การวิเคราะห์ปัจจัยสำคัญของเศรษฐกิจครัวเรือนด้วย การคัดเลือกคุณสมบัติ Gain Ratio Feature Selection

วชิราภรณ์ เจริญมา

โครงงานคอมพิวเตอร์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
สาขาวิชาคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสกลนคร
พ.ศ. 2564

สารบัญ

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

การฟื้นตัวในระยะข้างหน้าจะมีความแตกต่างกันในแต่ละกลุ่มธุรกิจและครัวเรือนที่สำคัญ หากการระบาดของโควิด-19 กินระยะเวลายาวนานขึ้น ความแตกต่างของการฟื้นตัว จะยิ่งทวีความรุนแรงมากขึ้น ซึ่งในท้ายที่สุดจะไปซ้ำเติมความเปราะบางของโครงสร้างเศรษฐกิจไทยที่เป็นต้นทุนอยู่เดิม (Rachot Leingchan. 2564: ออนไลน์) ซึ่งมหาวิทยาลัยราชภัฏสกลนครมีนโยบายและพันธกิจในการพัฒนาท้องถิ่นอย่างยั่งยืนบนพื้นฐานเศรษฐกิจพอเพียงเพื่อพัฒนาท้องถิ่น โดยมีโครงการน้อมนำศาสตร์พระราชาสู่การพัฒนาท้องถิ่น ยกระดับรายได้ซึ่งได้ดำเนินการมาตั้งแต่ปี 2561 จนถึงปัจจุบัน (มรสน. 2560: ออนไลน์) ในการพัฒนาสภาพทางเศรษฐกิจของคนในชุมชนซึ่งมหาวิทยาลัยราชภัฏสกลนคร ได้รับผิดชอบ 19 ตำบลในจังหวัดสกลนคร โดยมีการเก็บข้อมูลสภาพทางเศรษฐกิจไว้ในเป็นฐานข้อมูล ส่วนกลางของมหาวิทยาลัยราชภัฏกลนคร ซึ่งปัจจุบันยังไม่มีการวิเคราะห์ปัจจัยที่ส่งผลต่อสภาพเศรษฐกิจครัวเรือนที่จะสามารถสนับสนุนการตัดสินใจหรือการวางแผนการพัฒนาชุมชน

เทคโนโลยีในปัจจุบันมีหลากหลายเทคโนโลยีและมีหลากหลายศาสตร์ที่จะสามารถนำมาวิเคราะห์ข้อมูลขนาดใหญ่ได้โดยเฉพาะอย่างยิ่งหลักการทำเหมืองข้อมูลซึ่งเป็นหลักการในการวิเคราะห์เป็นหลักการของสถิติขั้นสูงที่จะสามารถวิเคราะห์ข้อมูลขนาดใหญ่ได้ โดยมีนักวิจัยหลากหลายศาสตร์ที่ได้นำเทคนิคการทำเหมืองข้อมูลนำมาประยุกต์ใช้งาน เช่น งานวิจัยของ นิภาพร ชนะมาร และพรรณี สิทธิเดช (2557) ได้ศึกษาการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนิสิตที่ศึกษาและสำเร็จการศึกษาแล้วในหลักสูตรเดียวกัน จำนวน 180 ระเบียน โดยใช้เทคนิคการคัดเลือกคุณสมบัติที่สำคัญ แล้วสร้างตัวแบบการพยากรณ์ด้วยเทคนิค BPNN และเทคนิค SVMs เมื่อทดลองใช้เทคนิคการรวมกลุ่ม ด้วยวิธี Bagging ร่วมกับ BPNN และ SVMs พบว่าผลการพยากรณ์ของ Baggingร่วมกับ BPNN (Bagging BPNN) มีค่าความผิดพลาดอยู่ใน ระดับต่ำสุด (RMSE=0.1051) งานของ วรายุทธ พลาศรี (2556) ได้ศึกษาการศึกษาปัจจัยที่มีผลต่อความยากจนของครัวเรือนในชนบท : กรณีศึกษาจังหวัดมหาสารคาม การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาสภาพเศรษฐกิจของครัวเรือนในชนบท สถานการณ์ความยากจน ลักษณะของครัวเรือนที่ยากจน และปัจจัยที่มีผลต่อความยากจนของครัวเรือนในชนบทจังหวัดมหาสารคาม โดยกลุ่มประชากรที่ใช้ในการศึกษาคือ ครัวเรือนที่อยู่ ในเขตพื้นที่ชนบทจังหวัดมหาสารคาม จำนวน 180,328 ครัวเรือน ขนาดกลุ่มตัวอย่างเท่ากับ 400 ครัวเรือน และงานวิจัยของ อัจจิมา มณฑาพันธุ์

(2562) งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเกี่ยวกับการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญเพื่อนำมาใช้ในการ ปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม โดยใช้วิธีการคัดเลือกคุณลักษณะจากเทคนิคต่าง ๆ จำนวน 7 เทคนิค ได้แก่เทคนิค Correlation Based Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection หลังจากคัดเลือกคุณลักษณะ ที่สำคัญจึงนำผลที่ได้จากแต่ละเทคนิคมาคำนวณหาค่าประสิทธิภาพในการพยากรณ์การเป็นมะเร็งเต้านมโดยการใช้ทคนิคซัพพอร์ต เวกเตอร์แมชชีน ผลการทดลองพบว่าร้อยละของความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านม จากจำนวนคุณลักษณะของข้อมูลทั้งหมด 30 คุณลักษณะเท่ากับ 91.39

จากที่กล่าวมาข้างต้นผู้วิจัยจึงมีความสนใจที่จะศึกษาวิเคราะห์ปัจจัยสำคัญทางเศรษฐกิจครัวเรือนด้วยวิธีการ Gain Ratio Feature Selection โดยใช้โปรแกรม RapidMiner เพื่อให้ได้ปัจจัยที่สำคัญสำหรับการวางแผนและการพัฒนาด้านเศรษฐกิจของชุมชนท้องถิ่นเพื่อให้สามารถมีคุณภาพชีวิตที่ดีขึ้น

1.2 วัตถุประสงค์ของโครงงาน

เพื่อวิเคราะห์ปัจจัยสำคัญของเศรษฐกิจครัวเรือนด้วยการคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection

1.3 ขอบเขตโครงงาน

1.3.1 ด้านข้อมูล

ข้อมูลที่ใช้ในการศึกษาครั้งนี้ คือข้อมูลประชากรจากภาคครัวเรือนเฉพาะครัวเรือนในเขตพื้นที่ชนบทของจังหวัดสกลนคร ซึ่งมี 20 หมู่บ้าน 12 ตำบล 12 อำเภอ โดยช่วงเวลาที่ทำการเก็บรวบรวมข้อมูลคือ ปี พ.ศ. 2563 – 2564 (สำนักวิทยบริการและเทคโนโลยีสารสนเทศ, 2563: ออนไลน์)

1.3.1.1 ประชากรที่ใช้ในการศึกษา ได้แก่ ครัวเรือนตำบลที่อยู่ในช่วงปี พ.ศ. 2561 - 2563 ได้มาจากข้อมูล 12 ตำบล ซึ่งมีจำนวน 17,933 ครัวเรือน

1.3.1.2 กลุ่มตัวอย่างที่ใช้ในการศึกษาและวิเคราะห์ข้อมูล ได้แก่ กลุ่มครัวเรือนบ้านในช่วงปี พ.ศ. 2561 - 2563 ได้มาจากการเลือกแบบเจาะจง (Purposive Sampling) จำนวน 3,233 ครัวเรือน

ข้อมูลจากฐานข้อมูล สภาพทางเศรษฐกิจครัวเรือนเป้าหมายตามโครงการจ้างงาน ประชาชนที่ได้รับผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า

2019 (COVID-19) โดยในฐานข้อมูลมีการเก็บข้อมูลจากฐานเศรษฐกิจชุมชนซึ่งมีการเก็บข้อมูลออกเป็น 10 ส่วน ดังนี้

ส่วนที่ 1 ข้อมูลทั่วไปครัวเรือน

ส่วนที่ 2 ทรัพย์สินของครัวเรือน

ส่วนที่ 3 อาชีพและรายได้ของครัวเรือน

ส่วนที่ 4 รายจ่ายของครัวเรือน

ส่วนที่ 5 หนี้สินของครัวเรือน

ส่วนที่ 6 ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19)

ส่วนที่ 7 การใช้เทคโนโลยีสารสนเทศ

ส่วนที่ 8 การเข้าร่วมการละเล่น การฟ้อน การรำ พิธีกรรมตามวิถีวัฒนธรรมชุมชน

ส่วนที่ 9 การเข้าร่วมโครงการที่ผ่านมาย้อนหลัง 3 ปี

ส่วนที่ 10 ข้อคิดเห็นและข้อเสนอแนะเพิ่มเติม และได้ศึกษางานวิจัยที่เกี่ยวข้องนำมาประยุกต์ใช้ในฐานข้อมูล Database โดยการใช้กระบวนการ CRIPS-DM (Cross Reference Industry Standard for Data Mining) ในการดำเนินงาน

1.3.2 ด้านเทคนิค

1.3.2.1 ใช้เทคนิค Gain Ratio Feature Selection เป็นเทคนิคที่ใช้ในการหาปัจจัยสำคัญ

1.3.2.2 เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เพื่อมาหาประสิทธิภาพเปรียบเทียบระหว่างปัจจัยเริ่มต้นและการหาปัจจัยสำคัญจากเทคนิค Gain Ratio Feature Selection มาเปรียบเทียบประสิทธิภาพจากนั้นจะได้มาซึ่งปัจจัยที่เหมาะสมที่สุด

1.3.2.3 ใช้กระบวนการ CRIPS-DM (Cross Reference Industry Standard for Data Mining) เป็นกระบวนการที่ใช้ในการดำเนินงานวิจัย

1.3.3 ด้านเครื่องมือในการพัฒนา

1.3.3.1 ซอฟต์แวร์

การศึกษาครั้งนี้ได้ทำการทดลองดำเนินการผ่านโปรแกรม RapidMiner Studio เวอร์ชัน 9.10 เป็นโปรแกรมที่ออกแบบมาสำหรับการวิเคราะห์ข้อมูล ของบริษัท RapidMiner (mypccrack 2565 : ออนไลน์)

1.3.3.2 ฮาร์ดแวร์

เครื่องคอมพิวเตอร์ Notebook Asus

-หน่วยประมวลผล ADM Ryzen 5 2500U with Radeon Vega Mobile Gfx 2.00 GHz

-หน่วยความจำหลัก (RAM): 8.00 GB

-ระบบปฏิบัติการ (OS): Windows 10 64-bit

1.4 แผนการดำเนินงาน

1.4.1 กำหนดหัวข้อและนำเสนอหัวข้อ

1.4.2 ค้นหาปัญหา โอกาสและเป้าหมาย

1.4.3 ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1.4.4 เสนอเค้าโครงโครงงาน

1.4.5 ศึกษาและวิเคราะห์ข้อมูล

1.4.6 ทำความเข้าใจข้อมูลและเตรียมข้อมูล

1.4.7 ดำเนินการพัฒนาโมเดล

1.4.8 ประเมินประสิทธิภาพการพัฒนาโมเดล

1.4.9 จัดทำเอกสารประกอบโครงงาน

1.4.10 นำเสนอโครงงานจบ

ตารางที่ 1.1 ระยะเวลาการดำเนินงาน

กิจกรรม	ม.ค	ก.พ	มี.ค	เม.ย	พ.ค
1.4.1 กำหนดหัวข้อและนำเสนอหัวข้อ	→				
1.4.2 ค้นหาปัญหา โอกาสและเป้าหมาย	→				
1.4.3 ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง	→				
1.4.4 เสนอเค้าโครงโครงงาน	→				
1.4.5 ศึกษาและวิเคราะห์ข้อมูล	→				
1.4.6 ทำความเข้าใจข้อมูลและเตรียมข้อมูล	→				
1.4.7 ดำเนินการพัฒนาโมเดล	→				
1.4.8 ประเมินประสิทธิภาพการพัฒนาโมเดล	→				
1.4.9 จัดทำเอกสารประกอบโครงงาน	→				
1.4.10 นำเสนอโครงงานจบ	→				

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ได้ปัจจัยที่สำคัญที่ส่งผลต่อสภาพทางเศรษฐกิจครัวเรือนเพื่อสนับสนุนหรือเป็นข้อมูลประกอบการตัดสินใจในการพัฒนาชุมชนท้องถิ่นสำหรับนักวิจัย

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ผู้วิจัยได้นำเสนอเนื้อหาที่เน้นถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง รวมถึงเอกสารและงานเขียนอื่นๆ ที่เกี่ยวข้องกับงานวิจัยโดยในบทนี้จะแบ่งเนื้อหาหลักๆ ออกเป็น 2 หัวข้อ ประกอบด้วย

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การทำเหมืองข้อมูล (Data Mining)

2.1.1.1 การทำเหมืองข้อมูล จำแนกออกเป็น 2 ประเภท

2.1.2 ต้นไม้ตัดสินใจ (Decision Tree)

2.1.3 Feature selection การคัดเลือกคุณสมบัติ

2.1.3.1 การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection

2.1.4 ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล

2.1.4.1 ความแม่นยำ (Accuracy)

2.1.5 CRISP-DM (Cross-Industry Standard Process For Data Mining)

2.1.6 โปรแกรม RapidMiner Studio

2.2 งานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

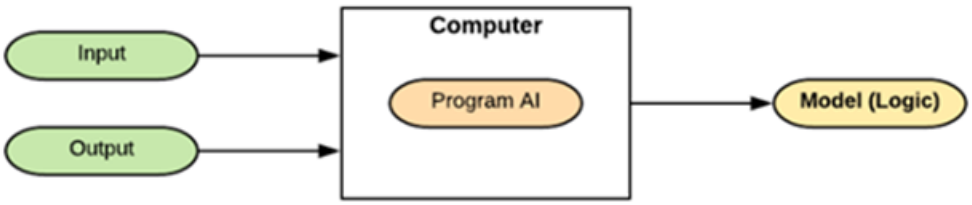
2.1.1 การทำเหมืองข้อมูล (Data Mining)

เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า เหมืองข้อมูล นั่นคือ เป็นการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ตต่าง ๆ โดยข้อมูลเหล่านี้จะเก็บจากรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้ง โดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining แล้วจะได้สิ่งที่เป็นประโยชน์ Data Mining เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า เหมืองข้อมูล นั่นคือ เป็นการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ตต่างๆ โดยข้อมูลเหล่านี้จะเก็บจากรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้ง โดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining แล้วจะได้สิ่งที่เป็นประโยชน์เช่น ลูกค้าส่วนใหญ่ที่ซื้อเบียร์มักจะซื้อผ้าอ้อมด้วย จะเห็นว่าข้อมูลนี้เป็นข้อมูลที่ไม่เคยคิดว่ามีความสัมพันธ์กัน และเมื่อได้ความรู้แบบนี้ก็อาจจะนำเป็นออกโปรโมชั่นหรือช่วยในการจัดวางชั้นสินค้า หรือเป็นแนวทางในการสั่งซื้อสินค้าในซุปเปอร์มาร์เก็ตต่อไปได้ นอกจากนี้ Data Mining ยังมีเทคนิคในการประยุกต์ใช้งานได้อย่างดี (หนึ่งหทัย ชัยอาภร.2559: ออนไลน์)

2.1.1.1 การทำเหมืองข้อมูล จำแนกออกเป็น 2 ประเภท คือ

1) Unsupervised Learning การสร้างโมเดลโดยใช้ข้อมูล input เพียงอย่างเดียวไม่มี target การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่อง โดยการสร้างโมเดลที่เหมาะสมกับข้อมูล การเรียนรู้แบบนี้แตกต่างจากการเรียนรู้แบบมีผู้สอน คือ จะไม่มีการระบุผลที่ต้องการหรือประเภทไว้ก่อน การเรียนรู้แบบนี้จะพิจารณาวัตถุเป็นเซตของตัวแปรสุ่ม แล้วจึงสร้างโมเดลความหนาแน่นร่วมของชุดข้อมูลการเรียนรู้แบบไม่มีผู้สอนสามารถนำไปใช้ร่วมกับการอนุมานแบบเบย์ เพื่อหาความน่าจะเป็นแบบมีเงื่อนไขของตัวแปรสุ่มโดยกำหนดตัวแปรที่เกี่ยวข้องให้ นอกจากนี้ยังสามารถนำไปใช้ในการบีบอัดข้อมูล ซึ่งโดยพื้นฐานแล้ว ขั้นตอนวิธีการบีบอัดข้อมูลจะขึ้นอยู่กับ การแจกแจงความน่าจะเป็นของข้อมูลไม่อย่างชัดแจ้งก็โดยปริยาย (สุพรรณ ฟ้าหยง. 2562)

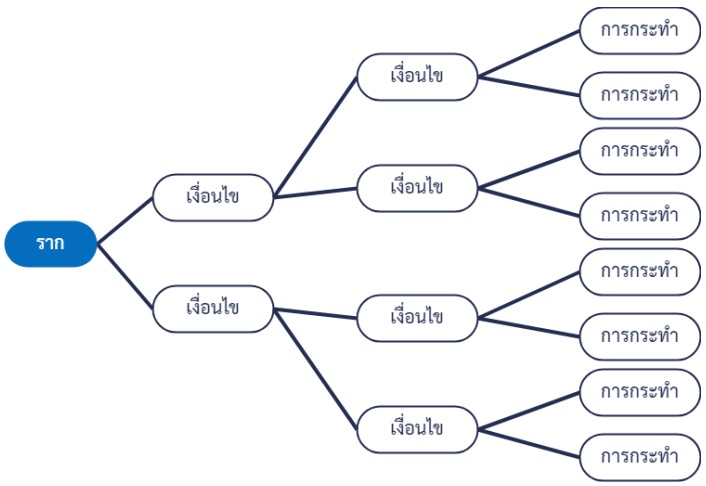
2) Supervised Learning เป็นการเรียนรู้ข้อมูลต่าง ๆ โดยมีผู้สอน อาศัยข้อมูลในการฝึกฝน เพื่อช่วยให้ตัวเทคโนโลยีสามารถเรียนรู้ผล และคาดคะเนผลลัพธ์ต่าง ๆ ได้อย่างแม่นยำมากยิ่งขึ้น โดยการเรียนรู้ในรูปแบบนี้มักถูกนำมาใช้งานในเชิงธุรกิจทั้งการคำนวณราคาบ้าน การคาดคะเนค่าเงิน หรือแม้แต่การวิเคราะห์ผลการแข่งขันต่าง ๆ เป็นต้น กระบวนการสร้าง model เรียกว่าการ “เทรน” ซึ่งสามารถกินเวลาได้ตั้งแต่หลักวินาทีจนถึงหลายๆ วัน แล้วแต่ความซับซ้อนของโจทย์ที่เราต้องการแก้ และพลังในการประมวลผลของเครื่องคอมพิวเตอร์ที่เราใช้เทรน



ภาพที่ 2.1 ภาพแสดงกระบวนการเทรน เพื่อให้ได้ model ที่ต้องการ
ที่มา : Phuri Chalermkiatsakul (2563: ออนไลน์)

2.1.2 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้การตัดสินใจ (decision tree) เป็นเครื่องมือที่ช่วยให้วิเคราะห์เหตุการณ์หรือสถานการณ์เพื่อการตัดสินใจได้อย่างเป็นระบบและรวดเร็ว ต้นไม้การตัดสินใจมีลักษณะเป็นกราฟรูปต้นไม้ ซึ่งแสดงที่ตั้งต้นที่มีรากและแขนงต่างๆแตกออกมาจากต้นไม้ไปในทิศทางเดียวจนกระทั่งนำไปสู่ข้อสรุปสำหรับการตัดสินใจได้ ต้นไม้การตัดสินใจมีประโยชน์ในการสรุปการตัดสินใจที่มีความซับซ้อนให้ง่ายต่อความเข้าใจ ปัจจุบันต้นไม้การตัดสินใจเป็นที่นิยมใช้ในงานหลายอย่าง เช่น การแพทย์ ธุรกิจ การเขียนโปรแกรม การสร้างเครื่องที่เรียนรู้ได้เอง การสร้างระบบผู้เชี่ยวชาญ ฯลฯ (ครรชิต มาลัยวงศ์.2553: ออนไลน์)



ภาพที่ 2.2 เทคนิคที่แนะนำ decision tree หรือต้นไม้ตัดสินใจ
ที่มา: ดัดแปลงจาก Nuthdanai wangpratham. (2564: ออนไลน์)

2.1.3 การคัดเลือกคุณสมบัติ (Feature Selection)

การคัดเลือกคุณสมบัติเป็นเทคนิคที่ช่วยลดจำนวนตัวแปรที่จะใช้ในตัวแบบพยากรณ์ อาจกระทำเพื่อเลือกตัวแปรที่ดีที่สุดเพียงตัวเดียว หรือเลือกกลุ่มของตัวแปรที่มีความสำคัญต่อการพยากรณ์ กระบวนการคัดเลือกคุณสมบัติเป็นกระบวนการที่สำคัญในการเตรียมข้อมูลของการทำเหมืองข้อมูล เพื่อทำให้การสร้างตัวแบบพยากรณ์มีประสิทธิภาพ เพราะจะช่วยลดมิติของข้อมูลและอาจช่วยให้การเรียนรู้วิธีการ พยากรณ์ดำเนินการได้เร็วขึ้นและมีประสิทธิภาพมากขึ้น ในงานวิจัยนี้ ทดลองใช้การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection

2.1.3.1 การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection เป็นวิธีคัดเลือกตัวแปรโดยมี หลักการเช่นเดียวกับการเลือกตัวแปรของการสร้างต้นไม้ตัดสินใจ เพื่อให้ได้ตัวแปรที่เป็นตัวแบ่งข้อมูล ออกเป็นกลุ่มย่อยที่มีสมาชิกภายในกลุ่มเป็นชนิดเดียวกันมากที่สุด (Homogeneous) ด้วยมาตรวัดการได้ ประโยชน์จากการแบ่งกลุ่มย่อยเรียกว่าอัตราส่วนเกน (Gain Ratio) ซึ่งเป็นอัตราส่วนของค่าเกน (Gain หรือ Information Gain) กับ ค่าสารสนเทศการแบ่งกลุ่ม (Split Info) อันเป็นการลดอิทธิพลของตัวแปรที่มีค่าหลายค่า ผลที่ได้รับจากการใช้เทคนิคนี้จะได้ลำดับของตัวแปรซึ่งตัวแปรที่อยู่ลำดับแรกๆ จะถือว่ามีอิทธิพลในการพยากรณ์ตัวแปรเป้าหมายมากกว่าตัวแปรในลำดับถัดไป ทำให้เราสามารถพิจารณาเลือกจำนวนตัวแปรที่เหมาะสมได้อย่างมีประสิทธิภาพ (Tan, Steinbach and Kumar. 2006; Asha, Manjunath and Jayaram. 2010) เกนเรโช (GR) เป็นการประเมินความน่าเชื่อถือของมิติข้อมูลโดยการวัด Gain Ratio ในแต่ละคลาสการคำนวณ GR โดยใช้ค่า SplitINFO ในสมการที 1 และการคำนวณค่าการวัด Gain Ratio ดังสมการที 2 (วีระยุธ พิมพาพร และพยุง มีสัจ 2557).

$$SplitINFO = \sum_{i=1}^{k} \frac{n_i}{n} log_2 \frac{ni}{n} \quad (1)$$

$$GainRatio = \frac{\Delta\, INFO}{SplitINFO} \quad (2)$$

2.1.4 ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล

ดังที่กล่าวไปแล้วว่าการนำโมเดลไปใช้งานจริงได้นั้นเราจำเป็นจะต้องทราบประสิทธิภาพของโมเดลเสียก่อน โดยทั่วไปแล้วจะมีตัววัดที่นิยมใช้กันในงานวิจัยและการทำงานต่างๆ อยู่ 5 ค่า (ธาดา จันตะคุณ, 2559) คือ

1) Precision เป็นการวัดค่าความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส

2) Recall เป็นการวัดค่าความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาส

3) F-measure เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส

4) ความแม่นยำ (Accuracy) เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส

ตารางที่ 2.1 Confusion Matrix ของข้อมูล Weather ซึ่งมี 2 คลาส

Predicted/Actual	Yes	No
Yes	TP	FP
No	FN	TN

จากในตารางที่ 1 ค่าที่แสดงในช่องต่าง ๆ ของตารางประกอบด้วย

1) True Positive (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งกำลังสนใจอยู่

2) True Negative (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งไม่ได้สนใจอยู่

3) False Positive (FP) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งกำลังสนใจอยู่

4) False Negative (FN) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งไม่ได้สนใจอยู่

หลังจากที่เราสร้างตาราง Confusion Matix ได้ดังตารางที่ 1 วิธีคำนวณค่า Precision, Recall, F-measure และ Accuracy

1) Precision เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \quad (3)$$

2) Recall เป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาส

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

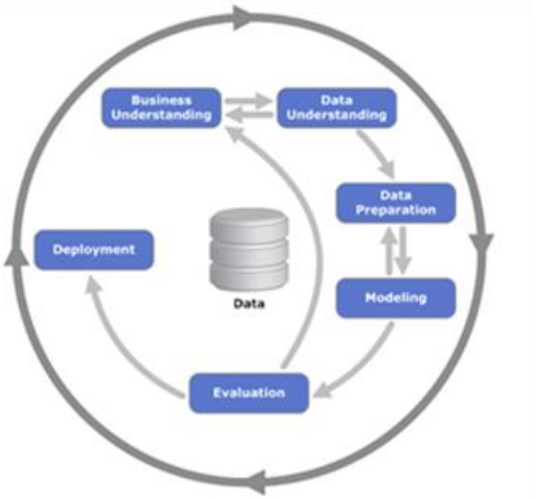
3) F-measure เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดล โดยพิจารณาแยกที่ละคลาส

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

4) Accuracy เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส คือ จำนวน True Positive ของทุกคลาสรวมกันได้เท่ากับ 6/10 = 60%

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

2.1.5 CRISP-DM (Cross-Industry Standard Process For Data Mining)



ภาพที่ 2.3 แสดงกระบวนการ Cross-industry standard process for data mining
ที่มา: Thapanee Boonchob (2564: ออนไลน์)

กระบวนการมาตรฐานที่ใช้สำหรับการทำเหมืองข้อมูล เพื่อทำการวิเคราะห์และนำไปใช้ประโยชน์ มีอยู่ 6 ขั้นตอน คือ

1) การทำความเข้าใจโจทย์ (Business Understanding) ขั้นตอนแรกมุ่งไปที่การทำความเข้าใจข้อมูลปัญหาและวัตถุประสงค์ของโครงการจากมุมมองข้อมูล จากนั้นแปลงปัญหาให้อยู่ในรูปของโจทย์สำหรับการวิเคราะห์ข้อมูล และวางแผนการดำเนินงานเบื้องต้น

2) การทำความเข้าใจข้อมูล (Data Understanding) ขั้นตอนนี้เริ่มต้นด้วยการรวบรวมข้อมูล จากนั้นทำความเข้าใจ ตรวจสอบคุณภาพ และเลือกข้อมูลที่เก็บรวบรวมมาว่าจะใช้ข้อมูลใดบ้างในการวิเคราะห์ขั้นตอนที่ 1 และ 2 สามารถทำกลับไปมาได้ เนื่องจากการทำความเข้าใจข้อมูลทำให้เราเข้าใจข้อมูลมากขึ้น และการเข้าใจข้อมูลก็ทำให้เราเข้าใจข้อมูลมากขึ้นเช่นกัน

3) การเตรียมข้อมูล (Data Preparation) ขั้นตอนการเตรียมข้อมูล หมายถึงขั้นตอนทั้งหมดที่จะทำเพื่อให้ข้อมูลดิบที่เรารวบรวมมา กลายเป็นข้อมูลสมบูรณ์ที่พร้อมจะเข้าสู่โมเดลในขั้นตอนที่ 4 เช่น การสร้างตาราง การลบข้อมูลที่ไม่ต้องการออก การแปลงข้อมูลให้อยู่ในรูปแบบที่ต้องการ

