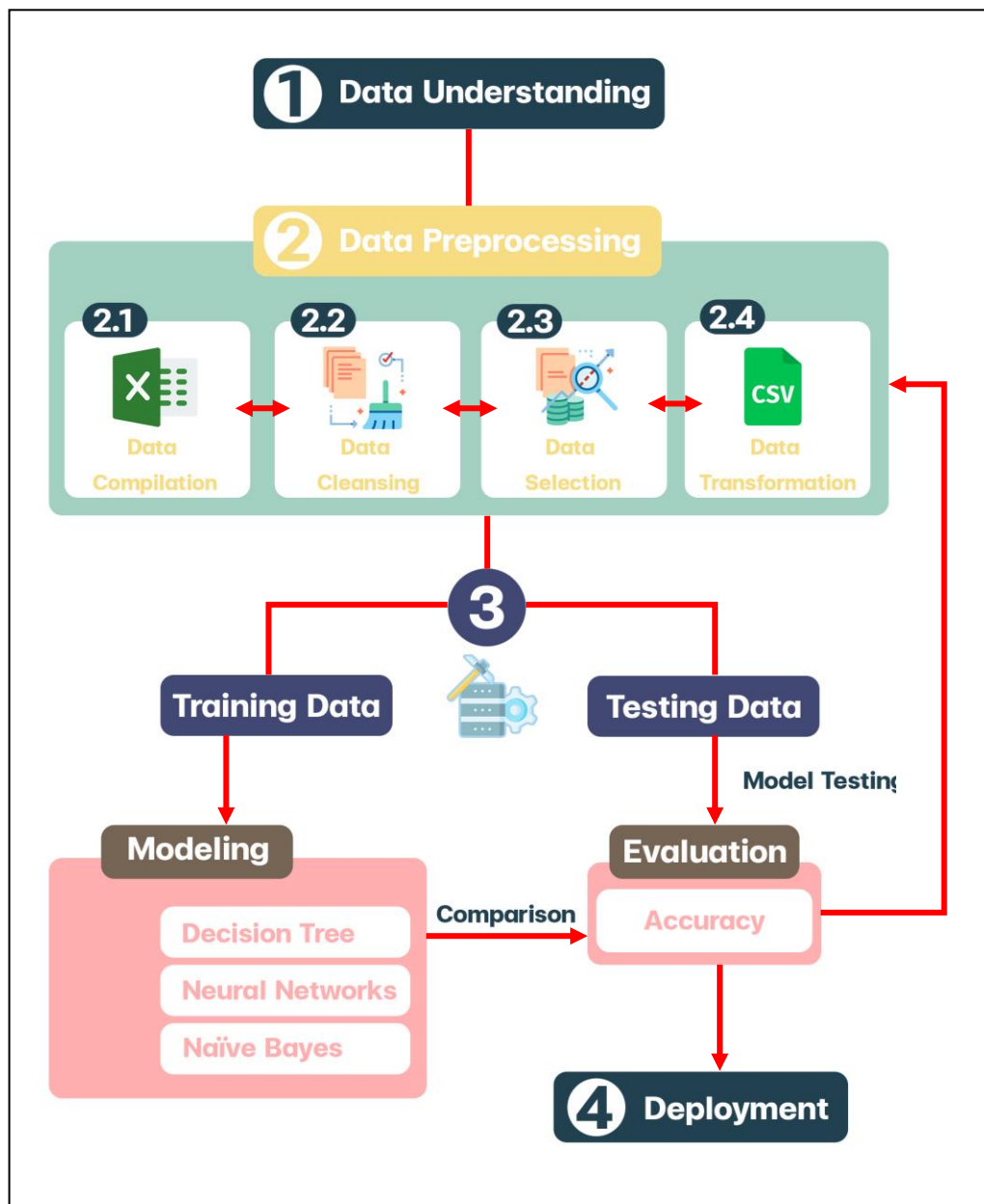


บทที่ 3

วิธีการดำเนินงาน

การวิจัยนี้ได้ดำเนินงานโดยประยุกต์ตามแนวทางในการทำเหมืองข้อมูล เป็นกระบวนการในการวิเคราะห์ข้อมูลและสร้างตัวแบบที่ได้รับความนิยมมากในปัจจุบัน เรียกว่า แนวคิดกระบวนการมาตรฐานอุตสาหกรรม หรือ CRISP-DM (Cross Reference Industry Standard for Data Mining) (Chapman et al. 2000) ผู้วิจัยได้กำหนดขั้นตอนของการดำเนินงาน ดังภาพที่ 3.1 รายละเอียดการทำงานแต่ละขั้นตอน มีดังนี้



ภาพที่ 3.1 กรอบการดำเนินงานวิจัย

จากภาพที่ 3.1 เป็นกระบวนการพัฒนาหาตัวแบบที่เหมาะสมกับการจำแนกประเภทข้อมูลสภาพทางเศรษฐกิจครัวเรือน โดยมีกระบวนการดังนี้

3.1 การทำความเข้าใจข้อมูล (Data Understanding)

