การเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกกับข้อมูลด้านการเงิน

ปรายฟ้า โกษศิริศิลป์

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ
ปีการศึกษา 2566

การเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกกับข้อมูลด้านการเงิน

ปรายฟ้า โกษศิริศิลป์

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ
ปีการศึกษา 2566

วิทยานิพนธ์ เรื่อง "การเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกกับข้อมูลด้านการเงิน"
ของ ปรายฟ้า โกษศิริศิลป์
ได้รับการพิจารณาให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติ

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการสอบวิทยานิพนธ์
(รองศาสตราจารย์ ดร.อัชฌา อระวีพร)

..... ประธานที่ปรึกษาวิทยานิพนธ์
(รองศาสตราจารย์ ดร.อนามัย นาอุดม)

..... กรรมการที่ปรึกษาวิทยานิพนธ์
(ผู้ช่วยศาสตราจารย์ ดร.จรัสศรี รุ่งรัตนาอุบล)

..... กรรมการผู้ทรงคุณวุฒิภายใน
(ผู้ช่วยศาสตราจารย์ ดร.กัลยา บุญหล้า)

อนุมัติ

.....
(รองศาสตราจารย์ ดร.กรองกาญจน์ ชูทิพย์)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกกับข้อมูลด้านการเงิน
ผู้วิจัย ปรายฟ้า โกษศิริศิลป์
ประธานที่ปรึกษา รองศาสตราจารย์ ดร.อนามัย นาอุดม
กรรมการที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.จรัสศรี รุ่งรัตนาอุบล
ประเภทสารนิพนธ์ วิทยานิพนธ์ วท.ม. สถิติ, มหาวิทยาลัยนเรศวร, 2566
คำสำคัญ การถดถอยลอจิสติกทวิภาค ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย เทคนิคนาอีฟเบย์ ข้อมูลด้านการเงิน ข้อมูลไม่สมดุล

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษากระบวนการทำงานและเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนก 3 เทคนิค ได้แก่ การถดถอยลอจิสติกทวิภาค เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย และเทคนิคนาอีฟเบย์ โดยใช้ชุดข้อมูลด้านการเงินที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและจำนวนของตัวแปรอิสระเชิงปริมาณแตกต่างกัน 3 ชุดข้อมูล ได้แก่ ชุดข้อมูลเครดิตเยอรมันที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ ชุดข้อมูลลูกค้าบัตรเครดิตที่มีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ และชุดข้อมูลการตลาดของธนาคารที่มีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ โดยศึกษาภายใต้ชุดข้อมูลตั้งต้นและชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม เทคนิคการสุ่มลด และเทคนิคการสุ่มผสมผสาน จากนั้นทำการทดสอบประสิทธิภาพด้วยหลักการ 5–Fold Cross-Validation โดยมีการวัดประสิทธิภาพตัวแบบการจำแนกจากค่าความแม่นยำ ค่าเรียกคืน ค่าความเที่ยง และค่าประสิทธิภาพโดยรวม ผลการวิจัยพบว่าการถดถอยลอจิสติกทวิภาคมีประสิทธิภาพดีที่สุดบนชุดข้อมูลเครดิตเยอรมันที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ และชุดข้อมูลการตลาดของธนาคารที่มีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ โดยมีค่าความแม่นยำเท่ากับ 76.00% และ 83.93% ตามลำดับ ในขณะที่เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพดีที่สุดบนชุดข้อมูลลูกค้าบัตรเครดิตที่มีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ โดยมีค่าความแม่นยำเท่ากับ 81.99%

Title COMPARISON OF CLASSIFICATION MODELS PERFORMANCE WITH FINANCIAL DATA

Author Praifa Kosasirisin

Advisor Associate Professor Anamai Na-udom, Ph.D.

Co-Advisor Assistant Professor Jaratsri Rungrattanaubol, Ph.D.

Academic Paper M.S. Thesis in Statistics - (Type A 2), Naresuan University, 2023

Keywords Binary Logistic Regression, Classification and Regression Tree, Naïve Bayes, Financial data, Imbalance dataset

ABSTRACT

This research aims to study the construction and compare the performance of three classification models, namely binary logistic regression, classification and regression tree, and Naïve Bayes techniques using three financial datasets with different numbers of qualitative and quantitative independent variables. The characteristics of three datasets are classified as the German credit dataset with higher number of qualitative than quantitative independent variables, Default of credit card client dataset with fewer qualitative than quantitative independent variables, and Bank marketing dataset with equal number of qualitative and quantitative independent variables, respectively. The study was employed under the original data set and the data set where the imbalance was adjusted using over sampling, under sampling, and hybrid methods. The performance of each classification technique was validated using the 5–Fold Cross-Validation technique and the efficiency comparison was performed by considering accuracy, recall, precision, and overall accuracy criteria. The results showed that Binary logistic regression performs best on the German credit dataset with a higher number of qualitative than quantitative independent variables and on a Bank marketing dataset with an equal number of qualitative and quantitative independent variables, with the accuracy of 76.00% and 83.93%, respectively. It was also found that the classification and regression tree technique performed best on Default of credit

card clients dataset with fewer qualitative than quantitative independent variables with an accuracy of 81.99%.

ประกาศคุณูปการ

ผู้วิจัยขอขอบพระคุณประธานที่ปรึกษาวิทยานิพนธ์ รองศาสตราจารย์ ดร.อนามัย นาอุดม เป็นอย่างสูงที่สละเวลาให้คำปรึกษาและให้คำแนะนำตลอดระยะเวลาการทำวิทยานิพนธ์ ขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.จรัสศรี รุ่งรัตนาอุบล กรรมการที่ปรึกษาวิทยานิพนธ์ ที่ให้คำปรึกษาเกี่ยวกับการจัดการข้อมูลด้วยวิธีการทำเหมืองข้อมูล ขอขอบคุณ รองศาสตราจารย์ ดร.อัชฌา อระวีพร ประธานกรรมการสอบวิทยานิพนธ์และกรรมการผู้ทรงคุณวุฒิภายนอก ที่ให้คำแนะนำแก้ไขส่วนที่บกพร่องของงานวิทยานิพนธ์เล่มนี้

ขอกราบขอบพระคุณคุณพ่อนพพล โกษศิริศิลป์ คุณแม่ประทุม มิ่งน้อย และขอบคุณเพื่อนร่วมรุ่นปริญญาโท และรุ่นพี่ปริญญาเอกทุกท่านที่ให้การสนับสนุนงานวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

ผู้วิจัยหวังเป็นอย่างยิ่งว่า งานวิจัยเรื่องการเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกกับข้อมูลด้านการเงิน จะเป็นประโยชน์ต่อประชาชนและบุคคลผู้สนใจที่จะนำงานวิจัยนี้ไปต่อยอดเพื่อพัฒนาองค์ความรู้ใหม่ ๆ ต่อไป

ปรายฟ้า โกษศิริศิลป์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

ในยุคปัจจุบันเป็นที่ทราบกันดีว่าเทคโนโลยีเข้ามามีบทบาทอย่างมากในทุกกิจกรรมของชีวิตประจำวัน เนื่องจากการที่เทคโนโลยีมีการพัฒนาและก้าวหน้าอย่างรวดเร็ว ทำให้หน่วยงานทั้งภาครัฐและเอกชนได้พยายามปรับเปลี่ยนแผนการดำเนินงานภายใต้สิ่งแวดล้อมที่มีความเคลื่อนไหวอย่างรวดเร็วนี้ ด้วยการนำเทคโนโลยีเข้ามาช่วยสนับสนุนการปฏิบัติงานด้านต่าง ๆ เช่น ด้านการแพทย์ ด้านการศึกษา ด้านวิทยาศาสตร์ ด้านการเกษตร ด้านการตลาด และในด้านอื่น ๆ อีกมากมาย โดยเฉพาะด้านการเงินที่ปัจจุบันต้องเผชิญกับความเปลี่ยนแปลงในหลายรูปแบบ เช่น ความต้องการของผู้บริโภคที่เปลี่ยนแปลงไปทุกวัน การเกิดขึ้นของเทคโนโลยีทางการเงินที่ก่อให้เกิดบริการทางการเงินใหม่ ๆ จากกลุ่มธุรกิจทั้งที่เป็นธนาคารและไม่เป็นธนาคาร หรือการทำธุรกรรมการเงินบนออนไลน์มากขึ้น ดังนั้นองค์กรทางการเงินจึงจำเป็นต้องมีการปรับตัวโดยการใช้ประโยชน์จากเทคโนโลยีเพื่อส่งเสริมศักยภาพในการให้บริการขององค์กรให้สามารถตอบสนองกลุ่มลูกค้าได้อย่างมีประสิทธิภาพ (Tiplawan, 2021) (สาครรัตน์ นักปราชญ์ และคัคนางค์ จามะริก, 2559)

การคิดค้นวิธีการและช่องทางในการทำธุรกรรมทางการเงินใหม่ ๆ ขององค์กรทางการเงินสามารถช่วยลดต้นทุนค่าธรรมเนียมและประหยัดเวลาในการทำธุรกรรมได้อย่างมาก ยกตัวอย่างเช่น ลูกค้าสามารถทำธุรกรรมทางการเงินและการลงทุนต่าง ๆ ไม่ว่าจะเป็นการโอนเงิน การถอนเงิน การซื้อหน่วยลงทุน การจ่ายบิล หรือการซื้อขายหลักทรัพย์ได้อย่างสะดวกและรวดเร็ว เมื่อมีการทำธุรกรรมทางการเงินมากขึ้น ข้อมูลทางการเงินจึงถูกสร้างในทุก ๆ วัน เมื่อข้อมูลเหล่านี้ถูกสะสมจนมีจำนวนมากและมีความหลากหลาย จึงเข้าข่ายของการเป็นข้อมูลขนาดใหญ่ (Big Data) ที่หมายถึงข้อมูลที่มีปริมาณมากมายมหาศาล มีขนาดใหญ่ มีรูปแบบของข้อมูลที่หลากหลายและซับซ้อน และยังคงเป็นเพียงข้อมูลดิบที่รอการนำมาวิเคราะห์เพื่อนำผลที่ได้มาสร้างมูลค่าทางธุรกิจ ข้อมูลเหล่านี้อาจไม่ได้อยู่ในรูปแบบที่องค์กรสามารถนำไปใช้ประโยชน์ได้ในทันที แต่อาจมีข้อมูลที่เป็นประโยชน์ต่อองค์กรบางอย่างแฝงอยู่ ซึ่งอาจต้องใช้กระบวนการการจัดการข้อมูลที่เหมาะสมเข้ามาช่วย ซึ่งการเป็นข้อมูลขนาดใหญ่นั้น ได้ถูกนิยามว่าจะมีคุณลักษณะสำคัญที่ต้องคำนึงถึง 5 ประการ (Ishwarappa & Anuradha, 2015) ได้แก่ ปริมาณ (Volume) ความหลากหลาย (Variety) ความเร็ว (Velocity) มูลค่า (Value) และความถูกต้อง (Veracity) คุณลักษณะเหล่านี้จัดเป็นอุปสรรคของการจัดการข้อมูลขนาดใหญ่ที่ทำให้การประมวลผลเป็นไปได้ยาก ไม่สามารถใช้ประโยชน์ได้ทันทีและต้องอาศัย

เทคโนโลยีที่ทันสมัย กระบวนการหนึ่งที่ได้รับความนิยมในการค้นหาสารสนเทศที่จะนำมาใช้งานข้อมูลขนาดใหญ่ก็คือ การทำเหมืองข้อมูล (Data Mining) การทำเหมืองข้อมูลเป็นการค้นหาข้อมูลที่มีประโยชน์จากแหล่งข้อมูลที่มีจำนวนมากมายมหาศาล เพื่อดึงข้อมูลที่มีประโยชน์มาทำการวิเคราะห์ค้นหารูปแบบหรือความสัมพันธ์ที่เกิดขึ้นในข้อมูล และจัดทำเป็นสารสนเทศเพื่อใช้ในการวางแผนบริหารจัดการหรือตัดสินใจเกี่ยวกับธุรกิจได้ (Usama et al., 1996) การทำเหมืองข้อมูลเป็นการบูรณาการศาสตร์จากงาน 3 ส่วน ได้แก่ สถิติ (Statistics) การเรียนรู้ของเครื่อง (Machine Learning) และฐานข้อมูล ซึ่งการเรียนรู้ของเครื่องสามารถเป็นที่นิยมอย่างมาก ในการเข้ามาช่วยจัดการกับข้อมูลขนาดใหญ่ ซึ่งจะแบ่งออกเป็น 3 ประเภท ได้แก่ การเรียนรู้แบบมีผู้สอน (Supervised Learning) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และการเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) โดยในการเรียนรู้ของเครื่องนั้น การเรียนรู้แบบมีผู้สอนมีการนำมาใช้งานกันอย่างแพร่หลาย เพราะเป็นการวิเคราะห์ข้อมูลที่มีตัวแปรตามกำกับไว้ ถ้าตัวแปรตามเป็นเชิงคุณภาพจะเกี่ยวข้องกับการวิเคราะห์การจำแนก ถ้าตัวแปรตามเป็นเชิงปริมาณจะเกี่ยวข้องกับการวิเคราะห์การถดถอย (Iqbal H, 2021) ซึ่งการวิเคราะห์การจำแนก (Classification) ถือเป็นการทำเหมืองข้อมูลประเภทหนึ่ง ที่ได้รับความนิยมในการนำไปแก้ปัญหาในหลากหลายสาขา และจากกระบวนการหารูปแบบ หรือหาความสัมพันธ์ของข้อมูล (Hanet al., 2011) ไม่ว่าสาขาใดก็ต้องการการวิเคราะห์ข้อมูลที่แม่นยำ และมีประสิทธิภาพที่สามารถนำไปใช้งานได้ในชีวิตจริง

ปัจจุบันองค์กรทางการเงินได้มีการนำกลไกการเรียนรู้แบบมีผู้สอนที่เกี่ยวกับการวิเคราะห์ตัวแบบการจำแนกมาประยุกต์ใช้ในองค์กรกันอย่างแพร่หลาย เช่น มีการนำข้อมูลพฤติกรรมการใช้งานโมบายแบงก์กิ้งของลูกค้ามาวิเคราะห์และคาดการณ์ผลิตภัณฑ์หรือโปรโมชันที่ลูกค้าน่าจะสนใจเพื่อนำเสนอให้กับลูกค้าแต่ละคน ประยุกต์ใช้กับการบริหารความเสี่ยงทางการเงิน ทำให้สามารถตรวจสอบข้อมูลที่เกิดขึ้นได้ตลอดเวลา โดยการวิเคราะห์ตัวแบบการจำแนกจะอาศัยชุดข้อมูลเรียนรู้เพื่อทำการพัฒนาตัวแบบการจำแนกและใช้ชุดข้อมูลทดสอบในการทดสอบตัวแบบการจำแนก โดยผลลัพธ์ของการเรียนรู้ของเครื่องมือสำหรับการพัฒนาตัวแบบการจำแนก คือการคาดการณ์ผลลัพธ์ที่เกิดจากข้อมูลที่ได้รับและนำผลลัพธ์ที่ได้ไปตรวจสอบกับชุดข้อมูลที่ใช้ในการทดสอบ ว่าตัวแบบการจำแนกที่ถูกพัฒนาขึ้นนั้น มีประสิทธิภาพที่ดีมากน้อยเพียงใด

การวิเคราะห์ตัวแบบการจำแนกจึงถูกนำมาประยุกต์ใช้ในการช่วยลดความเสี่ยงในกระบวนการทำงานและยังช่วยเพิ่มประสิทธิภาพในการทำงาน หรือนำไปประยุกต์ใช้ในตรวจจับการฉ้อโกง สามารถตรวจดูกิจกรรมหรือพฤติกรรมของบัญชีผู้ใช้งานทุกบัญชีได้อย่างรวดเร็ว และเมื่อพบ

สิ่งที่ผิดปกติไปจากรูปแบบการใช้งานของเจ้าของบัญชี ระบบจะทำการตรวจสอบและติดต่อไปยังเจ้าของบัญชีโดยทันที และในส่วนของงานวิจัยทางด้านการเงิน ได้มีผู้วิจัยหลายท่านศึกษาเกี่ยวกับการวิเคราะห์ตัวแบบการจำแนกบนชุดข้อมูลด้านการเงิน เช่น ข้อมูลบัญชีคู่ที่ตรงกันขององค์กรขนาดเล็กจากสหราชอาณาจักร (Irimia et al., 2015) ข้อมูลการประเมินความเสี่ยงจากการผิดนัดชำระหนี้ของสินเชื่อและการประเมินความเสี่ยงด้านเครดิต (Aida, 2017) ข้อมูลความเสี่ยงจากการผิดนัดชำระ (Begüm & Deniz, 2019) หรือข้อมูลธุรกรรมทางการเงินที่เป็นการฉ้อโกง (Fayaz et al., 2020) เป็นต้น จากงานวิจัยที่ใช้ข้อมูลบัญชีคู่ที่ตรงกันขององค์กรขนาดเล็กจากสหราชอาณาจักร พบว่า เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพดีที่สุด (Irimia et al., 2015) จากการใช้ตัวแบบการจำแนกมาประเมินข้อมูลความเสี่ยงจากการผิดนัดชำระหนี้ของสินเชื่อและการประเมินความเสี่ยงด้านเครดิต พบว่า เทคนิคนาอีฟเบย์ให้ประสิทธิภาพดีที่สุด (Aida, 2017) การนำตัวแบบการจำแนกที่ได้จากการวิเคราะห์ข้อมูลและนำตัวแบบการจำแนกที่ได้ไปใช้ในการทำนายข้อมูลความเสี่ยงจากการผิดนัดชำระ พบว่าการถดถอยลอจิสติกมีประสิทธิภาพดีที่สุดเช่นเดียวกัน (Begüm & Deniz, 2019) จากการนำตัวแบบการจำแนกที่ได้มาใช้กับข้อมูลธุรกรรมทางการเงินที่เป็นการฉ้อโกง พบว่า การถดถอยลอจิสติกมีประสิทธิภาพดีที่สุด (Fayaz et al., 2020) และจากงานวิจัยข้างต้นการนำชุดข้อมูลมาใช้ในงานวิจัยส่วนใหญ่จะไม่ได้คำนึงถึงจำนวนของตัวแปรอิสระเชิงคุณภาพและตัวแปรอิสระเชิงปริมาณ ทำให้ผู้วิจัยสนใจที่จะศึกษาประสิทธิภาพของตัวแบบการจำแนกกับชุดข้อมูลด้านการเงินที่มีลักษณะแตกต่างกัน ซึ่งในที่นี้หมายถึงชุดข้อมูลจะมีจำนวนของตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกัน และเนื่องจากชุดข้อมูลที่นำมาใช้ในการวิเคราะห์ข้อมูลเป็นชุดข้อมูลมีความไม่สมดุล (Imbalance data) ที่หมายถึง ข้อมูลที่มีจำนวนสมาชิกในกลุ่มหลักและกลุ่มรองไม่เท่ากันหรือไม่ใกล้เคียงกัน (เบญจภรณ์ และคณะ, 2557) ซึ่งอาจส่งผลต่อการจำแนกข้อมูล โดยจะมีการจำแนกข้อมูลของกลุ่มที่มีข้อมูลจำนวนมากได้อย่างแม่นยำ แต่ความถูกต้องและแม่นยำในการจำแนกประเภทข้อมูลของกลุ่มที่มีจำนวนข้อมูลน้อยจะลดลง (Farquad & Bose, 2012) มีนักวิจัยหลายท่านที่มีการนำชุดข้อมูลที่ไม่สมดุลมาทำให้เป็นชุดข้อมูลที่มีความสมดุล โดยใช้เทคนิคที่มีความหลากหลายเข้ามาช่วย เช่น การสุ่มตัวอย่างแบบง่าย การแบ่งกลุ่มข้อมูลแบบเคมีน การสุ่มเพิ่ม การสุ่มลด และการสุ่มผสมผสาน เป็นต้น

ดังนั้นงานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษากระบวนการทำงานและเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกกับชุดข้อมูลด้านการเงินที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและ จำนวนของตัวแปรอิสระเชิงปริมาณแตกต่างกัน 3 แบบ ได้แก่ ชุดข้อมูลเครดิตเยอรมัน ซึ่งมีจำนวนตัวแปร

