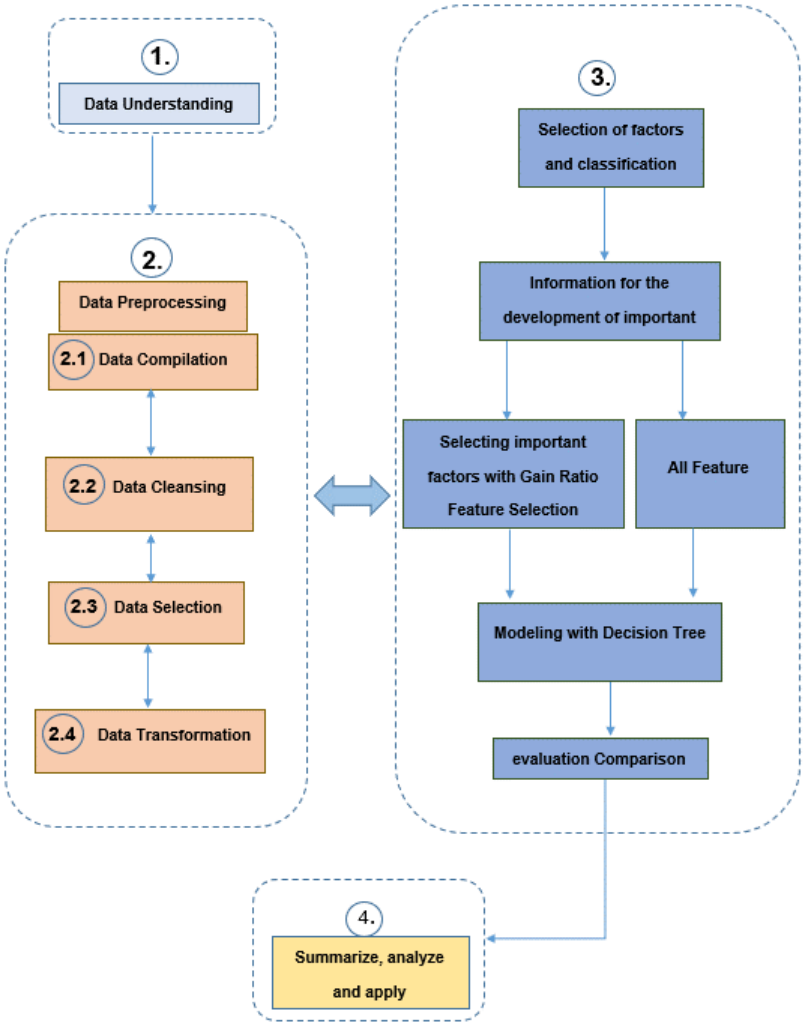
บทที่ 3
วิธีการดำเนินงาน

การดำเนินงานนี้ได้ใช้การประยุกต์ตามแนวทางในการทำเหมืองข้อมูลที่เรียกว่ากระบวนการมาตรฐานอุตสาหกรรม หรือ CRIPS-DM (Cross Reference Industry Standard for Data Mining) ที่ได้รับความนิยมมากในปัจจุบันซึ่งมีขั้นตอนการดำเนินงาน (chapman et al, 2000) ดังภาพที่ 3.1 รายละเอียดการทำงานแต่ละขั้นตอน มีดังนี้



ภาพที่ 3.1 กรอบการดำเนินงานวิจัย

จากภาพที่ 3.1 ด้วยปัจจัยทั้งหมดที่มี จะสามารถวิเคราะห์ปัจจัยที่ส่งผลต่อระดับเศรษฐกิจครัวเรือนได้ โดยมีกระบวนการดังนี้

3.1 การทำความเข้าใจข้อมูล (Data Understanding)

ข้อมูลที่ใช้ในการศึกษาครั้งนี้ คือข้อมูลประชากรจากภาคครัวเรือนเฉพาะครัวเรือน ของจังหวัดสกลนคร ซึ่งมี 12 หมู่บ้าน จำนวน 2,909 ครัวเรือน โดยช่วงเวลาที่ทำการเก็บรวบรวมข้อมูล คือ ปี พ.ศ. 2563 – 2564 และจากฐานข้อมูลเศรษฐกิจครัวเรือน (สำนักวิทยบริการและเทคโนโลยีสารสนเทศ, 2563: ออนไลน์) ในฐานข้อมูลนี้เป็นข้อมูลจากโครงการศาสตร์พระราชาซึ่งมีการเก็บข้อมูล 10 ส่วน ดังแสดงในภาพที่ 3.2 รวมทั้งหมด 178 ปัจจัย จำนวน 17,933 ระเบียน ผ่านการลดระเบียนครัวเรือนให้เหลือ 1,751 ระเบียน (ครัวเรือน) เนื่องจากแต่ละครัวเรือนมีสมาชิกแตกต่างกัน จึงใช้ขั้นตอนการแปลงรูปแบบข้อมูลให้เหลือครัวเรือนละ 1 ระเบียน

ส่วนที่ 1 ข้อมูลทั่วไปครัวเรือน จำนวน 31 ปัจจัย ได้แก่ ชื่อ-สกุล ผู้ให้ข้อมูล อายุสถานะในครอบครัว บ้านเลขที่ หมู่ที่ บ้าน ถนน ตำบล อำเภอ จังหวัด รหัสไปรษณีย์ โทรศัพท์บ้าน โทรศัพท์มือถือ สมาชิกในครอบครัว ที่ตั้งบ้าน ลักษณะบ้าน สภาพบ้านพัก ในครอบครัวมีบุตรหลานเรียนมหาวิทยาลัยราชภัฏสกลนครหรือไม่ รู้จักมหาวิทยาลัยราชภัฏสกลนครหรือไม่ หากมีบุตรหลานต้องการเสริมทักษะความรู้หรืออาชีพนอกเหนือจากหลักสูตรที่ศึกษาหรือไม่ สมาชิกคนที่ ชื่อ-สกุล สถานะภาพในครอบครัว สถานะภาพสมรส อายุ การศึกษา อาชีพ รายได้เฉลี่ย/เดือน ปัจจุบันทำงาน และโรคประจำตัวบทบาทในชุมชน

ส่วนที่ 2 ทรัพย์สินของครัวเรือน จำนวน 24 ปัจจัย

ส่วนที่ 2.1 ได้แก่ ลำดับทรัพย์สิน รายการทรัพย์สิน จำนวน ราคาสถานะ ยอดรวม (บาท) ดาวน์ (บาท) จำนวน (งวด) ค่างวด/เดือน (บาท) และการใช้ประโยชน์

ส่วนที่ 2.2 ได้แก่ ลำดับที่ดิน ขนาดพื้นที่ (ไร่/ตรว/ตรม) สิทธิในที่ดิน สภาพดิน ไร่ นา สวนและ อื่น ๆ

ส่วนที่ 2.3 ได้แก่ ชนิดสัตว์ เพศ วัตถุประสงค์การเลี้ยง ลักษณะการเลี้ยง แหล่งอาหาร และปัญหา/ความต้องการ

ส่วนที่ 3 อาชีพและรายได้ของครัวเรือน จำนวน 82 ปัจจัย

ส่วนที่ 3.1 มีข้อมูลดังนี้

- กระบวนการผลิตตามฤดูกาล ได้แก่ พันธุ์ จำนวน (ไร่) ผลผลิตต่อไร่ (กก.) แบ่งไว้ขาย (%) แบ่งไว้กิน (%) และแบ่งไว้ทำพันธุ์ (%)

- ต้นทุนการผลิต ได้แก่ อัตราต่อไร่ (บาท) พื้นที่ (ไร่) ต้นทุน (บาท) อัตราต่อไร่ (บาท) พื้นที่ (ไร่) และต้นทุน (บาท)

- ไร่ ได้แก่ แรงงานในครอบครัวที่ทำ (คน) ลำดับรายการ รายการชนิดพืช ค่าพันธุ์ ค่าเตรียมดิน ค่าปุ๋ยเคมี ค่าปุ๋ยอินทรีย์/ปุ๋ยคอก ค่ารักษาโรค/แมลง ค่าเก็บเกี่ยว ค่าขนส่งผลผลิต ค่าใช้จ่ายอื่น ๆ รวมค่าใช้จ่าย ผลผลิตต่อรอบการผลิต/ปี ราคาจำหน่ายผลผลิต/ปี เก็บไว้บริโภค/ปี (ร้อยละ) และนำไปขาย/ปี (ร้อยละ)