ข้อมูลที่ใช้ในการศึกษาครั้งนี้ คือข้อมูลประชากรจากภาคครัวเรือนเฉพาะครัวเรือนในเขตพื้นที่ชนบท ของจังหวัดสกลนคร ซึ่งมี 20 หมู่บ้าน 12 ตำบล 12 อำเภอ โดยช่วงเวลาที่ทำการเก็บรวบรวมข้อมูล คือ ปี พ.ศ. 2563 – 2564 และจากฐานข้อมูลสภาพทางเศรษฐกิจครัวเรือน (สำนักวิทยบริการและเทคโนโลยีสารสนเทศ, 2563: ออนไลน์) โดยในฐานข้อมูลนี้เป็นข้อมูลจากโครงการศาสตร์พระราชาส่งมีการเก็บข้อมูลออกเป็น 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 ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรนา

2019 (COVID - 19)

- ส่วนที่ 7 การใช้เทคโนโลยีสารสนเทศ
- ส่วนที่ 8 การเข้าร่วมการละเล่น การพักผ่อน การรำ พิธีกรรมตามวิถีวัฒนธรรม

ชุมชน

- ส่วนที่ 9 การเข้าร่วมโครงการที่ผ่านมาย้อนหลัง 3 ปี
- ส่วนที่ 10 ข้อคิดเห็นและข้อเสนอแนะเพิ่มเติม

3.2 การเตรียมข้อมูลสำหรับพัฒนาตัวแบบ (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 | พินนา | 305 |
| 7 | สร้างค้อ | 450 |
| 8 | วัฒนา | 99 |
| 9 | ม่วง | 336 |
| 10 | หนองสนม | 189 |
| 11 | บ้านแป้น | 211 |
| 12 | อุ่มจาน | 132 |
| รวม (ครัวเรือน) | | 2,909 |

เมื่อได้จำนวนครัวเรือนแล้วจากนั้นทำการคัดเลือกแอททริบิวต์สำหรับใช้สร้างตัวแบบการพยากรณ์ ซึ่งในจำนวนครัวเรือนเหล่านี้มีข้อมูลบางแอททริบิวต์ไม่สมบูรณ์ เช่น ค่าใช้จ่ายในการทำไร่ รายได้จากการจักรสาน ราคาจำหน่ายผลผลิต รายได้จากการทอผ้า ซึ่งได้ตัดแอททริบิวต์ออกไป จะได้แอททริบิวต์ทั้งหมด 15 แอททริบิวต์ ดังต่อไปนี้

ส่วนที่ 1 ข้อมูลทั่วไปครัวเรือน มีทั้งหมด 7 แอททริบิวต์ 2,909 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (นิภารัตน์ นักตรึงศ์, 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 นักตรึงศ์, 2561) ทั้งหมด 2 แอททริบิวต์ ได้แก่ มูลค่าทรัพย์สิน และวัตถุประสงค์การเลี้ยงสัตว์

ส่วนที่ 3 อาชีพและรายได้ของครัวเรือน มีทั้งหมด 68 แอททริบิวต์ 2,915 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน (สุวรรฐ แลสันกลาง, พิบูลย์ ขยโรว์สกุล, ฐิฎิกานต์ สุริยะสาร และชุตินิษฐ์ ปานคำ, 2563) ทั้งหมด 3 แอททริบิวต์ ได้แก่ ผลผลิต/ไร่ ต้นทุน และจำนวนไร่

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

ส่วนที่ 5 หนี้สินของครัวเรือน มีทั้งหมด 3 แอททริบิวต์ 2,909 ครัวเรือน ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจ

ครัวเรือน (นิภารัตน์ นักตริพงษ์, 2561: 196; สุวรัฐ แลสันกลาง, พิบูลย์ ขยโรว์สกุล, จุฬิกานต์ สุริยะสาร, และชุตินิษฐ์ ปานคำ, 2563: 40-43) ทั้งหมด 2 แอททริบิวต์ ได้แก่ แหล่งเงินทุน และปริมาณเงินทุน

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

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

ในส่วนที่ 8 การเข้าร่วมการเล่น การฟ้อน การรำ พิธีกรรมตามวิถีวัฒนธรรมชุมชน ส่วนที่ 9 การเข้าร่วมโครงการที่ผ่านมาอย่างน้อย 3 ปี และส่วนที่ 10 ข้อคิดเห็นและข้อเสนอแนะเพิ่มเติม ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลที่คาดว่าจะมีส่วนเกี่ยวข้องกับเศรษฐกิจครัวเรือน พบว่าทั้ง 3 ส่วน ไม่มีปัจจัยไหนที่ส่งผลกระทบต่อสภาพเศรษฐกิจครัวเรือน

จากข้อมูลครัวเรือนผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลให้เหมาะสมเพื่อนำมาใช้ในการสร้างตัวแบบการพยากรณ์ข้อมูลเศรษฐกิจครัวเรือน รวมได้ทั้งหมด 15 แอททริบิวต์ ดังแสดงในตารางที่ 3.2