อิสระเชิงคุณภาพมากกว่าเชิงปริมาณ ชุดข้อมูลลูกค้าบัตรเครดิต ซึ่งมีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ และชุดข้อมูลการตลาดของธนาคาร ซึ่งมีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณภายใต้ชุดข้อมูลตั้งต้นและชุดข้อมูลที่มีการปรับปรุงความไม่สมดุล โดยจะพัฒนาตัวแบบการจำแนกด้วย 3 เทคนิค ได้แก่ การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression) เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Classification and Regression Tree) และเทคนิคนาอีฟเบย์ (Naïve Bayes) โดยสร้างตัวแบบการจำแนกที่มีการแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบด้วย 5-Fold Cross-Validation และในส่วนของชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลให้สมดุลนั้น มีการใช้เทคนิคการสุ่ม 3 เทคนิค คือ การสุ่มเพิ่ม (Over sampling) การสุ่มลด (Under sampling) และการสุ่มผสมผสาน (Hybrid methods) โดยมีเกณฑ์ในการวัดประสิทธิภาพตัวแบบการจำแนก คือ ค่าความแม่นยำ (Accuracy) ค่าเรียกคืน (Recall) ค่าความเที่ยง (Precision) และค่าประสิทธิภาพโดยรวม (F1-Score)

1.2 จุดมุ่งหมายของการวิจัย

1.2.1 เพื่อศึกษากระบวนการทำงานของเทคนิคการจำแนกของการถดถอยลอจิสติกทวิภาค ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย และเทคนิคนาอีฟเบย์

1.2.2 เพื่อเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงินที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและจำนวนของตัวแปรอิสระเชิงปริมาณแตกต่างกัน

1.2.3 เพื่อเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงินภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุล

1.3 ขอบเขตการวิจัย

1.3.1 ขอบเขตด้านตัวแบบการจำแนก

งานวิจัยนี้มีจุดมุ่งหมายเพื่อศึกษากระบวนการทำงานและเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงิน โดยใช้เทคนิคการจำแนก 3 เทคนิค ได้แก่

1. การถดถอยลอจิสติกทวิภาค
2. เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย
3. เทคนิคนาอีฟเบย์

1.3.2 ขอบเขตด้านชุดข้อมูล

งานวิจัยใช้ชุดข้อมูลด้านการเงินที่มีจำนวนแปรอิสระตัวเชิงคุณภาพและเชิงปริมาณที่แตกต่างกันทั้งหมด 3 แบบ และมีตัวแปรตามเชิงคุณภาพ 1 ตัว แบ่งออกเป็น 2 กลุ่ม โดยคัดเลือกชุดข้อมูลมาจาก UCI Machine Learning Repository โดยแต่ละชุดข้อมูลมีรายละเอียดดังต่อไปนี้

1. ชุดข้อมูลเครดิตเยอรมัน (German credit dataset) จำนวน 1,000 แถว ซึ่งมีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ โดยมีจำนวนตัวแปรอิสระ 20 ตัวแปร ประกอบไปด้วยตัวแปรอิสระเชิงคุณภาพ 13 ตัวแปร และตัวแปรอิสระเชิงปริมาณ 7 ตัวแปร
2. ชุดข้อมูลลูกค้าบัตรเครดิต (Default of credit card clients dataset) จำนวน 30,000 แถว ซึ่งมีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ โดยมีจำนวนตัวแปรอิสระ 23 ตัวแปร ประกอบไปด้วยตัวแปรอิสระเชิงคุณภาพ 9 ตัวแปร และตัวแปรอิสระเชิงปริมาณ 14 ตัวแปร
3. ชุดข้อมูลการตลาดของธนาคาร (Bank marketing dataset) จำนวน 41,188 แถว ซึ่งมีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ โดยมีจำนวนตัวแปรอิสระ 20 ตัวแปร ประกอบไปด้วยตัวแปรอิสระเชิงคุณภาพ 10 ตัวแปร และตัวแปรอิสระเชิงปริมาณ 10 ตัวแปร

1.3.3 ขอบเขตด้านการปรับปรุงชุดข้อมูลที่มีความไม่สมดุล

งานวิจัยนี้มีการปรับปรุงชุดข้อมูลไม่สมดุลให้สมดุลโดยใช้เทคนิคการสุ่ม 3 เทคนิค ได้แก่ เทคนิคการสุ่มเพิ่ม (Over sampling) เทคนิคการสุ่มลด (Under sampling) และเทคนิคการสุ่มผสมผสาน (Hybrid methods)

1.3.4 ขอบเขตด้านการพัฒนาตัวแบบการจำแนก

งานวิจัยนี้ทำการแบ่งข้อมูลเป็นชุดข้อมูลเรียนรู้ (Training set) และชุดข้อมูลทดสอบ (Test set) โดยใช้การแบ่งข้อมูลด้วย 5-Fold Cross-Validation ซึ่งจะแบ่งข้อมูลออกเป็น 5 ชุด โดย 1 ชุดข้อมูลจะเป็นชุดข้อมูลทดสอบและ 4 ชุดที่เหลือเป็นชุดข้อมูลเรียนรู้ จากนั้นทำการสร้างตัวแบบการจำแนกและทดสอบทั้งหมด 5 รอบ

1.3.5 ขอบเขตเกณฑ์ที่ใช้วัดประสิทธิภาพตัวแบบการจำแนก

เกณฑ์ที่ใช้วัดประสิทธิภาพตัวแบบการจำแนกบนชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบมีทั้งหมด 4 เกณฑ์ ซึ่งประกอบไปด้วย

1. ค่าความแม่นยำ (Accuracy)
2. ค่าเรียกคืน (Recall)

3. ค่าความเที่ยง (Precision)
4. ค่าประสิทธิภาพโดยรวมถ่วงน้ำหนัก (F1-Score)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ได้รู้ถึงกระบวนการทำงานของเทคนิคการจำแนกของการถดถอยลอจิสติกทวิภาค ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย และเทคนิคนาอีฟเบย์

1.4.2 ได้รู้ถึงแนวทางในการเลือกใช้เทคนิคการจำแนกให้เหมาะสมกับชุดข้อมูลด้านการเงินที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและจำนวนของตัวแปรอิสระเชิงปริมาณแตกต่างกัน

1.4.3 ได้รู้ถึงแนวทางในการเลือกใช้เทคนิคการจำแนกให้เหมาะสมกับชุดข้อมูลด้านการเงินที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและจำนวนของตัวแปรอิสระเชิงปริมาณแตกต่างกันภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุล

1.5 นิยามศัพท์เฉพาะ

ข้อมูลขนาดใหญ่ คือข้อมูลที่มีปริมาณมหาศาลที่มีขนาดใหญ่จนยากต่อการประมวลผลเมื่อใช้เทคโนโลยีแบบดั่งเดิม โดยคุณสมบัติข้อมูลที่จะเป็นข้อมูลขนาดใหญ่ต้องมี 5 องค์ประกอบหลัก ได้แก่ ได้แก่ ปริมาณ (Volume) ความหลากหลาย (Variety) ความเร็ว (Velocity) มูลค่า (Value) และความถูกต้อง (Veracity) (Ishwarappa & Anuradha, 2015)

การทำเหมืองข้อมูล คือการค้นหาข้อมูลที่มีประโยชน์จากแหล่งข้อมูลที่มีจำนวนมากมายมหาศาล เพื่อดึงข้อมูลที่มีประโยชน์มาทำการวิเคราะห์ค้นหารูปแบบหรือความสัมพันธ์ที่เกิดขึ้นในข้อมูล และจัดทำเป็นสารสนเทศเพื่อใช้ในการวางแผนบริหารจัดการหรือตัดสินใจเกี่ยวกับธุรกิจได้ ซึ่งเป็นการบูรณาการศาสตร์จากงาน 3 ส่วน ได้แก่ สถิติ (Statistics) การเรียนรู้ของเครื่อง (Machine Learning) และฐานข้อมูล (Usama et al., 1996)

การเรียนรู้ของเครื่อง คือศาสตร์สำคัญในกระบวนการการทำเหมืองข้อมูล คือการทำให้คอมพิวเตอร์ สามารถเรียนรู้สิ่งต่าง ๆ และพัฒนาการทำงานให้ดีขึ้นได้ด้วยตัวเองจากข้อมูลและสภาพแวดล้อมที่ได้รับจากการเรียนรู้ของระบบ สามารถแบ่งเป็น 3 ประเภท ได้แก่ การเรียนรู้แบบมีผู้สอน (Supervised Learning) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และการเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) (Iqbal H, 2021)

การจำแนกข้อมูล คือวิธีการจำแนกกลุ่มข้อมูล โดยจะเรียนรู้ความสัมพันธ์ระหว่างตัวแปรอิสระ (Independent Variables) กับตัวแปรตาม (Dependent Variable) ซึ่งเป็นหนึ่งในวิธีที่ได้รับการศึกษาอย่างกว้างที่สุดในการทำเหมืองข้อมูล (Iqbal H, 2021)

การจัดกลุ่ม คือการแบ่งกลุ่มข้อมูล การวิเคราะห์แบ่งกลุ่ม เป็นวิธีการจัดกลุ่มข้อมูลที่มีลักษณะเหมือนกันไว้ในกลุ่มเดียวกันเป็นส่วนหลักของการการทำเหมืองข้อมูล การรู้จำแบบ การวิเคราะห์ภาพชีวสารสนเทศศาสตร์ การบีบอัดข้อมูล คอมพิวเตอร์กราฟิกส์ การเรียนรู้ของเครื่อง และใช้ในการวิเคราะห์ข้อมูลทางสถิติ

ข้อมูลไม่สมดุล คือข้อมูลที่มีจำนวนข้อมูลของบางกลุ่มมากกว่าจำนวนข้อมูลของอีกกลุ่มอยู่เป็นจำนวนมาก โดยแบ่งเป็นข้อมูลส่วนมาก (Majority Class) และข้อมูลส่วนน้อย (Minority Class) (Chawla et al., 2002)

เทคนิคการสุ่มเพิ่ม คือเทคนิคการเพิ่มข้อมูลที่อยู่ในคลาสส่วนน้อย โดยการสุ่มเพื่อเพิ่มข้อมูลให้กับคลาสส่วนน้อย ซึ่งอาจสุ่มเลือกข้อมูลจากข้อมูลเดิมหรือสร้างข้อมูลขึ้นมาใหม่จากตัวอย่างของข้อมูลเดิมก็ได้ (Nasritha et al., 2017)

เทคนิคการสุ่มลด คือเทคนิคที่ใช้ในการสุ่มลดจำนวนข้อมูลจากคลาสส่วนมาก เพื่อทำให้จำนวนข้อมูลระหว่างคลาสส่วนมากและคลาสส่วนน้อยมีจำนวนใกล้เคียงกันมากขึ้น (กีระชาติ สุขสุทธิ์, 2559)

เทคนิคการสุ่มผสมผสาน คือเทคนิคการนำการสุ่มเพิ่มและการสุ่มลดมาทำงานร่วมกัน โดยจะเป็นการสุ่มลดจำนวนข้อมูลจากคลาสส่วนมากและสุ่มเพิ่มจำนวนข้อมูลในคลาสส่วนน้อยให้จำนวนข้อมูลจากทั้งสองคลาสมีจำนวนใกล้เคียงกันหรือเท่ากัน (กีระชาติ สุขสุทธิ์, 2559)

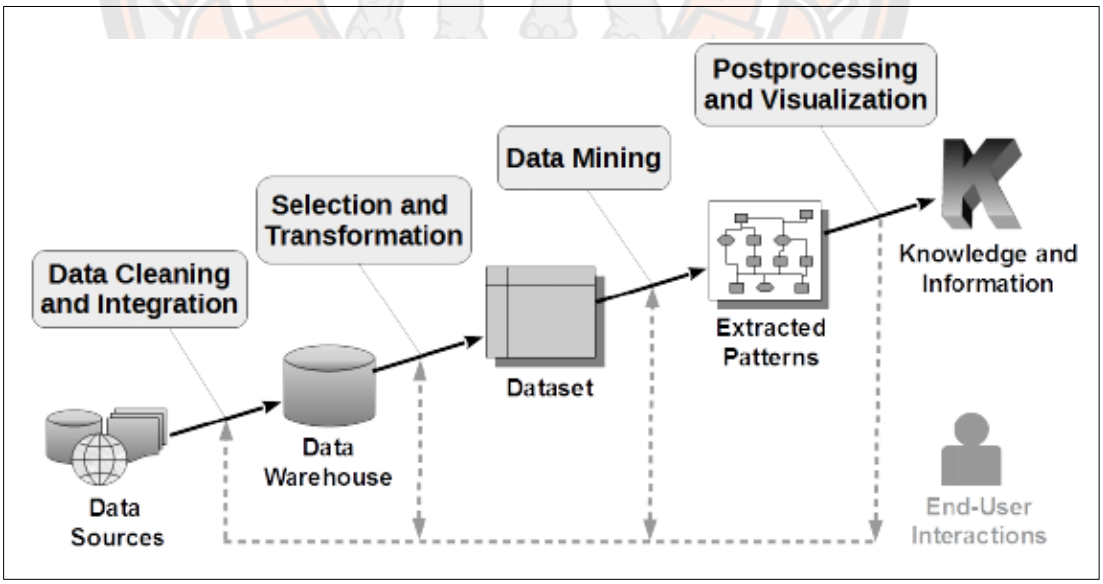
บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยครั้งนี้ผู้วิจัยต้องการศึกษากระบวนการทำงานและเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกทางด้านสถิติและทางด้านการเรียนรู้ของเครื่องโดยใช้ชุดข้อมูลด้านการเงิน โดยมีเอกสารและงานวิจัยที่เกี่ยวข้องดังนี้

2.1 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล คือ กระบวนการวิเคราะห์ข้อมูลขนาดใหญ่ เพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลส่วนใหญ่จะสามารถนำไปใช้ในหลากหลายสาขา เช่น ด้านการแพทย์ เพื่อความรวดเร็วในการวินิจฉัยโรค อีกทั้งในด้านธุรกิจ เพื่อช่วยในการตัดสินใจของนักลงทุน การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บ และตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายมาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้ จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล

กระบวนการของการทำเหมืองข้อมูล เป็นกระบวนการในการค้นหาลักษณะแฝงของข้อมูล (Pattern) ที่ซ่อนอยู่ในฐานข้อมูลโดยมีขั้นตอนดังภาพ 1



ภาพ 1 กระบวนการของการทำเหมืองข้อมูล

ที่มา : (Inthasone et al., 2014)

ขั้นตอนของการทำเหมืองข้อมูล มี 7 ขั้นตอนดังนี้

1. การกรองข้อมูล (Data Cleaning) เป็นขั้นตอนการคัดข้อมูลที่ไม่เกี่ยวข้องออก เนื่องจากข้อมูลมีความผิดปกติ เช่น ข้อมูลขาดหาย (Missing Value) ข้อมูลรบกวน (Noisy Data) ข้อมูลมีค่าผิดพลาด (Error) หรือมีค่าผิดปกติ (Outliers)

2. การรวบรวมข้อมูล (Data Integration) เป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน เพื่อลดความซ้ำซ้อน และความไม่สอดคล้องของข้อมูล เช่น มีข้อมูลในคลังข้อมูล (Data Warehouse) ในรูปแบบของดาต้าคิวบ์ (Data Cube) และมีข้อมูลในรูปแบบฐานข้อมูลเชิงสัมพันธ์ (Relational Database) จำเป็นต้องทำการรวมข้อมูลให้เป็นข้อมูลชุด

3. การคัดเลือกข้อมูล (Data Selection) เป็นขั้นตอนระบุถึงแหล่งข้อมูลที่จะนำมาใช้ในการทำเหมืองข้อมูล รวมถึงการนำข้อมูลที่ต้องการออกจากฐานข้อมูล เพื่อสร้างกลุ่มข้อมูลสำหรับพิจารณาในเบื้องต้น

4. การแปลงข้อมูล (Data Transformation) เป็นขั้นตอนการแปลงข้อมูลให้อยู่ในรูปแบบอย่างง่าย สำหรับนำมาใช้ในการทำเหมืองข้อมูล เช่น รูปแบบบรรทัดฐาน (Normalization)

5. การทำเหมืองข้อมูล (Data Mining) เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มี (Usama et al., 1996) โดยใช้เทคนิคต่าง ๆ ซึ่งสามารถแบ่งการวิเคราะห์เป็น 2 กลุ่มหลัก ได้แก่

- การวิเคราะห์เชิงพยากรณ์ (Predictive Analytics) เป็นกระบวนการการวิเคราะห์ข้อมูลที่มีอยู่ เพื่อพยากรณ์สิ่งที่จะเกิดขึ้นในอนาคต ที่สอดคล้องกับการเรียนรู้แบบมีผู้สอน (Supervised Learning) คือ การเรียนรู้ที่ทราบเป้าหมายที่ชัดเจนที่ต้องการศึกษา โดยการนำข้อมูลที่มีอยู่มาใช้ในการพยากรณ์ข้อมูลอนาคตที่ไม่ทราบผลลัพธ์ เพื่อให้เกิดความผิดพลาดน้อยที่สุด
- การวิเคราะห์เชิงบรรยาย (Descriptive Analytics) เป็นกระบวนการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) คือ การค้นหาลักษณะบางอย่างของข้อมูล ซึ่งไม่ทราบว่าข้อมูลนั้นจัดอยู่ในกลุ่มใด ใช้วิธีการวิเคราะห์ข้อมูลโดยอาศัยความเหมือน (Similarity) หรือความใกล้ชิด (Proximity) ของตัวแปรอิสระ เพื่อจัดข้อมูลที่เหมือนกัน หรือคล้ายคลึงกันมาไว้กลุ่มเดียวกัน

6. การประเมินผล (Pattern Evaluation) เป็นขั้นตอนประเมินผล นำเสนอองค์ความรู้ที่ได้เพื่อวิเคราะห์ แปลความหมาย และประเมินผลว่าผลลัพธ์นั้นเหมาะสมหรือตรงวัตถุประสงค์หรือไม่

7. การนำเสนอผลลัพธ์ (Knowledge Presentation) เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคการนำเสนอเพื่อให้เข้าใจง่าย

จากที่กล่าวไปแล้วข้างต้นว่าการทำเหมืองข้อมูลสามารถทำได้หลายรูปแบบ ทั้งนี้ขึ้นอยู่กับลักษณะของผลลัพธ์ของการทำเหมืองข้อมูล ในงานวิจัยเล่มนี้จะกล่าวถึงรูปแบบการทำเหมืองที่สัมพันธ์กับข้อมูลที่นำมาใช้ในงานวิจัย คือ วิธีการจำแนกประเภทของข้อมูล

2.2 การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression)

การถดถอยลอจิสติกเป็นวิธีการทางสถิติที่ใช้ในการศึกษาการพยากรณ์ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจต่อความน่าจะเป็นของการเกิดเหตุการณ์ที่ไม่สนใจ เมื่อมีตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มีค่าเพียง 2 ค่า จะเรียกว่า การถดถอยลอจิสติกแบบทวิภาค (Binary Logistic Regression) ในส่วนของตัวแปรอิสระอาจเป็นได้ทั้งตัวแปรเชิงปริมาณ ตัวแปรเชิงคุณภาพ หรือเป็นทั้งตัวแปรเชิงปริมาณและเชิงคุณภาพ (กัลยา วานิชย์บัญชา, 2555)

2.2.1. วัตถุประสงค์ของการถดถอยลอจิสติกทวิภาค

1. เพื่อศึกษาระดับความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระแต่ละตัวและศึกษาว่ามีตัวแปรใดบ้างที่สามารถอธิบายโอกาสที่จะทำให้เกิดเหตุการณ์ที่สนใจ และการไม่เกิดเหตุการณ์ที่สนใจ

2. เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่น่าสนใจจากตัวแบบที่เหมาะสม

2.2.2. ข้อตกลงเบื้องต้นของการถดถอยลอจิสติก

1. ตัวแปรตามเป็นข้อมูลเชิงคุณภาพ ซึ่งมีเพียง 2 ค่า เช่น 0 และ 1
2. ลอจิต (Logit) มีความสัมพันธ์เชิงเส้นกับตัวแปรอิสระ
3. ค่าเฉลี่ยของความคลาดเคลื่อนเป็นศูนย์ นั่นคือ $E(\epsilon_i) = 0$
4. ϵ_i และ ϵ_j เป็นอิสระกัน เมื่อ $i \neq j$ และ $i, j = 1, \dots, p$
5. ϵ_i และ X_i เป็นอิสระกัน
6. ตัวแปรอิสระไม่ควรมีความสัมพันธ์กันหรือไม่ควรเกิดปัญหาความสัมพันธ์เชิงเส้นพหุ (Multicollinearity)