- สวน ได้แก่ จำนวนพื้นที่ที่ทำ (ไร่) แรงงานในครอบครัวที่ทำ (คน) ลำดับรายการ รายการชนิดพืช ค่าพันธุ์ ค่าเตรียมดิน ค่าปุ๋ยเคมี ค่าปุ๋ยอินทรีย์/ปุ๋ยคอก ค่ารักษาโรค/แมลง ค้าจ้างแรงงาน ค่าเก็บเกี่ยว ค่าขนส่งผลผลิต ค่าใช้จ่ายอื่น ๆ รวมค่าใช้จ่าย ผลผลิตต่อรอบการผลิต/ปี ราคาจำหน่ายผลผลิต/ปี เก็บไว้บริโภค/ปี (ร้อยละ) และนำไปขาย/ปี (ร้อยละ)

ส่วนที่ 3.2 มีข้อมูลดังนี้

- การทอผ้า ได้แก่ มีการจัดตั้งกลุ่มหรือไม่ ชื่อกลุ่ม จำนวนสมาชิก รายได้ต่อเดือน ชนิดการย้อม ลำดับรายการ ชื่อผลิตภัณฑ์ จำนวน และราคา/ชิ้น

- จักรสาน ได้แก่ มีการจัดตั้งกลุ่มหรือไม่ ชื่อกลุ่ม จำนวนสมาชิก รายได้ต่อเดือน ชนิดการจักรสาน ลำดับรายการ ชื่อผลิตภัณฑ์ จำนวน และราคา/ชิ้น

- พืชผักสวนครัว ได้แก่ มีการจัดตั้งกลุ่มหรือไม่ ชื่อกลุ่ม จำนวนสมาชิก รายได้ต่อเดือน ช่องทางการตลาด ลำดับรายการ ชนิดพืชผักสวนครัว และพื้นที่ปลูก ปัญหา

- อาหารแปรรูป ได้แก่ มีการจัดตั้งกลุ่มหรือไม่ ชื่อกลุ่ม จำนวนสมาชิก รายได้ต่อเดือน ชนิดอาหาร ลำดับรายการ ชื่อผลิตภัณฑ์ จำนวน และราคา/ชิ้น

ส่วนที่ 4 รายจ่ายของครัวเรือน จำนวน 4 ปัจจัย ได้แก่ ลำดับรายการ รายการค่าใช้จ่าย บาท/เดือน และรวม/ปี

ส่วนที่ 5 หนี้สินของครัวเรือน จำนวน 4 ปัจจัย ได้แก่ ลำดับรายการ แหล่งเงินกู้ ปริมาณเงินกู้ และเงื่อนไข

ส่วนที่ 6 ผลกระทบจากสถานการณ์การระบาดของ COVID - 19 จำนวน 7 ปัจจัย ได้แก่ ได้รับผลกระทบหรือไม่ วีถีชีวิตประจำวัน อาชีพ รายได้ การศึกษา อื่น ๆ และต้องการความช่วยเหลือ

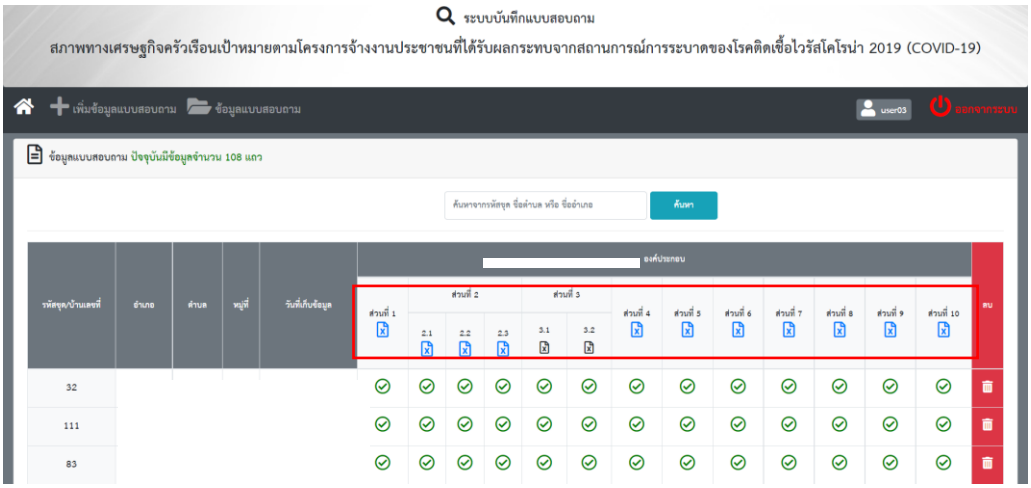
ส่วนที่ 7 การใช้เทคโนโลยีสารสนเทศ จำนวน 15 ปัจจัย ได้แก่ สมาร์ทโฟน (เครื่อง) คอมพิวเตอร์ตั้งโต๊ะ (เครื่อง) คอมพิวเตอร์โน๊ตบุ๊ค (เครื่อง) แท็บเล็ต ไอแพด ช่องทางรับรู้

ข้อมูลข่าวสาร ใช้คอมพิวเตอร์ทำกิจกรรมอะไร ครัวเรือนมีการใช้งานอินเตอร์เน็ต ใช้แพคเกจอินเตอร์เน็ตลักษณะใด ค่าใช้จ่ายอินเตอร์เน็ต/เดือน ใช้อินเตอร์เน็ตจากสถานที่ใด ใช้อินเตอร์เน็ตเพื่อขายสินค้าผ่านช่องทาง ซื้อสินค้าผ่านช่องทาง และข้อเสนอแนะ

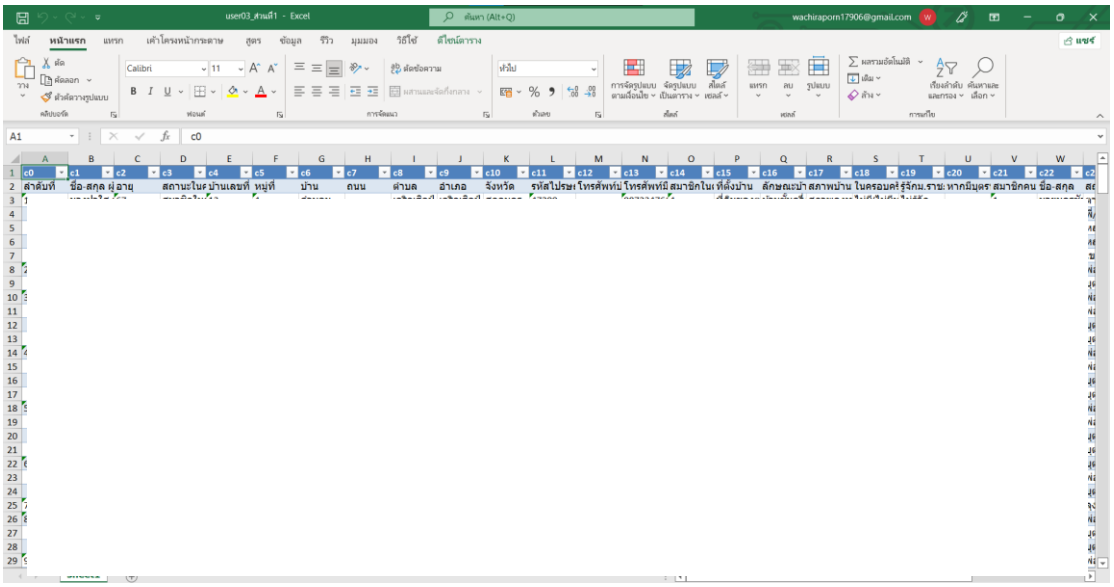
ส่วนที่ 8 การเข้าร่วมการละเล่น การฟ้อน การรำ พิธีกรรมตามวิถีวัฒนธรรมชุมชน จำนวน 5 ปัจจัย ได้แก่ ลำดับรายการ ชื่อ (การละเล่น/ประเพณี) จัดขึ้นในเดือน วัตถุประสงค์ บทบาท/หน้าที่

ส่วนที่ 9 การเข้าร่วมโครงการที่ผ่านมาย้อนหลัง 3 ปี จำนวน 5 ปัจจัย ได้แก่ ลำดับรายการ ชื่อ หน่วยงานที่ดำเนินการ สถานการณ์ปัจจุบัน และปัญหา/อุปสรรค

ส่วนที่ 10 ข้อคิดเห็นและข้อเสนอแนะเพิ่มเติม จำนวน 1 ปัจจัย ได้แก่ ข้อคิดเห็นและข้อเสนอแนะเพิ่มเติม