ตารางที่ 3.2 แสดงแอททริบิวต์ที่ส่งผลกระทบต่อสภาพเศรษฐกิจครัวเรือน

| ลำดับ | รายละเอียด |
|-----------------------------------------------------------------------------------------|----------------------------|
| ส่วนที่ 1 ข้อมูลทั่วไปครัวเรือน | |
| 1 | อายุ |
| 2 | อาชีพ |
| 3 | รายได้เฉลี่ย/เดือน |
| ส่วนที่ 2 ทรัพย์สินของครัวเรือน | |
| 4 | มูลค่าทรัพย์สิน |
| 5 | วัตถุประสงค์การเลี้ยงสัตว์ |
| ส่วนที่ 3 อาชีพและรายได้ของครัวเรือน | |
| 6 | ผลผลิต/ไร่ |
| 7 | ต้นทุน |
| 8 | จำนวนไร่ |
| ส่วนที่ 4 รายจ่ายของครัวเรือน | |
| 9 | ค่าใช้จ่าย/เดือน |
| ส่วนที่ 5 หนี้สินของครัวเรือน | |
| 10 | แหล่งเงินทุน |
| 11 | ปริมาณเงินทุน |
| ส่วนที่ 6 ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 (COVID - 19) | |
| 12 | ผลกระทบ |
| 13 | รายได้ลดลง |

ตารางที่ 3.2 แสดงแอททริบิวต์ที่ส่งผลต่อสภาพเศรษฐกิจครัวเรือน (ต่อ)

| ลำดับ | รายละเอียด |
|-----------------------------------|---------------------|
| 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 | มูลค่าทรัพย์สิน | Numeric |
| 7 | Animal Husbandry | วัตถุประสงค์การเลี้ยงสัตว์ | Nominal |
| 8 | Area | พื้นที่ก่อให้เกิดรายได้ | Numeric |
| 9 | Production Costs | ต้นทุนการผลิตการทำการเกษตร | Numeric |
| 10 | Product | ผลผลิตที่ได้จากการทำการเกษตร | 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 | ผลกระทบจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 (COVID - 19) | Nominal |
| 16 | Lower Income | รายได้ลดลงจากสถานการณ์การระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 (COVID - 19) | Nominal |
| 17 | Internet Use | การใช้อินเทอร์เน็ตที่ก่อให้เกิดรายได้ | Nominal |
| 18 | Sales Channel | ช่องทางการขายสินค้าที่ก่อให้เกิดรายได้ | Nominal |
| | Classification | การจัดหมวดหมู่ คลาสคำตอบ Low Income = รายได้น้อย Middle income = รายได้ปานกลาง High Income = รายได้สูง | Nominal |

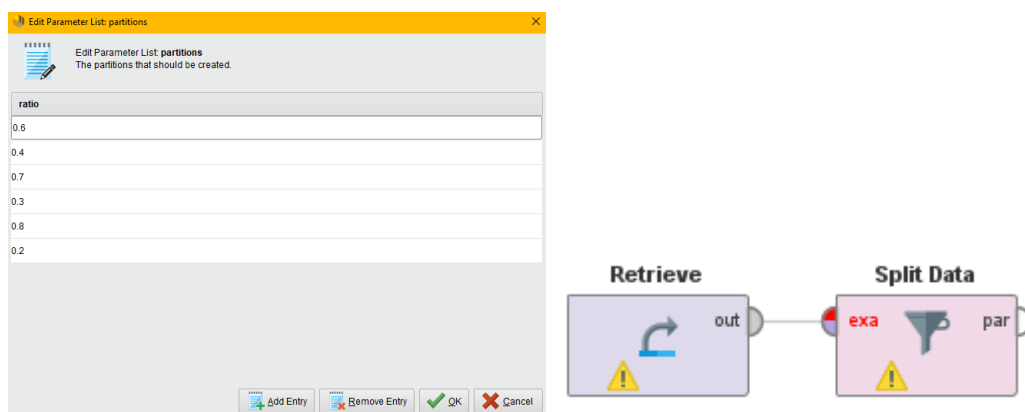
ตารางที่ 3.3 ข้อมูลเศรษฐกิจครัวเรือนที่ผ่านการทำความสะอาดและแปลงรูปแบบข้อมูล

| 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 (ระเบียบ) จึงทำการเปลี่ยนแปลงรูปแบบข้อมูลให้เหลือครัวเรือนละ 1 ระเบียบ ส่วนที่ 2 ได้มีการลดจำนวนระเบียบในครัวเรือน โดยการใช้สูตร SUM เพื่อหาผลบวกของทรัพย์สินครัวเรือนทั้งหมดให้เหลือ 1 ระเบียบ 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 หนี้ในระบบ และหนี้นอกระบบ ตัวแปรของ 2 แอททริบิวต์ คือ Yes/No และยังมีข้อมูลที่ขาดหายไปจึงพิจารณาจากค่าข้อมูลที่ปรากฏซ้ำกันมากที่สุดแล้วเติมค่าข้อมูลที่ขาดหายไป ส่วนที่ 6 ได้มีการเปลี่ยนรูปแบบตัวแปรของแอททริบิวต์ ผลกระทบ และรายได้ลดลง ส่วนที่ 7 ได้มีการเปลี่ยนรูปแบบตัวแปรของแอททริบิวต์ การใช้อินเทอร์เน็ต และช่องทางการขายสินค้า

