เหมาะสมกับการปลูกข้าว คือ ฤดูฝน และสิ่งที่ต้องคำนึงถึงปัจจัยย่อยอื่น ๆ เช่นปริมาณและการกระจายของน้ำฝน แสงที่มีผลต่อกระบวนการสังเคราะห์แสง อุณหภูมิที่มีผลต่อการงอกของเมล็ด การเจริญเติบโตของราก และการดูดธาตุอาหารของรากพืช (17) นอกจากนี้ปัจจัยที่ส่งผลต่อการผันแปรผลผลิต ยังรวมถึงความชื้นสัมพัทธ์ และอุณหภูมิ เป็นต้น เพื่อศึกษาหาปัจจัยที่เหมาะสมต่อการทำนายแนวโน้มของผลผลิตที่จะเกิดขึ้นในอนาคต ในงานวิจัยนี้ได้ใช้อัลกอริทึมแบบจำลองโครงข่ายประสาทเทียมเพื่อศึกษาความสัมพันธ์ของปัจจัยที่มีเหมาะสมต่อการทำนายผลผลิตข้าว จากการศึกษางานวิจัยของ อรนงค์ บุเกตุ (18) ที่ได้ทำการศึกษาการทำนายผลผลิตพืชไร่จากการพยากรณ์ปริมาณอ้อยของภาคตะวันออกเฉียงเหนือ ทำการประยุกต์ใช้แบบจำลองโครงข่ายประสาทเทียมที่มีกระบวนการเรียนรู้แบบแพร่ย้อนกลับ ที่อาศัยปัจจัยต่าง ๆ ที่มีความสัมพันธ์กับผลผลิตอ้อยซึ่งประกอบด้วย ข้อมูลปริมาณน้ำฝนเฉลี่ย อุณหภูมิเฉลี่ย ผลผลิตอ้อยเฉลี่ยต่อไร่ จึงได้มีการประยุกต์ใช้กับการพยากรณ์ผลผลิตข้าวโดยแบ่งการศึกษาเพิ่มเติมโดยทำการแบ่งออกเป็น 2 ปัจจัยเพื่อศึกษาหาปัจจัยที่เหมาะสมกับการพยากรณ์ ปัจจัยที่ 1 มีการใช้ปัจจัยนำเข้า ดังนี้ ปริมาณน้ำฝนเฉลี่ย อุณหภูมิเฉลี่ย ผลผลิตข้าวเฉลี่ยต่อไร่ ส่วนปัจจัยที่ 2 มีการเพิ่มปัจจัยนำเข้าเพื่อทำการทดสอบ ดังนี้ พื้นที่ปลูกข้าว พื้นที่เก็บเกี่ยว ปริมาณน้ำฝนเฉลี่ย อุณหภูมิเฉลี่ยสูงสุด อุณหภูมิเฉลี่ยต่ำสุด จำนวนวันที่ฝนตก ความชื้นสัมพัทธ์ และผลผลิตข้าวเฉลี่ยต่อไร่ จะเห็นได้ว่าโดยรวมปัจจัยที่ 2 มีประสิทธิภาพของการทำนายดีกว่า ดังนั้นในการวิจัยนี้จึงใช้ข้อมูลพื้นที่ปลูกข้าว พื้นที่เก็บเกี่ยว ปริมาณน้ำฝนเฉลี่ย อุณหภูมิเฉลี่ยสูงสุด อุณหภูมิเฉลี่ยต่ำสุด จำนวนวันที่ฝนตกในแต่ละเดือน ความชื้นสัมพัทธ์ และผลผลิตข้าวเฉลี่ยต่อไร่ เป็นปัจจัยที่มีความสัมพันธ์ต่อการทำนายผลผลิตข้าวที่มีประสิทธิภาพ โดยมีการทำการวิจัย 20 กรณีศึกษาดังแสดงด้วยรูปที่ 2