2.2.3. ตัวแบบการถดถอยลอจิสติกทวิภาค

การถดถอยลอจิสติกทวิภาค เป็นการศึกษาความสัมพันธ์ระหว่างความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจ $P(Y = 1 | x_i) = \pi(x_i)$ กับตัวแปรอิสระ โดยที่ $\pi(x_i)$ คือความน่าจะเป็นแบบมีเงื่อนไขที่ $Y = 1$ เมื่อ Y_i มีการแจกแจงแบบแบร์นูลลี ที่มีโอกาสเกิดความสำเร็จเท่ากับ $\pi(x_i)$ ซึ่งพบว่า $\pi(x_i)$ จะอยู่ในช่วง $(0,1)$ ตัวแบบการถดถอยลอจิสติกมีรูปแบบดังต่อไปนี้

$$P(Y=1|\,x_i)=\pi(x_i)=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}} \quad (1)$$

และ

$$P(Y=0|\,x_i)=1-\pi(x_i)=\frac{1}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}} \quad (2)$$

จากสมการ (1) และ (2) พบว่าความสัมพันธ์ไม่ได้อยู่ในรูปเชิงเส้น จึงมีการปรับให้อยู่ในรูปเชิงเส้นได้ดังนี้

กำหนดให้อัตราส่วนออดส์ (Odds Ratio) หมายถึงอัตราส่วนระหว่างโอกาสที่จะเกิดเหตุการณ์ที่สนใจ $(Y=1)$ กับความน่าจะเป็นที่จะไม่เกิดเหตุการณ์ที่สนใจ $(Y=0)$ ได้ดังนี้

$$Odds=\frac{P(Y=1)}{P(Y=0)}=\frac{\pi(x_i)}{1-\pi(x_i)} \quad (3)$$

ซึ่งสามารถเขียนความสัมพันธ์ให้อยู่ในรูปเชิงเส้น ซึ่งเรียกว่า ลอจิต (logit) ได้ดังนี้

$$ln(odds)=logit=log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p \quad (4)$$

เป้าหมายของการถดถอยลอจิสติก คือ การประมาณค่า $\beta_0,\beta_1,\ldots,\beta_p$ เนื่องจาก Y_i มีการแจกแจงแบบแบร์นูลลีซึ่งมีฟังก์ชันความน่าจะเป็น ดังนี้

$$P(Y=y)=p^y(1-p)^{1-y} \qquad ;\ y=0,1 \quad (5)$$

สำหรับตัวอย่างหน่วยที่ i จะได้ว่า

$$P(Y_i=y_i)=p^{y_i}(1-p)^{1-y_i} \qquad ;\ i=0,1 \quad (6)$$

เมื่อข้อมูลตัวอย่าง n หน่วย เป็นอิสระกัน ดังนั้นฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) คือ

$$L(\beta)=\prod_{i=1}^{n}\left(\pi(x_i)\right)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i} \quad (7)$$

ลอการิทึมธรรมชาติของฟังก์ชันภาวะน่าจะเป็นสูงสุด คือ

$$\begin{aligned}
l(\beta) &= ln\left(\mathrm{L}(\beta|y_1,\ldots,y_n)\right)\\
&= \sum_{i=1}^{n}\left[y_i ln\left(\pi(x_i)\right)+(1-y_i)ln\left(1-\pi(x_i)\right)\right]\\
&= \sum_{i=1}^{n}\left[y_i ln\left(\pi(x_i)\right)+ln\left(1-\pi(x_i)\right)-y_i ln\left(1-\pi(x_i)\right)\right]\\
&= \sum_{i=1}^{n}\left[y_i ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)+ln\left(1-\pi(x_i)\right)\right]
\end{aligned}$$

$$= \sum_{i=1}^{n}\left[y_i ln\left(\frac{e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}{1+e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}} \cdot \frac{1}{1-\frac{e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}{1+e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}} \right) + ln\left(1-\frac{e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}{1+e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}\right)\right]$$

ดังนั้น

$$l(\beta) = \sum_{i=1}^{n}\left[y_i(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p) - ln\left(1+e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}\right)\right] \quad (8)$$

การประมาณค่า $\beta_0, \beta_1, \ldots, \beta_p$ จะใช้วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood method) หรือประมาณค่า $\beta_0, \beta_1, \ldots, \beta_p$ ที่ทำให้ $l(\beta)$ มีค่ามากที่สุด โดยใช้เทคนิคการทำซ้ำ (Iteration techniques) ซึ่งส่วนใหญ่จะใช้โปรแกรมสำเร็จรูปทางสถิติในการประมาณค่า (ภัคสุภางค์ มาปรีดา, 2560)

2.2.4 การตรวจสอบความสัมพันธ์ของตัวแปรอิสระ

ในการถดถอยลอจิสติก ตัวแปรอิสระไม่ควรมีความสัมพันธ์กัน หรือไม่ควรเกิดปัญหาความสัมพันธ์เชิงเส้นพหุ ซึ่งการตรวจสอบความสัมพันธ์ของตัวแปรอิสระ สามารถพิจารณาจากค่า Variance Inflation Factor (VIF) โดยมีหลักเกณฑ์ในการพิจารณา คือ ถ้าค่า VIF มากกว่า 10 แสดงว่าตัวแปรอิสระนั้น ๆ มีความสัมพันธ์กันในระดับมาก หรือ เรียกกว่า ความสัมพันธ์เชิงเส้นพหุ (Myers, 1990)

$$VIF = \frac{1}{1-R_i^2} \quad (9)$$

หรือพิจารณาจากค่าความคลาดเคลื่อนยินยอม (Tolerance) ของตัวแปรอิสระ X_i เท่ากับ $1-R_i^2$ มีค่าอยู่ระหว่าง 0 ถึง 1 ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปรอิสระ X_i กับตัวแปรอิสระตัวอื่น คือ ถ้าค่าความคลาดเคลื่อนยินยอม มีค่าเข้าใกล้ 1 แสดงว่าตัวแปรอิสระ X_i นั้นมีความสัมพันธ์กับตัวแปรอื่นน้อย แต่ถ้าค่าความคลาดเคลื่อนยินยอมมีค่าเข้าใกล้ศูนย์ แสดงว่าตัวแปรอิสระ X_i มีความสัมพันธ์กับตัวแปรอิสระอื่นมาก นั่นคือเกิดความสัมพันธ์เชิงเส้นพหุ ถ้าค่าต่ำกว่า 0.1 แสดงว่ามีปัญหาเกี่ยวกับความสัมพันธ์เชิงเส้นพหุขั้นรุนแรง โดยคำนวณตามสมการ (10)

$$Tolerence = 1-R_i^2 = \frac{1}{VIF} \quad (10)$$

โดยที่ R_i^2 แทน สัมประสิทธิ์การกำหนด (Coefficient of Determination) ของการถดถอยของตัวแปรอิสระ X_i

2.3 เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Classification and Regression Tree)

ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีหลักการการเรียนรู้แบบมีผู้สอน (Supervised Learning) และเป็นเทคนิคการจำแนกข้อมูล (Classification) ที่มีลักษณะคล้ายต้นไม้ ประกอบไป

ด้วย โหนดราก (Root Node) โหนดภายใน (Internal Node) กิ่ง (Branch) และ ใบ (Leaf) เป็นอัลกอริทึมที่ใช้ในการสร้างต้นไม้ตัดสินใจ ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยเป็นต้นไม้ตัดสินใจที่มีโครงสร้างแบบไบนารี แต่ละโหนดแสดงถึงกลุ่มย่อยของข้อมูลที่ถูกสร้างโดยการแยกโหนดเป็นสองโหนดลูกซ้ำแล้วซ้ำอีก ดังนั้นต้นไม้ตัดสินใจที่ได้จะมีลักษณะกิ่งแยกออกจากโหนด 2 กิ่ง (Brieman et al., 1984) ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยเป็นเทคนิคการจำแนกที่สามารถใช้ได้ทั้งการจำแนกและการถดถอย ความแตกต่างระหว่างต้นไม้ตัดสินใจแบบจำแนก และต้นไม้ตัดสินใจแบบถดถอยนั้นจะขึ้นอยู่กับตัวแปรตาม ซึ่งต้นไม้ตัดสินใจแบบจำแนกจะใช้กับตัวแปรตามที่เป็นเชิงคุณภาพ ส่วนต้นไม้ตัดสินใจแบบถดถอยใช้กับตัวแปรตามที่เป็นค่าต่อเนื่อง

หลักการทำงานของต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีความเหมือนกัน แตกต่างกันที่ต้นไม้ตัดสินใจแบบจำแนกจะใช้ค่าการวัดความไม่บริสุทธิ์ (Gini impurity) ในการหาจุดที่ดีที่สุดในการแบ่งข้อมูล (Split point) ตามสมการ (11)

$$GINI_i = 1 - \sum_{j=1}^{k} p_{i,j}^2 \tag{11}$$

โดยที่ $p_{i,j}$ คือสัดส่วนจากจำนวนรายการข้อมูลในโหนดที่ i ของตัวแปรที่สนใจนั้นมีกี่รายการที่อยู่ในคลาส j

k คือจำนวนคลาสคำตอบ

โดยค่าการวัดความไม่บริสุทธิ์ค่ายิ่งต่ำยิ่งแบ่งข้อมูลได้ดี จากนั้นนำมาใช้คำนวณค่า Gini Split ของตัวแปร ตามสมการ (12) เพื่อคัดเลือกตัวแปรที่มีค่า Gini Split ต่ำที่สุดมาเป็นโหนดของต้นไม้

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI_i \tag{12}$$

โดยที่ n_i คือจำนวนข้อมูลของค่าที่ i ในตัวแปรที่สนใจ

n คือจำนวนข้อมูลทั้งหมดในตัวแปรที่สนใจ

สิ่งสำคัญในขั้นตอนการคำนวณที่จะได้มาซึ่งต้นไม้ตัดสินใจแบบจำแนกนั้น จะต้องทำการแตกกิ่งออกจากโหนด 2 กิ่ง หรือทำการแยกเป็นกิ่งซ้ายและกิ่งขวาเท่านั้น

และต้นไม้ตัดสินใจแบบถดถอยจะใช้ค่า Residual Sum of Squares ในการคัดเลือกโหนดที่ดีที่สุดเพื่อแตกกิ่ง ซึ่งในงานวิจัยนี้ผู้วิจัยได้เลือกใช้เทคนิคต้นไม้ตัดสินใจแบบจำแนกกับชุดข้อมูลด้านการเงิน

ขั้นตอนการสร้างต้นไม้ตัดสินใจ

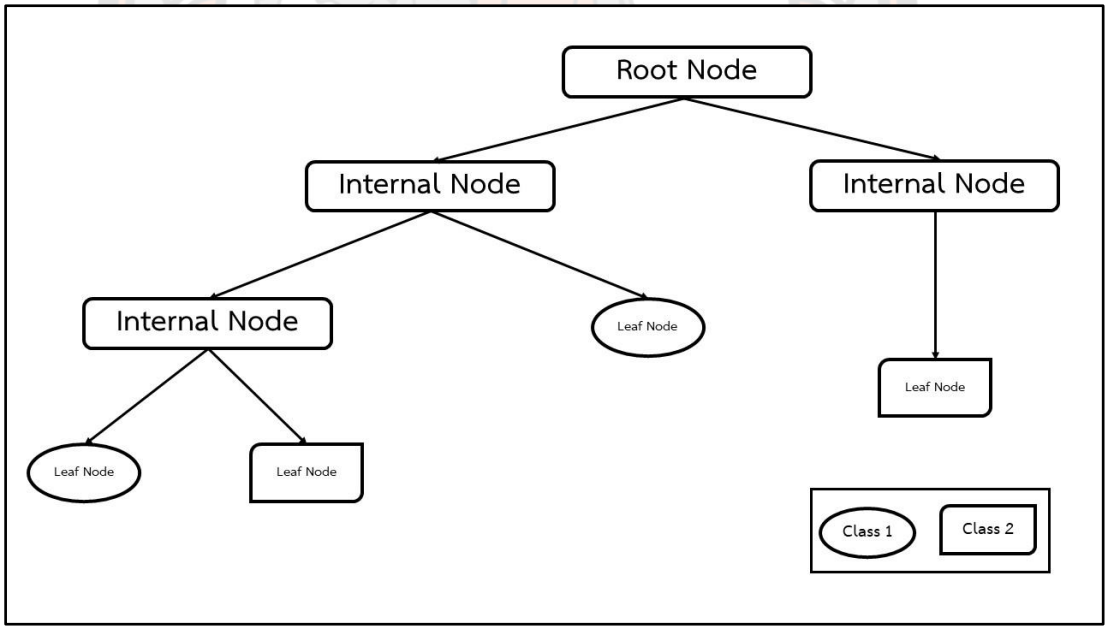
1. เริ่มต้นสร้างโหนดเพียงโหนดเดียวที่แสดงถึงชุดข้อมูล ถ้าภายในชุดข้อมูลมีตัวแปรตามเป็นตัวเดียวกันทั้งหมด ให้โหนดที่สร้างขึ้นมาเป็นใบ และกำหนดตัวแปรในใบด้วยตัวแปรตามนั้น

2. ถ้าภายในชุดข้อมูลมีหลายตัวแปร จะทำการเลือกตัวแปรอิสระที่มีความเหมาะสมที่สุดในการจำแนกชุดข้อมูล โดยวัดจากค่า Gini Split ของแต่ละตัวแปรอิสระและให้โหนดที่สร้างขึ้นมาเป็นโหนดรากและกำหนดตัวแปรในโหนดรากด้วยตัวแปรอิสระ

3. สร้างกิ่งออกมาจากโหนดรากด้วยค่าต่าง ๆ ที่เป็นไปได้ของโหนดรากนั้น และจำแนกชุดข้อมูลออกตามกิ่งต่าง ๆ ที่สร้างขึ้น

4. ทำวนซ้ำ เพื่อหาตัวแปรอิสระที่มีค่า Gini Split ที่น้อยที่สุดสำหรับข้อมูลที่ถูกแบ่งออกมาในแต่ละกิ่ง เพื่อนำตัวแปรอิสระมาสร้างโหนดตัดสินใจต่อไป โดยที่ตัวแปรอิสระที่ถูกเลือกมาเป็นโหนดแล้ว จะไม่ถูกเลือกอีก

5. ทำซ้ำจนกว่าจะได้ใบครบทุกกิ่งของต้นไม้ (หรือเข้าเงื่อนไขของการหยุดการแตกกิ่ง)



ภาพ 2 ต้นไม้ตัดสินใจด้วยเทคนิค CART

ตัวอย่างการคำนวณเทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

ตาราง 1 ตัวอย่างข้อมูลการประเมินความเสี่ยงเครดิตของลูกค้า

Customer	Saving	Assets	Income	Credit Risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	Medium	Low	25	Bad
4	Medium	High	50	Good
5	Low	High	100	Good
6	Medium	Low	25	Good
7	Low	High	25	Bad
8	Medium	Low	75	Good

จากชุดข้อมูลในตาราง 1 เมื่อพิจารณาตัวแปรอิสระ Income พบว่าตัวแปรเป็นเชิงปริมาณ ซึ่งการทำงานภายใต้ตัวแปรอิสระเชิงปริมาณ เนื่องจากต้นไม้ตัดสินใจแบบจำแนกมีโครงสร้างแบบไบนารี จึงทำการแบ่งข้อมูลด้วยค่าคงที่ออกเป็น 2 ส่วน ในที่นี้ขอยกตัวอย่างการแบ่งข้อมูลในตัวแปร Income ด้วยค่าคงที่ที่มีค่าน้อยกว่าหรือเท่ากับ 50 ให้เป็นคลาส 0 และค่าคงที่ที่มีค่ามากกว่า 50 ให้เป็นคลาส 1 ดังนั้นจะได้ตัวแปรอิสระ Income ≤ 50 และ Income > 50 ดังตาราง 2

ตาราง 2 ตัวอย่างข้อมูลการประเมินความเสี่ยงเครดิตของลูกค้าเมื่อปรับค่า Income

Customer	Saving	Assets	Income	Credit Risk
1	Medium	High	1	Good
2	Low	Low	0	Bad
3	Medium	Low	0	Bad
4	Medium	High	0	Good
5	Low	High	1	Good
6	Medium	Low	0	Good
7	Low	High	0	Bad
8	Medium	Low	1	Good

ขั้นตอนที่ 1 คำนวณค่าการวัดความไม่บริสุทธิ์ (Gini impurity) และคำนวณค่า Gini Split ตามลำดับ โดยพิจารณาจากกลุ่มคำตอบของตัวแปรตาม

ตัวแปร Saving

$$GINI_{Low} = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444$$

$$GINI_{Medium} = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.320$$

$$GINI_{split(Saving)} = \frac{3}{8}(0.444) + \frac{5}{8}(0.320) = 0.367$$

ตัวแปร Assets

$$GINI_{Low} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.500$$

$$GINI_{High} = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$$

$$GINI_{split(Assets)} = \frac{2}{8}(0.500) + \frac{4}{8}(0.375) = 0.438$$

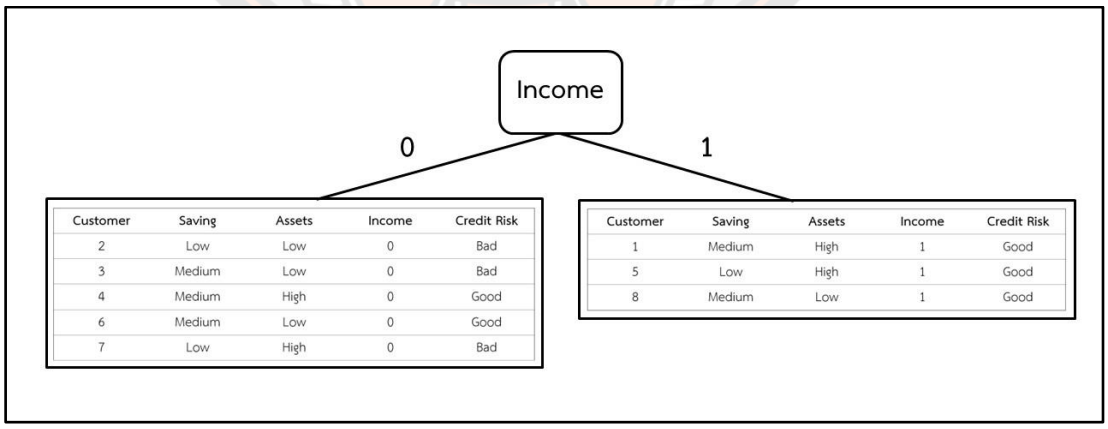
ตัวแปร Income

$$GINI_{0} = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.480$$

$$GINI_{1} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

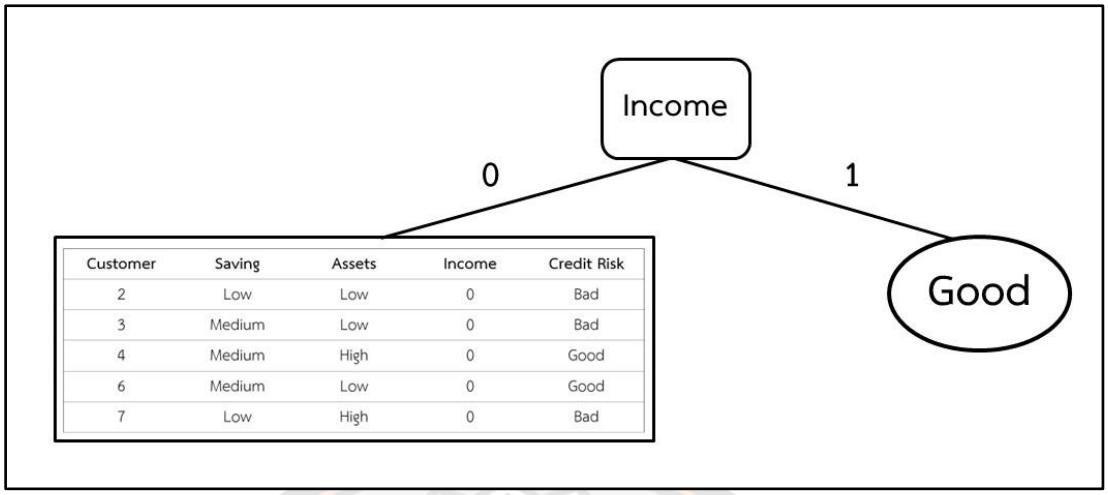
$$GINI_{split(Income)} = \frac{5}{8}(0.480) + \frac{3}{8}(0) = 0.300$$

จากขั้นตอนที่ 1 ตัวแปรอิสระที่มีค่า Gini Split ต่ำที่สุดคือ Income ดังนั้นเราจะเลือก Income เป็นโหนดราก จะได้ตัวแบบการจำแนกเบื้องต้นดังภาพ 3