3.3 การแบ่งชุดข้อมูลเพื่อใช้ในการสร้างตัวแบบ

ผู้วิจัยจะทำการทดสอบค่าความถูกต้องในการพยากรณ์ด้วยวิธี Cross Validation Test โดยทำการแบ่งข้อมูลออกเป็น 10 ส่วน (10-Fold Cross Validation) จากตัวแบบการพยากรณ์ที่ได้จากการใช้เทคนิคการจำแนกประเภทข้อมูล ด้วยตัวแบบต้นไม้ตัดสินใจ โครงข่ายประสาทเทียมแบบแพร่กลับ และนาอ็ฟเบย์ จะทำการทดสอบทั้ง 3 ตัวแบบให้ครบทั้ง 10 ส่วน จากนั้นจะทำการปรับปรุงวิธีการทดสอบให้มีความถูกต้องที่ดีขึ้น โดยการแบ่งชุดข้อมูลเพื่อใช้ในการสร้างตัวแบบแบ่งออกเป็น 2 ส่วน คือ 1) ข้อมูลเรียนรู้ (Training Data) 2) ข้อมูลทดสอบ (Testing Data) (Data Set = Training Set + Test Set) โดยจะรักษาสัดส่วนของข้อมูล และจะทำการสุ่มข้อมูลตามค่าสัดส่วนร้อยละ 60:40, 70:30 และ 80:20 ของข้อมูลจำนวน 1,751 ครัวเรือน ดังแสดงในภาพที่ 3.2 แล้วจะทำการทดสอบจากตัวแบบต้นไม้ตัดสินใจ โครงข่ายประสาทเทียมแบบแพร่กลับ และนาอ็ฟเบย์



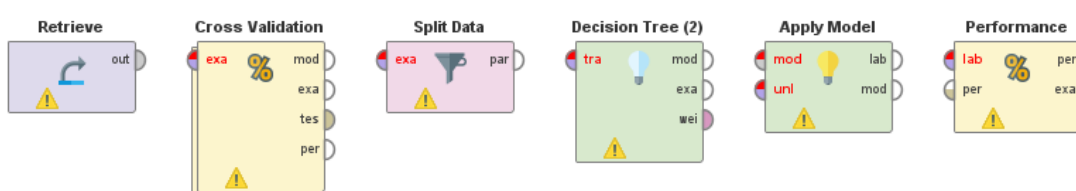
ภาพที่ 3.2 แสดงการแบ่งข้อมูลในโปรแกรม RapidMiner Studio

3.4 การสร้างตัวแบบ (Modeling)

ผู้วิจัยได้เลือกใช้โปรแกรม RapidMiner Studio เพื่อสร้างตัวแบบการพยากรณ์ที่เหมาะสมสำหรับการจำแนกสภาพเศรษฐกิจครัวเรือน โดยใช้ข้อมูลปัจจัย 18 ปัจจัย จำนวน 1,751 ครัวเรือน 18 แอททริบิวต์ และใช้เทคนิคการจำแนกประเภทข้อมูล ด้วยตัวแบบต้นไม้ตัดสินใจ โครงข่ายประสาทเทียมแบบแพร่กลับ และนาอ์ฟเบย์ หลังจากนั้นผู้วิจัยทำการวัดประสิทธิภาพของแบบจำลองทั้ง 3 แบบแล้วทำการนำแบบจำลองที่มีประสิทธิภาพที่ดีที่สุดไปใช้งาน

3.4.1 ตัวแบบต้นไม้ตัดสินใจ (Decision Tree)

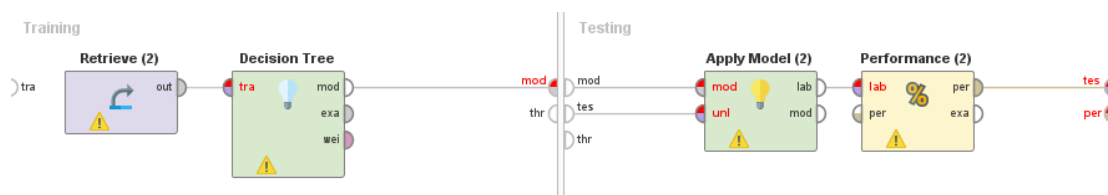
เริ่มต้นด้วยการนำข้อมูลที่ทำกรแบ่งข้อมูลออกเป็น 2 ส่วน เข้าสู่ตัวแบบจำลองดังแสดงในภาพที่ 3.3 ทำการกำหนดค่าความลึกของโนดใบ (Maximal Depth) มีค่าเท่ากับ 10 ดังแสดงในภาพที่ 3.5 ทำการพยากรณ์ความเหมาะสมของสภาพเศรษฐกิจครัวเรือน โดยมีแผนภาพต้นไม้ตัดสินใจจากการสร้างแบบจำลอง มีข้อมูลที่แบบจำลองได้พยากรณ์ออกมาด้วยแผนภาพต้นไม้ตัดสินใจ และได้ค่าประสิทธิภาพของแบบจำลองต้นไม้ตัดสินใจมีค่าความถูกต้อง



ภาพที่ 3.3 แสดงการนำข้อมูลเข้าสู่แบบแผนภาพต้นไม้ตัดสินใจ

3.4.1.1 Retrieve ข้อมูลสำหรับนำไปสร้างตัวแบบจากไฟล์ .CSV

3.4.1.2 Cross Validation ทำการทดสอบค่าความถูกต้องในการพยากรณ์ด้วยวิธี Cross Validation Test โดยทำการแบ่งข้อมูลออกเป็น 10 ส่วน (10-Fold Cross Validation) จากตัวแบบการพยากรณ์ที่ได้จากการใช้เทคนิคการจำแนกประเภทข้อมูล ด้วยตัวแบบต้นไม้ตัดสินใจ ดังแสดงในภาพที่ 3.4