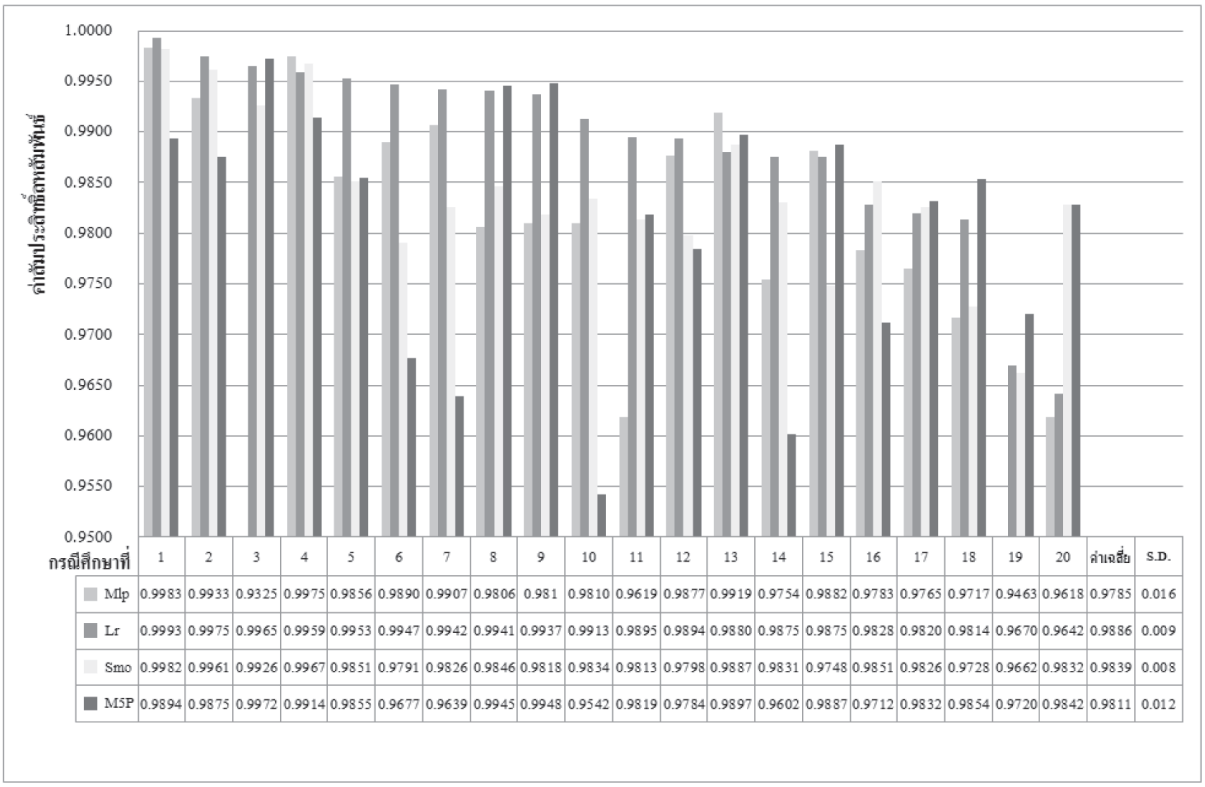


กรณีศึกษาที่	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	ค่าเฉลี่ย	SD
ปัจจัยที่ 2	0.9783	0.9907	0.991	0.9906	0.9933	0.9983	0.9975	0.9775	0.9754	0.9819	0.9877	0.9872	0.9960	0.9919	0.9919	0.9950	0.9817	0.9856	0.9663	0.9780	0.9868	0.01
ปัจจัยที่ 1	0.9720	0.9902	0.9845	0.9869	0.9895	0.9916	0.9904	0.9729	0.9741	0.9848	0.9747	0.9855	0.9922	0.9788	0.9852	0.9837	0.9812	0.9845	0.9602	0.9702	0.9817	0.01

รูปที่ 2. แผนภูมิแท่งแสดงค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างปัจจัยที่ 1 และปัจจัยที่ 2