4) การสร้างโมเดล (Modeling) ในขั้นตอนนี้ เราจะเลือกและทดสองสร้างโมเดลหลายๆแบบที่น่าจะสามารถแก้ไขปัญหาที่ต้องการได้ จากนั้นค่อยๆปรับค่าพารามิเตอร์ในแต่ละโมเดล เพื่อให้ได้โมเดลที่เหมาะสมที่สุดมาใช้ในการแก้ไขปัญหา

5) การวัดประสิทธิภาพของโมเดล (Evaluation) เราจะทำการวัดประสิทธิภาพของโมเดลที่ได้จากขั้นตอนที่ 4 เพื่อวัดว่าโมเดลมีประสิทธิภาพเพียงพอต่อการนำไปใช้งานแล้วหรือไม่ ซึ่งโมเดลแต่ละประเภทก็จะมีตัววัดประสิทธิภาพที่แตกต่างกันออกไป

6) การนำโมเดลไปใช้งานจริง (Deployment) เป็นการนำโมเดลที่เหมาะสมที่สุดไปใช้งานจริง เพื่อวิเคราะห์และแก้ปัญหาที่ต้องการ (Thapanee Boonchob.2563)

2.1.6 โปรแกรม RapidMiner Studio

ภาพที่ 2.4 โลโก้โปรแกรม RapidMiner
ที่มา: mypccrack (2565 : ออนไลน์)

RapidMiner คือซอฟต์แวร์ Data Science ใช้สำหรับการเตรียมข้อมูล การเรียนรู้เครื่อง การเรียนรู้ลึกการทำเหมืองข้อความ และการวิเคราะห์การทำนาย (Predictive analysis) เป็นซอฟต์แวร์ที่ช่วยในการจัดส่งข้อมูล และลดข้อผิดพลาดจนแทบจะไม่จำเป็นต้องเขียนโค้ดเพิ่ม แต่ที่ทำให้เป็นเครื่องมือที่ Data Scientist นิยมเลือกใช้เป็นเพราะว่า RapidMiner มีขั้นตอนพร้อมสำหรับการทำ Data mining (ขุดข้อมูล) และ Machine Learning ซึ่งรวมไปถึงการโหลดและการแปลงข้อมูล (ETL) การประมวลผลล่วงหน้าและการวาดภาพจากข้อมูล การวิเคราะห์เชิงพยากรณ์ และการสร้างแบบจำลองทางสถิติ การประเมินผลและการปรับใช้ ต่างๆ ล้วนเป็นสิ่งที่ Data Scientist จำเป็นต้องทำในการเข้าใจข้อมูลมากขึ้น ที่มา : (Achieve. Plus. 2563: ออนไลน์)

2.2 งานวิจัยที่เกี่ยวข้อง

นิภาพร ชนะมาร และพรรณี สิทธิเดช (2557) ได้ศึกษาการวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์ มีวัตถุประสงค์เพื่อการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการพยากรณ์ผลสัมฤทธิ์ ทางการเรียนของนิสิต โดยใช้เทคนิคการคัดเลือกคุณสมบัติที่สำคัญแล้วสร้างตัวแบบการพยากรณ์ด้วย เทคนิค BPNN และเทคนิค SVMs จากข้อมูลที่คัดเลือกซึ่งเป็นปัจจัยการเรียนรู้ที่สำคัญ ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลข้อมูลของนิสิตที่ศึกษาหลักสูตรปริญญา

ตรี สาขาวิชาวิทยาการ คอมพิวเตอร์ ฉบับปรับปรุง พ.ศ. 2548 จำนวน 180 ระเบียน ประกอบด้วย คุณสมบัติ 23 ตัวแปร แบ่งเป็น ตัวแปรอิสระ 22 ตัวแปร ผู้วิจัยได้ทดลองสร้างตัวแบบการพยากรณ์จากข้อมูลทั้งหมดที่มีตัวแปรอิสระ 22 ตัวแปร ด้วย เทคนิคBPNN และเทคนิคSVMs ได้ผลการพยากรณ์ที่มีค่ารากที่สองของกำลังสองของข้อผิดพลาด (Root Mean Square Error: RMSE) เท่ากับ 0.2444 และ 0.1246 ตามลำดับ หลังจากนั้น จึงทำการวิเคราะห์ปัจจัยการเรียนรู้ ด้วยการคัดเลือกคุณสมบัติที่สำคัญ โดยใช้เทคนิคการคัดเลือกคุณสมบัติ 3 วิธี ได้แก่ การคัดเลือกคุณสมบัติ แบบ Correlation-based Feature Selection การคัดเลือกคุณสมบัติแบบ Consistency-based Feature Selection และ การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection ผลการทดลองทั้งสามเทคนิคสามารถลดจำนวน ของคุณสมบัติจาก 22 ตัวแปร เหลือ 9ตัวแปร 10 ตัวแปร และ 11 ตัวแปร ตามลำดับ ผลของงานวิจัยนี้ให้ประโยชน์ในการ วิเคราะห์ปัจจัยการเรียนรู้และการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนิสิตซึ่งจะช่วยให้นิสิตสามารถ พยากรณ์ผลการเรียนของตนเอง และปรับพฤติกรรมการเรียน ได้เช่น การเพิ่มถอนรายวิชาให้เหมาะสมกับ ศักยภาพตนเอง

วรายุทธ พลาศรี (2556) ได้ศึกษาปัจจัยที่มีผลต่อความยากจนของครัวเรือนในชนบท : กรณีศึกษาจังหวัดมหาสารคาม. การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาสภาพเศรษฐกิจของครัวเรือนในชนบท สถานการณ์ความยากจน ลักษณะของครัวเรือนที่ยากจน และปัจจัยที่มีผลต่อความยากจนของครัวเรือนในชนบทจังหวัดมหาสารคาม โดยกลุ่มประชากรที่ใช้ในการศึกษาคือ ครัวเรือนที่อยู่ ในเขตพื้นที่ชนบทจังหวัดมหาสารคาม จำนวน 180,328 ครัวเรือน ขนาดกลุ่มตัวอย่างเท่ากับ 400 ครัวเรือน โดยได้เลือกวิธีการสุ่ม ตัวอย่างแบบหลายขั้นตอน (Multi-stage sampling method) และได้ทำการเก็บรวบรวมข้อมูลโดยการใช้แบบสอบถามเป็นเครื่องมือ หลังจากนั้นจึงนำข้อมูลที่ได้มาวิเคราะห์โดยการใช้สถิติพรรณนาและสถิติอนุมาน อนึ่งการศึกษาครั้งนี้ได้ใช้เส้นความยากจนของครัวเรือน ภาคตะวันออกเฉียงเหนือในเขตพื้นที่ชนบทปี2553ที่คำนวณโดยสำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติซึ่ง จากการคำนวณได้เส้นความยากจนเท่ากับ 1,565 บาทต่อคนต่อเดือน เป็นเกณฑ์ในการแบ่งกลุ่มครัวเรือนยากจนกับกลุ่มครัวเรือนที่ ไม่ยากจน ผลการศึกษาพบว่า ครัวเรือนในชนบทของจังหวัดมหาสารคามมีจำนวนสมาชิกในครัวเรือนเฉลี่ยครัวเรือนละ 4.16 คน มีจำนวนแรงงานในครัวเรือน เฉลี่ยครัวเรือนละ 3.15 คน และจำนวนสมาชิกที่มีรายได้ในครัวเรือนเฉลี่ยครัวเรือนละ 2.13 คน ระดับการศึกษาของหัวหน้าครัว เรือนของกลุ่มตัวอย่างส่วนใหญ่สำเร็จการศึกษาในระดับประถมศึกษาร้อยละ 49.3 อาชีพของหัวหน้าครัวเรือนส่วนใหญ่ประกอบอาชีพ เกษตรกรรมคิดเป็นร้อยละ 57.5 ครัวเรือนมีรายได้รวมเฉลี่ยครัวเรือนละ 16,036.89 บาทต่อเดือน และมีค่าใช้จ่ายสำหรับการอุปโภค บริโภค เฉลี่ยครัวเรือนละ 7,666 บาทต่อเดือน เมื่อคิดเป็นอัตราส่วนร้อยละของค่าใช้จ่ายอุปโภคบริโภคต่อรายได้จะเท่ากับ 47.80 มี หนี้สินเฉลี่ยครัวเรือนละ 187,530.38 บาท และครัวเรือนมีการเก็บออมคิดเป็นร้อยละ 76.5 ของจำนวนครัวเรือนตัวอย่างทั้งหมด สถานการณ์ความยากจนและลักษณะของครัวเรือนที่ยากจนพบว่า ครัวเรือนตัวอย่างในเขต

พื้นที่ชนบทของจังหวัดมหาสารคามมี สัดส่วนของครัวเรือนที่ยากจนคิดเป็นร้อยละ 31.2 โดยครัวเรือนที่ยากจนในเขตชนบทจะมีลักษณะร่วมคือ หัวหน้าครัวเรือนมีระดับ การศึกษาต่ำ มีครัวเรือนขนาดใหญ่ มีระดับรายได้ต่ำ มีขนาดพื้นที่ที่ใช้ในการประกอบอาชีพการเกษตรน้อย มีระดับความมั่งคั่งต่ำ และมีหนี้สิน ส่วนปัจจัยที่มีผลต่อความยากจนของครัวเรือน ได้แก่ ระดับการศึกษาของหัวหน้าครัวเรือน ขนาดของครัวเรือน ขนาดพื้นที่ที่ใช้ ในการประกอบอาชีพ ความมั่งคั่งและหนี้สินของครัวเรือน

ภัทร์พงศ์ พงศ์ภัทรกานต์, วิชัย พัวรุ่งโรจน์, คมยุทธ ไชยวงษ์, สุชาดา พรหมโคตร และปาริชาติ แสงระชัฏ (2560) ได้ศึกษาการใช้เทคนิคเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยในการใช้บริการห้องสมุดของนักศึกษา งานวิจัยนี้นำเสนอการทดสอบวิเคราะห์ปัจจัยในการใช้บริการห้องสมุดของนักศึกษา มหาวิทยาลัยราชภัฏเลย โดยใช้ข้อมูลการเข้าใช้บริการผ่านประตูอัตโนมัติในช่วงเดือนกุมภาพันธ์ถึง ตุลาคม 2559 ที่มี 9 ปัจจัยพื้นฐาน คือ วันที่เข้าใช้บริการ ช่วงเวลา เพศ คณะ ชั้นปี จังหวัดที่เกิด หมู่ เลือด จำนวนพี่น้อง และเกรดเฉลี่ยสะสม จำนวน 79,953 ชุดข้อมูล ทำการประมวลผลด้วยอัลกอริทึม C5.0, Neural Network และ CART เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของการคัดแยกข้อมูล ผลการศึกษา พบว่า อัลกอริทึม C5.0 ให้ค่าความถูกต้อง 97.78 % และใช้ระยะเวลาในการประมวลผล น้อยกว่าอัลกอริทึมที่นำมาเปรียบเทียบ ผลจากการวิเคราะห์ด้วยอัลกอริทึม C5.0 พบว่ามี 3 ปัจจัยที่มี อิทธิพลต่อการใช้บริการห้องสมุดของนักศึกษาที่ส่งผลตามคณะ คือ เกรดเฉลี่ยสะสม มีอิทธิพลสูงสุด ร้อยละ 93.8 เพศ มีอิทธิพลร้อยละ 6.0 และช่วงเวลา มีอิทธิพลร้อยละ 0.2 ซึ่งนำมาสร้างความสัมพันธ์ ได้ 21 ระดับ ซึ่งเป็นแนวทางในการประชาสัมพันธ์และส่งเสริมนักศึกษาเข้ามาใช้บริการห้องสมุดผ่าน คณะที่สังกัดได้ โดยเฉพาะเกรดเฉลี่ยมีผลอย่างมากในการเข้ามาใช้บริการ นักศึกษาที่มีเกรดสูงมี แนวโน้มการเข้าใช้ห้องสมุดมากกว่านักศึกษาที่มีเกรดต่ำ ดังนั้น ห้องสมุดควรเน้นไปที่การเปิดบริการ หรือเชิญชวนให้นักศึกษาที่มีเกรดน้อยเข้าห้องสมุดมากขึ้น ห้องสมุดควรคิดกิจกรรมส่งเสริมใหม่เพิ่ม มากขึ้นเพื่อให้นักศึกษามีความสนใจในการเข้าใช้ห้องสมุด

รัชพลกลัดชื่น และจรัญแสนราช (2561) การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา. การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการทำนายและคุณลักษณะที่มีต่อผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา โดยทำการศึกษาข้อมูลนักศึกษาระดับประกาศนียบัตรวิชาชีพ จำนวน 5,100 ระเบียน ตั้งแต่ปีการศึกษา 2550 -2559 9 สาขาวิชา 27 คุณลักษณะ โดยใช้เทคนิคการจำแนกข้อมูล 3 เทคนิค ได้แก่ Decision Tree : J48graft, Naïve Bayes และ Rule Induction ทำการเปรียบเทียบประสิทธิภาพตัวแบบการทำนายระหว่างการใช้คุณลักษณะทั้งหมดกับการเลือกคุณลักษณะแบบ forward select ทดสอบประสิทธิภาพตัวแบบทำนายด้วยวิธีการ 10-fold cross validation โดยใช้โปรแกรม Rapid Miner

Studio 8 จากนั้นนำผลการทดสอบประสิทธิภาพที่มีค่าความถูกต้องที่สูงที่สุด 2 ค่า มาทำการเปรียบเทียบด้วยวิธี T-Test ผลการศึกษาพบว่าการใช้เทคนิค Decision Tree : J48graft ด้วยการเลือกคุณลักษณะแบบ Forward Selection และ การเลือกคุณลักษณะทั้งหมด มีค่าความถูกต้องเท่ากับ 83.08% และ 81.71% ตามลำดับ และทดสอบด้วยวิธี T-Test พบว่าการทดสอบทั้งสองแบบมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 จากผลการเปรียบเทียบประสิทธิภาพในครั้งนี้ สามารถนำเทคนิค Decision Tree : J48graft ไปใช้ในการทำนายผลสัมฤทธิ์ทางการเรียน และเป็นแนวทางในการสอนเสริมหรือแนะแนวให้กับนักศึกษาต่อไป