ภาพที่ 3.2 ระบบฐานข้อมูลเศรษฐกิจครัวเรือน



ภาพที่ 3.3 ข้อมูลเศรษฐกิจครัวเรือนรูปแบบไฟล์ Excel

จากภาพที่ 3.3 เป็นการดาวน์โหลดข้อมูลออกมาจากระบบฐานข้อมูลเศรษฐกิจครัวเรือน ให้อยู่ในรูปแบบไฟล์ Excel เพื่อนำมาใช้ในขั้นตอนการเตรียมข้อมูลสำหรับวิเคราะห์ปัจจัยที่สำคัญ

3.2 การเตรียมข้อมูล (Data Preprocessing)

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

1.3.2.1 การรวบรวมข้อมูล

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

เมื่อได้จำนวนครัวเรือนแล้วจากนั้นคัดเลือกปัจจัย ซึ่งในจำนวนครัวเรือนเหล่านี้มีข้อมูลบางปัจจัยไม่สมบูรณ์ และไม่เกี่ยวข้องกับการหาระดับเศรษฐกิจครัวเรือน จึงได้ตัดปัจจัยเหล่านี้ออกไป จะได้ปัจจัยทั้งหมด 13 ปัจจัย ดังต่อไปนี้

ส่วนที่ 1 ข้อมูลทั่วไปครัวเรือน มีทั้งหมด 31 ปัจจัย 2,909 ครัวเรือน ผู้วิจัยได้วิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (นิภารัตน์ นักตรีพงศ์, 2561: 196; สมยศ ประจันบาล, 2548-2555: 5) ทั้งหมด 3 ปัจจัย ได้แก่ อายุ อาชีพ และรายได้เฉลี่ย/เดือน

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

ส่วนที่ 3 อาชีพ และรายได้ของครัวเรือน มีทั้งหมด 82 ปัจจัย 2,909 ครัวเรือน ผู้วิจัยได้วิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (สุวรัฐ แลสันกลาง, พิบูลย์ ชยโอว์สกุล, ฐิฏิกานต์ สุริยะสาร และชุตินิษฐ์ ปานคำ, 2563) ทั้งหมด 3 ปัจจัย ได้แก่ ผลผลิต/ไร่ ต้นทุน และจำนวนไร่

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

ส่วนที่ 5 หนี้สินของครัวเรือน มีทั้งหมด 4 ปัจจัย 2,909 ครัวเรือน ผู้วิจัยได้วิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (นิภารัตน์ นักตรีพงศ์, 2561: 196; สุวรัฐ แลสันกลาง, พิบูลย์ ชยโอว์สกุล, ฐิฏิกานต์ สุริยะสาร, และชุตินิษฐ์ ปานคำ, 2563: 40-43) ทั้งหมด 2 ปัจจัย ได้แก่ แหล่งเงินกู้ และปริมาณเงินกู้

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

ในส่วนที่ 6 ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรน่า 2019 ส่วนที่ 8 การเข้าร่วมการละเล่น การฟ้อน การรำ พิธีกรรมตามวิถีวัฒนธรรมชุมชน ส่วนที่ 9 การเข้าร่วมโครงการที่ผ่านมาย้อนหลัง 3 ปี และส่วนที่ 10 ข้อคิดเห็น และข้อเสนอแนะเพิ่มเติม ผู้วิจัยได้วิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน พบว่าทั้ง 3 ส่วน ไม่มีปัจจัยไหนที่ส่งผลต่อเศรษฐกิจครัวเรือน

จากข้อมูลครัวเรือนผู้วิจัยได้วิเคราะห์ข้อมูล และเตรียมข้อมูลให้เหมาะสมเพื่อนำมาใช้สร้างตัวแบบการวิเคราะห์ปัจจัยของข้อมูลเศรษฐกิจครัวเรือน รวมได้ทั้งหมด 13 ปัจจัย ดังแสดงในตารางที่ 3.2

ตารางที่ 3.2 แสดงปัจจัยที่ส่งผลต่อเศรษฐกิจครัวเรือน

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 การใช้เทคโนโลยีสารสนเทศ	
12	การใช้อินเทอร์เน็ต
13	ช่องทางการขายสินค้า

จากนั้นผู้วิจัยได้ทำความสะอาดข้อมูล และแปลงรูปแบบข้อมูล เพราะข้อมูลครัวเรือนทั้งหมดที่ได้เก็บมานั้นมีรูปแบบครัวเรือนที่ยังไม่สมบูรณ์ ซึ่งในงานวิจัยนี้จะเน้นและคัดเลือกเฉพาะข้อมูลครัวเรือนที่สมบูรณ์ ได้ครัวเรือนมาทั้งหมด จำนวน 1,751 ระเบียน (ครัวเรือน) แล้วทำให้ได้ปัจจัยในการสร้างตัวแบบจำนวน 16 ปัจจัย ดังแสดงในตารางที่ 3.3 เพื่อใช้ในการวิเคราะห์ปัจจัยที่เหมาะสม จากนั้นทำการแปลงรูปแบบข้อมูล ดังแสดงในตารางที่ 3.4, 3.5, 3.6

ตารางที่ 3.3 รายละเอียดของตัวแปรที่เป็นคุณลักษณะของกลุ่มตัวอย่างเศรษฐกิจครัวเรือน

ลำดับ	คุณลักษณะ	รายละเอียด	ชนิดข้อมูล
1	Education	วัยเรียน	Numeric
2	Working	วัยทำงาน	Numeric
3	Old	วัยสูงอายุ	Numeric
4	Occupation	อาชีพ	Nominal
5	AverageInY	รวมรายได้เฉลี่ย/ปี	Numeric
6	AssetValue	มูลค้าทรัพย์สิน	Numeric
7	AnimalHus	วัตถุประสงค์การเลี้ยงสัตว์	Nominal
8	Area	พื้นที่ก่อให้เกิดรายได้	Numeric
9	ProductCos	ต้นทุนการผลิตทำการเกษตร	Numeric
10	Product	ผลผลิตที่ได้จากการทำเกษตร	Numeric
11	TotalExpY	รวมค่าใช้จ่าย/ปี	Numeric
12	LoanB	หนี้ในระบบ	Nominal
13	LoanS	หนี้นอกระบบ	Nominal
14	TotalLia	รวมปริมาณหนี้สิน	Numeric
15	InternetUse	การใช้อินเทอร์เน็ตที่ก่อให้เกิดรายได้	Nominal
16	Sales Cha	ช่องทางการขายสินค้าที่ก่อให้เกิดรายได้	Nominal
17	Economic Threshold: ET	การจัดหมวดหมู่ คลาสคำตอบ Low Econ Lv = ระดับเศรษฐกิจรายได้น้อย Middle Econ Lv = ระดับเศรษฐกิจรายได้ปานกลาง High Econ Lv = ระดับเศรษฐกิจรายได้สูง	Nominal

ตารางที่ 3.4 ข้อมูลเศรษฐกิจครัวเรือนที่ผ่านการทำความสะอาด และแปลงรูปแบบคุณลักษณะเป็น Numeric, Nominal

ลำดับ	Education	Working	Old	Occupation	AssetValue	AnimalHus	Area	ProductCos	Product	LoanB	LoanS	TotalExpY	TotalLia	AverageInY	InternetUse	SalesCha
1	0	3	0	Agricultural	1567650	No	1200	11400	2800	Yes	No	145560	120000	444000	Yes	Yes
2	0	2	0	Agricultural	299750	Yes	1600	11400	2800	Yes	No	93600	140000	484000	Yes	Yes
3	1	4	1	Agricultural	437650	No	800	2400	2100	Yes	No	241880	90000	84000	Yes	Yes
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1749	2	2	0	Agricultural	1115500	No	600	7600	0	No	No	224400	0	590000	Yes	No
1750	0	1	0	Agricultural	381500	Yes	20000	16100	35200	Yes	No	39000	60000	108000	Yes	No
1751	0	1	0	Agricultural	310500	Yes	300	3000	9200	No	No	56400	0	272000	No	No

จากที่ได้ทำความสะอาดข้อมูล และแปลงรูปแบบข้อมูลดังแสดงในตารางที่ 3.4 พบว่า

ส่วนที่ 1 ได้มีการเปลี่ยนรูปแบบปัจจัย เช่น อายุ ได้ทำการแยกปัจจัยออกมาเป็น 3 ปัจจัย คือ วัยเรียน วัยทำงาน และวัยสูงอายุ เพราะบางครัวเรือนนั้นมีสมาชิกในครัวเรือนมากกว่า 1 คน (ระเบียน) จึงเปลี่ยนแปลงรูปแบบข้อมูลให้เหลือครัวเรือนละ 1 ระเบียน