ภาพ 3 เหตุการณ์การเลือกตัวแปรอิสระเป็นโหนดราก

จากภาพ 3 จะสังเกตได้ว่าชุดข้อมูลโหนดภายในจากกิ่ง 1 ข้อมูลจะเป็นคลาส Good ดังนั้นจะกำหนดให้ใบเป็น Good ดังภาพ 4

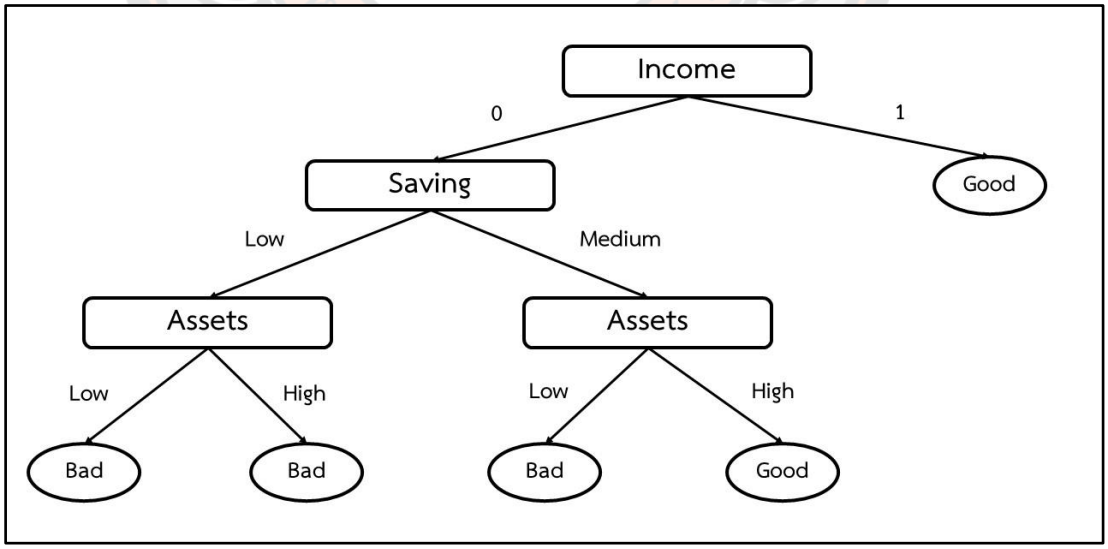


Customer	Saving	Assets	Income	Credit Risk
2	Low	Low	0	Bad
3	Medium	Low	0	Bad
4	Medium	High	0	Good
6	Medium	Low	0	Good
7	Low	High	0	Bad

ภาพ 4 โหนดภายในจากกิ่ง 1 เป็นคลาส Good

จากภาพ 4 โหนดรากคือ Income ประกอบไปด้วยกิ่ง 0 และ 1 โดยกิ่ง 1 มีใบคือ Good ในขณะที่กิ่ง 0 ต้องมีการพิจารณาโหนดภายในต่อไป

ขั้นตอนที่ 2 ทำซ้ำในขั้นตอนที่ 1 จนไม่สามารถแตกกิ่งได้อีกหรือหรือเข้าเงื่อนไขของการหยุดการแตกกิ่งที่ผู้ศึกษากำหนด จากนั้นหยุดการคำนวณ และจะได้ต้นไม้ตัดสินใจด้วยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยที่แตกกิ่งเสร็จสมบูรณ์ ดังภาพ 5



ภาพ 5 ต้นไม้ตัดสินใจที่แตกกิ่งเสร็จสมบูรณ์

2.4 เทคนิคนาอีฟเบย์ (Naïve Bayes)

เทคนิคนาอีฟเบย์ เป็นเทคนิคการจำแนกที่ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของทฤษฎีบทของเบย์ โดยกำหนดให้ $P(Y)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ Y และ $P(Y|X)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ Y เมื่อเกิดเหตุการณ์ X ก่อนหน้าแล้ว (Hong et al., 2021) โดยมีสมการความน่าจะเป็นแบบมีเงื่อนไข ดังนี้

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{13}$$

ในการวิเคราะห์ร่วมกับข้อมูลขนาดใหญ่ ซึ่งมีตัวแปรอิสระจำนวนมาก ดังนั้นสมการความน่าจะเป็นแบบมีเงื่อนไข มีดังนี้

$$P(Y_j|x_1, x_2, \dots, x_p) = \frac{P(x_1, x_2, \dots, x_p|Y_j)P(Y_j)}{P(x_1, x_2, \dots, x_p)} \tag{14}$$

ดังนั้นในสมการความน่าจะเป็นของการวิเคราะห์จำแนก จะใช้สมการดังต่อไปนี้

$$P(Y_j|x_1, x_2, \dots, x_p) = P(x_1|Y_j)P(x_2|Y_j)\cdots P(x_p|Y_j)P(Y_j) \tag{15}$$

ในงานวิจัยนี้จะจำแนกกลุ่มข้อมูลเพียง 2 กลุ่ม ดังนั้น $j = 0,1$ โดยในที่นี้ต้องการจำแนกคลาส (Y) ว่าคำตอบที่ได้ควรเป็นคลาสใด ซึ่งสามารถทำได้โดยการหาค่าของ Y ว่าคลาสใดให้ความน่าจะเป็นสูงที่สุด ซึ่งหมายถึงต้องการหาค่าที่ $Max\left(P(Y_j|x_1, x_2, \dots, x_p)\right)$ (เกรียงไกร ชัยมินทร์, 2557)

ต่อไปจะแสดงตัวอย่างการคำนวณการจำแนกโดยใช้เทคนิคนาอีฟเบย์ โดยตัวอย่างที่ใช้เป็นข้อมูลความเสี่ยงในการให้เครดิต ดังตาราง 3

ตาราง 3 ตัวอย่างข้อมูลการประเมินความเสี่ยงเครดิตของลูกค้า

Customer	Saving	Assets	Income	Credit Risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Low	25	Bad
4	Medium	High	50	Good
5	Low	High	100	Good
6	High	Low	25	Good
7	Low	Low	25	Bad
8	Medium	Low	75	Good

กำหนดให้ตัวอย่างที่ต้องการพยากรณ์ คือ $X = (High, Low, 65)$ จงแสดงการทำงานของเทคนิคนาอีฟเบย์ ว่าหากมีข้อมูลใหม่ที่มีคุณสมบัติดังกล่าวเข้ามา ค่าของ Credit Risk ควรจำแนกอยู่กลุ่มใด

ขั้นตอนที่ 1 คำนวณความน่าจะเป็นของแต่ละคลาส ดังนี้

$$P(Good) = \frac{5}{8} = 0.625, \quad P(Bad) = \frac{3}{8} = 0.375$$

ขั้นตอนที่ 2 คำนวณความน่าจะเป็นแต่ละตัวแปรได้ ดังต่อไปนี้

คำนวณความน่าจะเป็นของตัวแปร Saving = Highได้ดังนี้

$$P(High|Good) = \frac{1}{5} = 0.200, \quad P(High|Bad) = \frac{1}{3} = 0.333$$

คำนวณความน่าจะเป็นของตัวแปร Assets = Low ได้ดังนี้

$$P(Low|Good) = \frac{2}{5} = 0.400, \quad P(Low|Bad) = \frac{3}{3} = 1$$

ในกรณีที่มีตัวแปรอิสระเป็นเชิงคุณภาพ ในการคำนวณจะอาศัยแนวคิดของการแจกแจงแบบปกติ (Normal Distribution) ดังต่อไปนี้

$$P(Y_j|x_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

โดยที่ μ_{ij} คือ ค่าเฉลี่ยของข้อมูลตัวแปรอิสระที่ i และตัวแปรตามกลุ่มที่ j

σ_{ij} คือ ส่วนเบี่ยงเบนมาตรฐานตัวแปรอิสระที่ i และตัวแปรตามกลุ่มที่ j

ดังนั้นคำนวณความน่าจะเป็นของตัวแปร Income = 65 ได้ดังนี้

โดยที่ $\mu = 65$ (ค่าเฉลี่ย Income ที่ตอบ Good)

$\sigma = 25.495$ (ส่วนเบี่ยงเบนมาตรฐาน Income ที่ตอบ Good)

จะได้ $P(Income = 65|Good) = \frac{1}{\sqrt{2\pi}(25.495)} e^{-\frac{(65-65)^2}{2(25.495)^2}} = 0.0157$

โดยที่ $\mu = 33.33$ (ค่าเฉลี่ย Income ที่ตอบ Bad)

$\sigma = 11.785$ (ส่วนเบี่ยงเบนมาตรฐาน Income ที่ตอบ Bad)

$$P(Income = 65|Bad) = \frac{1}{\sqrt{2\pi}(11.785)} e^{-\frac{(65-33.33)^2}{2(11.785)^2}} = 0.000914$$

ขั้นตอนที่ 3 คำนวณหาความน่าจะเป็นเพื่อพยากรณ์การจำแนกกลุ่ม

คำนวณความน่าจะเป็นที่ Credit Risk จะเป็น Good และ Bad ตามลำดับ

$$\begin{aligned} P(Good|X) &= P(Good) \times P(Saving = High|Good) \times P(Assets = Low|Good) \\ &\quad \times P(Income = 65|Good) \\ &= 0.625 \times 0.2 \times 0.4 \times 0.0157 \\ &= 0.00785 \end{aligned}$$

$$
\begin{aligned}
P(Bad|X) &= P(Bad) \times P(Saving = High|Bad) \times P(Assets = Low|Bad) \\
&\quad \times P(Income = 65|Bad) \\
&= 0.375 \times 0.333 \times 1 \times 0.000914 \\
&= 0.000038
\end{aligned}
$$

จากการคำนวณพบว่า คำตอบของข้อมูลชุดใหม่จะจำแนก Credit Risk เป็น Good เนื่องจาก

$$P(Good|High, Low, 65) > P(Bad|High, Low, 65)$$

2.5 ข้อมูลไม่สมดุล (Imbalanced Data)

ข้อมูลไม่สมดุล คือ ข้อมูลที่มีจำนวนข้อมูลในกลุ่มหนึ่งมีจำนวนมากกว่าข้อมูลอีกกลุ่มหนึ่งเป็นจำนวนมาก ข้อมูลไม่สมดุลนั้นมีสาเหตุมาจากหลายปัจจัย เช่น เกิดจากลักษณะทางธรรมชาติของข้อมูลเอง ทางสาธารณสุขที่พบว่ามีผู้ป่วยที่ป่วยเป็นโรคน้อยกว่าผู้ที่มีสุขภาพแข็งแรงเป็นจำนวนมาก หรือข้อมูลการผลิตสินค้าในอุตสาหกรรมที่ผลิตครั้งละจำนวนมาก ซึ่งจำนวนสินค้าที่ดีมีมากกว่าสินค้าที่เสีย เป็นต้น นอกจากนี้ข้อมูลไม่สมดุลอาจเกิดจากการเก็บข้อมูลที่ผิดพลาดด้วยเช่นกัน (Chawla et al., 2014)

เนื่องจากขนาดข้อมูลของคลาสหนึ่งมีจำนวนแตกต่างกันกับอีกคลาสหนึ่งเป็นจำนวนมาก โดยแบ่งเป็นข้อมูลส่วนมาก (Majority Class) และข้อมูลส่วนน้อย (Minority Class) (Chawla et al., 2002) ข้อมูลที่ไม่สมดุลจะส่งผลต่อการจำแนกประเภทข้อมูลของกลุ่มข้อมูลส่วนน้อยได้ไม่ถูกต้อง หรือถูกต้องน้อย แต่ในขณะเดียวกันจะสามารถจำแนกประเภทข้อมูลของกลุ่มข้อมูลส่วนมากได้ถูกต้องกว่า

2.6 เทคนิคการสุ่มตัวอย่าง

งานวิจัยมีการแก้ปัญหาข้อมูลไม่สมดุล โดยมีการปรับปรุงข้อมูลในขั้นตอนก่อนที่จะมีการประมวลผล (Preprocessing) โดยการแก้ไขในระดับนี้จะแก้ไขกับข้อมูลโดยตรง โดยจะทำการปรับปรุงข้อมูลที่มีความไม่สมดุลให้กลายเป็นข้อมูลที่มีความสมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค ได้แก่

2.6.1 วิธีการสุ่มเพิ่ม (Over sampling)

เป็นเทคนิคที่ใช้ในการเพิ่มข้อมูลที่อยู่ในคลาสส่วนน้อย โดยการสุ่มเพื่อเพิ่มข้อมูลให้คลาสส่วนน้อย โดยการสุ่มเลือกข้อมูลจากข้อมูลเดิมหรือสร้างข้อมูลขึ้นมาใหม่จากตัวอย่างของข้อมูลเดิม (Nasritha et al., 2017)

- Random Over Sampler: ROS วิธีนี้เป็นการเพิ่มข้อมูลในคลาสส่วนน้อยโดยใช้การสุ่มจากชุดข้อมูลเดิม การสุ่มนี้อาจจะต้องระวังเรื่องปัญหาการซ้ำซ้อนของ

ข้อมูลที่เกิดจากการสุ่มซ้ำ ซึ่งจะส่งผลให้ตัวแบบการจำแนกเกิดการ Overfitting ได้

- Synthetic Minority Oversampling Technique: SMOTE วิธีนี้เป็นการเพิ่มข้อมูลในคลาสส่วนน้อย โดยมีการสังเคราะห์ข้อมูลใหม่จากชุดข้อมูลเดิม ทำให้ข้อมูลในคลาสส่วนน้อยมีการกระจายตัวที่ดีขึ้น และช่วยให้ค่าความสำคัญของข้อมูลไม่สูญหาย ซึ่งช่วยลดปัญหาการเกิด Overfitting ได้

2.6.2 วิธีการสุ่มลด (Under sampling)

เป็นเทคนิคที่ใช้ในการการปรับปรุงข้อมูลให้มีความสมดุลด้วยวิธีการสุ่มลดจำนวนข้อมูลจากคลาสส่วนมากลง เพื่อทำให้จำนวนข้อมูลระหว่างคลาสส่วนมากและคลาสส่วนน้อยมีจำนวนใกล้เคียงกันมากขึ้น (กีระชาติ สุขสุทธิ์, 2559)

2.6.3 วิธีการสุ่มผสมผสาน (Hybrid method)

เป็นการนำวิธีการสุ่มเพิ่มและวิธีการสุ่มลดมาทำงานร่วมกัน โดยวิธีนี้จะเป็นการสุ่มลดจำนวนข้อมูลจากคลาสส่วนมากและสุ่มเพิ่มข้อมูลในคลาสส่วนน้อย ให้จำนวนข้อมูลจากทั้งสองคลาสมีจำนวนใกล้เคียงกันหรือเท่ากัน (กีระชาติ สุขสุทธิ์, 2559)

2.7 การพัฒนาตัวแบบจำแนก

การพัฒนาตัวแบบการจำแนกนั้นเป็นการนำชุดข้อมูลที่มีอยู่มาเรียนรู้ โดยขั้นตอนแรกจะต้องนำชุดข้อมูล (Dataset) มาแบ่งออกเป็น 2 ชุด ได้แก่ ชุดข้อมูลเรียนรู้ (Training set) และชุดข้อมูลทดสอบ (Test set) จากนั้นนำชุดข้อมูลเรียนรู้ที่ได้ไปสร้างตัวแบบการจำแนกและนำชุดข้อมูลทดสอบมาทดสอบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลทดสอบ ดังนั้นชุดข้อมูลทดสอบมีจุดประสงค์เพื่อใช้ในประเมินประสิทธิภาพของตัวแบบการจำแนก

โดยทั่วไปการแบ่งข้อมูลสามารถทำได้โดยการสุ่มอิสระ เพื่อแบ่งข้อมูลออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบตามสัดส่วนต่าง ๆ ตามที่ผู้ศึกษาต้องการ เช่น แบ่งข้อมูลออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ 80:20 และ 70:30 เป็นต้น โดยชุดข้อมูลเรียนรู้จะมีจำนวนมากกว่าชุดข้อมูลทดสอบเสมอ

การแบ่งชุดข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบจัดเป็นสิ่งที่สำคัญมากกับการประเมินประสิทธิภาพตัวแบบการจำแนกด้วยเทคนิคต่าง ๆ เพื่อให้มีความละเอียดและรอบครอบกว่าวิธีการแบ่งชุดข้อมูลดังกล่าวก่อนหน้าและเพื่อให้ได้ตัวแบบการจำแนกที่มีประสิทธิภาพ ในงานวิจัยจะใช้การแบ่งข้อมูลด้วย k-Fold Cross-Validation โดยกลไกการทำงานของวิธีตรวจสอบไขว้จะมีการแบ่ง

ชุดข้อมูลเป็น k กลุ่มและนำข้อมูลจำนวน k-1 กลุ่มมาเป็นชุดข้อมูลเรียนรู้ โดยหนึ่งกลุ่มที่เหลือจะเป็นชุดข้อมูลทดสอบ และจะมีการวนสับเปลี่ยนกลุ่มเหล่านี้ เพื่อพัฒนาตัวแบบการจำแนกและทดสอบทั้งหมด k รอบดังแสดงในภาพที่ 6 ซึ่งเป็นการทำงานของ 5-Fold Cross-Validation โดยชุดข้อมูล 100% จะถูกแบ่งออกเป็น 5 กลุ่ม กลุ่มละ 20% จากนั้นในรอบที่ 1 จะให้ชุดข้อมูลที่ 1 เป็นชุดข้อมูลทดสอบ และชุดที่ 2-5 เป็นชุดข้อมูลเรียนรู้ จากนั้นทำการสร้างตัวแบบการจำแนกและทดสอบทั้งหมด 5 รอบ

	ชุดที่ 1	**ชุดที่ 2**	**ชุดที่ 3**	**ชุดที่ 4**	**ชุดที่ 5**
รอบที่ 1	Test	Train	Train	Train	Train
รอบที่ 2	Train	Test	Train	Train	Train
รอบที่ 3	Train	Train	Test	Train	Train
รอบที่ 4	Train	Train	Train	Test	Train
รอบที่ 5	Train	Train	Train	Train	Test

ภาพ 6 โครงสร้างการทำงานของ 5-Fold Cross-Validation

2.8 เกณฑ์การวัดประสิทธิภาพของตัวแบบการจำแนก

การวัดประสิทธิภาพของตัวแบบการจำแนกเป็นเครื่องมือตรวจสอบประสิทธิภาพของตัวแบบการจำแนก เพื่อตรวจสอบว่าตัวแบบการจำแนกมีความแม่นยำตามเกณฑ์ที่ใช้วัดประสิทธิภาพมากน้อยเพียงใด โดยเกณฑ์ที่ใช้วัดประสิทธิภาพของตัวแบบการจำแนกในวิจัยนี้มีด้วยกัน 4 เกณฑ์ ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าเรียกคืน (Recall) ค่าความเที่ยง (Precision) และค่าประสิทธิภาพโดยรวมถ่วงน้ำหนัก (F1-Score) โดยเกณฑ์ทั้ง 4 เกณฑ์ที่เลือกใช้จะอ้างอิงการคำนวณค่าที่ได้จากเมทริกซ์ความสับสน (Confusion Matrix) ดังต่อไปนี้

2.8.1 เมทริกซ์ความสับสน (Confusion Matrix)

เมทริกซ์ความสับสน เป็นตารางที่ใช้ในการประเมินผลลัพธ์การทำนายของตัวแบบการจำแนก โดยนำผลลัพธ์ของตัวแบบการจำแนกเปรียบเทียบกับผลลัพธ์จริง หรือจากผลเฉลยที่ทราบค่าแท้จริงอยู่ก่อนแล้ว การประเมินผลลัพธ์การทำนายของตัวแบบ สามารถวัดจากผลลัพธ์ที่ได้จากการ

จำแนกกลุ่ม ได้แก่ ค่าความถูกต้องเชิงบวก (True Positive) ค่าความถูกต้องเชิงลบ (True Negative) ค่าความผิดพลาดเชิงบวก (False Positive) และค่าความผิดพลาดเชิงลบ (False Negative) แสดงดังตาราง 4

ตาราง 4 เมทริกซ์ความสับสน (Confusion Matrix)

ค่าที่แท้จริง	**ค่าที่ทำนายได้ (Predicted Class)**	
(Actual Class)	**Positive**	**Negative**
Positive	TP	FN
Negative	FP	TN

อธิบายรายละเอียดจากตารางได้ดังนี้

TP (True Positive) คือ จำนวนครั้งที่การจำแนกข้อมูลซึ่งมีค่าแท้จริงอยู่ใน Class Yes และมีการทำนายว่าอยู่ใน Class Yes (ทำนายถูกต้อง)