สมใจ ตามแต่รัมย์ (2560) ได้ศึกษาปัจจัยที่ส่งผลต่อระดับความสำเร็จหมู่บ้านเศรษฐกิจพอเพียงต้นแบบอำเภอพานทองจังหวัดชลบุรี. มีวัตถุประสงค์เพื่อศึกษาระดับความสำเร็จของหมู่บ้านเศรษฐกิจพอเพียงต้นแบบ อำเภอพานทอง เพื่อศึกษาปัจจัยที่ส่งผลต่อระดับความสำเร็จของหมู่บ้านเศรษฐกิจพอเพียงต้นแบบ อำเภอพานทอง จังหวัดชลบุรีประชากรที่ใช้ในการศึกษาครั้งนี้คือประชาชน ที่อาศัยอยู่จริงในหมู่บ้านเศรษฐกิจพอเพียงต้นแบบอำเภอพานทองทั้ง 10 หมู่บ้าน จำนวนประชากร 7,342คน และกลุ่มตัวอย่างจำนวน 380คน เครื่องมือที่ใช้ในการเก็บ รวบรวมข้อมูลการวิจัยครั้งนี้คือแบบสอบถาม และสถิติที่ใช้ในการวิเคราะห์ข้อมูลได้แก่ค่าความถี่ร้อยละค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐาน การวิเคราะห์ถดถอยพหุคูณ ด้วยวิธีแบบขั้นตอน ผลการศึกษา พบว่า ระดับความสำเร็จหมู่บ้านเศรษฐกิจพอเพียงต้นแบบอำเภอพานทองจังหวัดชลบุรีอยู่บนระดับมาก

ประเสริฐ บัวทอง (2560) ได้ศึกษาปัจจัยที่มีผลต่อการตัดสินใจปลูกทุเรียนของเกษตรกรในตำบลอ่างคีรี อำเภอมะขาม จังหวัดจันทบุรี. มีวัตถุประสงค์ ประการแรกเพื่อศึกษาปัจจัยส่วนบุคคลที่มีผล ต่อการตัดสินใจปลูกทุเรียนของเกษตรกรในพื้นที่ ตำบลอ่างคีรี อำเภอมะขาม จังหวัดจันทบุรีและประการที่สองเพื่อศึกษาปัจจัยทางเศรษฐกิจและปัจจัยด้านกายภาพที่มีผลต่อการตัดสินใจปลูกทุเรียนของเกษตรกรในพื้นที่ ตำบลอ่างคีรีอำเภอมะขาม จังหวัดจันทบุรี โดยมีกลุ่มตัวอย่าง คือ เกษตรกรสวนทุเรียนที่ขึ้นทะเบียนการปลูกทุเรียนในตำบลอ่างคีรี อำเภอมะขามจังหวัดจันทบุรี จำนวน 300 ครัวเรือน และสัมภาษณ์เชิงลึก จำนวน 10 ครัวเรือน ผลการวิจัยเป็นไปตามสมมติฐานที่ตั้งไว้พบว่า เกษตรกรส่วนใหญ่เป็นเพศชายอายุ 31-40 ปี ระดับการศึกษาประถมศึกษารายได้ต่อปีของครอบครัว750,000-1,000,000 บาท มีจำนวนแรงงานในการปลูกทุเรียน 6-10 คน ต้นทุนในการปลูกทุเรียน 100,001-200,000 บาท สายพันธุ์ทุเรียนที่ปลูกหมอนทอง ลักษณะของดินเป็นดินร่วน ลักษณะพื้นที่เป็นที่ราบสูงขนาดของแหล่งน้ำ มีขนาดตั้ง แต่1ไร่ลงมาการคมนาคมสะดวกสบายเป็นช่วงขนาดพื้นที่ปลูกทุเรียนมีขนาด 26-50ไร่ ประสบการณ์ในการปลูกทุเรียน 3-6 ปีการสัมภาษณ์เกษตรกรส่วนใหญ่ให้เหตุผลว่า ทำไมถึงตัดสินใจปลูกทุเรียน เพราะทุเรียนเป็นผลไม้ที่มีความต้องการของตลาดสูง

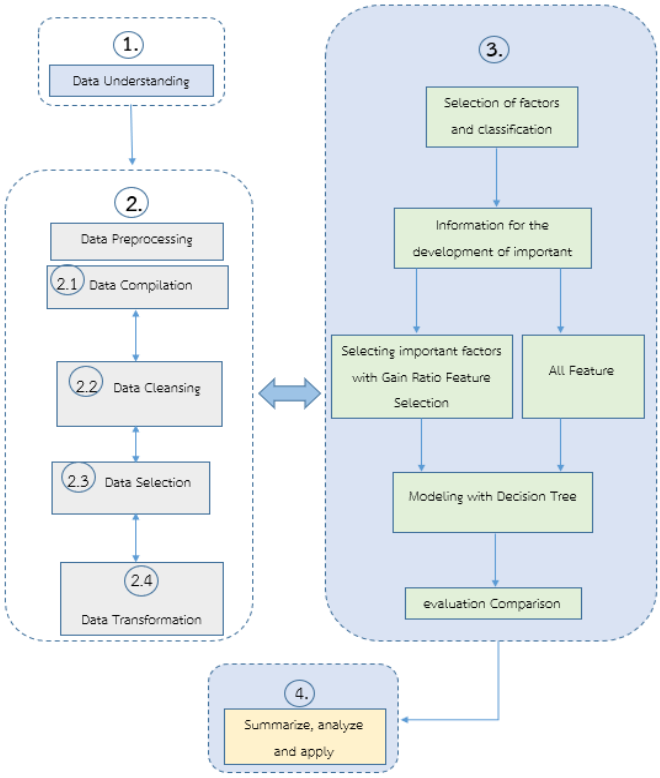
วีระยุธพิมพาพร และพยุงมีสัจ (2557) ได้ศึกษาการวิเคราะห์องค์ประกอบของชุดข้อมูลที่ซับซ้อนด้วยวิธีการเลือกคุณลักษณะสำคัญแบบพลวัต มีวัตถุประสงค์เพื่อศึกษากระบวนการ

วิเคราะห์องค์ประกอบ (Factor Analysi:) บนชุดข้อมูลที่ทับซ้อนด้วยวิธีการ เลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS) โดยประยุกต์ใช้กระบวนการเลือกตัวแป (Feature Selection) และการวิเคราะห์กลุ่ม (Clustering analysis) ข้อมูลในการประมวลผลเกิดจากกิจกรรมต่าง ๆ ในระบบการเรียนออนไลน์ (e-Learning) โดยเน้นปัจจัยที่ส่งผลโดยตรงต่อผลสัมฤทธิ์ทางการเรียน ผลการวิจัยพบว่าประสิทธิภาพโดยรวมของกระบวนการวิเคราะห์องค์ประกอบ โดยใช้อัลกอริทึมการเลือกลักษณะสำคัญ แบบพลวัต ให้ค่าความถูกต้องสูงสุดที่ 45.17% โดยใช้ 3 ตัวแปร สำหรับกระบวนการวิเคราะห์องค์ประกอบโดยวิธีการคำนวณหาค่า GAIN ของข้อมูลด้วย Information Gain และ Gain ratio ให้ค่าความถูกต้องสูงสุดที่ 44.80% โดยใช้ตัวแปร 7 ตัวแปร จากผลการวิจัยสามารถสรุปได้ว่า อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัตมีค่าความถูกต้องสูงกว่า และใช้จำนวนตัวแปรที่น้อยกว่า วิธีการคำนวณหาค่า GAIN ของข้อมูลด้วย Information Gain และ Gain ratio

อัจจิมา มณฑาพันธุ์(2562) งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเกี่ยวกับการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญเพื่อนำมาใช้ในการ ปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม โดยใช้วิธีการคัดเลือกคุณลักษณะจากเทคนิคต่าง ๆ จำนวน 7 เทคนิค ได้แก่เทคนิค Correlation Based Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio เทคนิค Chi-Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection หลังจากคัดเลือกคุณลักษณะ ที่สำคัญจึงนำผลที่ได้จากแต่ละเทคนิคมาคำนวณหาค่าประสิทธิภาพในการพยากรณ์การเป็นมะเร็งเต้านมโดยการใช้ทคนิคซัพพอร์ต เวกเตอร์แมชชีน ผลการทดลองพบว่าร้อยละของความถูกต้องในการพยากรณ์การเป็นมะเร็งเต้านม จากจำนวนคุณลักษณะของ ข้อมูลทั้งหมด 30 คุณลักษณะเท่ากับ 91.39 ขณะที่เทคนิค Evolutionary Selection ให้ผลดีที่สุดโดยสามารถลดคุณลักษณะ ที่สำคัญเหลือเพียง 16 คุณลักษณะ และให้ผลการวัดค่าความถูกต้องในการพยากรณ์ได้ดีถึงร้อยละ 95.26

บทที่ 3
วิธีดำเนินงานวิจัย

การดำเนินงานนี้ได้ใช้การประยุกต์ตามแนวทางในการทำเหมืองข้อมูลที่เรียกว่ากระบวนการมาตรฐานอุตสาหกรรม หรือ CRIPS-DM (Cross Reference Industry Standard for Data Mining) ที่ได้รับความนิยมมากในปัจจุบันซึ่งมีขั้นตอนการดำเนินงาน (chapman et al. 2000) ดังดังภาพที่ 3.1



ภาพที่ 3.1 กรอบการดำเนินงานวิจัย

จากภาพที่ 3.1 จากปัจจัยทั้งหมดที่มี จะสามารถหาปัจจัยที่ส่งผลต่อระดับเศรษฐกิจครัวเรือนได้

3.1 การทำความเข้าใจข้อมูล (Data Understanding)

ข้อมูลที่ใช้ในการศึกษาครั้งนี้ คือข้อมูลประชากรจากภาคครัวเรือนเฉพาะครัวเรือนในเขตพื้นที่ชนบท ของจังหวัดสกลนคร ซึ่งมี 20 หมู่บ้าน 12 ตำบล 12 อำเภอ โดยช่วงเวลาที่ทำการเก็บรวบรวมข้อมูล คือ ปี พ.ศ. 2563 – 2564 และจากฐานข้อมูลสภาพทางเศรษฐกิจครัวเรือน (สำนักวิทย¬บริการและเทคโนโลยีสารสนเทศ, 2563: ออนไลน์) โดยในฐานข้อมูลนี้เป็นข้อมูลจากโครงการศาสตร์พระราชาซึ่งมีการเก็บข้อมูลออกเป็น 10 ส่วน รวมทั้งหมด 136 แอททริบิวต์ ได้มา 17,933 ครัวเรือน ดังนี้

ส่วนที่ 1 ข้อมูลทั่วไปครัวเรือน

ส่วนที่ 2 ทรัพย์สินของครัวเรือน

ส่วนที่ 3 อาชีพและรายได้ของครัวเรือน

ส่วนที่ 4 รายจ่ายของครัวเรือน

ส่วนที่ 5 หนี้สินของครัวเรือน

ส่วนที่ 6 ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19)

ส่วนที่ 7 การใช้เทคโนโลยีสารสนเทศ

ส่วนที่ 8 การเข้าร่วมการละเล่น การฟ้อน การรำ พิธีกรรมตามวิถีวัฒนธรรมชุมชน

ส่วนที่ 9 การเข้าร่วมโครงการที่ผ่านมาย้อนหลัง 3 ปี

ส่วนที่ 10 ข้อคิดเห็นและข้อเสนอแนะเพิ่มเติม

3.2 การเตรียมข้อมูล (Data Preprocessing)

การเตรียมข้อมูลก่อนการประมวลผลเป็นขั้นตอนสำคัญในกระบวนการทำเหมืองข้อมูล ซึ่งหากกระบวนการเตรียมข้อมูลไม่ได้ทำอย่างรอบคอบแล้ว จะทำให้ไม่ได้ชุดข้อมูลที่เป็นตัวแทนที่เหมาะสมสำหรับการสร้างโมเดลการทำนายซึ่งจะทำให้ผลลัพธ์การทำนายที่ได้ไม่มีความแม่นยำ ดังนั้นการเตรียมข้อมูลจึงเป็นขั้นตอนที่มีความสำคัญมาก ซึ่งประกอบด้วย 4 ขั้นตอน ได้แก่ การรวบรวมข้อมูล (Data Compilation) การทำความสะอาดข้อมูล (Data Cleansing) การคัดเลือกข้อมูล (Data Selection) และการเปลี่ยนแปลงรูปแบบของข้อมูล (Data Transformation)

1.3.2.1 การรวบรวมข้อมูล (Data Compilation)

ในส่วนนี้ใช้ข้อมูลเศรษฐกิจครัวเรือนในช่วงปี พ.ศ. 2561-2563 ที่สามารถวิเคราะห์ข้อมูล ได้มาจากการเลือกแบบเจาะจง (Purposive Sampling) จำนวน 2,909 ครัวเรือน ดังตัวอย่างแสดงข้อมูลตามตารางที่ 3.1

ตารางที่ 3.1 จำนวนข้อมูลครัวเรือนที่ได้มาจากการเลือกแบบเจาะจง

ลำดับที่	ตำบล	จำนวนครัวเรือน
1	ค้อเขียว	102
2	แพด	120
3	โคกศิลา	93
4	ท่าก้อน	354
5	นาหัวบ่อ	518
6	พันนา	305
7	สร้างค้อ	450
8	วัฒนา	99
9	ม่วง	336
10	หนองสนม	189
11	บ้านแป้น	211
12	อุ่มจาน	132
รวม (ครัวเรือน)		2,909

เมื่อได้จำนวนครัวเรือนแล้วจากนั้นทำการคัดเลือกแอททริบิวต์ สำหรับใช้สร้างตัวแบบการพยากรณ์ ซึ่งในจำนวนครัวเรือนเหล่านี้มีข้อมูลบางแอททริบิวต์ไม่สมบูรณ์ เช่น ค่าใช้จ่ายในการทำไร่ รายได้จากการจักรสาน ราคาจำหน่ายผลผลิต รายได้จากการทอผ้า ซึ่งได้ตัดแอททริบิวต์ออกไป จะได้แอททริบิวต์ทั้งหมด 15 แอททริบิวต์ ดังต่อไปนี้