ส่วนที่ 2 ได้มีการลดจำนวนระเบียนในครัวเรือน โดยการใช้สูตร SUM เพื่อหาผลบวกของทรัพย์สินครัวเรือนทั้งหมดให้เหลือ 1 ระเบียน (ครัวเรือน)

ส่วนที่ 3 ได้เปลี่ยนแปลงหน่วยไร่เป็นงาน เปลี่ยนจากหน่วยงานให้เป็นหน่วยตารางวา เช่น 1 ไร่ = 4 งาน 4 งาน = 400 ตารางวา เพราะจะได้ง่ายต่อการนำเข้าโปรแกรม

ส่วนที่ 4 ได้เปลี่ยนแปลงข้อมูลค่าใช้จ่าย/เดือนของครัวเรือน ให้เป็นรายจ่ายเฉลี่ย/ปี โดยการใช้สูตร (ค่าใช้จ่ายแต่ละคน * 12 นำมาบวกกัน) จะได้ค่าใช้จ่ายเฉลี่ย/ปี ของครัวเรือน

ส่วนที่ 5 ได้มีการเปลี่ยนรูปแบบปัจจัยของแหล่งเงินกู้ แยกออกมาเป็น 2 ปัจจัย ได้แก่ หนี้ในระบบ และหนี้นอกระบบ ตัวแปรของ 2 ปัจจัย คือ เป็นหนี้ = Yes ไม่เป็นหนี้ = No และยังมีข้อมูลที่ขาดหายไปจึงพิจารณาจากค่าข้อมูลที่ปรากฏซ้ำกันมากที่สุดแล้วเติมค่าข้อมูลที่ขาด

ส่วนที่ 7 ได้มีการเปลี่ยนรูปแบบตัวแปรของปัจจัย การใช้อินเทอร์เน็ต ตัวแปรได้แก่ ใช้อินเทอร์เน็ต = Yes ไม่ใช้อินเทอร์เน็ต = No และช่องทางการขายสินค้า ตัวแปรได้แก่ ขายสินค้าบนออนไลน์ = Yes ไม่ได้ขายสินค้าบนออนไลน์ = No

จากนั้นปรับค่าคุณลักษณะของข้อมูลบางปัจจัย ดังแสดงในตารางที่ 3.6 ได้แก่ รวมรายได้เฉลี่ย/ปี พื้นที่ก่อให้เกิดรายได้ มูลค้าทรัพย์สิน ต้นทุนการผลิต ผลผลิต รวมค่าใช้จ่าย/ปี รวมปริมาณหนี้สิน ดังแสดงในตารางที่ 3.5 ให้อยู่ในช่วงค่าน้อยสุด และค่ามากสุดตามที่กำหนด ซึ่งนิยมใช้ค่าน้อยสุดเป็น 0 และ ค่ามากสุดเป็น 1 บางครั้งนิยมเรียกวิธีการนี้ว่า การทำข้อมูลให้เป็นปกติแบบ 0-1 (0-1 Normalization) (นิภาพร ชนะมาร 2560: ดุษฎีนิพนธ์) แสดงดังสมการที่ 3.1

$$V_i' = \frac{V_i - MIN_V}{MAX_V - MIN_V} \quad \text{----------(3.1)}$$

โดยที่ v'_i หมายถึง ค่าใหม่ของคุณลักษณะตัวที่ i ของข้อมูล v,
v_i หมายถึง ค่าของคุณลักษณะตัวที่ i ของข้อมูล v เดิม,
MAX_v หมายถึง ค่าที่มากที่สุดของคุณลักษณะนั้น

MIN_v หมายถึง ค่าที่น้อยที่สุดของคุณลักษณะนั้น

ยกตัวอย่างการหาค่า Normalization ของปัจจัยรวมรายได้เฉลี่ย/ปี (AverageInY) แสดงวิธีการดังนี้

$$v_i' = \frac{444000-20250}{947200-20250}$$

$$v_i' = 0.457$$

- ค่า 444000 คือ ตัวแปรในปัจจัยรวมรายได้เฉลี่ย/ปี แต่ละระเบียน และเป็นค่าที่เราต้องการหาคุณลักษณะใหม่
- ค่า 20250 คือ ตัวแปรในปัจจัยรวมรายได้เฉลี่ย/ปี แต่ละระเบียน โดยการหาค่าเฉลี่ยที่น้อยที่สุด
- ค่า 947200 คือ ตัวแปรปัจจัยรวมรายได้เฉลี่ย/ปี แต่ละระเบียน โดยการหาค่าเฉลี่ยมากที่สุด

เมื่อได้ข้อมูลมาแล้วผู้วิจัยจะทำการคัดเลือกปัจจัยที่สำคัญด้วยเทคนิค Gain Ratio โดยจะนำข้อมูลจากตารางที่ 3.5 เพื่อนำมาเปรียบเทียบ All Feature

ตารางที่ 3.5 การปรับค่าคุณลักษณะของข้อมูลเศรษฐกิจครัวเรือนบางปัจจัย

ลำดับ	Education	Working	Old	Occupation	AssetValue	AnimalHus	Area	ProductCos	Product	LoanB	LoanS	TotalLia	TotalExpY	AverageInY	InternetUse	SalesCha
1	0	3	0	Agricultural	0.129	No	0.010	0.070	0.001	Yes	No	0.017	0.027	0.457	Yes	Yes
2	0	2	0	Agricultural	0.025	Yes	0.020	0.070	0.001	Yes	No	0.019	0.017	0.500	Yes	Yes
3	1	4	1	Agricultural	0.036	No	0.010	0.020	0.001	Yes	No	0.013	0.045	0.069	Yes	Yes
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1749	2	2	0	Agricultural	0.092	No	0.010	0.050	0.000	No	No	0.000	0.042	0.615	Yes	No
1750	0	1	0	Agricultural	0.031	Yes	0.220	0.100	0.015	Yes	No	0.008	0.007	0.095	Yes	No
1751	0	1	0	Agricultural	0.026	Yes	0.000	0.020	0.004	No	No	0.000	0.010	0.272	No	No

3.3 การแบ่งชุดข้อมูล

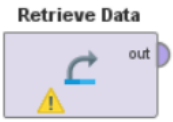
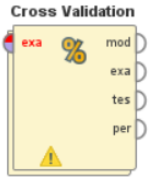
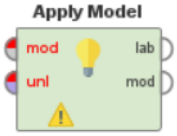
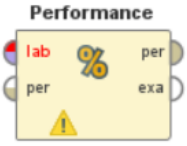
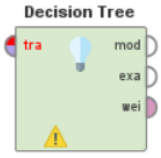
ในงานวิจัยนี้ได้มีการแบ่งข้อมูลเป็นแบบเปอร์เซ็นต์ โดยจะรักษาสัดส่วนของข้อมูล และจะทำการสุ่มข้อมูล (Random) ตามค่าสัดส่วนร้อยละ 60:40, 70:30 และ 80:20 ของข้อมูลจำนวน 1,751 ระเบียน (ครัวเรือน) เพื่อใช้ในการวิเคราะห์ปัจจัยแบ่งออกเป็น 2 ส่วน คือ 1) ข้อมูลเรียนรู้ (Training Data) 2) ข้อมูลทดสอบ (Testing Data) ดังแสดงในตารางที่ 3.6

ตารางที่ 3.6 การแบ่งชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ

เปอร์เซ็นต์	ข้อมูลสำหรับเรียนรู้	ข้อมูลสำหรับทดสอบ
60:40	1051	700
70:30	1226	525
80:20	1400	351

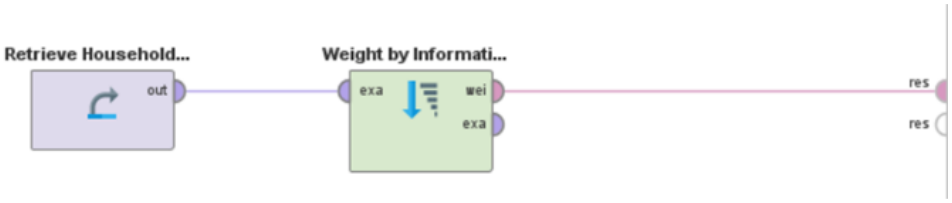
การทดลองนี้ผู้วิจัยได้ทำการทดลองเปรียบเทียบประสิทธิภาพของตัวแบบ Decision Tree จำนวน 1,751 ระเบียน (ครัวเรือน) ใช้ข้อมูลปัจจัยทั้งหมด (All Feature) ดังตารางที่ 3.5 และปัจจัยสำคัญที่คัดเลือกด้วยวิธีการ Gain Ratio ดังตารางที่ 3.8 เพื่อใช้ในการทดสอบข้อมูลแต่ละรอบจะมีตัวโอเปอเรเตอร์ (Operator) ที่เกี่ยวข้อง ดังแสดงในตารางที่ 3.7