จากรูปที่ 2 จะเห็นว่าจากการศึกษาปัจจัยที่มีผลต่อการทำนายซึ่งเป็นปัจจัยทางธรรมชาติ และเป็นปัจจัยที่ควบคุมไม่ได้ ผลของการศึกษาปัจจัยที่ 1 ในบางกรณีศึกษามีค่าใกล้เคียงและมีค่าสูงกว่าปัจจัยที่ 2 โดยปัจจัยที่ 1 มีค่าเฉลี่ยของสัมประสิทธิ์สหสัมพันธ์โดยรวมเป็น 0.9817 แต่ในภาพรวมของกรณีศึกษาในปัจจัยที่ 2 มีค่าสูงกว่าซึ่งมีค่าเฉลี่ยของสัมประสิทธิ์สหสัมพันธ์โดยรวมเป็น 0.9868 ดังนั้นจะเห็นได้ว่าปัจจัยที่ 2 มีค่าความสัมพันธ์ที่ดีกว่าปัจจัยที่ 1 จึงเป็นปัจจัยที่เหมาะสมกับการพยากรณ์ผลผลิตข้าว

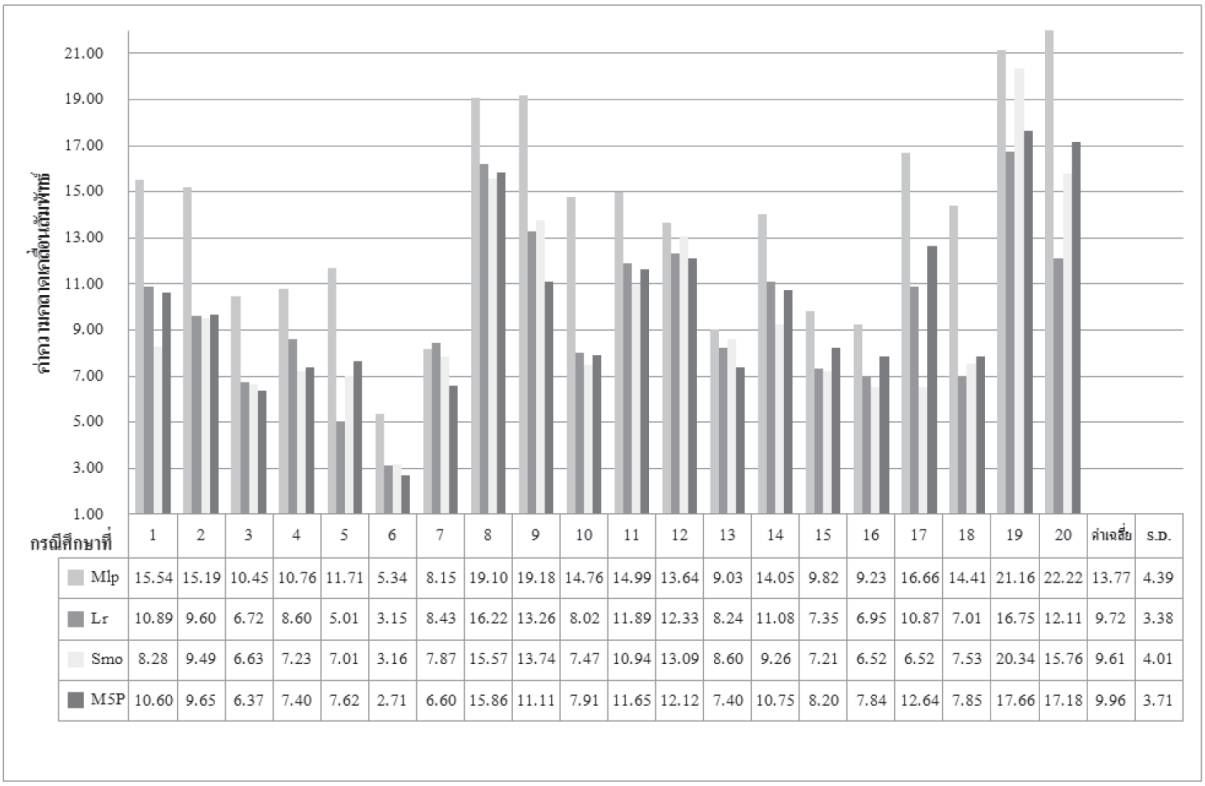
เมื่อได้ปัจจัยที่เหมาะสม ได้ทำการประยุกต์ใช้เทคนิคเหมืองข้อมูล คือ แบบจำลองโครงข่ายประสาทเทียม ทำการเปรียบเทียบกับอีก 3 เทคนิค คือ ซัพพอร์เวคเตอร์แมชชีน การถดถอยเชิงเส้น และต้นไม้การตัดสินใจ เพื่อนำมาสร้างแบบจำลองโดยเลือกอัลกอริทึมการเรียนรู้ที่ใช้ตามลำดับ คือ multilayer