ภาพที่ 3.4 แสดงโอเปอเรเตอร์ Cross Validation เพื่อสร้างตัวแบบ

3.4.1.3 Split Data การแบ่งชุดข้อมูลเพื่อใช้ในการสร้างตัวแบบ แบ่งออกเป็น 2 ส่วน คือ 1) ข้อมูลเรียนรู้ (Training Data) 2) ข้อมูลทดสอบ (Testing Data) (Data Set = Training Set + Test Set) โดยจะรักษาสัดส่วนของข้อมูล และจะทำการสุ่มข้อมูลตามค่า สัดส่วนร้อยละ 60:40, 70:30 และ 80:20 ของข้อมูลจำนวน 1,751 ครึ่งเรือน

3.4.1.4 Decision Tree ตัวแบบต้นไม้ตัดสินใจที่ใช้ในการพยากรณ์ การจำแนกประเภทข้อมูล

| Parameters | |
|------------------------------------------------------|----------|
| Decision Tree (2) (Decision Tree) | |
| criterion | accuracy |
| maximal depth | 10 |
| <input checked="" type="checkbox"/> apply pruning | |
| confidence | 0.1 |
| <input checked="" type="checkbox"/> apply prepruning | |
| minimal gain | 0.01 |
| minimal leaf size | 2 |
| minimal size for split | 4 |
| number of prepruning alternatives | 3 |

ภาพที่ 3.4 แสดงการกำหนดค่าความลึกของโหนดใบ

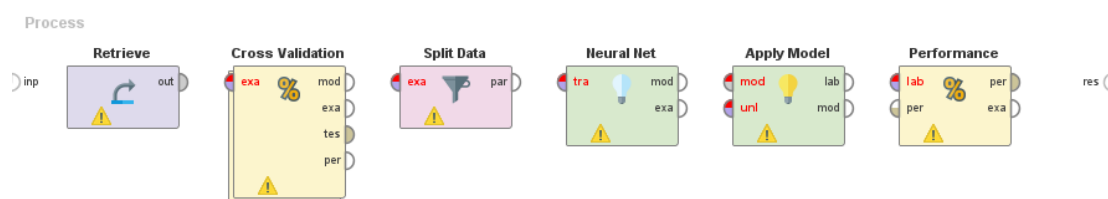
3.4.1.5 Apply Model การนำโมเดลไปใช้งาน เป็นการนำโมเดลที่สร้างได้ไปใช้ในการพยากรณ์หรือหาคำตอบให้กับข้อมูลใหม่ซึ่งยังไม่รู้คลาสคำตอบ

3.4.1.6 Performance การประเมินประสิทธิภาพ แสดงรายการค่าเกณฑ์ประสิทธิภาพ เกณฑ์ประสิทธิภาพเหล่านี้กำหนดโดยอัตโนมัติเพื่อให้เหมาะสมกับประเภทงานการเรียนรู้

3.4.2 ตัวแบบโครงข่ายประสาทเทียมแบบแพร่กลับ (Backpropagation

Neural Network: BPNN)

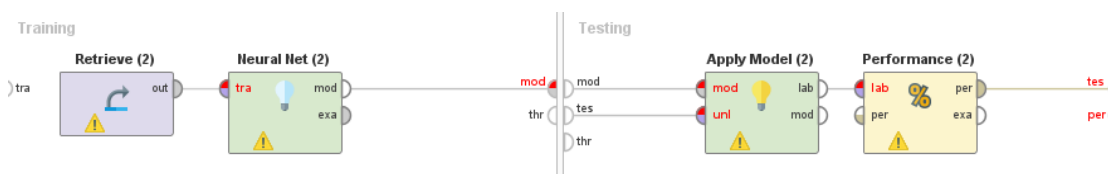
เริ่มต้นด้วยการนำข้อมูลที่ทำการแบ่งข้อมูลออกเป็น 2 ส่วน เข้าสู่ตัวแบบจำลองแสดงดังภาพที่ 3.5 ทำการปรับตั้งค่า จำนวนชั้นซ่อน (Hidden Layer Node) คือ 13, 6 ดังแสดงในภาพที่ 3.7 และทำการปรับตั้งค่า พารามิเตอร์ ของโครงข่ายประสาทเทียมแบบแพร่กลับ ให้กลับโปรแกรม ดังนี้ การแบ่งข้อมูล (Number of Folds) 10 ส่วน จำนวนรอบที่ใช้ในการฝึกสอน (Training Cycles) 1,000 อัตราการเรียนรู้ (Learning Rate) 0.1-0.5 และ ค่าสัมประสิทธิ์โมเมนตัม (Momentum) 0.1-0.5 ดังแสดงในภาพที่ 3.8



ภาพที่ 3.5 แสดงการนำข้อมูลเข้าสู่แบบจำลองโครงข่ายประสาทเทียมแบบแพร่กลับ

3.4.1.1 Retrieve ข้อมูลสำหรับนำไปสร้างตัวแบบจากไฟล์ .CSV

3.4.1.2 Cross Validation ทำการทดสอบค่าความถูกต้องในการพยากรณ์ด้วยวิธี Cross Validation Test โดยทำการแบ่งข้อมูลออกเป็น 10 ส่วน (10-Fold Cross Validation) จากตัวแบบการพยากรณ์ที่ได้จากการใช้เทคนิคการจำแนกประเภทข้อมูล ด้วยตัวแบบต้นไม่ตัดสินใจ ดังแสดงในภาพที่ 3.6