ส่วนที่ 1 ข้อมูลทั่วไปครัวเรือน มีทั้งหมด 7 แอททริบิวต์ 2,909 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (นิภารัตน์ นักตรีพงศ์, 2561: 196; สมยศ ประจันบาล, 2548-2555: 5) ทั้งหมด 3 แอททริบิวต์ ได้แก่ อายุ อาชีพ และรายได้เฉลี่ย/เดือน

ส่วนที่ 2 ทรัพย์สินของครัวเรือน มีทั้งหมด 24 แอททริบิวต์ 2,909 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (นิภารัตน์ นักตรีพงศ์, 2561) ทั้งหมด 2 แอททริบิวต์ ได้แก่ มูลค้าทรัพย์สิน และวัตถุประสงค์การเลี้ยงสัตว์

ส่วนที่ 3 อาชีพและรายได้ของครัวเรือน มีทั้งหมด 68 แอททริบิวต์ 2,909 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (สุวรัฐ แลสันกลาง, พิบูลย์ ชยโอว์สกุล, ฐิฏิกานต์ สุริยะสาร และชุตินิษฐ์ ปานคำ, 2563) ทั้งหมด 3 แอททริบิวต์ ได้แก่ ผลผลิต/ไร่ ต้นทุน และจำนวนไร่

ส่วนที่ 4 รายจ่ายของครัวเรือน มีทั้งหมด 3 แอททริบิวต์ 2,909 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (นิภารัตน์ นักตรีพงศ์, 2561: 196) ทั้งหมด 1 แอททริบิวต์ ได้แก่ ค่าใช้จ่าย/เดือน

ส่วนที่ 5 หนี้สินของครัวเรือน มีทั้งหมด 3 แอททริบิวต์ 2,909 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (นิภารัตน์ นักตรีพงศ์, 2561: 196; สุวรัฐ แลสันกลาง, พิบูลย์ ชยโอว์สกุล, ฐิฏิกานต์ สุริยะสาร, และชุตินิษฐ์ ปานคำ, 2563: 40-43) ทั้งหมด 2 แอททริบิวต์ ได้แก่ แหล่งเงินกู้ และปริมาณเงินกู้

ส่วนที่ 6 ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19) มีทั้งหมด 8 แอททริบิวต์ 2,909 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (นิภารัตน์ นักตรีพงศ์, 2561: 196) ทั้งหมด 2 แอททริบิวต์ ได้แก่ ผลกระทบ และรายได้ลดลง

ส่วนที่ 7 การใช้เทคโนโลยีสารสนเทศ มีทั้งหมด 15 แอททริบิวต์ 2,909 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (อัครนันท์ คิดสม, 2561: 97-98) ทั้งหมด 2 แอททริบิวต์ ได้แก่ การใช้อินเทอร์เน็ต และช่องทางการขายสินค้า

ในส่วนที่ 8 การเข้าร่วมการละเล่น การฟ้อน การรำ พิธีกรรมตามวิถีวัฒนธรรมชุมชน ส่วนที่ 9 การเข้าร่วมโครงการที่ผ่านมาย้อนหลัง 3 ปี และส่วนที่ 10 ข้อคิดเห็นและข้อเสนอแนะเพิ่มเติม ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน พบว่าทั้ง 3 ส่วน ไม่มีปัจจัยไหนที่ส่งผลต่อสภาพเศรษฐกิจครัวเรือน

จากข้อมูลครัวเรือนผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลให้เหมาะสมเพื่อนำมาใช้ในการสร้างตัวแบบการพยากรณ์ข้อมูลเศรษฐกิจครัวเรือน รวมได้ทั้งหมด 15 แอททริบิวต์ ดังแสดงในตารางที่ 3.2

ตารางที่ 3.2 แสดงแอททริบิวต์ที่ส่งผลต่อสภาพเศรษฐกิจครัวเรือน

ลำดับ	รายละเอียด
ส่วนที่ 1 ข้อมูลทั่วไปครัวเรือน	
1	อายุ
2	อาชีพ
3	รายได้เฉลี่ย/เดือน
ส่วนที่ 2 ทรัพย์สินของครัวเรือน	
4	มูลค้าทรัพย์สิน
5	วัตถุประสงค์การเลี้ยงสัตว์
ส่วนที่ 3 อาชีพและรายได้ของครัวเรือน	
6	ผลผลิต/ไร่
7	ต้นทุน
8	จำนวนไร่
ส่วนที่ 4 รายจ่ายของครัวเรือน	
9	ค่าใช้จ่าย/เดือน
ส่วนที่ 5 หนี้สินของครัวเรือน	
10	แหล่งเงินกู้
11	ปริมาณเงินกู้
ส่วนที่ 6 ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19)	
12	ผลกระทบ
13	รายได้ลดลง
ส่วนที่ 7 การใช้เทคโนโลยีสารสนเทศ	
14	การใช้อินเทอร์เน็ต
15	ช่องทางการขายสินค้า

จากนั้นผู้วิจัยได้ทำการทำความสะอาดข้อมูล (Data Cleansing) และแปลงรูปแบบข้อมูล (Data Transformation) เพราะข้อมูลครัวเรือนทั้งหมดที่ได้ทำการเก็บมานั้นมีรูปแบบครัวเรือนที่ยังไม่สมบูรณ์ ซึ่งในงานวิจัยนี้จะเน้นและคัดเลือกเฉพาะข้อมูลครัวเรือนที่สมบูรณ์จำนวน 1,751 ครัวเรือน แล้วทำให้ได้แอททริบิวต์ ในการสร้างตัวแบบจำนวน 18 แอททริบิวต์ เพื่อใช้ในการสร้างตัวแบบการพยากรณ์ที่เหมาะสม จากนั้นทำการแปลงรูปแบบข้อมูล ดังแสดงในตารางที่ 3.3

ตารางที่ 3.3 รายละเอียดของตัวแปรที่เป็นคุณลักษณะของกลุ่มตัวอย่างสภาพเศรษฐกิจครัวเรือน

ลำดับ	คุณลักษณะ	รายละเอียด	ชนิดข้อมูล
1	Education Age	วัยเรียน	Numeric
2	Working Age	วัยทำงาน	Numeric
3	Old Age	วัยสูงอายุ	Numeric
4	Occupation	อาชีพ	Nominal
5	Average Income/Year	รวมรายได้เฉลี่ย/ปี ของครัวเรือน	Numeric
6	Asset Value	มูลค้าทรัพย์สิน	Numeric
7	Animal Husbandry	วัตถุประสงค์การเลี้ยงสัตว์	Nominal
8	Area	พื้นที่ก่อให้เกิดรายได้	Numeric
9	Production Costs	ต้นทุนการผลิตการทำการเกษตร	Numeric
10	Product	ผลผลิตที่ได้จากการทำเกษตร	Numeric
11	Total Expenses/Year	รวมค่าใช้จ่าย/ปี ของครัวเรือน	Numeric
12	Loan Bank	หนี้ในระบบ	Nominal
13	Loan Shark	หนี้ในระบบ	Nominal
14	Total Liabilities	รวมปริมาณหนี้สินของครัวเรือน	Numeric
15	Effect	ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19)	Nominal
16	Lower Income	รายได้ลดลงจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19)	Nominal
17	Internet Use	การใช้อินเทอร์เน็ตที่ก่อให้เกิดรายได้	Nominal
18	Sales Channel	ช่องทางการขายสินค้าที่ก่อให้เกิดรายได้	Nominal
19	Classification	การจัดหมวดหมู่ คลาสคำตอบ Low Income = รายได้น้อย Middle income = รายได้ปานกลาง High Income = รายได้สูง	Nominal

ตารางที่ 3.3 ข้อมูลเศรษฐกิจครัวเรือนที่ผ่านการทำความสะอาดและแปลงรูปแบบข้อมูล

ลำดับ	Education Age	Working Age	Old Age	...	Lower Income	Internet Use	Sales Channel
1	0	3	0	...	Yes	Yes	Yes
2	0	2	0	...	Yes	Yes	Yes
3	1	4	1	...	Yes	Yes	Yes
1,749	2	2	0	...	Yes	Yes	No
1,750	0	1	0	...	Yes	Yes	No
1,751	0	1	0	...	Yes	Yes	No

จากที่ได้ทำความสะอาดข้อมูล (Data Cleansing) และแปลงรูปแบบข้อมูล (Data Transformation) ดังแสดงในตารางที่ 3.3 พบว่า ในส่วนที่ 1 ได้มีการเปลี่ยนรูปแบบแอททริบิวต์ เช่น อายุ ได้ทำการแยกแอททริบิวต์ออกมาเป็น 3 แอททริบิวต์ คือ วัยเรียน วัยทำงาน และวัยสูงอายุ เพราะบางครัวเรือนนั้นมีสมาชิกในครัวเรือนมากกว่า 1 คน (ระเบียน) จึงทำการเปลี่ยนแปลงรูปแบบข้อมูลให้เหลือครัวเรือนละ 1 ระเบียน ส่วนที่ 2 ได้มีการลดจำนวนระเบียนในครัวเรือน โดยการใช้สูตร SUM เพื่อหาผลบวกของทรัพย์สินครัวเรือนทั้งหมดให้เหลือ 1 ระเบียน ส่วนที่ 3 ได้เปลี่ยนแปลงหน่วยจากงาน ให้เป็นหน่วยไร่ เช่น 4 งาน = 1 ไร่ เพราะจะได้ง่ายต่อการนำเข้าโปรแกรม ส่วนที่ 4 ได้เปลี่ยนแปลงข้อมูลค่าใช้จ่าย/เดือนของครัวเรือนให้เป็นรายจ่ายเฉลี่ย/ปี โดยการใช้สูตร (ค่าใช้จ่ายแต่ละคน * 12 นำมาบวกกัน) จะได้ค่าใช้จ่ายเฉลี่ย/ปี ของครัวเรือน ส่วนที่ 5 ได้มีการเปลี่ยนรูปแบบแอททริบิวต์ของแหล่งเงินกู้ แยกออกมาเป็น 2 แอททริบิวต์ ได้แก่ หนี้ในระบบและหนี้นอกระบบ ตัวแปรของ 2 แอททริบิวต์ คือ Yes/No ส่วนที่ 6 ได้มีการเปลี่ยนรูปแบบตัวแปรของแอททริบิวต์ ผลกระทบ และรายได้ลดลง ส่วนที่ 7 ได้มีการเปลี่ยนรูปแบบตัวแปรของแอททริบิวต์ การใช้อินเทอร์เน็ต และช่องทางการขายสินค้า

3.3 การคัดเลือกคุณสมบัติ

หลังจากขั้นตอน Transform ข้อมูล คลีนเสร็จ จะได้ข้อมูลมาใช้ในการสร้างคลาส สำหรับการหาปัจจัยที่สำคัญ ผู้วิจัยได้นำข้อมูลจาก All Feature ที่ได้จากขั้นตอน Data Preprocessing โดยใช้ข้อมูลปัจจัย 19 ปัจจัย จำนวน 1,751 ครัวเรือน 15 แอททริบิวต์ นำไปเข้าโปรแกรม RapidMiner Studio และหาปัจจัยที่สำคัญ ด้วยเทคนิค Gain Ratio Feature Selection เมื่อได้ปัจจัยที่คัดเลือกมาแล้วจะนำไปสู่ขั้นตอนการสร้างโมเดล

ตารางที่ 3.4 All Feature มีปัจจัยทั้งหมด 19 ปัจจัย จำนวน 1,751 ครัวเรือน 15 แอททริบิวต์

ลำดับ	คุณลักษณะ	รายละเอียด	ชนิดข้อมูล
1	Education Age	วัยเรียน	Numeric
2	Working Age	วัยทำงาน	Numeric
3	Old Age	วัยสูงอายุ	Numeric
4	Occupation	อาชีพ	Nominal
5	Average Income/Year	รวมรายได้เฉลี่ย/ปี ของครัวเรือน	Numeric
6	Asset Value	มูลค้าทรัพย์สิน	Numeric
7	Animal Husbandry	วัตถุประสงค์การเลี้ยงสัตว์	Nominal
8	Area	พื้นที่ก่อให้เกิดรายได้	Numeric
9	Production Costs	ต้นทุนการผลิตการทำการเกษตร	Numeric
10	Product	ผลผลิตที่ได้จากการทำเกษตร	Numeric
11	Total Expenses/Year	รวมค่าใช้จ่าย/ปี ของครัวเรือน	Numeric
12	Loan Bank	หนี้ในระบบ	Nominal
13	Loan Shark	หนี้ในระบบ	Nominal
14	Total Liabilities	รวมปริมาณหนี้สินของครัวเรือน	Numeric
15	Effect	ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19)	Nominal
16	Lower Income	รายได้ลดลงจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19)	Nominal
17	Internet Use	การใช้อินเทอร์เน็ตที่ก่อให้เกิดรายได้	Nominal
18	Sales Channel	ช่องทางการขายสินค้าที่ก่อให้เกิดรายได้	Nominal
19	Classification	การจัดหมวดหมู่ คลาสคำตอบ Low Income = รายได้น้อย Middle income = รายได้ปานกลาง High Income = รายได้สูง	Nominal

นำไปเข้าโปรแกรม RapidMiner Studio และหาปัจจัยที่สำคัญ ด้วยเทคนิค Gain Ratio Feature Selection ดังภาพที่ 3.1 ได้ปัจจัยที่สำคัญมา 16 ปัจจัย ดังตารางที่ 3.5 เมื่อได้ปัจจัยที่คัดเลือกมาแล้วจะนำไปสู่ขั้นตอนการสร้างโมเดล