perceptron, SMOreg, linear regression และ M5P มาสร้างเป็นแบบจำลองเพื่อใช้ในการทำนายผลผลิตข้าว ผลของการทดสอบข้อมูลแสดงในรูปที่ 3 จะเป็นการทดสอบวัดค่าสัมประสิทธิ์สหสัมพันธ์เพื่อใช้วัดทิศทางและระดับของความสัมพันธ์ระหว่างข้อมูลในการพยากรณ์ จะเห็นได้ว่าในแต่อัลกอริทึมมีค่าใกล้เคียงกันมาก แต่เมื่อดูในภาพรวม linear regression มีค่าสัมประสิทธิ์สหสัมพันธ์ในการทำนายโดยรวมมีค่าเฉลี่ยเท่ากับ 0.9886 ซึ่งเป็นแบบจำลองที่ดีที่สุดของการทดสอบและยังมีค่าค่าส่วนเบี่ยงเบนมาตรฐานเข้าใกล้ 0 แสดงว่าไม่ค่อยมีการกระจายของข้อมูล ซึ่งค่าของข้อมูลโดยรวมไม่แตกต่างกันมาก รองลงมาคือ SMOreg และต้นไม้การตัดสินใจ มีค่าประสิทธิ์สหสัมพันธ์ในการทำนายโดยรวมมีค่าเฉลี่ยเท่ากับ 0.9839 และ 0.9811 ตามลำดับ ส่วน multilayer perceptron แสดงผลลัพธ์สัมประสิทธิ์สหสัมพันธ์ในการทำนายโดยรวมมีค่าเฉลี่ยน้อยที่สุด เท่ากับ 0.9785