FP (False Positive) คือ จำนวนครั้งที่การจำแนกข้อมูลซึ่งมีค่าแท้จริงอยู่ใน Class No แต่มีการทำนายว่าอยู่ใน Class Yes (ทำนายผิด คำตอบจริง No แต่ทำนาย Yes)

TN (True Negative) คือ จำนวนครั้งที่การจำแนกข้อมูลซึ่งมีค่าแท้จริงอยู่ใน Class No และมีการทำนายว่าอยู่ใน Class No (ทำนายถูกต้อง)

FN (False Negative) คือ จำนวนครั้งที่การจำแนกข้อมูลซึ่งมีค่าแท้จริงอยู่ใน Class Yes แต่มีการทำนายว่าอยู่ใน Class No (ทำนายผิด คำตอบจริง Yes แต่ทำนายว่า No)

2.8.2. เกณฑ์วัดประสิทธิภาพของตัวแบบการจำแนก

โดยเกณฑ์วัดประสิทธิภาพที่ใช้ในงานวิจัยทั้ง 4 เกณฑ์ ได้แก่

1) ค่าความแม่นยำ (Accuracy: Acc) คือจำนวนข้อมูลจำแนกถูกต้องต่อจำนวนข้อมูลทั้งหมด ดังสมการ (16)

$$\text{Acc} = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

2) ค่าเรียกคืน (Recall: Re) คือ ค่าที่อธิบายถึงความถูกต้องของผลการทำนายของกลุ่มข้อมูลที่กำลังพิจารณาอยู่เมื่อเทียบกับผลของความเป็นจริง สามารถคำนวณได้ตามสมการ (17) และ (18)

$$Re(Yes) = \frac{TP}{TP + FN} \tag{17}$$

และ

$$Re(No) = \frac{TN}{TN + FP} \tag{18}$$

3) ค่าความเที่ยง (Precision: Pre) คือ ค่าที่อธิบายถึงความถูกต้องของกลุ่มข้อมูลที่กำลังพิจารณาเมื่อเทียบกับผลลัพธ์ของการทำนาย สามารถคำนวณได้ตามสมการ (19) และ (20)

$$Pre(Yes) = \frac{TP}{TP + FP} \tag{19}$$

และ

$$Pre(No) = \frac{TN}{TN + FN} \tag{20}$$

4) ค่าประสิทธิภาพโดยรวม (F1-Score: F1) คือค่าที่แสดงประสิทธิภาพที่ได้จากการนำค่าการเรียกคืน (Recall) กับค่าความเที่ยง (Precision) มารวมเป็นค่าเดียว ดังสมการ (21), (22) และ (23)

$$F1(Yes) = \frac{2 \times Pre(Yes) \times Re(Yes)}{Pre(Yes) + Re(Yes)} \tag{21}$$

และ

$$F1(No) = \frac{2 \times Pre(No) \times Re(No)}{Pre(No) + Re(No)} \tag{22}$$

จะได้

$$F1 = \frac{n_{Yes}}{N}\big(F1(Yes)\big) + \frac{n_{No}}{N}\big(F1(No)\big) \tag{23}$$

โดยที่ n_{Yes} แทนจำนวนของค่าจริงบวก

n_{No} แทนจำนวนของค่าจริงลบ

N แทนจำนวนของค่าทั้งหมด

2.9 งานวิจัยที่เกี่ยวข้อง

Irimia Dieguez et al. (2015) เปรียบเทียบประสิทธิภาพการทำนายของวิธีการแบบไม่ใช้พารามิเตอร์ ได้แก่ ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Classification and Regression Tree) และการถดถอยลอจิสติก (Logistic Regression) โดยการใช้ชุดการจับคู่บัญชีที่ตรงกันของ

วิสาหกิจขนาดย่อม ชุดข้อมูลจัดทำโดยหน่วยงานสินเชื่อของสหราชอาณาจักร ในช่วงปี พ.ศ. 2542 ถึง พ.ศ. 2551 ซึ่งมีจำนวนข้อมูทั้งหมด 39,710 ชุด มีการใช้ตัวแปรเกี่ยวกับเศรษฐกิจมหภาค ตัวแปรทางการเงินและที่ไม่ใช่ตัวแปรทางการเงินรวมทั้งหมด 25 ตัวแปร โดยใช้ค่าพื้นที่ใต้กราฟ ROC (Area Under ROC Curve: AUC) และค่าคาดหวังของฟังก์ชันระหว่างคลาสค่าความจริงและคลาสค่าทำนาย (Expected Misclassification Cost: EMC) เป็นเกณฑ์ในการวัดประสิทธิภาพของตัวแบบการจำแนก ผลการวิจัยพบว่า ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีค่า AUC และค่า EMC เท่ากับ 0.7% (-3.1%) และ 20.86% (7.49%) ในตัวอย่างชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบตามลำดับ ดังนั้น ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพดีกว่าการถดถอยลอจิสติก

Jie and Jennifer (2016) ศึกษาการทำนายจำนวนเงินที่เกินกำหนดโดยใช้การถดถอยลอจิสติก (Logistic Regression) เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) ในงานวิจัยกำหนดค่า K เท่ากับ 5 (5-Nearest Neighbor) และเทคนิคป่าสุ่ม (Random Forest) ชุดข้อมูลที่ใช้ในการวิเคราะห์จัดทำโดย Equifax ซึ่งมีตัวแปรทางการเงิน 305 ตัวแปร แต่มีตัวแปร 23 ตัวแปรที่เกี่ยวข้องกับการเกินกำหนดค้างชำระ มีจำนวนข้อมูลจากธุรกิจที่ไม่ซ้ำกันมากกว่า 11,787,287 แถว โดยเป็นข้อมูลตั้งแต่ปี พ.ศ. 2549 ถึง พ.ศ. 2557 โดยใช้ค่าพื้นที่ใต้กราฟ ROC (Area Under ROC Curve: AUC) เป็นเกณฑ์ในการวัดประสิทธิภาพของตัวแบบการจำแนก ผลการวิจัยพบว่า เทคนิคเพื่อนบ้านใกล้ที่สุดมีค่า AUC เท่ากับ 95.37% ส่วนเทคนิคป่าสุ่มมีค่า AUC เท่ากับ 82.01% และการถดถอยลอจิสติกมีค่า AUC เท่ากับ 96.3% ดังนั้นการถดถอยลอจิสติกเป็นวิธีที่ดีที่สุดในการใช้ทำนายจำนวนเงินที่เกินกำหนด

Aida Krichene (2017) ศึกษาเกี่ยวกับการคาดการณ์การผิดนัดชำระของเงินกู้ระยะสั้นสำหรับธนาคารพาณิชย์ตูนิเซีย เพื่อให้ผู้ให้บริการสินเชื่อสามารถวิเคราะห์หรือทำความเข้าใจระดับความเสี่ยงของลูกค้าสินเชื่อได้ โดยนำเทคนิคนาอีฟเบย์ (Naïve Bayes) มาใช้หาตัวแบบการจำแนกเพื่อใช้ประเมินความเสี่ยงจากการผิดนัดชำระหนี้ของสินเชื่อและการประเมินความเสี่ยงด้านเครดิตของลูกค้า ในงานวิจัยมีการใช้ทั้งตัวแปรที่เกี่ยวกับการเงินและตัวแปรที่ไม่เกี่ยวกับการเงินรวมทั้งหมด 24 ตัวแปร โดยตัวแปร 22 ตัวแปร เป็นตัวแปรที่เกี่ยวกับการเงินและ 2 ตัวแปรที่เหลือ เป็นตัวแปรที่ไม่เกี่ยวกับการเงิน โดยมีการใช้ชุดข้อมูลของสินเชื่อที่ธนาคารพาณิชย์แห่งหนึ่งมอบให้กับบริษัทอุตสาหกรรมในตูนิเซียในปี พ.ศ. 2546, พ.ศ. 2547, พ.ศ. 2548 และ พ.ศ. 2549 โดยใช้ค่าพื้นที่ใต้

กราฟ ROC (Area Under ROC Curve: AUC) เป็นเกณฑ์ในการวัดประสิทธิภาพของตัวแบบการจำแนก ผลการวิจัยพบว่า เทคนิคนาอีฟเบย์มีประสิทธิภาพดีโดยมีค่า AUC อยู่ที่ร้อยละ 69

Begüm and Deniz (2019) ศึกษาตัวแบบการจำแนกที่ได้จากการวิเคราะห์ข้อมูล และนำตัวแบบการจำแนกที่ได้ไปใช้ในการทำนายความเสี่ยงจากการผิดนัดชำระ เพื่อหลีกเลี่ยงปัญหาการชำระเงินที่อาจเกิดขึ้น และเพื่อลดปัญหาในการขยายเครดิต โดยใช้ชุดข้อมูลจากการสำรวจของสถาบันสถิติแห่งตุรกีในปี 2558 จำนวน 20,275 หน่วย ใช้ตัวแปรทั้งหมด 13 ตัวแปรรวมคลาส และใช้เทคนิคการจำแนก 6 เทคนิค ได้แก่ เทคนิคนาอีฟเบย์ (Naïve Bayes) ข่ายงานเบย์ (Bayesian Network) ต้นไม้ตัดสินใจ (J48) เทคนิคป่าสุ่ม (Random Forest) โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron) และ การถดถอยลอจิสติก (Logistic Regression) ในการสร้างตัวแบบ โดยใช้ WEKA 3.9 เป็นเครื่องมือในการวิเคราะห์ และใช้ รากที่สองของค่าเฉลี่ยความผิดพลาดกำลังสอง (Root Mean Square Error: RMSE) พื้นที่เส้นโค้ง ROC (Receiver Operating Characteristic Area) ค่าความแม่นยำ (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และ ค่าประสิทธิภาพ (F-measure) เป็นเกณฑ์ในการวัดประสิทธิภาพ ผลการวิจัยพบว่า การถดถอยลอจิสติกเป็นเทคนิคที่มีประสิทธิภาพดีที่สุดโดยมีค่า RMSE เท่ากับ 0.342 ค่าความแม่นยำเท่ากับ 83.108% พื้นที่เส้นโค้ง ROC เท่ากับ 0.843 ค่าความเที่ยงเท่ากับ 0.822 ค่าเรียกคืนเท่ากับ 0.831 และค่าประสิทธิภาพเท่ากับ 0.82

Barış and Derviş (2020) ใช้เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Networks: ANN) อัลกอริทึมการจำแนก C5.0 และต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Classification and Regression Tree: CART) เพื่อทำนายความสำเร็จและความล้มเหลวทางการเงินของ 126 ธุรกิจที่ดำเนินการใน BIST (Borsa İstanbul) ภาคอุตสาหกรรมการผลิต ชุดข้อมูลอยู่ในช่วงปี พ.ศ. 2549 ถึง พ.ศ. 2552 ในการศึกษาใช้จำนวนตัวแปรเชิงปริมาณ 25 ตัวแปร และจำนวนตัวแปรเชิงคุณภาพ 4 ตัวแปร ผลการวิจัยพบว่า ความแม่นยำในการจำแนกโดยรวมจากสูงสุดไปต่ำสุดของ 3 ปี ก่อนปีแห่งความสำเร็จ-ล้มเหลว (สำหรับปี พ.ศ. 2549) คือ 84.21% สำหรับ CART, 81.58% สำหรับ ANN และ 76.32% สำหรับ C5.0 ตามลำดับ ความแม่นยำในการจำแนกโดยรวมจากค่าสูงสุดไปต่ำสุดของ 2 ปี ก่อนปีที่ประสบความสำเร็จ-ล้มเหลว (สำหรับปี พ.ศ. 2550) คือ 86.84% สำหรับ CART, 84.21% สำหรับ ANN, 78.95% สำหรับ C5.0 ตามลำดับ และความแม่นยำในการจำแนกโดยรวมจากสูงสุดไปต่ำสุดของ 1 ปี ก่อนปีที่ประสบความสำเร็จ-ล้มเหลว (สำหรับปี พ.ศ. 2551) คือ 92.11% สำหรับ CART, 92.11% สำหรับ ANN และ 86.84% สำหรับ C5.0 ตามลำดับ พบว่า ตัวแบบการ

จำแนกที่ได้จากเทคนิคโครงข่ายประสาทเทียมและต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีความมีประสิทธิภาพที่ใกล้เคียงกัน

Fayaz et al. (2020) ศึกษาการนำตัวแบบการจำแนกที่ได้จากเทคนิคการจำแนก 3 เทคนิค ได้แก่ การถดถอยลอจิสติก (Logistic Regression) เทคนิคนาอีฟเบย์ (Naïve Bayes) และเทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) มาคาดการณ์การทำธุรกรรมทางการเงินที่เป็นการฉ้อโกง ซึ่งข้อมูลที่นำมาใช้ในการวิเคราะห์นั้น เป็นข้อมูลที่ไม่สมดุล จึงมีการนำเทคนิคการสุ่มตัวอย่างซ้ำมาประยุกต์ใช้กับข้อมูล เพื่อให้ได้ผลลัพธ์ของการวิเคราะห์ที่ดีขึ้น การศึกษานี้ใช้ชุดข้อมูลที่มีสัดส่วนที่แตกต่างกัน 3 ชุด และใช้เทคนิคการสุ่มลดสำหรับชุดข้อมูลที่ไม่สมดุล จากการศึกษาพบว่า การถดถอยลอจิสติกให้ค่าความแม่นยำ ค่าความไว ค่าความจำเพาะ ค่าความเที่ยง ค่าประสิทธิภาพ และพื้นที่ใต้เส้นโค้งดีกว่าเทคนิคนาอีฟเบย์ และเทคนิคเพื่อนบ้านใกล้ที่สุด แสดงให้เห็นว่าการถดถอยลอจิสติกมีประสิทธิภาพดีที่สุด

Yuzhen and Jingqiao (2021) ใช้วิธีการถดถอยลอจิสติกวิเคราะห์ลักษณะที่แตกต่างกันของลูกค้าที่แตกต่างกันในการซื้อผลิตภัณฑ์ทางการเงินและปรับปรุงความถูกต้องของการระบุตัวตนของลูกค้า โดยพิจารณาจากการเปรียบเทียบดัชนีค่าประสิทธิภาพ วิธีการจำแนกประเภทด้วยการวิเคราะห์ความสัมพันธ์ภายในระหว่างการสื่อสารของที่ปรึกษาผลิตภัณฑ์กับลูกค้า สินเชื่อส่วนบุคคล และเงินฝาก ซึ่งตัวแบบการจำแนกที่ได้จากวิธีการถดถอยลอจิสติกสามารถคาดการณ์พฤติกรรมการซื้อผลิตภัณฑ์ทางการเงินและปรับปรุงความถูกต้องของการระบุตัวตนของลูกค้าได้อย่างแม่นยำ

ตาราง 5 แสดงผลสรุปงานวิจัยที่เกี่ยวข้องด้านเทคนิคการจำแนกและจำนวนของตัวแปรอิสระ

งานวิจัย	ปี	เทคนิคการจำแนก				จำนวนตัวแปรอิสระ	
		LR	CART	NB	อื่น ๆ	เชิงคุณภาพ	เชิงปริมาณ
Irimia, A et al.	2015	✓	✓*			-	-
Jie and Jennifer	2016	✓*			✓	-	-
Aida Krichene	2017			✓*		-	-
Begüm and Deniz	2019	✓*		✓	✓	-	-
Barış and Derviş	2020		✓*		✓	4	25
Fayaz et al.	2020	✓*		✓	✓	0	28
Yuzhen and Jingqiao	2021	✓*			✓	-	-

* เทคนิคการจำแนกที่ดีที่สุด

บทที่ 3
วิธีดำเนินการวิจัย

งานวิจัยนี้มีจุดมุ่งหมายเพื่อศึกษากระบวนการทำงานของตัวแบบการจำแนก ได้แก่ การถดถอยลอจิสติกทวิภาค ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย และเทคนิคนาอีฟเบย์ จากนั้นทำการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงินที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกันทั้งหมด 3 แบบ และเพื่อเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงินในชุดข้อมูลตั้งต้นและชุดข้อมูลที่มีการปรับปรุงข้อมูลไม่สมดุล โดยมีขั้นตอนในการดำเนินงานวิจัยดังต่อไปนี้

3.1 ข้อมูลที่ใช้ในการวิจัย

งานวิจัยนี้ได้คัดเลือกชุดข้อมูลจาก UCI Machine Learning Repository โดยข้อมูลแต่ละชุดมีลักษณะดังต่อไปนี้

3.1.1 ชุดที่ 1 ชุดข้อมูลเครดิตเยอรมัน (German credit dataset)

เป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ โดยมีจำนวนข้อมูลทั้งหมด 1,000 แถว ซึ่งประกอบด้วยจำนวนตัวแปรอิสระ 20 ตัวแปร แบ่งเป็นตัวแปรอิสระเชิงคุณภาพ 13 ตัวแปร และตัวแปรอิสระเชิงปริมาณ 7 ตัวแปร และมีจำนวนตัวแปรตามเชิงคุณภาพ 1 ตัวซึ่งแบ่งออกเป็น 2 กลุ่มได้แก่ การจัดประเภทลูกค้าที่ดีจำนวน 700 แถว และการจัดประเภทลูกค้าที่ไม่ดีจำนวน 300 แถว

3.1.2 ชุดที่ 2 ชุดข้อมูลลูกค้าบัตรเครดิต (Default of credit card clients dataset)

เป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ โดยมีจำนวนข้อมูลทั้งหมด 30,000 แถว ซึ่งประกอบด้วยจำนวนตัวแปรอิสระ 23 ตัวแปร แบ่งเป็นตัวแปรอิสระเชิงคุณภาพ 9 ตัวแปร และตัวแปรอิสระเชิงปริมาณ 14 ตัวแปร และมีจำนวนตัวแปรตามเชิงคุณภาพ 1 ตัวซึ่งแบ่งออกเป็น 2 กลุ่มได้แก่ ลูกค้าไม่ผิดนัดชำระเดือนหน้าจำนวน 23,364 แถว และลูกค้าผิดนัดชำระเดือนหน้าจำนวน 6,636 แถว

3.1.3 ชุดที่ 3 ชุดข้อมูลการตลาดของธนาคาร (Bank marketing dataset)

มีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ โดยมีจำนวนข้อมูลทั้งหมด 41,188 แถว ซึ่งประกอบด้วยจำนวนตัวแปรอิสระ 20 ตัวแปร แบ่งเป็นตัวแปรอิสระเชิงคุณภาพ 10 ตัวแปร และตัวแปรอิสระเชิงปริมาณ 10 ตัวแปร และมีจำนวนตัวแปรตามเชิงคุณภาพ 1 ตัวซึ่งแบ่งออกเป็น 2 กลุ่มได้แก่ ลูกค้าสมัครฝากประจำจำนวน 36,548 แถว และลูกค้าไม่สมัครฝากประจำจำนวน 4,640 แถว

ตาราง 6 แสดงรายละเอียดชุดข้อมูลด้านการเงินทั้ง 3 ชุด

ชุดข้อมูล	จำนวนตัวแปรอิสระ		ค่าของตัวแปรตาม		สัดส่วน
	เชิงคุณภาพ	เชิงปริมาณ	0	1	ความสมดุล
เครดิตเยอรมัน	13	7	300	700	1:2.33
ข้อมูลลูกค้าบัตรเครดิต	9	14	6,636	23,364	1:3.55
การตลาดของธนาคาร	10	10	4,640	36,548	1:8.09

จากตารางที่ 6 พบว่า ชุดข้อมูลด้านการเงินทั้ง 3 ชุด ไม่มีข้อมูลผิดปกติจากการบันทึกข้อมูลสูญหาย (Missing Data) ดังนั้นสามารถนำชุดข้อมูลทั้ง 3 ชุดไปทำการวิเคราะห์ต่อไปได้

3.2 เครื่องมือที่ใช้ในการวิจัย

3.2.1 ใช้โปรแกรม Microsoft Excel ในการคัดกรองข้อมูลเบื้องต้น

3.2.2 ใช้โปรแกรม RStudio เวอร์ชัน 4.1.0 ในการปรับปรุงชุดข้อมูลไม่สมดุลให้สมดุลและสร้างตัวแบบการจำแนกทั้ง 3 เทคนิค ดังนี้

1) ปรับปรุงชุดข้อมูลไม่สมดุลให้สมดุลโดยใช้แพ็กเกจ ROSE โดยแบ่งเป็น 3 เทคนิค ดังนี้