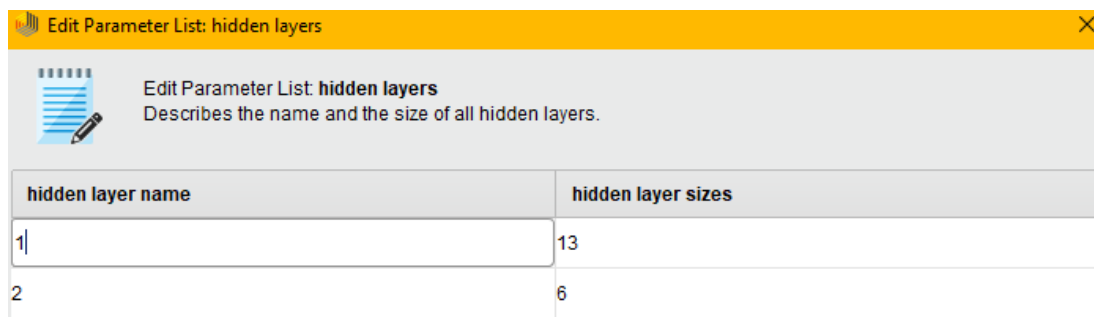
ภาพที่ 3.6 แสดงโอเปอเรเตอร์ Cross Validation เพื่อสร้างตัวแบบ

3.4.1.3 Split Data การแบ่งชุดข้อมูลเพื่อใช้ในการสร้างตัวแบบ แบ่งออกเป็น 2 ส่วน คือ 1) ข้อมูลเรียนรู้ (Training Data) 2) ข้อมูลทดสอบ (Testing Data) (Data Set = Training Set + Test Set) โดยจะรักษาสัดส่วนของข้อมูล และจะทำการสุ่มข้อมูลตามค่า สัดส่วนร้อยละ 60:40, 70:30 และ 80:20 ของข้อมูลจำนวน 1,751 ครั้วเรือน

3.4.1.4 Neural Network ตัวแบบโครงข่ายประสาทเทียมแบบแพร่กลับที่ใช้ในการพยากรณ์การจำแนกประเภทข้อมูล

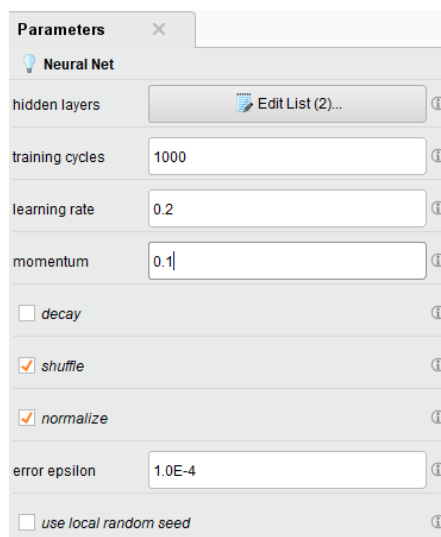
3.4.1.5 Apply Model การนำโมเดลไปใช้งาน เป็นการนำโมเดลที่สร้างได้ไปใช้ในการพยากรณ์หรือหาคำตอบให้กับข้อมูลใหม่ซึ่งยังไม่รู้คลาสคำตอบ

3.4.1.6 Performance การประเมินประสิทธิภาพ แสดงรายการค่าเกณฑ์ประสิทธิภาพ เกณฑ์ประสิทธิภาพเหล่านี้กำหนดโดยอัตโนมัติเพื่อให้เหมาะสมกับประเภทงานการเรียนรู้



| hidden layer name | hidden layer sizes |
|-------------------|--------------------|
| 1 | 13 |
| 2 | 6 |

ภาพที่ 3.7 แสดงการตั้งค่าจำนวน Layer ในแบบจำลองโครงข่ายประสาทเทียมแบบแพร่กลับ

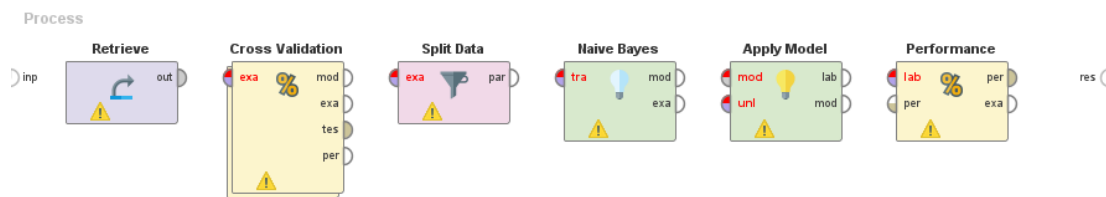


ภาพที่ 3.8 แสดงการนำข้อมูลเข้าสู่แบบจำลองโครงข่ายประสาทเทียมแบบแพร่กลับ

หลังจากทำการปรับตั้งค่าตัวแบบจำลองแล้ว จากนั้นทำการใช้งานแบบจำลองโครงข่ายประสาทเทียมแบบแพร่กลับ เพื่อพยากรณ์ความเหมาะสมของสภาพเศรษฐกิจครัวเรือน และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้องเป็นร้อยละ