กรณีศึกษาที่	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	ค่าเฉลี่ย	S.D.
Mlp	0.9983	0.9933	0.9325	0.9975	0.9856	0.9890	0.9907	0.9806	0.981	0.9810	0.9619	0.9877	0.9919	0.9754	0.9882	0.9783	0.9765	0.9717	0.9463	0.9618	0.9785	0.016
Lr	0.9993	0.9975	0.9965	0.9959	0.9953	0.9947	0.9942	0.9941	0.9937	0.9913	0.9895	0.9894	0.9880	0.9875	0.9875	0.9828	0.9820	0.9814	0.9670	0.9642	0.9886	0.009
Smo	0.9982	0.9961	0.9926	0.9967	0.9851	0.9791	0.9826	0.9846	0.9818	0.9834	0.9813	0.9798	0.9887	0.9831	0.9748	0.9851	0.9826	0.9728	0.9662	0.9832	0.9839	0.008
M5P	0.9894	0.9875	0.9972	0.9914	0.9855	0.9677	0.9639	0.9945	0.9948	0.9542	0.9819	0.9784	0.9897	0.9602	0.9887	0.9712	0.9832	0.9854	0.9720	0.9842	0.9811	0.012

รูปที่ 3. แผนภูมิแท่งแสดงค่าสัมประสิทธิ์สหสัมพันธ์ของแต่ละอัลกอริทึม

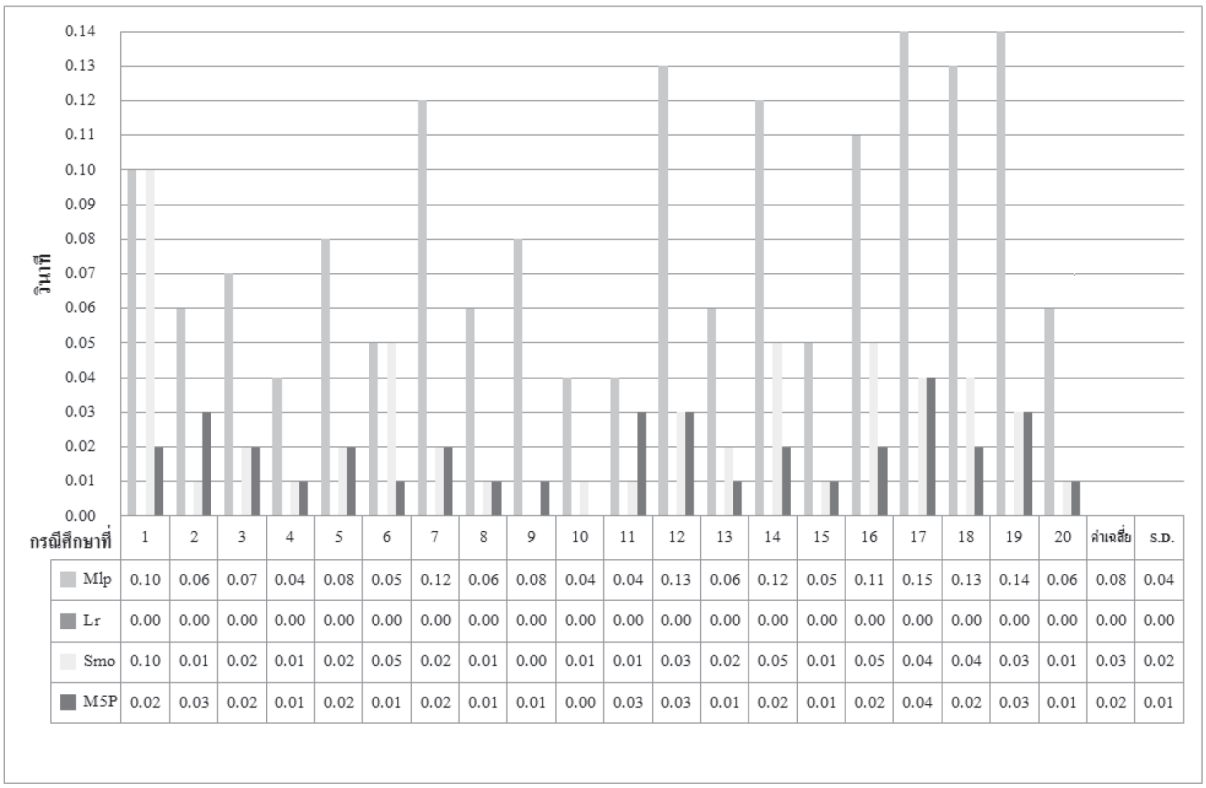
หมายเหตุ Mlp คือ multilayer perceptron, lr คือ linear regression, Smo คือ SMOreg และ M5P คือ M5P



กรณีศึกษาที่	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	ค่าเฉลี่ย	S.D.
Mlp	15.54	15.19	10.45	10.76	11.71	5.34	8.15	19.10	19.18	14.76	14.99	13.64	9.03	14.05	9.82	9.23	16.66	14.41	21.16	22.22	13.77	4.39
Lr	10.89	9.60	6.72	8.60	5.01	3.15	8.43	16.22	13.26	8.02	11.89	12.33	8.24	11.08	7.35	6.95	10.87	7.01	16.75	12.11	9.72	3.38
Smo	8.28	9.49	6.63	7.23	7.01	3.16	7.87	15.57	13.74	7.47	10.94	13.09	8.60	9.26	7.21	6.52	6.52	7.53	20.34	15.76	9.61	4.01
M5P	10.60	9.65	6.37	7.40	7.62	2.71	6.60	15.86	11.11	7.91	11.65	12.12	7.40	10.75	8.20	7.84	12.64	7.85	17.66	17.18	9.96	3.71