ตารางที่ 3.7 โอเปอเรเตอร์ (Operator) ที่เกี่ยวข้องในกระบวนการวิเคราะห์ปัจจัยสำคัญ

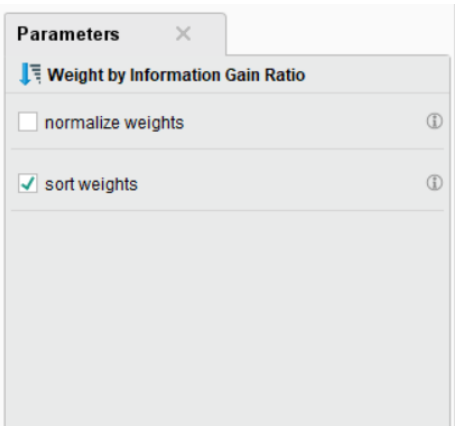
โอเปอเรเตอร์ (Operator)		รายละเอียด
	Read Excel	ใช้สำหรับในการอ่านไฟล์เอกสาร Excel
	Weight by Information Gain Ratio	ใช้สำหรับวิเคราะห์ปัจจัยที่สำคัญ
	Split Data	ใช้สำหรับแบ่งชุดข้อมูลเพื่อใช้ในการวิเคราะห์ปัจจัยสำคัญ
	Cross Validation	ใช้สำหรับแบ่งข้อมูลสำหรับการวิเคราะห์ปัจจัยสำคัญ และทดสอบแบบจำลอง แบบ K-fold Cross-Validation
	Apply Model	ใช้สำหรับในการพยากรณ์ชุดข้อมูลทดสอบ (Testing Data)
	Performance	ใช้สำหรับแสดงตัวชี้วัดของการวิเคราะห์ปัจจัยสำคัญ
	Decision Tree	ใช้สำหรับสร้างต้นไม้ตัดสินใจเพื่อทดสอบปัจจัย

3.4 การคัดเลือกคุณสมบัติด้วย Gain Ratio

หลังจากขั้นตอนการแปลงข้อมูล จะได้ข้อมูลมาใช้ในการวิเคราะห์ปัจจัยที่สำคัญ ผู้วิจัยได้นำข้อมูลจาก All Feature ที่ได้จากขั้นตอน Data Preprocessing โดยใช้ข้อมูลปัจจัย 16 ปัจจัย จำนวน 1,751 ครัวเรือน นำไปเข้าโปรแกรม RapidMiner Studio และวิเคราะห์ปัจจัยที่สำคัญ ด้วยเทคนิค Gain Ratio ดังภาพที่ 3.4 และกำหนดค่า Parameters เป็นค่า Default ดังภาพที่ 3.5 ได้ปัจจัยดังภาพที่ 3.6 และแสดงปัจจัยที่สำคัญ 10 ปัจจัย โดยคัดเลือกปัจจัยค่า weight จากค่าที่ 0.06 เป็นต้นไป (ปะพาดา ณ วิเชียร, ภาคภูมิ มันแอ, ญาณพัฒน ชูชื่น และ และสุภาวดี มากอน. 2563) ดังตารางที่ 3.8



ภาพที่ 3.4 ตัวอย่างการ Feature Selection ด้วยเทคนิค Gain Ratio



ภาพที่ 3.5 แสดงการกำหนด Parameters

attribute	weight
SalesCha	0.004
AnimalHus	0.005
LoanS	0.014
ProductionCos	0.014
Product	0.018
Area	0.026
LoanB	0.090
Occupation	0.105
InternetUse	0.159
TotalExpY	0.207
AssetValue	0.207
TotalLia	0.207
Old	0.226
Education	0.226
Working	0.226
AverageInY	0.644

ภาพที่ 3.6 ข้อมูลการ Feature Selection ด้วยเทคนิค Gain Ratio

ตารางที่ 3.8 รายละเอียดของตัวแปร Gain Ratio Feature Selection ที่เป็นคุณลักษณะของกลุ่มตัวอย่างเศรษฐกิจครัวเรือน

ลำดับ	คุณลักษณะ	รายละเอียด	ชนิดข้อมูล
1	AverageInY	รวมรายได้เฉลี่ย/ปี	Numeric
2	Working	วัยทำงาน	Numeric
3	Old	วัยสูงอายุ	Numeric
4	Education	วัยเรียน	Numeric
5	TotalExpY	รวมค่าใช้จ่าย/ปี	Numeric
6	InternetUse	การใช้อินเทอร์เน็ตที่ก่อให้เกิดรายได้	Nominal
7	Occupation	อาชีพ	Nominal
8	TotalLia	รวมปริมาณหนี้สิน	Numeric
9	AssetValue	มูลค้าทรัพย์สิน	Numeric
10	LoanB	หนี้ในระบบ	Nominal

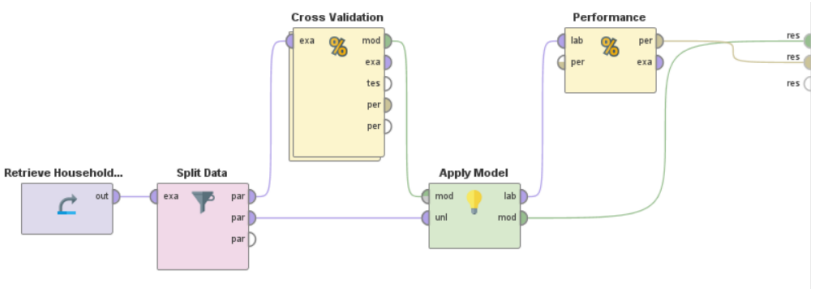
เมื่อได้ปัจจัยที่คัดเลือกมาแล้วจะนำไปสู่ขั้นตอนการสร้างตัวแบบ นำไปเข้าโปรแกรม RapidMiner Studio และเปรียบเทียบระหว่าง All Feature และปัจจัยสำคัญที่คัดเลือกด้วยวิธีการ Gain Ratio

3.5 การสร้างโมเดล (Modeling)

เมื่อจัดเตรียมข้อมูลสำหรับการเรียนรู้และข้อมูลสำหรับการทดสอบเสร็จสิ้นจะเริ่มกระบวนการวิเคราะห์ปัจจัยสำคัญ การวิเคราะห์ข้อมูลจะเป็นการวิเคราะห์ผ่านโปรแกรม RapidMiner Studio ในขั้นตอนนี้จะไม่มีการเปรียบเทียบกับเทคนิคอื่น ๆ แต่จะเปรียบเทียบข้อมูลที่ได้มาจากขั้นตอน Data Preprocessing (All Feature) กับข้อมูลที่ได้มาจากขั้นตอน Feature Selection แล้วจะทำการสร้างตัวแบบด้วยเทคนิค Decision Tree การวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Feature Selection กับ All Feature ว่าข้อมูลไหนได้ประสิทธิภาพที่ดีที่สุด เพื่อใช้ในการทดสอบข้อมูลแต่ละรอบจะมีโอเปอเรเตอร์ (Operator) ที่เกี่ยวข้อง ดังตารางที่ 3.7

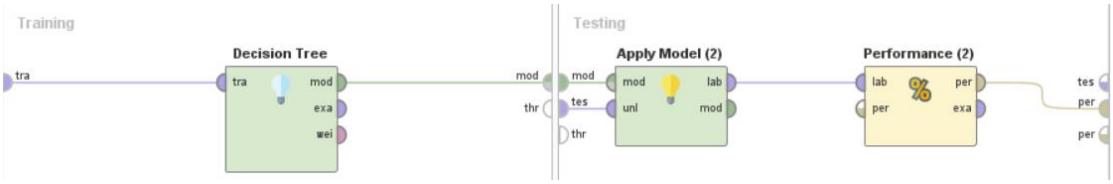
แสดงการวิเคราะห์ปัจจัยสำคัญสำหรับข้อมูลเศรษฐกิจครัวเรือน โดยใช้เครื่องมือสำหรับวิเคราะห์ข้อมูล RapidMiner Studio ในการสร้างตัวแบบ และทดสอบ