- เทคนิคการสุ่มเพิ่ม ใช้ฟังก์ชัน ovun.sample() ซึ่งมีพารามิเตอร์ที่สำคัญคือ method กำหนดค่าเป็น over
- เทคนิคการสุ่มลด ใช้ฟังก์ชัน ovun.sample() ซึ่งมีพารามิเตอร์ที่สำคัญคือ method กำหนดค่าเป็น under
- เทคนิคการสุ่มผสมผสาน ใช้ฟังก์ชัน ovun.sample() ซึ่งมีพารามิเตอร์ที่สำคัญคือ method กำหนดค่าเป็น both

2) สร้างตัวแบบด้วยการถดถอยลอจิสติกทวิภาคโดยใช้ฟังก์ชัน glm()

3) สร้างตัวแบบด้วยต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยโดยใช้แพ็กเกจ rpart และฟังก์ชัน rpart()

4) สร้างตัวแบบด้วยเทคนิคนาอีฟเบย์โดยใช้แพ็กเกจ naivebayes และฟังก์ชัน naive_bayes()

3.3 วิธีวิเคราะห์และจัดเตรียมข้อมูล

การวิจัยนี้มีจุดมุ่งหมายเพื่อเปรียบเทียบตัวแบบการจำแนกโดยใช้เทคนิคทางสถิติและเทคนิคการเรียนรู้ของเครื่องทั้งหมด 3 เทคนิค ภายใต้ชุดข้อมูลทางการเงินที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและจำนวนตัวแปรอิสระเชิงปริมาณที่แตกต่างกัน 3 แบบ ได้แก่ ชุดข้อมูลเครดิตเยอรมัน (German credit dataset) ชุดข้อมูลลูกค้าบัตรเครดิต (Default of credit card clients dataset) ชุดข้อมูลการตลาดของธนาคาร (Bank marketing dataset) เนื่องจากชุดข้อมูลที่นำมาใช้ในการวิเคราะห์เป็นข้อมูลที่มีความไม่สมดุลสูง จึงต้องมีการปรับปรุงชุดข้อมูลไม่สมดุลให้สมดุล โดยแบ่งการรายงานเป็น 4 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 ศึกษารายละเอียดแต่ละตัวแปรในชุดข้อมูล

ชุดข้อมูลที่ 1 ชุดข้อมูลเครดิตเยอรมัน

ตาราง 7 แสดงรายละเอียดชุดข้อมูลเครดิตเยอรมัน

ตัวแปร	ความหมาย	ลักษณะของข้อมูล
X_1	สถานะของบัญชี	เชิงคุณภาพ
X_2	ระยะเวลา	เชิงปริมาณ
X_3	ประวัติเครดิต	เชิงคุณภาพ
X_4	วัตถุประสงค์	เชิงคุณภาพ
X_5	วงเงิน	เชิงปริมาณ
X_6	บัญชีออมทรัพย์/พันธบัตร	เชิงคุณภาพ
X_7	ระยะเวลาการจ้างงาน	เชิงคุณภาพ
X_8	อัตราการผ่อนชำระ(%)	เชิงปริมาณ
X_9	สถานภาพสมรสและเพศ	เชิงคุณภาพ
X_{10}	ลูกหนี้อื่น/ผู้ค้ำประกัน	เชิงคุณภาพ
X_{11}	ปัจจุบันอยู่อาศัยกี่ปี	เชิงปริมาณ
X_{12}	คุณสมบัติ	เชิงคุณภาพ
X_{13}	อายุ	เชิงปริมาณ
X_{14}	แผนการผ่อนชำระ	เชิงคุณภาพ
X_{15}	ที่อยู่อาศัย	เชิงคุณภาพ
X_{16}	จำนวนสินเชื่อในธนาคาร	เชิงปริมาณ
X_{17}	สถานะการทำงาน	เชิงคุณภาพ

ตัวแปร	ความหมาย	ลักษณะของข้อมูล
X_{18}	จำนวนผู้มีหน้าที่บำรุงรักษา	เชิงปริมาณ
X_{19}	โทรศัพท์	เชิงคุณภาพ
X_{20}	แรงงานต่างด้าว	เชิงคุณภาพ
Y	การจัดประเภทเครดิตลูกค้า (ดี = 0, ไม่ดี = 1)	เชิงคุณภาพ

ชุดข้อมูลที่ 2 ชุดข้อมูลลูกค้าบัตรเครดิต

ตาราง 8 แสดงรายละเอียดชุดข้อมูลลูกค้าบัตรเครดิต

ตัวแปร	ความหมาย	ลักษณะของข้อมูล
X_1	จำนวนเครดิตที่กำหนด	เชิงปริมาณ
X_2	เพศ	เชิงปริมาณ
X_3	ระดับการศึกษา	เชิงปริมาณ
X_4	สถานภาพสมรส	เชิงคุณภาพ
X_5	อายุ	เชิงปริมาณ
X_6	การชำระหนี้เดือน ก.ย.	เชิงปริมาณ
X_7	การชำระหนี้เดือน ส.ค.	เชิงคุณภาพ
X_8	การชำระหนี้เดือน ก.ค.	เชิงปริมาณ
X_9	การชำระหนี้เดือน มิ.ย.	เชิงคุณภาพ
X_{10}	การชำระหนี้เดือน พ.ค.	เชิงคุณภาพ
X_{11}	การชำระหนี้เดือน เม.ย.	เชิงคุณภาพ
X_{12}	เงินเรียกเก็บเดือน ก.ย.	เชิงคุณภาพ
X_{13}	เงินเรียกเก็บเดือน ส.ค.	เชิงปริมาณ
X_{14}	เงินเรียกเก็บเดือน ก.ค.	เชิงปริมาณ
X_{15}	เงินเรียกเก็บเดือน มิ.ย.	เชิงปริมาณ
X_{16}	เงินเรียกเก็บเดือน พ.ค.	เชิงปริมาณ
X_{17}	เงินเรียกเก็บเดือน เม.ย.	เชิงปริมาณ
X_{18}	เงินที่ชำระเดือน ก.ย.	เชิงปริมาณ
X_{19}	เงินที่ชำระเดือน ส.ค.	เชิงปริมาณ
X_{20}	เงินที่ชำระเดือน ก.ค.	เชิงปริมาณ
X_{21}	เงินที่ชำระเดือน มิ.ย.	เชิงปริมาณ

ตัวแปร	ความหมาย	ลักษณะของข้อมูล
X_{22}	เงินที่ชำระเดือน พ.ค.	เชิงปริมาณ
X_{23}	เงินที่ชำระเดือน เม.ย.	เชิงปริมาณ
Y	ผิดนัดชำระเดือนหน้า (ไม่ใช่ = 0, ใช่ = 1)	เชิงคุณภาพ

ชุดข้อมูลที่ 3 ชุดข้อมูลการตลาดของธนาคาร

ตาราง 9 แสดงรายละเอียดชุดข้อมูลการตลาดของธนาคาร

ตัวแปร	ความหมาย	ลักษณะของข้อมูล
X_1	อายุ	เชิงปริมาณ
X_2	งาน	เชิงคุณภาพ
X_3	สถานภาพสมรส	เชิงคุณภาพ
X_4	การศึกษา	เชิงคุณภาพ
X_5	การมีเครดิต	เชิงคุณภาพ
X_6	ที่อยู่อาศัย	เชิงคุณภาพ
X_7	เงินกู้	เชิงคุณภาพ
X_8	ผู้ติดต่อ	เชิงคุณภาพ
X_9	เดือนที่ติดต่อล่าสุด	เชิงคุณภาพ
X_{10}	วันที่ติดต่อล่าสุด	เชิงคุณภาพ
X_{11}	ระยะเวลาติดต่อล่าสุด	เชิงปริมาณ
X_{12}	แคมเปญ	เชิงปริมาณ
X_{13}	จำนวนวันหลังได้รับการติดต่อ	เชิงปริมาณ
X_{14}	จำนวนผู้ติดต่อก่อนแคมเปญ	เชิงปริมาณ
X_{15}	ผลลัพธ์ของแคมเปญ	เชิงคุณภาพ
X_{16}	อัตราการเปลี่ยนแปลงการจ้างงาน	เชิงปริมาณ
X_{17}	ดัชนีราคาผู้บริโภค	เชิงปริมาณ
X_{18}	ดัชนีความเชื่อมั่นผู้บริโภค	เชิงปริมาณ
X_{19}	อัตรา Euribor 3 เดือน	เชิงปริมาณ
X_{20}	จำนวนพนักงาน	เชิงปริมาณ
Y	ลูกค้าสมัครฝากประจำ (ใช่ = 0, ไม่ใช่ = 1)	เชิงคุณภาพ

ขั้นตอนที่ 2 ทำการแปลงข้อมูลของตัวแปรอิสระก่อนนำไปสร้างตัวแบบการจำแนก โดยแปลงค่าตัวแปรให้อยู่ในรูปปกติ (Normalization) ซึ่งมีค่าอยู่ในช่วง 0 ถึง 1 คำนวณได้ดังนี้

$$X^* = \frac{X - min(X)}{max(X) - min(X)} \quad (24)$$

โดยที่ X^* คือ ค่าที่ได้หลังการแปลงค่าให้อยู่ในรูปปกติ

X คือ ค่าข้อมูลที่ต้องการแปลง

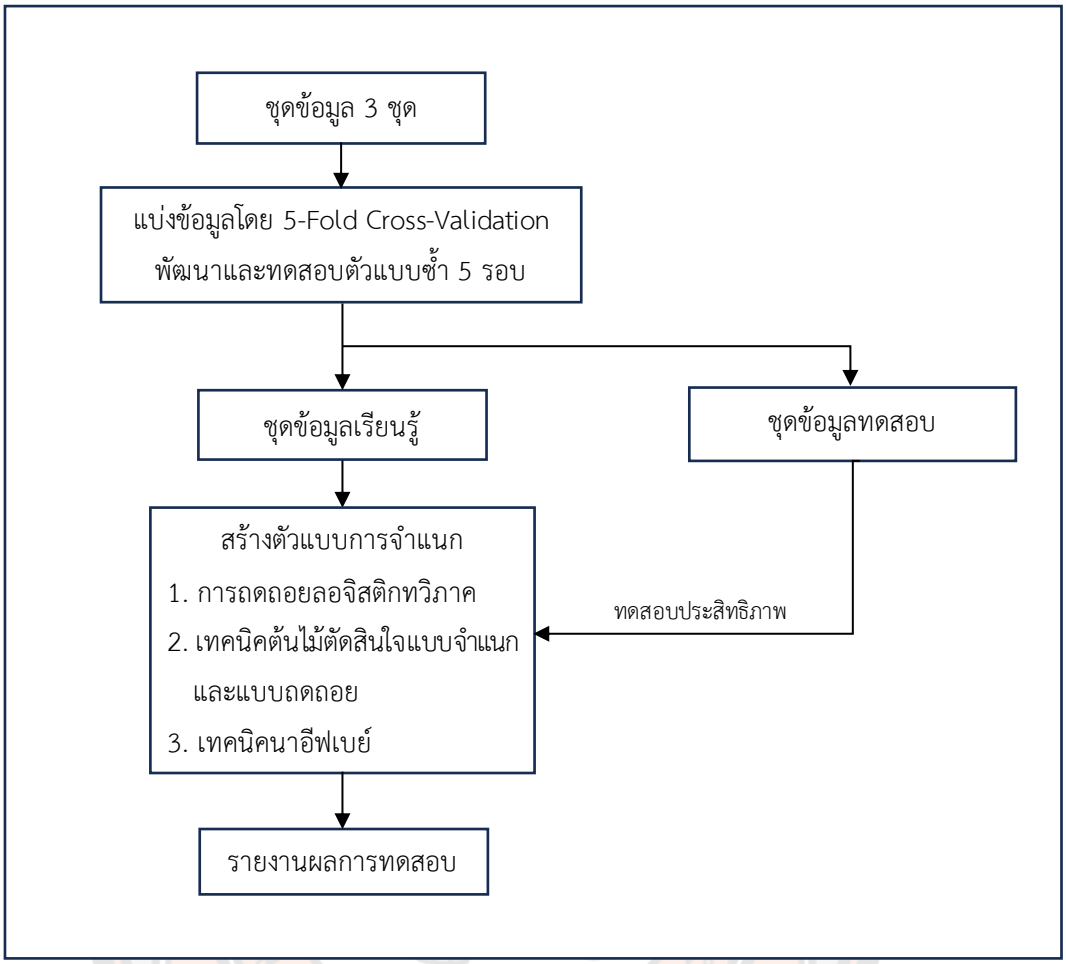
$min(X)$ คือ ค่าข้อมูลที่น้อยที่สุดในตัวแปร

$max(X)$ คือ ค่าข้อมูลที่มากที่สุดในตัวแปร

ขั้นตอนที่ 3 พัฒนาตัวแบบการจำแนกด้วยการเลือกใช้เทคนิคการจำแนก 3 เทคนิค ได้แก่ การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression) เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Classification and Regression Tree) และเทคนิคนาอีฟเบย์ (Naïve Bayes) ซึ่งมีรายละเอียดดังนี้

1. สร้างตัวแบบด้วยการถดถอยลอจิสติกทวิภาคโดยใช้ฟังก์ชัน glm()
2. สร้างตัวแบบด้วยต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยโดยใช้แพ็กเกจ rpart และฟังก์ชัน rpart() ซึ่งกำหนดพารามิเตอร์ที่สำคัญ ดังนี้
   - minsplit คือจำนวนขั้นต่ำของค่าสังเกตที่มีในโหนด กำหนดค่า minsplit เท่ากับ 20 (Deepika Singh, 2020)
   - minbucket คือจำนวนขั้นต่ำของค่าสังเกตที่โหนดสุดท้าย กำหนดค่า $\text{minbucket} = \frac{\text{minsplit}}{3}$ ในที่นี้เท่ากับ 7 (Deepika Singh, 2020)
3. สร้างตัวแบบด้วยเทคนิคนาอีฟเบย์โดยใช้แพ็กเกจ naivebayes และฟังก์ชัน naive_bayes()

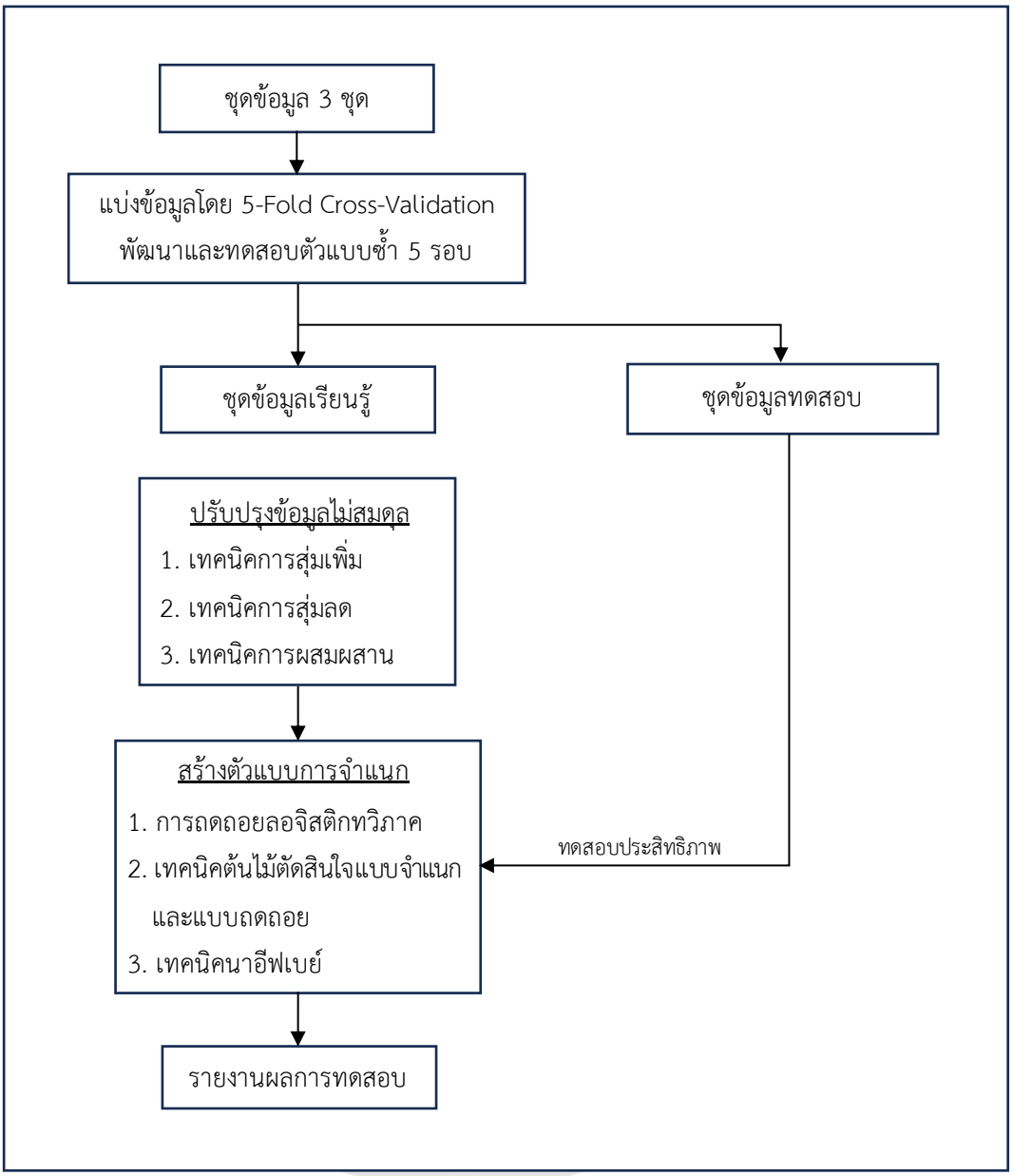
จากนั้นจะทำการแบ่งชุดข้อมูลเรียนรู้เป็น 80 เปอร์เซ็นต์ และชุดข้อมูลทดสอบเป็น 20 เปอร์เซ็นต์ จากข้อมูลทั้งหมด 100 เปอร์เซ็นต์ และนำหลักการ 5-Fold Cross-Validation มาพัฒนาตัวแบบการจำแนก โดยนำชุดข้อมูลเรียนรู้ (4 ชุดข้อมูล) มาพัฒนาตัวแบบการจำแนกและนำชุดข้อมูลทดสอบ (1 ชุดข้อมูลที่เหลือ) มาทดสอบประสิทธิภาพของตัวแบบการจำแนก ซึ่งจะทำการพัฒนาและทดสอบประสิทธิภาพของตัวแบบการจำแนกทั้งหมด 5 รอบ ซึ่งจะใช้ค่าความแม่นยำ (Accuracy) ค่าเรียกคืน (Recall) ค่าความเที่ยง (Precision) และค่าประสิทธิภาพโดยรวมถ่วงน้ำหนัก (F1-Score) เป็นเกณฑ์ในการวัดประสิทธิภาพของตัวแบบการจำแนก



ภาพ 7 ผังกระบวนการสร้างตัวแบบการจำแนกบนชุดข้อมูลตั้งต้น

ขั้นตอนที่ 4 ทำการปรับปรุงชุดข้อมูลไม่สมดุลให้สมดุล โดยใช้เทคนิคการสุ่ม 3 เทคนิค คือ เทคนิคการสุ่มเพิ่ม (Over sampling) เทคนิคการสุ่มลด (Under sampling) และเทคนิคการสุ่มผสมผสาน (Hybrid method) ซึ่งใช้แพ็คเกจ ROSE จากโปรแกรม RStudio ในการปรับปรุงชุดข้อมูลไม่สมดุลให้สมดุล โดยมีการใช้ฟังก์ชั่นในแต่ละเทคนิค ดังนี้

1. เทคนิคการสุ่มเพิ่ม ใช้ฟังก์ชัน ovun.sample() โดยใช้พารามิเตอร์ method ที่กำหนดค่าเป็น over
2. เทคนิคการสุ่มลด ใช้ฟังก์ชัน ovun.sample() โดยใช้พารามิเตอร์ method ที่กำหนดค่าเป็น under
3. เทคนิคการสุ่มผสมผสาน ใช้ฟังก์ชัน ovun.sample() โดยใช้พารามิเตอร์ method ที่กำหนดค่าเป็น both



ภาพ 8 ผังกระบวนการสร้างตัวแบบการจำแนกบนชุดข้อมูลที่ปรับปรุงความไม่สมดุล

ในการทำการปรับปรุงชุดข้อมูลไม่สมดุลให้สมดุลในแต่ละชุดนั้น มีรายละเอียดดังนี้

ชุดข้อมูลที่ 1 ชุดข้อมูลเครดิตเยอรมัน

จากข้อมูลจำนวน 1,000 แถว มีจำนวนตัวแปรตามเชิงคุณภาพ 1 ตัวซึ่งแบ่งออกเป็น 2 กลุ่ม ได้แก่ การจัดประเภทลูกค้าที่ดีจำนวน 700 แถว และการจัดประเภทลูกค้าที่ไม่ดีจำนวน 300 แถว พบว่าอัตราส่วนของข้อมูลการจัดประเภทลูกค้าที่ดีต่อข้อมูลการจัดประเภทลูกค้าที่ไม่ดีเท่ากับ 2.33 : 1 ทำให้เกิดความไม่สมดุลของข้อมูลอย่างมาก ดังนั้นเพื่อลดอัตราความไม่สมดุลของข้อมูล ผู้วิจัยใช้วิธีการสุ่มตัวอย่าง 3 เทคนิค ดังนี้