รูปที่ 4. แผนภูมิแท่งแสดงค่าความคลาดเคลื่อนสัมพัทธ์ของแต่ละอัลกอริทึม

หมายเหตุ Mlp คือ multilayer perceptron, lr คือ linear regression, Smo คือ SMOreg และ M5P คือ M5P



กรณีศึกษาที่	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	ค่าเฉลี่ย	S.D.
Mlp	0.10	0.06	0.07	0.04	0.08	0.05	0.12	0.06	0.08	0.04	0.04	0.13	0.06	0.12	0.05	0.11	0.15	0.13	0.14	0.06	0.08	0.04
Lr	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Smo	0.10	0.01	0.02	0.01	0.02	0.05	0.02	0.01	0.00	0.01	0.01	0.03	0.02	0.05	0.01	0.05	0.04	0.04	0.03	0.01	0.03	0.02
M5P	0.02	0.03	0.02	0.01	0.02	0.01	0.02	0.01	0.01	0.00	0.03	0.03	0.01	0.02	0.01	0.02	0.04	0.02	0.03	0.01	0.02	0.01

รูปที่ 5. แผนภูมิแท่งแสดงเวลาการประมวลผลของแต่ละอัลกอริทึม

หมายเหตุ Mlp คือ multilayer perceptron, lr คือ linear regression, Smo คือ SMOreg และ M5P คือ M5P

จากผลของการทำนายในส่วนของค่าความคลาดเคลื่อนสัมพัทธ์ที่แสดงความผิดพลาดของการทำนาย ดังแสดงในรูปที่ 4 จะเห็นได้ว่าอัลกอริทึมที่มีข้อผิดพลาดน้อยที่สุดของการทดสอบคือ SMOreg มีค่าความผิดพลาดเฉลี่ยโดยรวมน้อยที่สุด เท่ากับ 9.61 รองลงมาคือ linear regression และ MP5 มีค่าค่าความคลาดเคลื่อนสัมพัทธ์เฉลี่ยใกล้เคียงกัน 9.72 และ 9.96 ตามลำดับ ในขณะที่ multilayer perceptron ให้ผลลัพธ์ของความผิดพลาดเฉลี่ยที่เกิดขึ้นมากที่สุด เท่ากับ 13.77 ดังนั้นจะเห็นได้ว่าในการทดสอบนี้จึงไม่ควรนำมาใช้ในการทำนายผลผลิต ในส่วนของการพิจารณาเรื่องของเวลาที่ใช้ในการประมวลผล จะเห็นได้ว่า linear regression ใช้เวลาน้อยที่สุดมีค่าเฉลี่ยเป็น 0 วินาที ดังแสดงในรูปที่ 5 แต่ multilayer perceptron กลับใช้เวลาเฉลี่ยมากที่สุดมีค่าเป็น 0.08 วินาที ดังนั้นจะเห็นได้ว่าการทำนายผลผลิตข้าวที่มีประสิทธิภาพมากที่สุดในภาพรวมของการทดสอบนี้ คือ linear regression เนื่องจากเป็นอัลกอริทึมที่มีทั้งความสามารถในการทำนายข้อมูลที่ถูกต้องความแม่นยำและรวดเร็วของการทำนายอาศัยปัจจัยเหมาะสมต่อการทำนายดังนี้ ข้อมูลพื้นที่ปลูกข้าว พื้นที่เก็บเกี่ยว ปริมาณน้ำฝนเฉลี่ย อุณหภูมิเฉลี่ยสูงสุด อุณหภูมิเฉลี่ยต่ำสุด จำนวนวันที่ฝนตกในแต่ละเดือน ความชื้นสัมพัทธ์ และผลผลิตข้าวเฉลี่ยต่อไร่ ในอนาคตสามารถนำแบบจำลองที่ได้ไปประยุกต์กับวิธีการทำนายผลผลิตของพืชไร่ชนิดอื่น ๆ ได้เพื่อทำนายผลผลิตทางการเกษตร