3.5.1 การสร้างตัวแบบด้วยข้อมูล All Feature

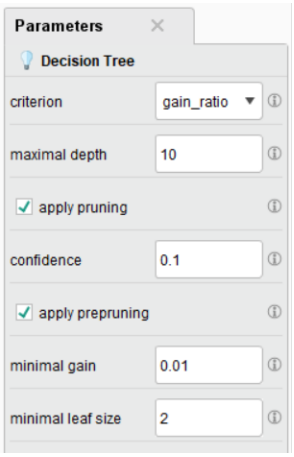


ภาพที่ 3.7 ตัวอย่างการสร้างตัวแบบด้วยข้อมูล All Feature ด้วยเทคนิค Decision Tree

เริ่มต้นด้วยการนำข้อมูลที่แบ่งข้อมูลออกเป็น 2 ส่วน เข้าสู่แบบจำลอง ดังแสดงในภาพที่ 3.7 ภายในโอเปอเรเตอร์ Cross Validation เป็นการวิเคราะห์ปัจจัย ดังแสดงในภาพที่ 3.8 กำหนดค่าความลึกของโหนดใบ (Maximal Depth) มีค่าเท่ากับ 10 ดังแสดงในภาพที่ 3.9 ทำการวิเคราะห์ปัจจัยที่สำคัญที่เหมาะสมกับข้อมูลเศรษฐกิจครัวเรือน โดยมีแผนภาพ Decision Tree ตัวอย่างการประเมินประสิทธิภาพของแบบจำลอง Decision Tree ดังแสดงในภาพที่ 3.7 – 3.14



ภาพที่ 3.8 การทดสอบประสิทธิภาพของเทคนิค Decision Tree



ภาพที่ 3.9 แสดงการกำหนดค่า Maximal Depth

นำข้อมูล All Feature สร้างตัวแบบด้วยเทคนิค Decision Tree ข้อมูลตามค่าสัดส่วนร้อยละ 70:30 โดยใช้ 5-Fold Cross Validation จะได้ผลลัพธ์ดังภาพที่ 3.10

accuracy: 98.86%

	true High econ lv	true Low econ lv	true Middle econ lv	class precision
pred. High econ lv	271	1	4	98.19%
pred. Low econ lv	0	90	1	98.90%
pred. Middle econ lv	0	0	158	100.00%
class recall	100.00%	98.90%	96.93%	

ภาพที่ 3.10 แสดงค่าความถูกต้องของเทคนิค Decision Tree การวิเคราะห์ปัจจัยด้วย All Feature

จากภาพที่ 3.10 ผลที่ได้ เป็นตัวอย่างผลของชุดข้อมูลเรียนรู้ และข้อมูลทดสอบ 70:30 และภาพนี้เป็นการวัดความถูกต้องของตัวแบบ โดยพิจารณารวมทุกคลาส และในตารางภาพมีคลาสคำตอบอยู่ 3 ค่า คือ High econ lv, Middle econ lv และ Low econ lv ฉะนั้นตาราง Confusion Matrix นี้จะสร้างได้เป็นตารางขนาด 3*3 โดยข้อมูลด้านคอลัมน์คือ คลาสที่

อยู่ในข้อมูลเรียนรู้ (Actual) และข้อมูลในแนวนอนคือ คลาสที่ตัวแบบจำการวิเคราะห์ปัจจัย ได้ (Predicted) แสดงวิธีการดังนี้

$$\text{Accuracy} = \frac{((271 + 90 + 158) + (0 + 0 + 0))}{525}$$

$$= \frac{519}{525}$$

$$=0.9886$$

$$= 98.86\%$$

True Positive: TP คือ จำนวนข้อมูลที่ถูกว่าเป็นคลาสClassification = High econ lv มีจำนวน 217

False Positive: FP คือ จำนวนข้อมูลที่ผิดมาเป็นคลาส Classification = Middle econ lv มีจำนวน 4

True Negative: TN คือ จำนวนข้อมูลที่ถูกว่าเป็นคลาส Classification = Low econ lv มีจำนวน 0

True Positive: TP คือ จำนวนข้อมูลที่ถูกว่าเป็นคลาสClassification = Middle econ lv มีจำนวน 378

False Positive: FP คือ จำนวนข้อมูลที่ผิดมาเป็นคลาส Classification = High econ lv มีจำนวน 2

True Negative: TN คือ จำนวนข้อมูลที่ถูกว่าเป็นคลาส Classification = Low econ lv มีจำนวน 0

True Positive: TP คือ จำนวนข้อมูลที่ถูกว่าเป็นคลาสClassification = Low econ lv มีจำนวน 212

True Negative: TN คือ จำนวนข้อมูลที่ผิดมาเป็นคลาส Classification = High econ lv มีจำนวน 0

True Negative: TN คือ จำนวนข้อมูลที่ถูกว่าเป็นคลาส Classification = Middle econ lv มีจำนวน 0

นำข้อมูล All Feature สร้างตัวแบบด้วยเทคนิค Decision Tree ข้อมูลตามค่าสัดส่วนร้อยละ 70:30 โดยใช้ 5-Fold Cross Validation จะได้ผลลัพธ์ดังภาพที่ 3.11

weighted_mean_precision: 99.03%, weights: 1, 1, 1

	true High econ lv	true Low econ lv	true Middle econ lv	class precision
pred. High econ lv	271	1	4	98.19%
pred. Low econ lv	0	90	1	98.90%
pred. Middle econ lv	0	0	158	100.00%
class recall	100.00%	98.90%	96.93%	

ภาพที่ 3.11 แสดงค่าความแม่นยำของเทคนิค Decision Tree การวิเคราะห์ปัจจัยด้วย All Feature

จากภาพที่ 3.11 เป็นการวัดความแม่นยำของตัวแบบ โดยพิจารณาแยกทีละคลาส แสดงวิธีการดังนี้

$$\text{Precision (High)} = \frac{271}{271+1+4}$$

$$= 0.9819$$

$$\text{Precision (Middle)} = \frac{158}{158+0}$$

$$= 1$$

$$\text{Precision (Low)} = \frac{90}{90+1}$$

$$= 0.9890$$

$$(0.9819 + 1 + 0.9890)/3 = 0.9903$$

$$= 99.03\%$$

นำข้อมูล All Feature สร้างตัวแบบด้วยเทคนิค Decision Tree ข้อมูลตามค่าสัดส่วนร้อยละ 70:30 โดยใช้ 5-Fold Cross Validation จะได้ผลลัพธ์ดังภาพที่ 3.12

weighted_mean_recall: 98.61%, weights: 1, 1, 1

	true High econ lv	true Low econ lv	true Middle econ lv	class precision
pred. High econ lv	271	1	4	98.19%
pred. Low econ lv	0	90	1	98.90%
pred. Middle econ lv	0	0	158	100.00%
class recall	100.00%	98.90%	96.93%	

ภาพที่ 3.12 แสดงค่าความระลึกของเทคนิค Decision Tree การวิเคราะห์ปัจจัยด้วย All Feature

จากภาพที่ 3.12 เป็นการสัดความถูกต้องของตัวแบบ โดยพจารณาแยกทีละคลาส แสดงวิธีการดังนี้

$$\text{Recall (High)} = \frac{271}{271+0}$$

$$= 1$$

$$\text{Recall (Middle)} = \frac{158}{158+4+1}$$

$$= 0.9693$$

$$\text{Recall (Low)} = \frac{90}{90+1}$$

$$= 0.9890$$

$$(1 + 0.9693 + 0.9890) /3 = 0.9861$$

$$= 98.61\%$$

Precision	Recall	P*R	PR*2	P+R	F-Measure
99.03	98.61	9765.348	19530.7	197.64	98.82

ภาพที่ 3.13 แสดงค่าถ่วงดุลของของเทคนิค Decision Tree การวิเคราะห์ปัจจัยด้วย All Feature

จากภาพที่ 3.13 เป็นการวัดค่าความแม่นยำ และค่าถ่วงดุล พร้อมกันของตัวแบบ โดยพิจารณาแยกทีละคลาส ดังแสดงวิธีการดังนี้

$$\text{F-measure} = 2\frac{(99.03*99.61)}{99.03+99.61}$$

$$= 98.82\%$$

root_mean_squared_error

root_mean_squared_error: 0.084 +/- 0.000

ภาพที่ 3.14 แสดงค่ารากที่สองของค่าความคลาดเคลื่อนเฉลี่ยของเทคนิค Decision Tree การวิเคราะห์ปัจจัยด้วย All Feature