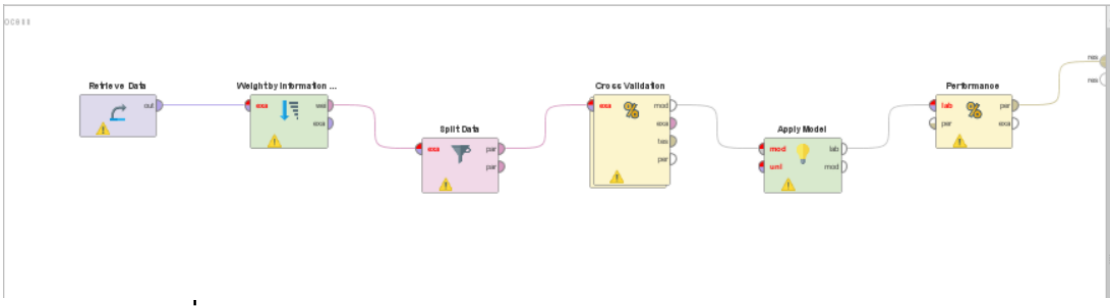
ภาพที่ 3.2 ตัวอย่างการ Feature Selection

ตารางที่ 3.5 ตัวอย่าง ผลของการ หาปัจจัยที่สำคัญ ด้วยเทคนิค Gain Ratio Feature Selection คัดเลือกมาทั้งหมด 16 ปัจจัย

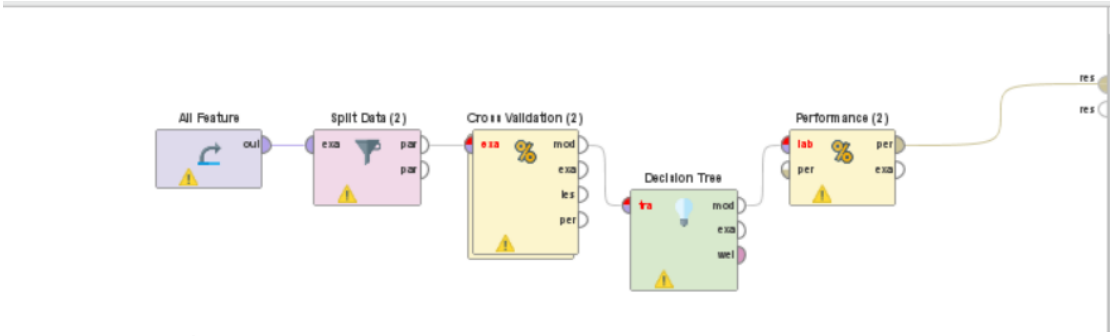
ลำดับ	คุณลักษณะ	รายละเอียด	ชนิดข้อมูล
1	Occupation	อาชีพ	Nominal
2	Average Income/Year	รวมรายได้เฉลี่ย/ปี ของครัวเรือน	Numeric
3	Asset Value	มูลค้าทรัพย์สิน	Numeric
4	Animal Husbandry	วัตถุประสงค์การเลี้ยงสัตว์	Nominal
5	Area	พื้นที่ก่อให้เกิดรายได้	Numeric
6	Production Costs	ต้นทุนการผลิตการทำการเกษตร	Numeric
7	Product	ผลผลิตที่ได้จากการทำเกษตร	Numeric
8	Total Expenses/Year	รวมค่าใช้จ่าย/ปี ของครัวเรือน	Numeric
9	Loan Bank	หนี้ในระบบ	Nominal
10	Loan Shark	หนี้ในระบบ	Nominal
11	Total Liabilities	รวมปริมาณหนี้สินของครัวเรือน	Numeric
12	Effect	ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19)	Nominal
13	Lower Income	รายได้ลดลงจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 (COVID - 19)	Nominal
14	Internet Use	การใช้อินเทอร์เน็ตที่ก่อให้เกิดรายได้	Nominal
15	Sales Channel	ช่องทางการขายสินค้าที่ก่อให้เกิดรายได้	Nominal
16	Classification	การจัดหมวดหมู่ คลาสคำตอบ Low Income = รายได้น้อย Middle income = รายได้ปานกลาง High Income = รายได้สูง	Nominal

3.4 การสร้างโมเดล (Modeling)

ในขั้นตอนนี้จะไม่มีการเปรียบเทียบกับเทคนิคอื่นๆ แต่จะเปรียบเทียบข้อมูลที่ได้มาจากขั้นตอน Data Preprocessing กับข้อมูลที่ได้มาจากขั้นตอน Feature Selection ดังภาพที่ 3.2 แล้วจะเอาเข้าโมเดลของ Decision Tree การหาปัจจัยด้วยเทคนิค Gain Ratio Feature Selection กับ All Feature ดังภาพที่ 3.3 ว่าข้อมูลไหนได้ประสิทธิ์ภาพที่ดีที่สุด



ภาพที่ 3.3 ตัวอย่างการหาปัจจัยด้วยเทคนิค Gain Ratio Feature Selection ด้วยโมเดล Decision Tree



ภาพที่ 3.4 ตัวอย่างการหาปัจจัยด้วย All Feature ด้วยโมเดล Decision Tree

3.5 การวัดประสิทธิภาพของโมเดล (Evaluation)

เลือกใช้วิธีประเมินประสิทธิภาพของการพยากรณ์ด้วยค่าความถูกต้อง (Accuracy) จะทำการทดสอบค่าความถูกต้องในการพยากรณ์ด้วยวิธี Cross Validation Test โดยทำการแบ่งข้อมูลออกเป็น 5 ส่วน (5-fold Cross Validation) และ 10 ส่วน (10-fold Cross Validation) และจะทำการสุ่มข้อมูลตามค่าสัดส่วนร้อยละ 60:40, 70:30 และ 80:20 ของข้อมูลจำนวน 1,751 ครัวเรือน แล้วจะทำการทดสอบระหว่าง All Feature และ ข้อมูลที่ได้มาจากขั้นตอน Feature Selection ด้วยโมเดล Decision Tree จากผลลัพธ์ปัจจัยที่ได้จากการเปรียบเทียบระหว่าง All Feature และ ข้อมูลที่ได้มาจากขั้นตอน Feature Selection

3.6 นำไปใช้งาน (Deployment)

เป็นการนำปัจจัยสำคัญของเศรษฐกิจครัวเรือนที่เหมาะสมที่สุดไปใช้งานจริง เพื่อวิเคราะห์และแก้ปัญหาที่ต้องการ สำหรับสนับสนุนหรือเป็นข้อมูลประกอบการตัดสินใจในการวิจัยในลำดับต่อไป

บรรณานุกรม

กองนโยบายและแผน. (2560). **แผนยุทธศาสตร์มหาวิทยาลัยราชภัฏสกลนคร ระยะ 20 ปี** (พ.ศ. 2560 –2579). สืบค้นเมื่อ 10 มกราคม 2565 จากเว็บไซต์ : https://www.snru.ac.th /wpcontent/uploads/2018/02/-60-79.pdf?fbclid=IwAR3ac-E3KYNQb

กิตติคุณ แสงนิล และประสพชัย พสุนนท. (2561). "ความน่าเชื่อถือ ความถูกต้อง ความแม่นยำ และความเที่ยงตรง" ความสอดคล้องในวิธีการและความคลาดเคลื่อนจากการวัดของการวิจัยทางด้านสรีรวิทยาการออกกำลังกาย. **Veridian E-Journal, Science and Technology Silpakorn University.** 5(6) หน้า 1-19.

ครรชิต มาลัยวงศ์. (2553).**ต้นไม้การตัดสินใจ.** สืบค้นเมื่อ 14 มกราคม 2565 จากเว็บไซต์ https://kb.hsri.or.th/dspace/handle/11228/2964?locale-attribute=th

ณัฐพร เห็นเจริญเลิศ. (2558**). การวิเคราะห์ ข้อมูลด้วยเทคนิค Data Mining โดยซอฟต์แวร์ RapidMiner Studio 6 (ขั้นพื้นฐานและปานกลาง).** สืบค้นเมื่อ 9 มกราคม 2564 จากเว็บไซต์ https://www.stou.ac.th/Schools/sst/main/KM/KM%20Post/58/RapidMiner.pdf?fbclid=IwAR

ธนพล สราญจิตร์. (2558). ปัญหาความยากจนในสังคมไทย. **วารสารวิชาการมหาวิทยาลัยอีสเทิร์นเอเชีย ฉบับสังคมศาสตร์และมนุษศาสตร์**, 5(2), 12-21.

นิภาพร ชนะมาร และพรรณี สิทธิเดช. (2557). การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์. **วารสารมหาวิทยาลัยราชภัฏสกลนคร.** 6(12), 46 หน้า.

นิภารัตน์ นักตรีพงศ์. (2561). เปรียบเทียบการกระจายรายได้ของครัวเรือน กรณีศึกษา หมู่ 2 บ้านทะเลน้อยและหมู่ 7 บ้านหัวป่าเขียว ตำบลทะเลน้อย อำเภอควนขนุน จังหวัดพัทลุง. **วารสารวิชาการมหาวิทยาลัยธนบุรี**, 12(29), 194-202.

นิเวศ จิระวิชิตชัย. (2553). **การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ.**สืบค้นเมื่อ 11 มกราคม 2565 จากเว็บไซต์ http://ssruir.ssru.ac.th /handle/ssruir/377.

ประเสริฐ บัวทอง. (2560). **ปัจจัยที่มีผลต่อการตัดสินใจปลูกทุเรียนของเกษตรกรในตำบลอ่างคีรี อำเภอมะขามจังหวัดจันทบุรี**. วิทยานิพนธ์หลักสูตรบริหารธุรกิจมหาบัณฑิต, มหาวิทยาลัยบูรพา, 78 หน้า.

พิมเพ็ญ พรเฉลิมพงศ์. **accuracy-ความถูกต้อง-ความแม่นยำ.** สืบค้นเมื่อ 20 มกราคม 2565 จากเว็บไซต์http://www.foodnetworksolution.com/wiki/word/4289/accuracy

ภัทร์พงศ์ พงศ์ภัทรกานต์, วิชัย พัวรุ่งโรจน์, คมยุทธ ไชยวงษ์, สุชาดา พรหมโคตร และ ปาริชาติ แสง

ระชัฏ. (2560). การใช้เทคนิคเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยในการใช้บริการห้องสมุดของนักศึกษา.**PULINET Journal**;4(2), 11-18.

ภานุพงศ์ สุขสุวรรณ. (2562). **Model Evaluation, Model Optimization and Deployment.** สืบค้น 9 มกราคม 2565, จาก https://medium.com/tni-university/model-evaluation-and-deployment-848f33e9b395

ภูริพัทร์ ทองคำ. (2559). **อัลกอริทึมแบบรวมสำหรับการเลือกคุณสมบัติของข้อมูล.** วิทยานิพนธ วิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยธรรมศาสตร์, 119 หน้า.

มหาวิทยาลัยราชภัฏสกลนคร. (2563). **ระบบบันทึกแบบสอบถามสภาพทางเศรษฐกิจครัวเรือนเป้าหมายตามโครงการจ้างงาน ประชาชนที่ได้รับผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า2019 (COVID-19).** สืบค้น 9 มกราคม 2565, จาก http://qcovid.snru.ac.th/Default.aspx?fbclid

รัชพล กลัดชื่น และจรัญ แสนราช. (2561). การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา. **Research Journal**

วีระยุธ พิมพาพร และพยุง มีสัจ. (2557). การวิเคราะห์องค์ประกอบของชุดข้อมูลที่ซับซ้อนด้วยวิธีการเลือกคุณลักษณะสำคัญแบบพลวัต. **วารสารศรีปทุมปริทัศน์.** 6(6). หน้า 90-99.

สมใจ ตามแต่รัมย์. (2560). **ปัจจัยที่ส่งผลต่อระดับความสำเร็จหมู่บ้านเศรษฐกิจพอเพียงต้นแบบอำเภอพานทองจังหวัดชลบุรี.** วิทยานิพนธ์หลักสูตรรัฐประศาสนศาสตรมหาบัณฑิต, มหาวิทยาลัยบูรพา, 95 หน้า.

สมยศ ประจันบาล. (2548-2555). **ปัจจัยที่มีอิทธิพลต่อการเปลี่ยนแปลงการใช้จ่ายของครัวเรือนไทย** (วิทยานิพนธ์ปริญญามหาบัณฑิต). คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนาบริหารศาสตร์. 4-5 หน้า.

สุพรรณ ฟ้าหยง. (2562). **Machine Learning 4 ประเภท.** สืบค้นเมื่อ 13 มกราคม 2565 จากเว็บไซต์: http://codeonthehill.com/machine-learning-types.

สุวรัฐ แลสันกลาง พิบูลย์ ชยโอว์สกุล ฐิฏิกานต์ สุริยะสาร ชุตินิษฐ์ ปานคำ. (2563). การบริหารจัดการหนี้สินครัวเรือนแบบมีส่วนร่วมจังหวัดลำปาง. **วารสารวิชาการเครือข่ายบัณฑิตศึกษามหาวิทยาลัยราชภัฏภาคเหนือ**, 10(2), 31-44.

หนึ่งหทัย ชัยอาภร. (2559). **การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง.** สืบค้นเมื่อ 14 มกราคม 2565จากเว็บไซต์ https://erp.mju.ac.th/acticleDetail.aspx?qid=551.

อัครนันท์ คิดสม. (2561). ความสามารถในการจ่ายค่าบริการอินเทอร์เน็ตความเร็วสูง แบบประจำที่ของครัวเรือนไทย. **วารสารบริหารธรุกิจเทคโนโลยีมหานคร**, 15(2), 97-98.

อัจจิมา มณฑาพันธุ์. (2562). การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะทีสำคัญในการปรับปรุงการ

พยากรณ์มะเร็งเต้านม. **Royal Thai Air Force Medical Gazette.** 65(2). หน้า49-56

Achieve.Plus.(2563). **Rapidminer เสกคนไม่มีพื้นฐานให้เป็นเซียน.** สืบค้นเมื่อ 9 มกราคม 2565 จากเว็บไซต์https://medium.com/achieve-space/rapidminer-99-9bf6ab20d1aa

Mpcrkadmn. (2565). **MY PC Crack Full Version Free Download**. สืบค้นเมื่อ 9 มกราคม 2565จากเว็บไซต์ https://mypccrack.com/rapidminer-studio-

Nuthdanai wangpratham. (2564). **Predictive Modeling**. สืบค้นเมื่อ 15 มกราคม 2565 จาก เว็บไซต์https://nutdnuy.medium.com/predictive-modeling-f9881b3e4c02

Panupong Suksuwan. (2016). **Model Evaluation, Model Optimization and Deployment.**สืบค้นเมื่อ 14 มกราคม 2565 จากเว็บไซต์ Phuri Chalermkiatsakul. https://phuri.medium.com/supervised-learning

Rachot Leingchan. (2564). **เศรษฐกิจไทยจะเป็นอย่างไร หากเราต้องอยู่กับโควิด-19 ไปตลอดกาล.**สืบค้นเมื่อ 17 มกราคม 2565 จากเว็บไซต์ https://www.krungsri.com/th/ research/research-intelligence/ri-covid- recovery-2021? fbclid.