1. เทคนิคการสุ่มเพิ่ม (Over sampling)

ทำการสุ่มเพิ่มจำนวนข้อมูลของกลุ่ม 0 (240) ให้มีขนาดเท่ากับกลุ่ม 1 (560) จากการสุ่มเพิ่ม ดังนั้นจำนวนข้อมูลของกลุ่ม 0 (540) มีขนาดใกล้เคียงกับกลุ่ม 1 (560)

2. เทคนิคการสุ่มลด (Under sampling)

ทำการสุ่มลดจำนวนข้อมูลของกลุ่ม 1 (560) ให้มีขนาดเท่ากับกลุ่ม 0 (240) จากการสุ่มลด ดังนั้นจำนวนข้อมูลของกลุ่ม 1 (221) มีขนาดใกล้เคียงกับกลุ่ม 0 (240)

3. เทคนิคการสุ่มผสมผสาน (Hybrid method)

ทำการสุ่มผสมผสานจำนวนข้อมูลของกลุ่ม 1 (560) และกลุ่ม 0 (240) โดยทำการสุ่มลดข้อมูลจากกลุ่มส่วนมากและสุ่มเพิ่มข้อมูลในกลุ่มส่วนน้อย จากการสุ่มผสมผสาน ดังนั้นจำนวนข้อมูลของกลุ่ม 0 (389) มีขนาดใกล้เคียงกับกลุ่ม 1 (411)

ชุดข้อมูลที่ 2 ชุดข้อมูลลูกค้าบัตรเครดิต

จากข้อมูลจำนวน 30,000 แถว มีจำนวนตัวแปรตามเชิงคุณภาพ 1 ตัวซึ่งแบ่งออกเป็น 2 กลุ่มได้แก่ ลูกค้าไม่ผิดนัดชำระเดือนหน้าจำนวน 23,364 แถว และลูกค้าผิดนัดชำระเดือนหน้าจำนวน 6,636 แถว พบว่าอัตราส่วนของข้อมูลลูกค้าไม่ผิดนัดชำระเดือนหน้าต่อข้อมูลลูกค้าผิดนัดชำระเดือนหน้าเท่ากับ 3.55 : 1 ทำให้เกิดความไม่สมดุลของข้อมูลอย่างมาก ดังนั้นเพื่อลดอัตราความไม่สมดุลของข้อมูล ผู้วิจัยใช้วิธีการสุ่มตัวอย่าง 3 เทคนิค ดังนี้

1. เทคนิคการสุ่มเพิ่ม (Over sampling)

ทำการสุ่มเพิ่มจำนวนข้อมูลของกลุ่ม 0 (5,308) ให้มีขนาดเท่ากับกลุ่ม 1 (18,692) จากการสุ่มเพิ่ม ดังนั้นจำนวนข้อมูลของกลุ่ม 0 (18,308) มีขนาดใกล้เคียงกับกลุ่ม 1 (18,377)

2. เทคนิคการสุ่มลด (Under sampling)

ทำการสุ่มลดจำนวนข้อมูลของกลุ่ม 1 (18,692) ให้มีขนาดเท่ากับกลุ่ม 0 (5,308) จากการสุ่มลด ดังนั้นจำนวนข้อมูลของกลุ่ม 1 (5,248) มีขนาดใกล้เคียงกับกลุ่ม 0 (5,280)

3. เทคนิคการสุ่มผสมผสาน (Hybrid method)

ทำการสุ่มผสมผสานจำนวนข้อมูลของกลุ่ม 1 (18,692) และกลุ่ม 0 (5,308) โดยทำการสุ่มลดข้อมูลจากกลุ่มส่วนมากและสุ่มเพิ่มข้อมูลในกลุ่มส่วนน้อย จากการสุ่มผสมผสาน ดังนั้นจำนวนข้อมูลของกลุ่ม 0 (11,746) มีขนาดใกล้เคียงกับกลุ่ม 1 (11,911)

ชุดข้อมูลที่ 3 ชุดข้อมูลการตลาดของธนาคาร

จากข้อมูลจำนวน 41,188 แถว มีจำนวนตัวแปรตามเชิงคุณภาพ 1 ตัวซึ่งแบ่งออกเป็น 2 กลุ่มได้แก่ ลูกค้าสมัครฝากประจำจำนวน 36,548 แถว และลูกค้าไม่สมัครฝากประจำจำนวน 4,640 แถว พบว่าอัตราส่วนของข้อมูลลูกค้าสมัครฝากประจำต่อข้อมูลลูกค้าไม่สมัครฝากประจำเท่ากับ 8.09 : 1 ทำให้เกิดความไม่สมดุลของข้อมูลอย่างมาก ดังนั้นเพื่อลดอัตราความไม่สมดุลของข้อมูล ผู้วิจัยใช้วิธีการสุ่มตัวอย่าง 3 เทคนิค ดังนี้

1. เทคนิคการสุ่มเพิ่ม (Over sampling)

ทำการสุ่มเพิ่มจำนวนข้อมูลของกลุ่ม 0 (3,762) ให้มีขนาดเท่ากับกลุ่ม 1 (29,239) จากการสุ่มเพิ่ม ดังนั้นจำนวนข้อมูลของกลุ่ม 0 (29,305) มีขนาดใกล้เคียงกับกลุ่ม 1 (29,239)

2. เทคนิคการสุ่มลด (Under sampling)

ทำการสุ่มลดจำนวนข้อมูลของกลุ่ม 1 (29,239) ให้มีขนาดเท่ากับกลุ่ม 0 (3,762) จากการสุ่มลด ดังนั้นจำนวนข้อมูลของกลุ่ม 1 (3,730) มีขนาดใกล้เคียงกับกลุ่ม 0 (3,762)

3. เทคนิคการสุ่มผสมผสาน (Hybrid method)

ทำการสุ่มผสมผสานจำนวนข้อมูลของกลุ่ม 1 (29,239) และกลุ่ม 0 (3,762) โดยทำการสุ่มลดข้อมูลจากกลุ่มส่วนมากและสุ่มเพิ่มข้อมูลในกลุ่มส่วนน้อย จากการสุ่มผสมผสาน ดังนั้นจำนวนข้อมูลของกลุ่ม 0 (16,393) มีขนาดใกล้เคียงกับกลุ่ม 1 (16,608)

หลังจากทำการปรับปรุงชุดข้อมูลไม่สมดุลให้สมดุลแล้ว จากนั้นให้ทำซ้ำในขั้นตอนที่ 3 ซึ่งจะทำการพัฒนาและทดสอบประสิทธิภาพของตัวแบบการจำแนกกับชุดข้อมูลที่ทำการปรับปรุงชุดข้อมูลไม่สมดุลให้สมดุล และจะรายงานผลการทดสอบประสิทธิภาพของตัวแบบการจำแนกต่อไป

3.4 การแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ

งานวิจัยนี้มีการใช้ชุดข้อมูลด้านการเงินจำนวน 3 ชุด ได้แก่ ชุดข้อมูลเครดิตเยอรมัน ชุดข้อมูลลูกค้าบัตรเครดิต และชุดข้อมูลการตลาดของธนาคาร มาสร้างตัวแบบการจำแนก โดยจะทำการแบ่งข้อมูลออกเป็นชุดข้อมูลเรียนรู้ (Training set) และชุดข้อมูลทดสอบ (Test set) ซึ่งใช้การแบ่งข้อมูลด้วย k-Fold Cross-Validation ที่กำหนดค่า k เท่ากับ 5 (5-Fold Cross-Validation) โดยชุดข้อมูล 100% จะถูกแบ่งออกเป็น 5 กลุ่ม กลุ่มละ 20% จากนั้นในรอบที่ 1 จะให้ชุดข้อมูลที่ 1 เป็นชุดข้อมูลทดสอบ และชุดที่ 2-5 เป็นชุดข้อมูลเรียนรู้ จากนั้นทำการสร้างตัวแบบการจำแนกและทดสอบทั้งหมด 5 รอบดังภาพ 6 เมื่อทำการแบ่งข้อมูลเรียบร้อยแล้ว ดังนั้นชุดข้อมูลที่ได้จะเป็นชุดข้อมูลตั้งต้น ซึ่งมีจำนวนชุดข้อมูลทั้งหมด 3 ชุด

เนื่องจากชุดข้อมูลด้านการเงินทั้ง 3 ชุดที่นำมาใช้ เป็นชุดข้อมูลที่มีความไม่สมดุลสูงมาก ดังนั้นเพื่อลดอัตราความไม่สมดุลของข้อมูล ผู้วิจัยจึงใช้เทคนิคการสุ่ม 3 เทคนิค ได้แก่ เทคนิคการสุ่มเพิ่ม (Over sampling) เทคนิคการสุ่มลด (Under sampling) และเทคนิคการสุ่มผสมผสาน (Hybrid method) เมื่อทำการสุ่มข้อมูลจากชุดข้อมูลทั้ง 3 ชุด โดยใช้เทคนิคการสุ่มดังกล่าว ดังนั้นชุดข้อมูลที่ได้จะเป็นชุดข้อมูลจากการสุ่มเพิ่ม การสุ่มลด และการสุ่มผสมผสาน ซึ่งมีจำนวนชุดข้อมูลทั้งหมด 9 ชุด จากนั้นนำชุดข้อมูลจำนวน 12 ชุดที่ได้ นำไปสร้างตัวแบบการจำแนกต่อไป

3.5 ขั้นตอนการพัฒนาตัวแบบการจำแนก

ในหัวข้อนี้ผู้วิจัยจะนำชุดข้อมูลจำนวน 12 ชุด มาสร้างตัวแบบการจำแนก โดยพารามิเตอร์ของการสร้างตัวแบบการจำแนกด้วย 3 เทคนิค มีดังนี้

3.5.1 การถดถอยลอจิสติกทวิภาคโดยใช้ฟังก์ชัน glm()

3.5.2 สร้างตัวแบบด้วยต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยโดยใช้แพ็กเกจ rpart และฟังก์ชัน rpart() ซึ่งกำหนดพารามิเตอร์ที่สำคัญ ดังนี้

1) minsplit คือจำนวนขั้นต่ำของค่าสังเกตที่มีในโหนด กำหนดค่า minsplit เท่ากับ 20 (Deepika Singh, 2020)

2) minbucket คือจำนวนขั้นต่ำของค่าสังเกตที่โหนดสุดท้าย กำหนดค่า $\mathbf{minbucket} = \frac{\mathbf{minsplit}}{3}$ ในที่นี้เท่ากับ 7 (Deepika Singh, 2020)

3.5.3 สร้างตัวแบบด้วยเทคนิคนาอีฟเบย์โดยใช้แพ็กเกจ naivebayes และฟังก์ชัน naive_bayes()

3.6 การประเมินประสิทธิภาพตัวแบบการจำแนก

การวัดประสิทธิภาพเป็นเครื่องมือตรวจสอบประสิทธิภาพของตัวแบบการจำแนกโดยเกณฑ์ทั้ง 4 เกณฑ์ที่เลือกใช้จะอ้างอิงการคำนวณจากค่าที่ได้จากเมทริกซ์ความสับสน (Confusion Matrix) ดังนี้

1. ค่าความแม่นยำ (Accuracy)
2. ค่าเรียกคืน (Recall)
3. ค่าความเที่ยง (Precision)
4. ค่าประสิทธิภาพโดยรวมถ่วงน้ำหนัก (F1-Score)

บทที่ 4
ผลการวิจัย

ในบทนี้ผู้วิจัยจะนำเสนอผลการวิจัยที่ได้จากการพัฒนาตัวแบบการจำแนกของชุดข้อมูล ด้านการเงิน โดยใช้เทคนิคการจำแนกทั้งหมด 3 เทคนิค ได้แก่ การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression) ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Classification and Regression Tree) และเทคนิคนาอีฟเบย์ (Naïve Bayes) และใช้ชุดข้อมูลด้านการเงินที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณแตกต่างกัน 3 ชุด ได้แก่ ชุดข้อมูลเครดิตเยอรมัน ชุดข้อมูลลูกค้าบัตรเครดิต และชุดข้อมูลการตลาดของธนาคาร จากนั้นนำชุดข้อมูลมาสร้างตัวแบบการจำแนก โดยแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบด้วยวิธี 5-Fold Cross-Validation เนื่องจากชุดข้อมูลทั้ง 3 ชุด มีความไม่สมดุลอยู่มาก จึงมีการปรับปรุงชุดข้อมูลที่มีความไม่สมดุลให้สมดุลโดยใช้เทคนิคการสุ่ม 3 เทคนิค คือ การสุ่มเพิ่ม (Over sampling) การสุ่มลด (Under sampling) และการสุ่มผสมผสาน (Hybrid method) โดยมีเกณฑ์ที่ใช้วัดประสิทธิภาพตัวแบบการจำแนก ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และค่าประสิทธิภาพโดยรวม (F1-Score) ซึ่งได้มีการกำหนดสัญลักษณ์ที่ใช้ในการวิเคราะห์ ดังนี้

Acc	แทน	ค่าความแม่นยำ
CART	แทน	ตัวแบบการจำแนกของเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย
FN	แทน	ค่าความผิดพลาดเชิงลบ
FP	แทน	ค่าความผิดพลาดเชิงบวก
F1	แทน	ค่าประสิทธิภาพโดยรวม
LR	แทน	ตัวแบบการจำแนกการถดถอยลอจิสติกทวิภาค
NB	แทน	ตัวแบบการจำแนกของเทคนิคนาอีฟเบย์
Pre	แทน	ค่าความเที่ยง
Re	แทน	ค่าการเรียกคืน
TN	แทน	ค่าความถูกต้องเชิงลบ
TP	แทน	ค่าความถูกต้องเชิงบวก

โดยการนำเสนอผลการวิจัยครั้งนี้ จะแบ่งการนำเสนอผลการวิเคราะห์ข้อมูลออกเป็น 2 หัวข้อหลัก ได้แก่

1. ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงินที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกันทั้งหมด 3 แบบ

2. ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงินภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม

4.1 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงินที่มีจำนวนตัวแปรอิสระเชิงคุณภาพและจำนวนตัวแปรอิสระเชิงปริมาณที่แตกต่างกันทั้งหมด 3 แบบ

ในหัวข้อนี้ผู้วิจัยจะนำเสนอผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงินทั้งหมด 3 ชุด ได้แก่ ชุดข้อมูลเครดิตเยอรมัน ชุดข้อมูลเริ่มต้นของลูกค้าบัตรเครดิต และชุดข้อมูลการตลาดของธนาคาร ซึ่งแต่ละชุดข้อมูลจะมีจำนวนตัวแปรอิสระเชิงคุณภาพและเชิงปริมาณที่แตกต่างกันทั้งหมด 3 แบบ ได้แก่ ชุดข้อมูลเครดิตเยอรมัน (ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ) ชุดข้อมูลลูกค้าบัตรเครดิต (ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ) และชุดข้อมูลการตลาดของธนาคาร (ชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ) ซึ่งมีการพัฒนาตัวแบบการจำแนกโดยใช้การถดถอยลอจิสติกทวิภาค เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย และเทคนิคนาอีฟเบย์ ซึ่งจะแสดงผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลทั้ง 3 ชุด ดังต่อไปนี้

4.1.1 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมัน

ในการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกของเทคนิคการจำแนกทั้ง 3 เทคนิคบนชุดข้อมูลเครดิตเยอรมันที่เป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณนั้นจะแสดงผลลัพธ์ดังตารางที่ 10

ตาราง 10 แสดงผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมันภายใต้ชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ

เทคนิคการจำแนก	ชุดข้อมูลเรียนรู้				ชุดข้อมูลทดสอบ			
	Pre (%)	Re (%)	F1 (%)	Acc (%)	Pre (%)	Re (%)	F1 (%)	Acc (%)
LR	80.53	89.21	76.30	77.35	79.69	88.29	74.82	76.00

CART	83.31	93.25	81.21	82.18	77.69	86.29	71.60	73.00
NB	80.74	84.25	74.63	75.03	80.07	83.29	73.41	73.80

จากตารางที่ 10 เมื่อพิจารณาผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมันที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ ภายใต้ชุดข้อมูลเรียนรู้ พบว่า เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพดีที่สุดด้วยค่าความแม่นยำ (Acc) เท่ากับ 82.18% รองลงมาคือการถดถอยลอจิสติกทวิภาค ซึ่งมีค่าความแม่นยำ (Acc) เท่ากับ 77.35% และเทคนิคนาอีฟเบย์ที่มีค่าความแม่นยำ (Acc) น้อยที่สุด ซึ่งมีค่าเท่ากับ 75.03%

ในขณะที่ภายใต้ชุดข้อมูลทดสอบ พบว่า การถดถอยลอจิสติกทวิภาคมีประสิทธิภาพดีที่สุดด้วยค่าความแม่นยำ (Acc) เท่ากับ 76.00% รองลงมาคือเทคนิคนาอีฟเบย์ ซึ่งมีค่าความแม่นยำ (Acc) เท่ากับ 73.80% และเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยที่มีค่าความแม่นยำ (Acc) น้อยที่สุด ซึ่งมีค่าเท่ากับ 73.00%

4.1.2 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลลูกค้าบัตรเครดิต

ในการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกของเทคนิคการจำแนกทั้ง 3 เทคนิคบนชุดข้อมูลลูกค้าบัตรเครดิตที่เป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณนั้น จะแสดงผลลัพธ์ดังตารางที่ 11

ตาราง 11 แสดงผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลลูกค้าบัตรเครดิตภายใต้ชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ

เทคนิคการจำแนก	**ชุดข้อมูลเรียนรู้**				**ชุดข้อมูลทดสอบ**			
	Pre (%)	**Re (%)**	**F1 (%)**	**Acc (%)**	**Pre (%)**	**Re (%)**	**F1 (%)**	**Acc (%)**
LR	82.46	95.33	77.59	80.57	82.42	95.26	77.48	80.50
CART	83.41	95.92	79.35	81.96	83.40	95.92	79.37	81.99
NB	87.18	78.73	75.63	74.42	87.22	78.71	75.62	74.44

จากตารางที่ 11 เมื่อพิจารณาผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกในชุดข้อมูลลูกค้าบัตรเครดิตที่มีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ ภายใต้ชุดข้อมูลเรียนรู้

พบว่า เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพดีที่สุดด้วยค่าความแม่นยำ (Acc) เท่ากับ 81.96% รองลงมาคือการถดถอยลอจิสติกทวิภาค ซึ่งมีค่าความแม่นยำ (Acc) เท่ากับ 80.57% และเทคนิคนาอีฟเบย์ที่มีค่าความแม่นยำ (Acc) น้อยที่สุด ซึ่งมีค่าเท่ากับ 74.42%

ในขณะที่ภายใต้ชุดข้อมูลทดสอบ พบว่า เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพดีที่สุดด้วยค่าความแม่นยำ (Acc) เท่ากับ 81.99% รองลงมาคือการถดถอยลอจิสติกทวิภาค ซึ่งมีค่าความแม่นยำ (Acc) เท่ากับ 80.50% และเทคนิคนาอีฟเบย์ที่มีค่าความแม่นยำ (Acc) น้อยที่สุด ซึ่งมีค่าเท่ากับ 74.44%

4.1.3 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลการตลาดของธนาคาร

ในการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกของเทคนิคการจำแนกทั้ง 3 เทคนิคบนชุดข้อมูลการตลาดของธนาคารที่เป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ จะแสดงผลลัพธ์ดังตารางที่ 12

ตาราง 12 แสดงผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลการตลาดของธนาคารภายใต้ชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ

เทคนิคการจำแนก	ชุดข้อมูลเรียนรู้				ชุดข้อมูลทดสอบ			
	Pre (%)	Re (%)	F1 (%)	Acc (%)	Pre (%)	Re (%)	F1 (%)	Acc (%)
LR	93.46	97.56	90.99	91.78	91.30	91.07	82.96	83.93
CART	94.88	97.00	92.35	92.68	84.03	53.77	N/A*	50.41
NB	94.82	88.59	86.79	85.57	94.28	74.71	69.63	72.08