จากภาพที่ 3.14 คือค่าคลาดเคลื่อนกำลังสองเฉลี่ยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย เป็นการถอดรากที่สอง จะทำให้ได้ค่าบวกลบ สำหรับการบอกค่า Error ของตัวแบบ แสดงวิธีการดังนี้

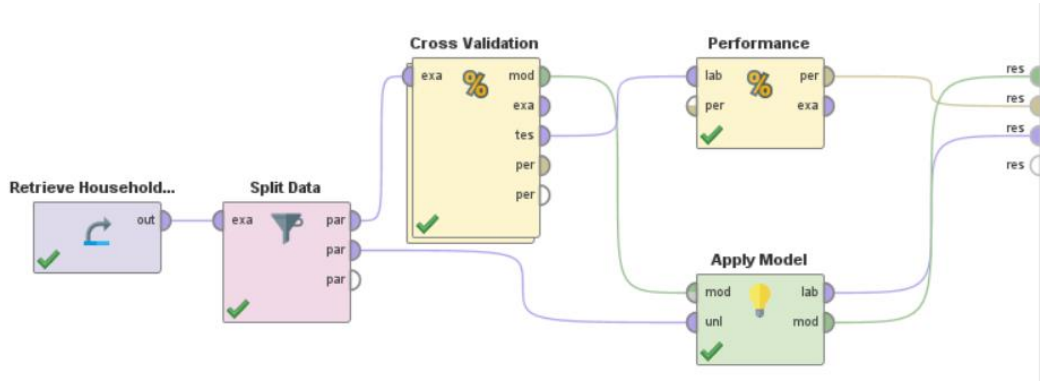
$$\text{RMSE} = \sqrt{\frac{((271-276)+(90-91)+(0))^2}{525}}$$

$$= \sqrt{0.685714286}$$

$$= 0.084$$

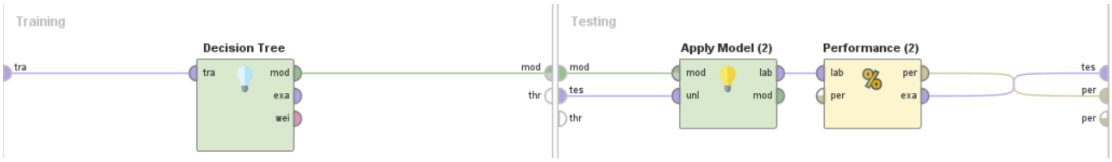
3.5.2 การสร้างตัวแบบด้วยข้อมูล Gain Ratio Feature Selection

จากภาพที่ 3.15 แสดงให้เห็นถึงการวิเคราะห์ปัจจัยที่สำคัญ ด้วยปัจจัยที่ได้มาจาก เทคนิค Gain Ratio ซึ่งภายในโอเปอเรเตอร์ Cross Validation เป็นการวิเคราะห์ปัจจัยด้วยตัวแบบ Decision Tree

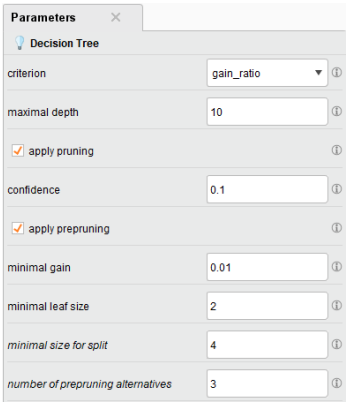


ภาพที่ 3.15 ตัวอย่างการสร้างตัวแบบด้วยข้อมูล Gain Ratio ด้วยเทคนิค Decision Tree

เริ่มต้นด้วยการนำข้อมูลที่แบ่งข้อมูลออกเป็น 2 ส่วน เข้าสู่แบบจำลอง ดังแสดงในภาพที่ 3.16 ภายในโอเปอเรเตอร์ Cross Validation เป็นการวิเคราะห์ปัจจัย ดังแสดงในภาพที่ 3.17 กำหนดค่าความลึกของโหนดใบ (Maximal Depth) มีค่าเท่ากับ 10 ดังแสดงในภาพที่ 3.18 ทำการวิเคราะห์ความเหมาะสมของข้อมูลเศรษฐกิจครัวเรือน โดยมีแผนภาพ Decision Tree ตัวอย่างการประเมินประสิทธิภาพของเทคนิค Decision Tree ดังแสดงในภาพที่ 3.15 – 3.22



ภาพที่ 3.16 การทดสอบประสิทธิภาพของเทคนิค Decision Tree



ภาพที่ 3.17 แสดงการกำหนดค่า Maximal Depth

นำข้อมูล การวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio สร้างตัวแบบด้วยเทคนิค Decision Tree ข้อมูลตามค่าสัดส่วนร้อยละ 70:30 โดยใช้ 5-Fold Cross Validation จะได้ผลลัพธ์ดัง ภาพที่ 3.18

accuracy: 99.51%

	true High econ lv	true Middle econ lv	true Low econ lv
pred. High econ lv	630	4	0
pred. Middle econ lv	2	378	0
pred. Low econ lv	0	0	212

ภาพที่ 3.18 การสร้างตัวแบบด้วยข้อมูล Gain Ratio แสดงค่าความถูกต้องของเทคนิค Decision Tree

จากภาพที่ 3.18 ผลที่ได้ เป็นตัวอย่างผลของชุดข้อมูลเรียนรู้ และข้อมูลทดสอบ 70:30 และภาพนี้เป็นการวัดความถูกต้องของตัวแบบ โดยพิจารณารวมทุกคลาส และในตารางภาพมีคลาสคำตอบอยู่ 3 ค่า คือ High econ lv, Middle econ lv และ Low econ lv ฉะนั้นตาราง Confusion Matrix นี้จะสร้างได้เป็นตารางขนาด 3*3 โดยข้อมูลด้านคอลัมน์คือ คลาสที่อยู่ในข้อมูลเรียนรู้ (Actual) และข้อมูลในแนวนอนคือ คลาสที่ตัวแบบจำการวิเคราะห์ปัจจัย ได้ (Predicted) แสดงวิธีการดังนี้

$$\text{Accuracy} = \frac{((630 + 378 + 212) + (0 + 0 + 0))}{1226}$$

$$= \frac{1220}{1226}$$

$$= 0.9951$$

$$= 99.51\%$$

True Positive: TP คือ จำนวนข้อมูลที่พยากรณ์ถูกว่าเป็นคลาส Classification = High econ lv มีจำนวน 630

False Positive: FP คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาส Classification = Middle econ lv มีจำนวน 4

True Negative: TN คือ จำนวนข้อมูลที่พยาพรณ์ถูกว่าเป็นคลาส Classification = Low econ lv มีจำนวน 0

True Positive: TP คือ จำนวนข้อมูลที่พยากรณ์ถูกว่าเป็นคลาส Classification = Middle econ lv มีจำนวน 378

False Positive: FP คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาส Classification = High econ lv มีจำนวน 2

True Negative: TN คือ จำนวนข้อมูลที่พยาพรณ์ถูกว่าเป็นคลาส Classification = Low econ lv มีจำนวน 0

True Positive: TP คือ จำนวนข้อมูลที่พยากรณ์ถูกว่าเป็นคลาส Classification = Low econ lv มีจำนวน 212

True Negative: TN คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาส Classification = High econ lv มีจำนวน 0

True Negative: TN คือ จำนวนข้อมูลที่พยาพรณ์ถูกว่าเป็นคลาส Classification = Middle econ lv มีจำนวน 0

นำข้อมูล การวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio สร้างตัวแบบด้วยเทคนิค Decision Tree ข้อมูลตามค่าสัดส่วนร้อยละ 70:30 โดยใช้ 5-Fold Cross Validation จะได้ผลลัพธ์ดัง ภาพที่ 3.19

weighted_mean_precision: 99.61%, weights: 1, 1, 1

	true High econ lv	true Middle econ lv	true Low econ lv
pred. High econ lv	630	4	0
pred. Middle econ lv	2	378	0
pred. Low econ lv	0	0	212

ภาพที่ 3.19 การสร้างตัวแบบด้วยข้อมูล Gain Ratio แสดงค่าความแม่นยำของเทคนิค Decision