3.4.3 ตัวแบบนาอ์เบย์ (Naive Bayes)

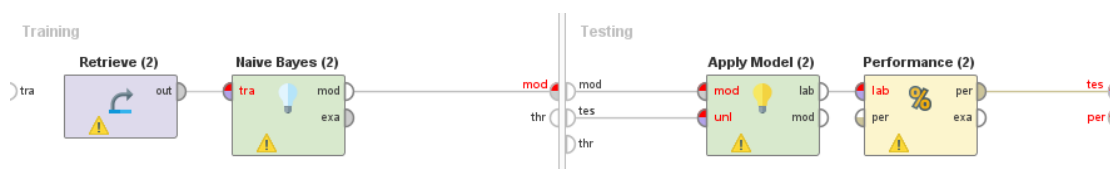
เริ่มต้นด้วยการนำข้อมูลทำการแบ่งข้อมูลออกเป็น 2 ส่วนแล้ว เข้าสู่ตัวแบบจำลองดังแสดงในภาพที่ 3.9 ทำการพยากรณ์ความเหมาะสมของสภาพเศรษฐกิจครัวเรือน และวัดประสิทธิภาพของแบบจำลองวิธีการเรียนรู้แบบอย่างง่ายเพื่อให้ได้ค่าความถูกต้องเป็นร้อยละ



ภาพที่ 3.9 แสดงการนำข้อมูลเข้าสู่แบบวิธีการเรียนรู้แบบอย่างง่าย

3.4.1.1 Retrieve ข้อมูลสำหรับนำไปสร้างตัวแบบจากไฟล์ .CSV

3.4.1.2 Cross Validation ทำการทดสอบค่าความถูกต้องในการพยากรณ์ด้วยวิธี Cross Validation Test โดยทำการแบ่งข้อมูลออกเป็น 10 ส่วน (10-Fold Cross Validation) จากตัวแบบการพยากรณ์ที่ได้จากการใช้เทคนิคการจำแนกประเภทข้อมูล ด้วยตัวแบบต้นไม้ตัดสินใจ ดังแสดงในภาพที่ 3.10



ภาพที่ 3.10 แสดงโอเปอเรเตอร์ Cross Validation เพื่อสร้างตัวแบบ

3.4.1.3 Split Data การแบ่งชุดข้อมูลเพื่อใช้ในการสร้างตัวแบบ แบ่งออกเป็น 2 ส่วน คือ 1) ข้อมูลเรียนรู้ (Training Data) 2) ข้อมูลทดสอบ (Testing Data) (Data Set = Training Set + Test Set) โดยจะรักษาสัดส่วนของข้อมูล และจะทำการสุ่มข้อมูลตามค่าสัดส่วนร้อยละ 60:40, 70:30 และ 80:20 ของข้อมูลจำนวน 1,751 ครึ่งเรือน

3.4.1.4 Naive Bayes ตัวแบบนาอิวเบย์ที่ใช้ในการพยากรณ์การจำแนกประเภทข้อมูล

3.4.1.5 Apply Model การนำโมเดลไปใช้งาน เป็นการนำโมเดลที่สร้างได้ไปใช้ในการพยากรณ์หรือหาคำตอบให้กับข้อมูลใหม่ซึ่งยังไม่รู้คลาสคำตอบ

3.4.1.6 Performance การประเมินประสิทธิภาพ แสดงรายการค่าเกณฑ์ประสิทธิภาพ เกณฑ์ประสิทธิภาพเหล่านี้กำหนดโดยอัตโนมัติเพื่อให้เหมาะสมกับประเภทงานการเรียนรู้

จากการทดลองในข้างต้นได้ทำการทดลองแบบจำลองทั้งหมด 3 แบบด้วยกันคือ ต้นไม้ตัดสินใจ โครงข่ายประสาทเทียมแบบแพร่กลับ และนาอิวเบย์ พบว่าแบบจำลองต่าง ๆ ให้ค่าความถูกต้องหรือให้ประสิทธิภาพดังนี้ ต้นไม้ตัดสินใจ ให้ค่าความถูกต้องร้อยละ..... โครงข่ายประสาทเทียมแบบแพร่กลับ ให้ค่าความถูกต้องร้อยละ..... และนาอิวเบย์ ให้ค่าความถูกต้องร้อยละ..... เห็นได้ว่าตัวแบบ..... ให้ค่าความถูกต้องที่ดีที่สุดโดยให้ค่าความถูกต้องอยู่ที่ร้อยละ.....