Rajamangala University of Technology Thanyaburi. 17(1). หน้า 1-10.

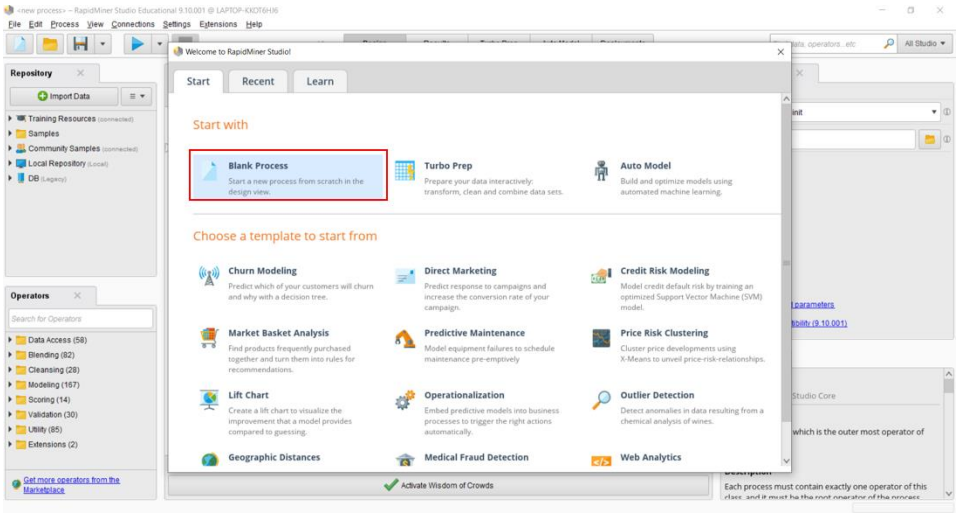
วรายุทธ พลาศร. (2555). การศึกษาปัจจัยที่มีผลต่อความยากจนของครัวเรือนในชนบท: กรณีศึกษาจังหวัดมหาสารคาม. **วารสารมหาวิทยาลัยราชภัฏมหาสารคาม.** 7(1). หน้า 29-38.

Thapanee Boonchob. (2563). **เข้าใจ CRISP-DM ฉบับเร่งรัด.** สืบค้นเมื่อ 11 มกราคม 2565 จากเว็บไซต์https://kamboonchob.medium.com/94-b0913050198f

ภาคผนวก

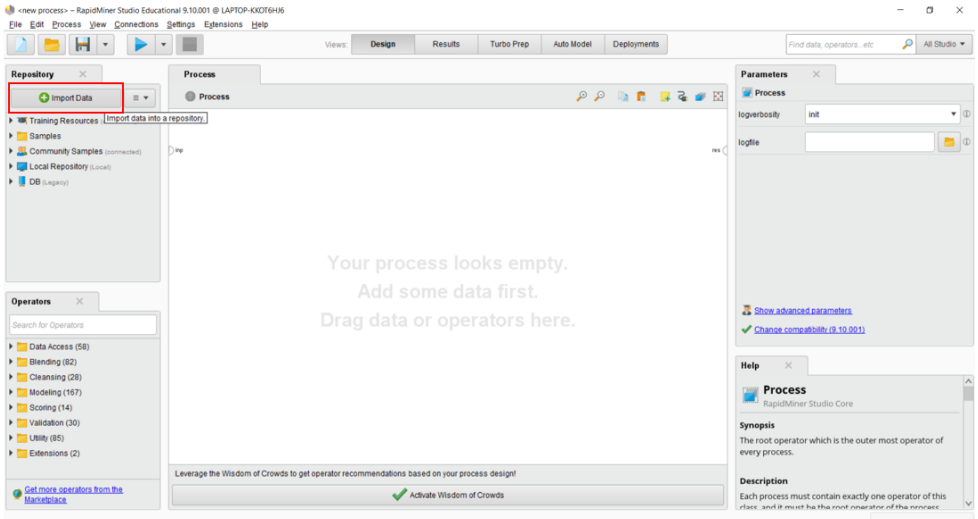
แสดงขั้นตอนการใช้งานโปรแกรม

1) เมื่อเปิดโปรแกรมมาแล้วจะพบหน้าต่าง Welcome to RapidMiner Studio จะพบ option ต่างๆให้เลือกใช้ คลิก Blank Process เพื่อเริ่มต้นการทำงาน



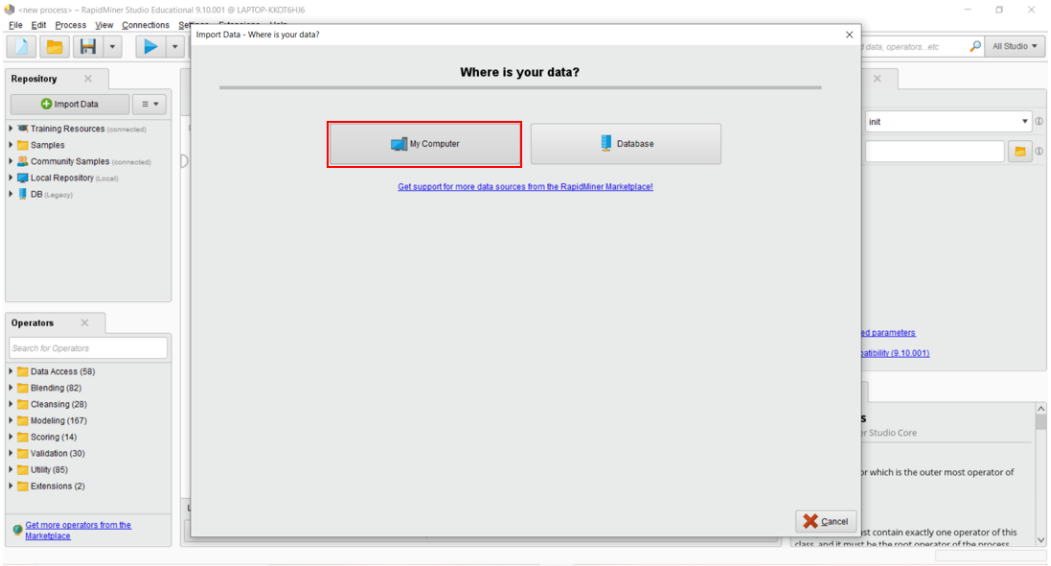
ภาพที่ 1 เริ่มต้นเข้าใช้งานโปรแกรม

2) ต่อมาคลิก import Data เพื่อนำข้อมูลเข้าโปรแกรม



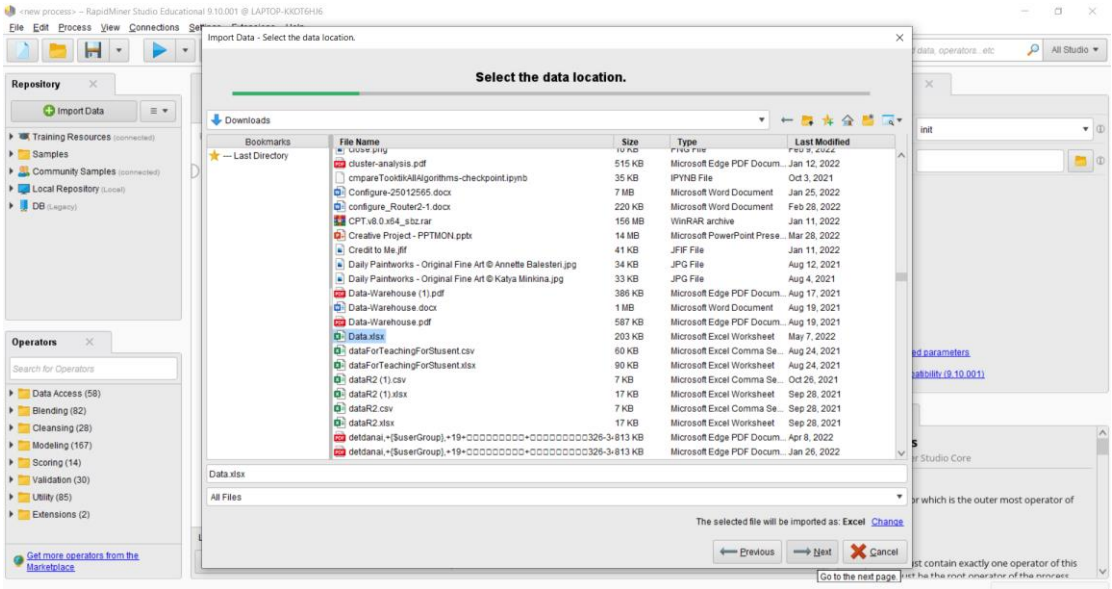
ภาพที่ 2 import Data เพื่อนำข้อมูลเข้าโปรแกรม

3) แล้วคลิก My Computer เพื่อนำข้อมูลในเครื่องเข้าโปรแกรม



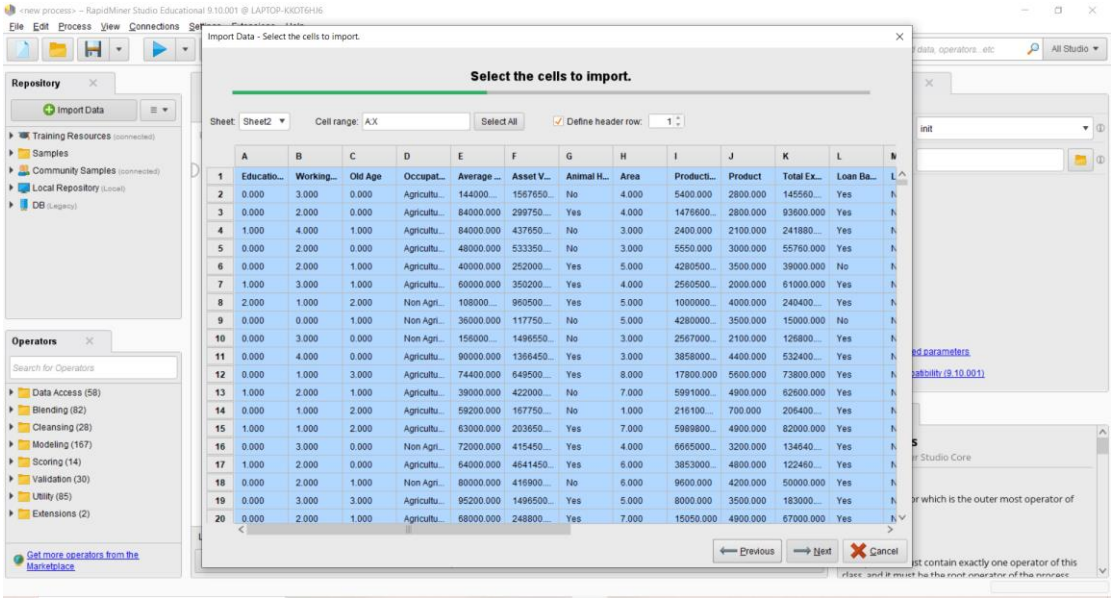
ภาพที่ 3 คลิก My Computer เพื่อนำข้อมูลในเครื่องเข้าโปรแกรม

4) เลือกไฟล์ข้อมูลที่จะนำเข้าโปรแกรม แล้วกด Next



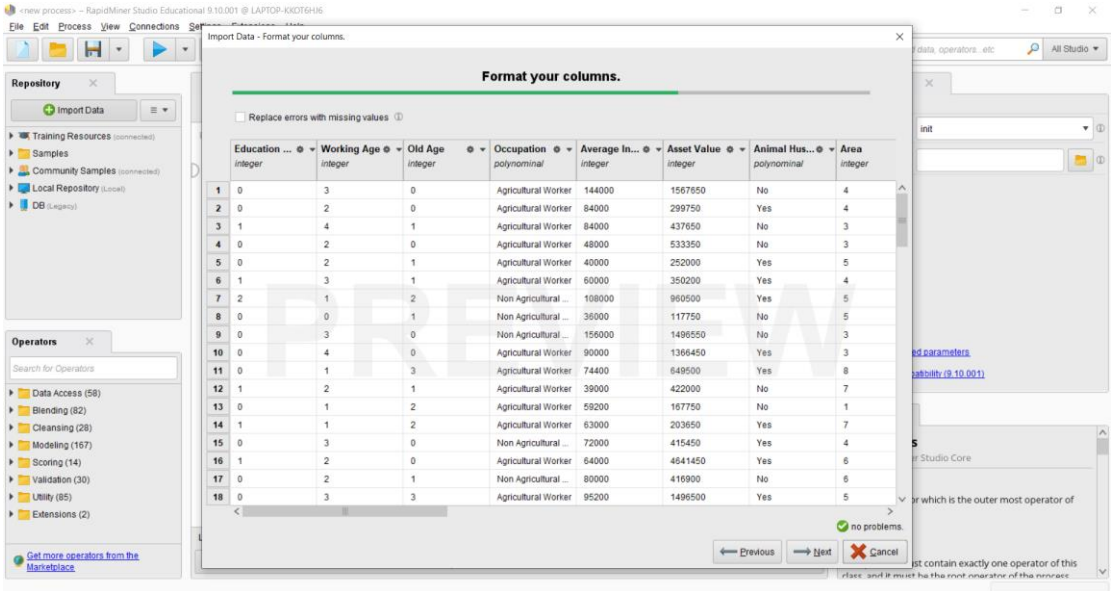
ภาพที่ 4 เลือกไฟล์ที่ต้องการจะนำเข้าโปรแกรม