จากภาพที่ 3.19 เป็นการวัดความแม่นยำของตัวแบบ โดยพิจารณาแยกทีละคลาส แสดงวิธีการดังนี้

$$\text{Precision (High)} = \frac{630}{630 + 4}$$
$$= 0.9937$$

$$\text{Precision (Middle)} = \frac{378}{378 + 2}$$
$$= 0.9947$$

$$\text{Precision (Low)} = \frac{212}{212 + 0}$$
$$= 1$$
$$(0.9937 + 0.9947 + 1)/3 = 0.9961$$
$$= 99.61\%$$

weighted_mean_recall: 99.55%, weights: 1, 1, 1

	true High econ lv	true Middle econ lv	true Low econ lv
pred. High econ lv	630	4	0
pred. Middle econ lv	2	378	0
pred. Low econ lv	0	0	212

ภาพที่ 3.20 การสร้างตัวแบบด้วยข้อมูล Gain Ratio แสดงค่าความระลึกของเทคนิค Decision

จากภาพที่ 3.20 เป็นการสัดความถูกต้องของตัวแบบ โดยพจารณาแยกทีละคลาส แสดงวิธีการดังนี้

$$\text{Recall (High)} = \frac{630}{630 + 2}$$
$$= 0.9968$$

$$\text{Recall (Middle)} = \frac{378}{378 + 4}$$
$$= 0.9895$$

$$\text{Recall (Low)} = \frac{212}{212+0}$$
$$= 1$$

$$(0.9968 + 0.9895 + 1) /3 = 0.9955$$
$$= 99.55\%$$

Precision	Recall	P*R	PR*2	P+R	F-Measure
99.61	99.55	9916.1755	19832.351	199.16	99.58

ภาพที่ 3.21 การสร้างตัวแบบด้วยข้อมูล Gain Ratio แสดงค่าถ่วงดุลของเทคนิค Decision Tree

จากภาพที่ 3.21 เป็นการวัดค่าความแม่นยำ และค่าถ่วงดุล พร้อมกันของตัวแบบ โดยพิจารณาแยกทีละคลาส ดังแสดงวิธีการดังนี้

$$\text{F-measure} = 2\frac{(99.61 * 99.55)}{99.61 + 99.55}$$

$$= 99.58\%$$

root_mean_squared_error

root_mean_squared_error: 0.062 +/- 0.000

ภาพที่ 3.22 การสร้างตัวแบบด้วยข้อมูล Gain Ratio แสดงค่ารากที่สองของค่าความคลาดเคลื่อนเฉลี่ยของเทคนิค Decision Tree

จากภาพที่ 3.22 คือค่าคลาดเคลื่อนกำลังสองเฉลี่ยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย เป็นการถอดรากที่สอง จะทำให้ได้ค่าบวกลบ สำหรับการบอกค่า Error แสดงวิธีการดังนี้

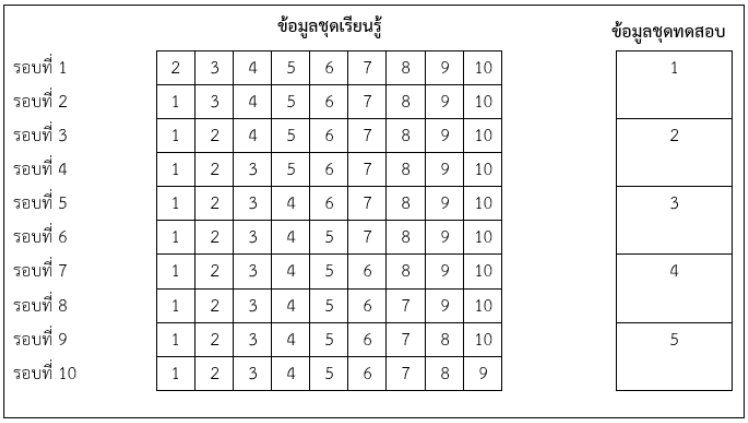
$$\text{RMSE} = \sqrt{\frac{((630 - 634) + (378 - 380) + (0))^2}{1400}}$$

$$= \sqrt{0.003844}$$

$$= 0.062$$

3.6 การวัดประสิทธิภาพของโมเดล (Evaluation)

ในงานวิจัยนี้จะทำการทดสอบค่าความถูกต้องในการวิเคราะห์ปัจจัยด้วยวิธี Cross Validation Test โดยทำการแบ่งสัดส่วนทดสอบประสิทธิ์ภาพด้วยวิธี 5-fold Cross Validation ดังภาพที่ 3.23 และ วิธี 10-fold Cross Validation ดังภาพที่ 3.24 ของข้อมูลจำนวน 1,751 ระเบียน แล้วจะทำการทดสอบระหว่าง All Feature และ ข้อมูลที่ได้มาจากขั้นตอน Feature Selection ด้วยเทคนิค Decision Tree จากผลลัพธ์ปัจจัยที่ได้จากการเปรียบเทียบระหว่าง All Feature และ ข้อมูลที่ได้มาจากขั้นตอน Feature Selection



ภาพที่ 3.23 การแบ่งสัดส่วนทดสอบประสิทธิ์ภาพด้วยวิธี 5-Fold Cross Validation

	ข้อมูลชุดเรียนรู้									ข้อมูลชุดทดสอบ
รอบที่ 1	2	3	4	5	6	7	8	9	10	1
รอบที่ 2	1	3	4	5	6	7	8	9	10	2
รอบที่ 3	1	2	4	5	6	7	8	9	10	3
รอบที่ 4	1	2	3	5	6	7	8	9	10	4
รอบที่ 5	1	2	3	4	6	7	8	9	10	5
รอบที่ 6	1	2	3	4	5	7	8	9	10	6
รอบที่ 7	1	2	3	4	5	6	8	9	10	7
รอบที่ 8	1	2	3	4	5	6	7	9	10	8
รอบที่ 9	1	2	3	4	5	6	7	8	10	9
รอบที่ 10	1	2	3	4	5	6	7	8	9	10

ภาพที่ 3.24 การแบ่งสัดส่วนทดสอบประสิทธิ์ภาพด้วยวิธี 10-Fold Cross Validation

การคำนวณประสิทธิภาพของตัวแบบจำลอง สามารถคำนวณได้จากตาราง Confusion Matrix ซึ่งเป็นตารางสรุปจำนวนข้อมูลที่ตัวแบบมีการวิเคราะห์ปัจจัยได้อย่างถูกต้องและไม่ถูกต้อง

ตารางที่ 3.9 The Confusion Matrix

ค่าที่ทำการพยากรณ์ได้ (Predicted Class)	**ค่าที่แท้จริง (Actual Class)**	
	Class YES	**Class NO**
Class YES	True Positive: TP	False Negative: FN
Class NO	False Positive: FP	True Negative: TN

ทำการประเมินประสิทธิภาพของการวิเคราะห์ปัจจัย ค่าความถูกต้อง (Accuracy) และวัดประสิทธิภาพด้วยการวัดความถูกต้องของตัวแบบ (Recall), การวัดความแม่นยำของตัวแบบ

(Precision), การวัดค่า Precision และ Recall พร้อมกันของตัวแบบ (F-measure) (root mean square error, RMSE) ค่าคลาดเคลื่อนกำลังสองเฉลี่ยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ดังแสดงในสมการที่ (3.2) (3.3) (3.4) (3.5) และ (3.6) ตามลำดับ

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{----------(3.2)}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \text{----------(3.3)}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \text{----------(3.4)}$$

$$\text{F-measure} = 2\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \text{----------(3.5)}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{Error})^2} \text{----------(3.6)}$$

โดยที่ TP คือ ค่าที่พยากรณ์ถูกต้อง (ข้อมูลบอกว่าจริง พยากรณ์ว่าจริง)

TN คือ ค่าที่พยากรณ์ถูกต้อง (ข้อมูลบอกว่าไม่จริง พยากรณ์ว่าไม่จริง)

FP คือ ค่าที่พยากรณ์ไม่ถูกต้อง (ข้อมูลบอกว่าจริง พยากรณ์ว่าไม่จริง)

FN คือ ค่าที่พยากรณ์ไม่ถูกต้อง (ข้อมูลบอกว่าไม่จริง พยากรณ์ว่าจริง)

3.7 นำไปใช้งาน (Deployment)

เป็นการนำปัจจัยสำคัญของข้อมูลเศรษฐกิจครัวเรือนที่เหมาะสมที่สุดไปใช้งานจริง เพื่อวิเคราะห์และแก้ปัญหาที่ต้องการ สำหรับสนับสนุนหรือเป็นข้อมูลประกอบการตัดสินใจในการวิจัยในลำดับต่อไป