3.5 การประเมินตัวแบบ (Evaluation)

ในงานวิจัยนี้ใช้การทดสอบแบบไขว้ทบแบบ 10 ส่วน (10-Fold Cross Validation) แล้วทำการทดสอบเพื่อประเมินประสิทธิภาพ การวัดค่าประสิทธิภาพของเทคนิควิธีต่าง ๆ จะต้องทำการเลือกข้อมูลสำหรับเรียนรู้ (Training Data) และข้อมูลสำหรับทดสอบ (Testing Data) และเลือกใช้วิธีแบบสุ่มเลือกแบ่งข้อมูลโดยจะรักษาสัดส่วนของข้อมูล และจะทำการสุ่มข้อมูลตามค่าสัดส่วนร้อยละ 60:40, 70:30 และ 80:20 ต่อจากนั้นให้นำข้อมูลบางส่วนมาทำการเรียนรู้ และนำข้อมูลบางส่วนมาทำการทดสอบแบบจำลองที่ได้จากการเรียนรู้ โดยในการทำงานจะทำการเลือกสุ่มข้อมูลออกเป็น k ชุด ในการทดลองครั้งแรก ข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดเรียนรู้ ในการทดลองครั้งที่ 2 ข้อมูลชุดที่ 2 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดเรียนรู้ ทำจนกระทั่งข้อมูลทุกชุดได้ถูกนำมาเป็นข้อมูลชุดทดสอบและข้อมูลชุดเรียนรู้ ซึ่งจะมีการทดลองทั้งหมด k ครั้ง ในงานวิจัยนี้ได้เลือกใช้ค่า $k = 10$ ดังแสดงในภาพที่ 11

| | 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 11 10-Fold Cross Validation

การคำนวณประสิทธิภาพของตัวแบบจำลอง สามารถคำนวณได้จากตาราง Confusion Matrix ซึ่งเป็นตารางสรุปจำนวนข้อมูลที่ตัวแบบมีการจำแนกได้อย่างถูกต้องและไม่ถูกต้อง

ตารางที่ 3.4 The Confusion Matrix

| ค่าที่แท้จริง (Actual Class) | ค่าที่ทำนายได้ (Predicted Class) | |
|---------------------------------|----------------------------------|--------------------|
| | Class YES | Class NO |
| Class YES | True Positive: TP | False Negative: FN |
| Class NO | False Positive: FP | True Negative: TN |

แล้วทำการประเมินประสิทธิภาพของการพยากรณ์โดยใช้เกณฑ์การวัดประสิทธิภาพของตัวแบบเรียนรู้ด้วยวิธี Predictive Model ซึ่งประกอบด้วยค่าความถูกต้อง (Accuracy) ค่าความ

แม่นยำ (Precision: P) ค่าความระลึก (Recall: R) และค่าถ่วงดุล (F-Measure) ซึ่งมีค่าอยู่ระหว่าง 0 – 1 หมายถึงประสิทธิภาพดี ดังแสดงในสมการที่ (1) (2) (3) และ (4) ตามลำดับ

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F - measure = 2 \frac{PR}{P+R} \quad (4)$$

โดยที่

TP คือ ค่าที่พยากรณ์ถูกต้อง (ข้อมูลบอกว่าจริง พยากรณ์ว่าจริง)

TN คือ ค่าที่พยากรณ์ถูกต้อง (ข้อมูลบอกว่าไม่จริง พยากรณ์ว่าไม่จริง)

FP คือ ค่าที่พยากรณ์ไม่ถูกต้อง (ข้อมูลบอกว่าจริง พยากรณ์ว่าไม่จริง)

FN คือ ค่าที่พยากรณ์ไม่ถูกต้อง (ข้อมูลบอกว่าไม่จริง พยากรณ์ว่าจริง)

ในงานวิจัยนี้ได้เลือกใช้วิธีประเมินประสิทธิภาพของการพยากรณ์ด้วยค่าความถูกต้อง โดยใช้ข้อมูลสำหรับเรียนรู้ทดลองปรับค่าพารามิเตอร์ที่เหมาะสมของทั้งตัวแบบต้นไม้ตัดสินใจ โครงข่ายประสาทเทียมแบบแพร่กลับ และนาอ็ฟเบย์ ได้นำมาทดลองกับข้อมูลชุดทดสอบเปรียบเทียบกับประสิทธิภาพการทำงานของข้อมูลทั้งสองชุดเพื่อป้องกันการเกิด Over-Fitting นอกจากนี้ยังได้ทำการทดสอบผลการทดลองกับข้อมูลอีกชุดหนึ่ง คือ ชุดตรวจสอบ เพื่อเพิ่มความเชื่อมั่นของตัวจำแนกประเภท ดังแสดงในตารางที่ 3.5

ตารางที่ 3.4 ค่าความถูกต้องของข้อมูลชุดทดสอบ และชุดตรวจสอบ

| ข้อมูล | Decision Tree | BPNN | Naive Bayes |
|------------|---------------|-----------|-------------|
| ชุดทดสอบ | ค่าร้อยละ | ค่าร้อยละ | ค่าร้อยละ |
| ชุดตรวจสอบ | ค่าร้อยละ | ค่าร้อยละ | ค่าร้อยละ |

3.6 การนำไปใช้งาน (Deployment)

หลังจากทำการประเมินผลตัวจำแนกของข้อมูลชุดเรียนรู้ข้อมูลชุดทดสอบ และข้อมูลชุดตรวจสอบเรียบร้อยแล้วได้ผล สามารถนำตัวแบบที่ได้สร้างขึ้นมาใช้ประโยชน์จริงในการจำแนกประเภทข้อมูลสภาพเศรษฐกิจครัวเรือน สำหรับสนับสนุนหรือเป็นข้อมูลประกอบการตัดสินใจในการวิจัยในลำดับต่อไป