4. สรุป

ไทยกำลังประสบปัญหาผลผลิตข้าวต่อไร่น้อยกว่าที่ควรจะเป็น ในขณะที่พื้นที่เพาะปลูกลดลง หากมีเทคโนโลยีสารสนเทศมาช่วยสนับสนุนการตัดสินใจ ส่งผลความเสี่ยงของการปลูกข้าวที่ได้ผลผลิตต่อไร่ที่ลดลง จากการศึกษาพบว่าเนื่องจากปัจจัยที่เลือกทำนายที่มีความแตกต่างกันก็จะส่งผลให้ค่าความผิดพลาดในการทำนายแตกต่างด้วยเช่นกัน สำหรับงานวิจัยนี้ได้จัดกลุ่มข้อมูลให้ความใกล้ชิดของข้อมูลแคบลงเพื่อลดค่าความผิดพลาดที่เกิดขึ้น ซึ่งผลของการวิจัยพบว่า 8 ปัจจัยนำเข้าที่เหมาะสม

ต่อการทดสอบนี้ มีดังนี้ ข้อมูลพื้นที่ปลูกข้าว พื้นที่เก็บเกี่ยว ปริมาณน้ำฝนเฉลี่ย อุณหภูมิเฉลี่ยสูงสุด อุณหภูมิเฉลี่ยต่ำสุด จำนวนวันที่ฝนตกในแต่ละเดือน ความชื้นสัมพัทธ์ และผลผลิตข้าวเฉลี่ยต่อไร่ นำปัจจัยนำเข้าดังกล่าวมาทดสอบประสิทธิภาพของแต่ละเทคนิคของอัลกอริทึม ผลการศึกษากลับพบว่าแบบจำลองโครงข่ายประสาทเทียม มีค่าความผิดพลาดเฉลี่ยมากที่สุด ในการทำนายผลผลิตข้าวนี้ นี้จึงไม่เหมาะกับการที่จะนำมาใช้ แต่ในทางตรงกันข้าม ซัพพอร์เวคเตอร์แมชชีน และการถดถอยเชิงเส้น กลับมีค่ามีค่าประสิทธิ์สหสัมพันธ์ในการทำนายเฉลี่ยและความผิดพลาดเฉลี่ยใกล้เคียงกัน แต่ในการทำนายที่มีประสิทธิภาพควรจะต้องมีความสามารถทำนายข้อมูลที่ถูกต้อง แม่นยำ และรวดเร็ว การถดถอยเชิงเส้นจะเป็นอัลกอริทึมที่มีประสิทธิภาพโดยรวมดีที่สุดสำหรับการทดสอบนี้ ไม่ว่าจะเป็นค่าสัมประสิทธิ์สหสัมพันธ์ของการทำนาย ค่าความคลาดเคลื่อนสัมพัทธ์ และเวลาที่ใช้ในการประมวลผล เนื่องมากจากการทำงานของอัลกอริทึมการถดถอยเชิงเส้นเหมาะที่จะนำไปใช้ตีความหรือวิเคราะห์ข้อมูลเชิงตัวเลข ต่อไปสามารถนำแบบจำลองที่ได้จากการทำนายผลไปพัฒนาสร้างระบบเทคโนโลยีสารสนเทศเพื่อทำนายผลผลิตข้าวต่อไป โดยระบบที่ได้สามารถนำข้อมูลของการทำนายไปใช้ประกอบการตัดสินใจหรือให้คำแนะนำการปลูกข้าวแก่เกษตรกรได้เพื่อส่งผลให้ได้ผลผลิตที่สูงขึ้นและลดความเสี่ยงที่จะปลูกข้าวได้ผลผลิตต่ำ ซึ่งงานวิจัยในขั้นต่อไปนี้คือ การนำข้อมูลที่ได้ทำการศึกษาและจากผลการวิจัยที่ได้กล่าวมาแล้วข้างต้น ไปสร้างแบบจำลองที่มีการบูรณาการร่วมกันกับตัวแทนปัญญา (intelligent agent) (19) เพื่อให้ระบบทำนายผลผลิตข้าวสามารถมีการเรียนรู้เพิ่มขึ้นและทำงานได้ด้วยตัวเองเพื่อทำให้การทำนายหรือการเอาไปใช้งานประกอบการตัดสินใจมีประสิทธิภาพมากยิ่งขึ้น