5) เมื่อเลือกแล้วจะพบข้อมูลทั้งหมดที่จะเข้าโปรแกรมเมื่อเช็คข้อมูลเสร็จแล้วกด Next



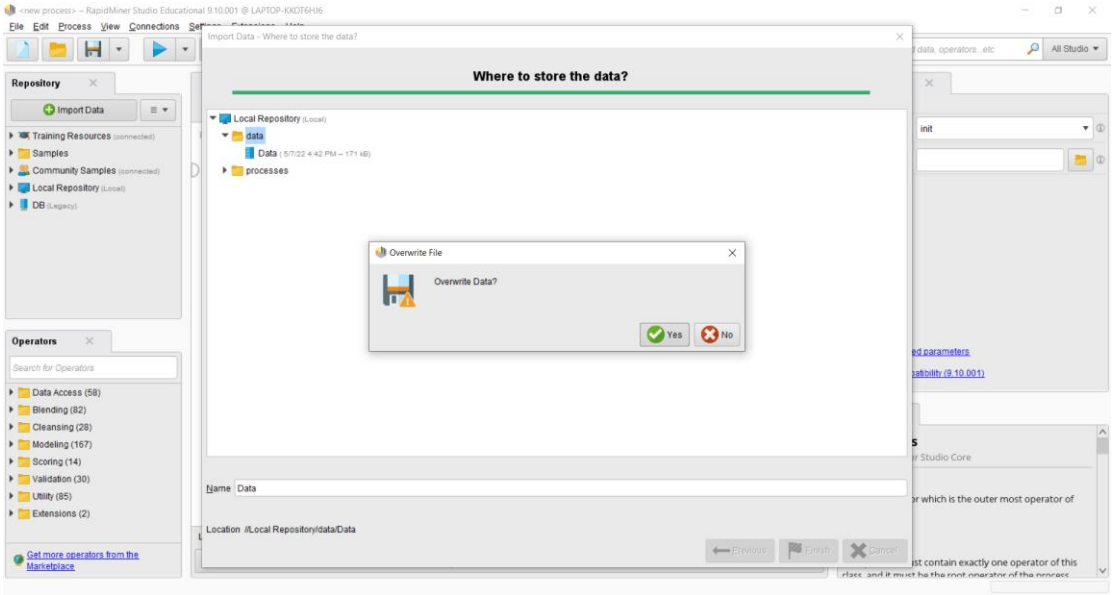
ภาพที่ 5 คัดเลือกข้อมูลที่จะนำเข้าโปรแกรม

6) เช็คข้อมูลที่จะเข้าโปรแกรมข้อมูลส่วนที่ผิดหรือไม่



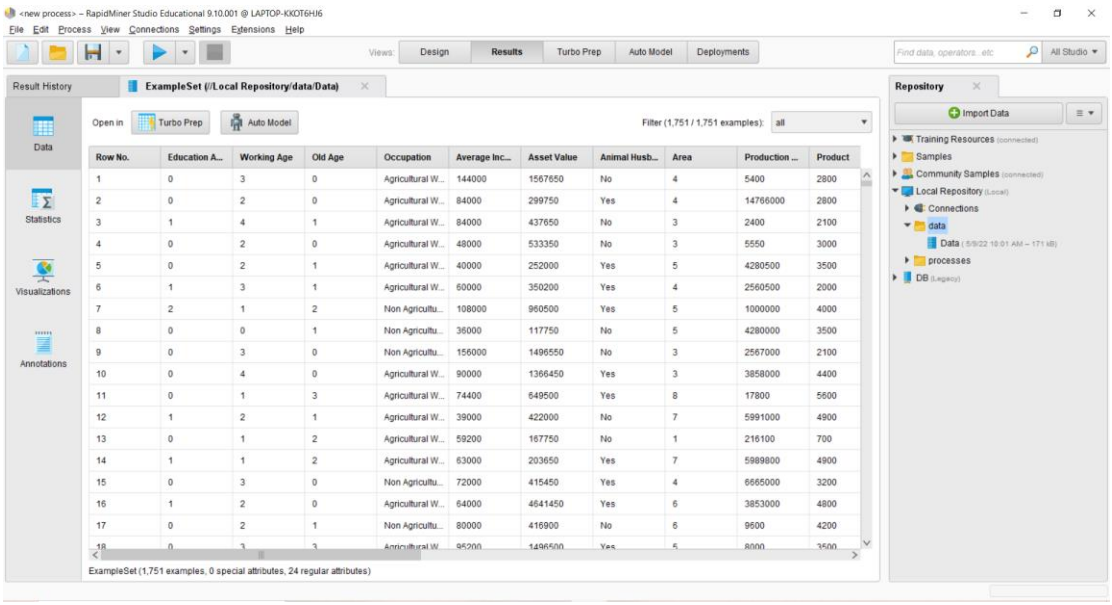
ภาพที่ 6 การเช็คข้อมูล

7) เมื่อเสร็จแล้วเลือกโฟลเดอร์ที่จะบันทึกแล้วกด Finish



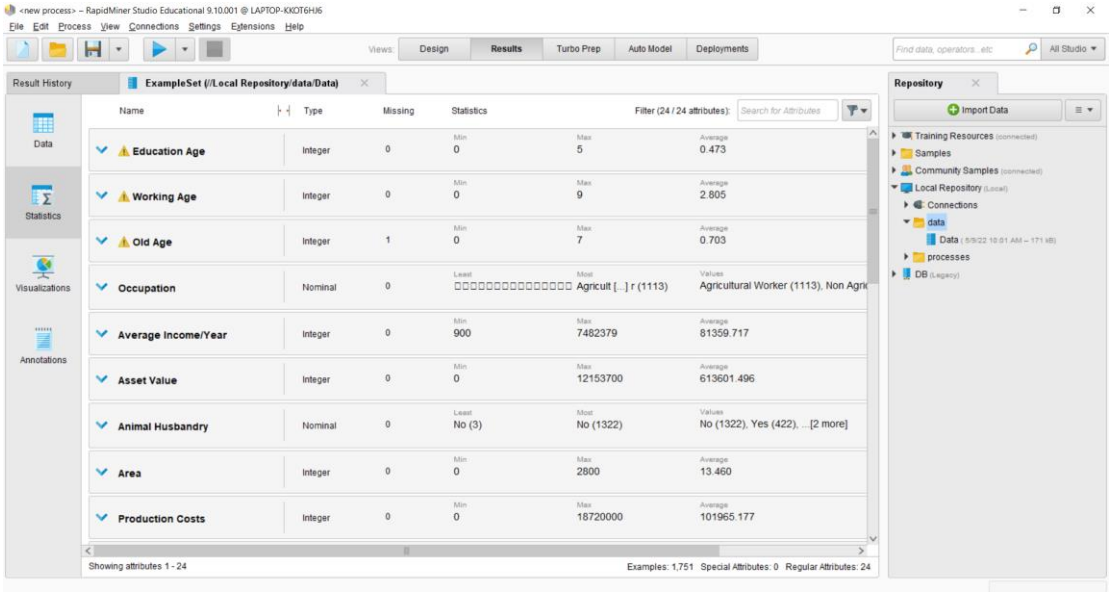
ภาพที่ 7 เลือกบักทึกไฟล์แล้ว Save

8) จากนั้นจะขึ้นหน้าข้อมูลแอททริบิวต์



ภาพที่ 8 แสดงข้อมูลแอททริบิ้ว

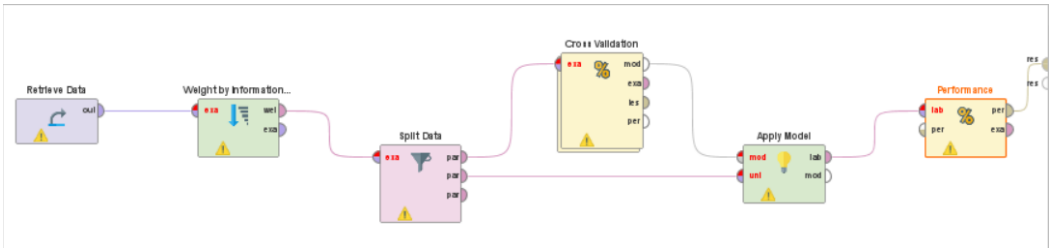
9) จากนั้นจะขึ้นหน้าข้อมูลทั้งหมด



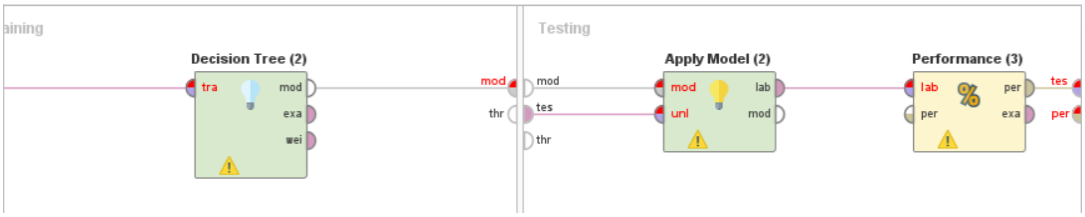
ภาพที่ 9 แสดงข้อมูลทั้งหมด

ขั้นตอนการคัดเลือกปัจจัยสำคัญ

1) จะทำการดึงข้อมูลเข้ามาหลังจากนั้นจะนำ โมเดล Gain Ratio Feature Selection มาเพื่อคัดเลือกปัจจัยที่สำคัญสำหรับข้อมูลเศรษฐกิจครัวเรือน ข้อมูลจะถูกบ่งออกเป็น 2 ส่วนใหญ่ ใน Split Data โดยทำการแบ่งข้อมูลออกเป็น 5 ส่วน (5-fold Cross Validation) และ 10 ส่วน (10-Fold Cross Validation) และจะทำการสุ่มข้อมูลตามค่าสัดส่วนร้อยละ 60:40, 70:30 และ 80:20 ใน Cross Validation และทำการคัดเลือกปัจจัยด้วย Decision Tree

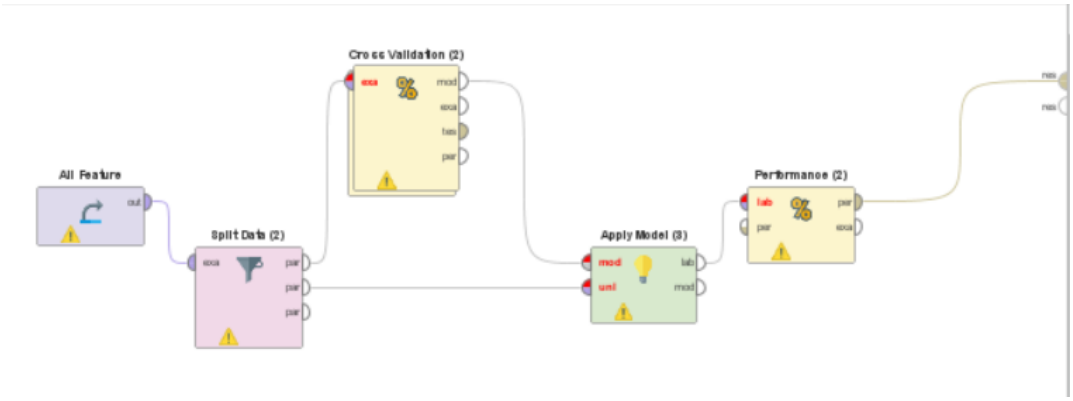


ภาพที่ 10 นำข้อมูลเข้าโมเดล Gain Ratio Feature Selection

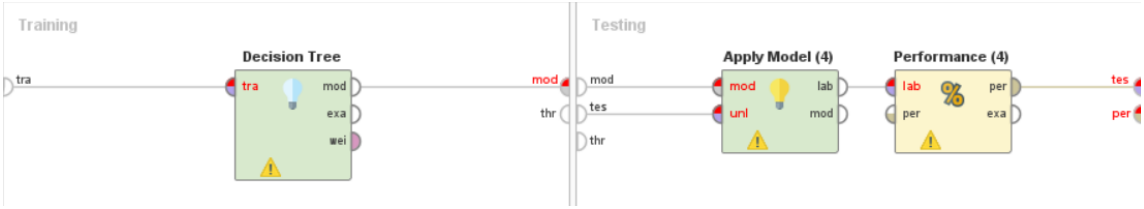


ภาพที่ 11 โอเปอเรเตอร์การคัดเลือกปัจจัยสำคัญ

2) จะทำการดึงข้อมูล All Feature เข้ามาหลังจากนั้นจะนำมาเพื่อคัดเลือกปัจจัยที่สำคัญสำหรับข้อมูลเศรษฐกิจครัวเรือน ข้อมูลจะถูกบ่งออกเป็น 2 ส่วนใหญ่ๆใน Split Data โดยทำการแบ่งข้อมูลออกเป็น 5 ส่วน (5-Fold Cross Validation) และ 10 ส่วน (10-Fold Cross Validation) และจะทำการสุ่มข้อมูลตามค่าสัดส่วนร้อยละ 60:40, 70:30 และ 80:20 ใน Cross Validation และทำการคัดเลือกปัจจัยด้วย Decision Tree



ภาพที่ 12 นำ All Feature เข้าโมเดลเพื่อหาปัจจัยสำคัญ



ภาพที่ 13 โอเปอเรเตอร์การคัดเลือกปัจจัยสำคัญ

ประวัติผู้จัดทำ

ผู้จัดทำ	นางสาววชิราภรณ์ เจริญมา
วันเดือนปีเกิด	วันพุธ ที่ 6 เดือนธันวาคม 2543
การศึกษา	ประกาศนียบัตรมัธยมศึกษาตอนต้น โรงเรียนร่มเกล้าสกลนคร ประกาศนียบัตรมัธยมศึกษาตอนปลาย โรงเรียนร่มเกล้าสกลนคร ปริญญาตรี (วท.บ.วิทยาการคอมพิวเตอร์) มหาวิทยาลัยราชภัฏสกลนคร
คติประจำตัว	เหนื่อยวันนี้ วันหน้าเหนื่อยกว่าเดิม
อีเมล์	wachiraporn.ja62@snru.ac.th
เบอร์โทรศัพท์	0968091381