5. กิตติกรรมประกาศ

ขอขอบพระคุณส่วนปฏิบัติการข้อมูลการเกษตร สังกัดสำนักงานเศรษฐกิจการเกษตร และสำนักพัฒนาอุตุนิยมวิทยา กรมอุตุนิยมวิทยาที่ได้สนับสนุนการให้ข้อมูล

6. เอกสารอ้างอิง

(1) United States Department of Agriculture. Foreign Agricultural Service: Grain:World Markets and Trade. [Internet] 2013 Aug 1 [updated 2013 Aug; cited 2013 Aug 8]. Available from: www.fas.usda.gov/psdonline/circulars/grain.pdf.

(2) Pisanwanich A. A Comparison of Rice Production and Trade Competitiveness between Thailand and Vietnam in the ASEAN Market. University of the Thai Chamber of Commerce Journal. Vol. 31 No. 2011 April – June 2; 158-170.

(3) Konphang A.. Paddy price forecasting using data mining technique. The 23rd Nation graduate research conference. 2011; 82-87.

(4) Raorane A.A., Kulkarni R.V. Data Mining: An effective tool for yield estimation in the agricultural sector. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). 1. 2012; 75-79.

(5) Pinyosri D., Chiewchanwattana S., Sunat K. Sugarcane Yield Estimation Modeling using Artificial Neural Network. Proceedings of The 13rd. National Computer Science and Engineering Conference: NCSEC2009. 2009 November 4-6; 221-225.

(6) Larose D.T. Discovery Knowledge in Data : An Introduction to Data Mining. New York : John Wiley. 2005; 30-33.

(7) Piriyakul M. Missing data replacement models in Social Science research =simulation study of simple models. 2005; [cited 2013 Aug 10]. Available from: ThaiLIS Digital Collection (ThaiLIS).

(8) Jaikla R. Rice yield prediction using Fuzzy Inference System and Support Vector Machine. 2009; [cited 2013 Aug 10]. Available from: ThaiLIS Digital Collection (ThaiLIS).

(9) Kijsirikul B. Data Mining Algorithms, the final report on the Joint Government and Private Sectors, Bangkok,: Chulalongkorn University Press, 2004; 18-27.

(10) Witten I. H. Frank E. Data Mining : Practical Machine Learning Tools and Techniques. 2nd ed. California : Morgan Kaufmann Publishers. 2005; 256-233: 408: 413-414.

(11) Kantanantha N. Forecasting by Causal Methods. Engineering Journal . Volume 4 Issue 1. 2012; 33-38.

(12) Thinsungnoen T., Kulnawin K., Thinsungnoen M. Applying data mining technique to discover the factors contributing to students'inability to graduate on time. Payap University Research Symposium 2012. 2012 Febuary 17; 34-42.

(13) Han J., Kamber M. Data Mining Concepts and Techniques. 2nd ed. California : Morgan Kaufmann Publishers. 2006; 5-9:47-96:363.

(14) Alpaydin E. Introduction To Machine Learning. MIT press. 2006; 331-334.

(15) Cios K., Pedrycz W., Swiniarski, R., Kurgan L. Data Mining: A Knowledge Discovery Approach. New York : Springer. 2007; 360-361.

(16) Shahnawaz M., Saxena K. A Comparative Study of Various Regression Model for Data Farming. International Journal of Wisdom Based Computing, Vol. 2(1). 2012 April; 29-34.

(17) Iamsupasit N. Plants and environment. The course lectures the economics plant units 1-7. Department of Agriculture Extensional and Co-Op. Sukhothai Thammathirat Open University. 2004; 149-219.

(18) Bukate O. Sugarcane forecasting model of northeast by artificial neural network. KKU Science Journal Volume 41 Number 1; 213-225.

(19) Kulnawin K., Longpradit , P. Design of Quality Information Technology Model to Predict Crop Yield by Six Sigma technique. RSU-UTCC-NDSI-TCC graduate student conference 2011. 2011 September; 1159-1170.