* มี 1 ชุดทดสอบที่ TN (True Negative) เท่ากับ 0

จากตารางที่ 12 เมื่อพิจารณาผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกในชุดข้อมูลการตลาดของธนาคารที่มีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณภายใต้ชุดข้อมูลเรียนรู้ พบว่า เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพดีที่สุดด้วยค่าความแม่นยำ (Acc) เท่ากับ 92.68% รองลงมาคือการถดถอยลอจิสติกทวิภาค ซึ่งมีค่าความแม่นยำ (Acc) เท่ากับ 91.78% และเทคนิคนาอีฟเบย์ที่มีค่าความแม่นยำ (Acc) น้อยที่สุด ซึ่งมีค่าเท่ากับ 85.57%

ในขณะที่ภายใต้ชุดข้อมูลทดสอบ พบว่า การถดถอยลอจิสติกทวิภาคมีประสิทธิภาพดีที่สุด ด้วยค่าความแม่นยำ (Acc) เท่ากับ 83.93% รองลงมาคือเทคนิคนาอีฟเบย์ ซึ่งมีค่าความแม่นยำ (Acc) เท่ากับ 72.08% และเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยที่มีค่าความแม่นยำ (Acc) น้อยที่สุด ซึ่งมีค่าเท่ากับ 50.41%

ตาราง 13 แสดงผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมัน ชุดข้อมูลลูกค้าบัตรเครดิต และชุดข้อมูลการตลาดของธนาคาร ภายใต้ชุดข้อมูลทดสอบ

เทคนิคการจำแนก	ชุดข้อมูลเครดิตเยอรมัน				ชุดข้อมูลลูกค้าบัตรเครดิต				ชุดข้อมูลการตลาดของธนาคาร			
	Pre (%)	Re (%)	F1 (%)	Acc (%)	Pre (%)	Re (%)	F1 (%)	Acc (%)	Pre (%)	Re (%)	F1 (%)	Acc (%)
LR	79.69	88.29	74.82	76.00	82.42	95.26	77.48	80.50	91.30	91.07	82.96	83.93
CART	77.69	86.29	71.60	73.00	83.40	95.92	79.37	81.99	84.03	53.77	N/A*	50.41
NB	80.07	83.29	73.41	73.80	87.22	78.71	75.62	74.44	94.28	74.71	69.63	72.08

* มี 1 ชุดทดสอบที่ TN (True Negative) เท่ากับ 0

จากตารางที่ 13 เป็นการพิจารณาประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลทดสอบ พบว่า การถดถอยลอจิสติกทวิภาคมีประสิทธิภาพดีที่สุดบนชุดข้อมูลเครดิตเยอรมัน ซึ่งมีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ และชุดข้อมูลการตลาดของธนาคาร ซึ่งมีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ

ในขณะที่เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพดีที่สุดบนชุดข้อมูลลูกค้าบัตรเครดิต ซึ่งมีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ

ตาราง 14 แสดงผลการจำแนกประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมัน ชุดข้อมูลลูกค้าบัตรเครดิต และชุดข้อมูลการตลาดของธนาคาร ภายใต้ชุดข้อมูลทดสอบ

ชุดข้อมูล	เทคนิคการจำแนก	Y=0			Y=1			Acc (%)
		Pre (%)	Re (%)	F1 (%)	Pre (%)	Re (%)	F1 (%)	
เครดิตเยอรมัน (1:2.33)	LR	63.65	47.33	54.02	79.69	88.29	83.74	76.00
	CART	57.15	42.00	48.01	77.69	86.29	81.71	73.00
	NB	57.83	51.67	54.31	80.07	83.29	81.59	73.80
ลูกค้าบัตรเครดิต	LR	63.00	28.40	39.13	82.42	95.26	88.38	80.50

ชุดข้อมูล	เทคนิคการจำแนก	Y=0			Y=1			Acc (%)
		Pre (%)	Re (%)	F1 (%)	Pre (%)	Re (%)	F1 (%)	
(1:3.55)	CART	69.80	33.20	44.97	83.40	95.92	89.22	81.99
	NB	44.11	59.18	50.53	87.22	78.71	82.74	74.44
การตลาดของ	LR	42.47	27.68	23.99	91.30	91.07	90.44	83.93
ธนาคาร	CART	23.51	23.37	N/A*	84.03	53.77	58.25	50.41
(1:8.09)	NB	57.44	51.20	32.77	94.28	74.71	74.31	72.08

* มี 1 ชุดทดสอบที่ TN (True Negative) เท่ากับ 0

จากตารางที่ 14 แสดงประสิทธิภาพการจำแนกของค่าคำตอบของกลุ่ม Y=0 และ Y=1 ซึ่งแสดงด้วยค่าเรียกคืน (Re) บนชุดข้อมูลทั้ง 3 ชุด พบว่า แม้การถดถอยลอจิสติกทวิภาคและเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยจะให้ค่าความแม่นยำที่สูงที่สุด แต่เมื่อพิจารณาค่าเรียกคืน (Re) เมื่อ Y=0 พบว่า ตัวแบบการจำแนกจากเทคนิคทั้งสองมีค่าเรียกคืนที่ต่ำ แต่จะมีค่าเรียกคืนที่สูงเมื่อ Y=1 ทั้งนี้เพราะ คำตอบ Y=1 มีจำนวนมากกว่า Y=0 ทำให้การทำนายข้อมูลของกลุ่มส่วนน้อยมีประสิทธิภาพที่ต่ำ โดยปัญหาเหล่านี้เกิดจากการที่ข้อมูลมีความไม่สมดุลอยู่มาก ซึ่งอาจทำให้ค่าประสิทธิภาพการจำแนกกลุ่มน้อยมีประสิทธิภาพที่ต่ำ ทำให้ผู้วิจัยได้นำวิธีการปรับปรุงความไม่สมดุลมาใช้ในการพัฒนาตัวแบบการจำแนก

ซึ่งจากตารางที่ 14 พบว่า เทคนิคนาอีฟเบย์สามารถจำแนกคำตอบในกลุ่ม Y=0 หรือกลุ่มส่วนน้อยได้ดีกว่าการถดถอยลอจิสติกทวิภาค และเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

4.2 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลด้านการเงินภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุล

จากตารางที่ 14 จะเห็นว่าปัญหาของข้อมูลไม่สมดุล อาจส่งผลต่อค่าเกณฑ์การวัดประสิทธิภาพ เนื่องจากตัวแบบการจำแนกมักจะทำนายคำตอบเป็นข้อมูลกลุ่มส่วนมาก และทำให้ประสิทธิภาพในการทำนายคำตอบเป็นข้อมูลกลุ่มส่วนน้อยค่อนข้างต่ำ โดยค่าความแม่นยำมักจะขึ้นอยู่กับประสิทธิภาพการจำแนกในการทำนายข้อมูลกลุ่มส่วนมาก

ในหัวข้อนี้หลังจากนำเทคนิคการปรับปรุงข้อมูลให้สมดุลมาใช้ ผู้วิจัยจึงนำเสนอผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุง

ความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค คือ การสุ่มเพิ่ม การสุ่มลด และการสุ่มผสมผสาน บนชุดข้อมูลด้านการเงินในชุดข้อมูลทดสอบ ทั้งหมด 3 ชุด ซึ่งมีรายละเอียดดังหัวข้อต่อไปนี้

4.2.1 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมันภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุล

ตาราง 15 แสดงผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมันภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค (ชุดข้อมูลทดสอบ)

เทคนิคการจำแนก	ปรับปรุงความไม่สมดุล	Y=0			Y=1			Acc (%)	Change*
		Pre (%)	Re (%)	F1 (%)	Pre (%)	Re (%)	F1 (%)		
LR	ตั้งต้น	63.65	47.33	54.02	79.69	88.29	83.74	76.00	-
	สุ่มเพิ่ม	51.42	68.67	58.77	84.33	72.14	77.74	71.10	-0.0645
	สุ่มลด	49.69	72.67	58.99	85.41	68.43	75.96	69.70	-0.0829
	สุ่มผสมผสาน	50.15	69.33	58.14	84.25	70.29	76.59	70.00	-0.0789
CART	ตั้งต้น	57.15	42.00	48.01	77.69	86.29	81.71	73.00	-
	สุ่มเพิ่ม	51.10	57.33	53.53	80.76	76.00	78.10	70.40	-0.0356
	สุ่มลด	49.01	69.67	56.84	82.77	65.77	72.89	66.78	-0.0852
	สุ่มผสมผสาน	45.37	68.67	54.58	82.56	64.14	72.14	65.50	-0.1027
NB	ตั้งต้น	57.83	51.67	54.31	80.07	83.29	81.59	73.80	-
	สุ่มเพิ่ม	51.94	68.00	58.73	84.18	72.71	77.93	71.30	-0.0339
	สุ่มลด	50.96	66.67	57.64	83.45	72.14	77.32	70.50	-0.0447
	สุ่มผสมผสาน	50.56	67.00	57.56	83.43	71.57	77.00	70.20	-0.0488

$$* \text{ Change} = \frac{Acc_{\text{ชุดข้อมูลตั้งต้น}} - Acc_{\text{ชุดข้อมูลปรับปรุงความไม่สมดุล}}}{Acc_{\text{ชุดข้อมูลตั้งต้น}}}$$

จากตารางที่ 15 พบว่า

1. เมื่อพิจารณาการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมันภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค จะเห็นว่า การนำเทคนิคการสุ่มเพิ่ม การสุ่มลด และการสุ่มผสมผสานมาใช้นั้น ไม่ได้ทำให้ค่าความแม่นยำ (Acc) ของตัวแบบการจำแนกใดมีค่าดีขึ้น

2. เมื่อพิจารณาจากค่าเรียกคืน (Re) พบว่า การนำเทคนิคการสุ่ม 3 เทคนิคมาใช้ จะทำให้มีการทำนายคำตอบในกลุ่ม Y=0 หรือกลุ่มส่วนน้อยมีค่าความถูกต้องที่มากขึ้น

3. เมื่อพิจารณาจากอัตราการเปลี่ยนแปลงของค่าความแม่นยำพบว่า การปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่มมีประสิทธิภาพดีกว่าเทคนิคการสุ่มลด และเทคนิคการสุ่มผสมผสาน โดยมีอัตราการเปลี่ยนแปลงของค่าความแม่นยำในชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่มเมื่อเปรียบเทียบกับชุดข้อมูลตั้งต้นของตัวแบบการจำแนกด้วยการถดถอยลอจิสติกทวิภาค เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย และเทคนิคนาอีฟเบย์เท่ากับ -0.0645%, -0.0356% และ -0.0339% ตามลำดับ

4.2.2 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลลูกค้าบัตรเครดิตภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุล

ตาราง 16 แสดงผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลลูกค้าบัตรเครดิตภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค (ชุดข้อมูลทดสอบ)

เทคนิคการจำแนก	ปรับปรุงความไม่สมดุล	Y=0			Y=1			Acc (%)	Change*
		Pre (%)	Re (%)	F1 (%)	Pre (%)	Re (%)	F1 (%)		
LR	ตั้งต้น	63.00	28.40	39.13	82.42	95.26	88.38	80.50	-
	สุ่มเพิ่ม	47.94	57.74	52.36	87.28	82.14	84.63	76.78	-0.0462
	สุ่มลด	47.82	57.78	52.31	87.28	82.05	84.58	76.72	-0.0470
	สุ่มผสมผสาน	47.66	57.79	52.21	87.27	81.92	84.50	76.62	-0.0482
CART	ตั้งต้น	69.80	33.20	44.97	83.40	95.92	89.22	81.99	-
	สุ่มเพิ่ม	46.93	57.88	51.82	87.21	81.40	84.20	76.23	-0.0703
	สุ่มลด	46.93	57.88	51.82	87.21	81.40	84.20	76.23	-0.0703
	สุ่มผสมผสาน	46.93	57.88	51.82	87.21	81.40	84.20	76.23	-0.0703
NB	ตั้งต้น	44.11	59.18	50.53	87.22	78.71	82.74	74.44	-
	สุ่มเพิ่ม	42.63	63.06	50.87	87.89	75.90	81.45	73.10	-0.0180
	สุ่มลด	43.19	61.90	50.86	87.71	76.85	81.91	73.59	-0.0114
	สุ่มผสมผสาน	42.45	62.66	50.60	87.78	75.88	81.39	72.99	-0.0195

* $\text{Change} = \frac{Acc_{\text{ชุดข้อมูลตั้งต้น}} - Acc_{\text{ชุดข้อมูลปรับปรุงความไม่สมดุล}}}{Acc_{\text{ชุดข้อมูลตั้งต้น}}}$

จากตารางที่ 16 พบว่า

1. เมื่อพิจารณาการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลลูกค้าบัตรเครดิตภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค จะเห็นว่า การนำเทคนิคการสุ่มเพิ่ม การสุ่มลด และการสุ่มผสมผสานมาใช้นั้น ไม่ได้ทำให้ค่าความแม่นยำ (Acc) ของตัวแบบการจำแนกใดมีค่าดีขึ้น

2. เมื่อพิจารณาจากค่าเรียกคืน (Re) พบว่า การนำเทคนิคการสุ่ม 3 เทคนิคมาใช้ จะทำให้มีการทำนายคำตอบในกลุ่ม Y=0 หรือกลุ่มส่วนน้อยมีค่าความถูกต้องที่มากขึ้น

3. เมื่อพิจารณาจากอัตราการเปลี่ยนแปลงของค่าความแม่นยำพบว่า ตัวแบบการจำแนกด้วยการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่มมีประสิทธิภาพดีที่สุด โดยมีอัตราการเปลี่ยนแปลงของค่าความแม่นยำในชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่มเมื่อเปรียบเทียบกับชุดข้อมูลตั้งต้นของตัวแบบการจำแนกด้วยการถดถอยลอจิสติกทวิภาคเท่ากับ -0.0462% ส่วนตัวแบบการจำแนกด้วยเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มทั้ง 3 เทคนิคมีประสิทธิภาพดีเหมือนกันโดยมีอัตราการเปลี่ยนแปลงของค่าความแม่นยำในชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิคเมื่อเปรียบเทียบกับชุดข้อมูลตั้งต้นของตัวแบบการจำแนกด้วยเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยเท่ากับ -0.0703% และตัวแบบการจำแนกด้วยเทคนิคนาอีฟเบย์บนชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลดมีประสิทธิภาพดีที่สุด โดยมีอัตราการเปลี่ยนแปลงของค่าความแม่นยำในชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลดเมื่อเปรียบเทียบกับชุดข้อมูลตั้งต้นของตัวแบบการจำแนกด้วยเทคนิคนาอีฟเบย์เท่ากับ -0.0114%

4.2.3 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลการตลาดของธนาคารภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุล

ตาราง 17 แสดงผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลการตลาดของธนาคารภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค (ชุดข้อมูลทดสอบ)

เทคนิคการจำแนก	ปรับปรุงความไม่สมดุล	Y=0			Y=1			Acc (%)	Change*
		Pre (%)	Re (%)	F1 (%)	Pre (%)	Re (%)	F1 (%)		
LR	ตั้งต้น	42.47	27.68	23.99	91.30	91.07	90.44	83.93	-
	สุ่มเพิ่ม	31.09	69.52	41.00	95.21	74.25	81.96	73.70	-0.1219
	สุ่มลด	30.81	69.66	40.71	95.27	73.88	81.68	73.38	-0.1257
	สุ่มผสมผสาน	30.90	69.91	40.74	95.34	73.43	81.25	73.02	-0.1300
CART	ตั้งต้น	23.97	25.24	7.18	84.03	53.88	58.35	50.65	-
	สุ่มเพิ่ม	33.22	38.85	17.49	88.90	60.71	61.90	58.30	0.1510
	สุ่มลด	30.06	34.80	13.86	88.42	60.73	61.69	57.83	0.1418
	สุ่มผสมผสาน	53.36	24.41	20.22	69.45	80.99	62.34	59.02	0.1653
NB	ตั้งต้น	57.44	51.20	32.77	94.28	74.71	74.31	72.08	-
	สุ่มเพิ่ม	55.92	60.40	37.62	95.37	69.81	71.05	68.77	-0.0459
	สุ่มลด	55.97	61.32	38.49	95.47	69.49	70.85	68.60	-0.0483
	สุ่มผสมผสาน	55.90	60.57	37.73	79.11	65.65	55.20	53.88	-0.2525

$$* \text{ Change} = \frac{Acc_{\text{ชุดข้อมูลตั้งต้น}} - Acc_{\text{ชุดข้อมูลปรับปรุงความไม่สมดุล}}}{Acc_{\text{ชุดข้อมูลตั้งต้น}}}$$

จากตารางที่ 17 พบว่า

1. เมื่อพิจารณาการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลการตลาดของธนาคารภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค จะเห็นว่า การนำเทคนิคการสุ่มเพิ่ม การสุ่มลด และการสุ่มผสมผสานมาใช้นั้น ไม่ได้ทำให้ค่าความแม่นยำ (Acc) ของตัวแบบการจำแนกใดมีค่าดีขึ้น

2. เมื่อพิจารณาจากค่าเรียกคืน (Re) พบว่า การนำเทคนิคการสุ่ม 3 เทคนิคมาใช้ จะทำให้มีการทำนายคำตอบในกลุ่ม Y=0 หรือกลุ่มส่วนน้อยมีค่าความถูกต้องที่มากขึ้น

3. เมื่อพิจารณาจากอัตราการเปลี่ยนแปลงของค่าความแม่นยำพบว่า ตัวแบบการจำแนกด้วยการถดถอยลอจิสติกทวิภาค และเทคนิคนาอีฟเบย์บนชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วย

เทคนิคการสุ่มเพิ่มมีประสิทธิภาพดีที่สุด โดยมีอัตราการเปลี่ยนแปลงของค่าความแม่นยำในชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่มเมื่อเปรียบเทียบกับชุดข้อมูลตั้งต้นของตัวแบบการจำแนกด้วยการถดถอยลอจิสติกทวิภาค และเทคนิคนาอีฟเบย์เท่ากับ -0.1219% และ -0.0459% ตามลำดับ ส่วนตัวแบบการจำแนกด้วยเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสานมีประสิทธิภาพดีที่สุด โดยมีอัตราการเปลี่ยนแปลงของค่าความแม่นยำในชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสานเมื่อเปรียบเทียบกับชุดข้อมูลตั้งต้นของตัวแบบการจำแนกด้วยเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยเท่ากับ 0.1653%

ตาราง 18 แสดงผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมัน ชุดข้อมูลลูกค้าบัตรเครดิต และชุดข้อมูลการตลาดของธนาคารภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มทั้ง 3 เทคนิค (ชุดข้อมูลทดสอบ)

เทคนิคการจำแนก	**ปรับปรุงความไม่สมดุล**	**ชุดข้อมูลเครดิตเยอรมัน**		**ชุดข้อมูลเริ่มต้นของลูกค้า**		**ชุดข้อมูลเครดิตเยอรมัน**	
		Acc (%)	**Change***	**Acc (%)**	**Change***	**Acc (%)**	**Change***
LR	ตั้งต้น	76.00	-	80.50	-	83.93	-
	สุ่มเพิ่ม	71.10	-0.0645	76.78	-0.0462	73.70	-0.1219
	สุ่มลด	69.70	-0.0829	76.72	-0.0470	73.38	-0.1257
	สุ่มผสมผสาน	70.00	-0.0789	76.62	-0.0482	73.02	-0.1300
CART	ตั้งต้น	73.00	-	81.99	-	50.65	-
	สุ่มเพิ่ม	70.40	-0.0356	76.23	-0.0703	58.30	0.1510
	สุ่มลด	66.78	-0.0852	76.23	-0.0703	57.83	0.1418
	สุ่มผสมผสาน	65.50	-0.1027	76.23	-0.0703	59.02	0.1653
NB	ตั้งต้น	73.80	-	74.44	-	72.08	-
	สุ่มเพิ่ม	71.30	-0.0339	73.10	-0.0180	68.77	-0.0459
	สุ่มลด	70.50	-0.0447	73.59	-0.0114	68.60	-0.0483
	สุ่มผสมผสาน	70.20	-0.0488	72.99	-0.0195	83.93	-

* $\text{Change} = \frac{Acc_{\text{ชุดข้อมูลตั้งต้น}} - Acc_{\text{ชุดข้อมูลปรับปรุงความไม่สมดุล}}}{Acc_{\text{ชุดข้อมูลตั้งต้น}}}$

จากตารางที่ 18 เมื่อพิจารณาจากอัตราการเปลี่ยนแปลงของค่าความแม่นยำในชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มทั้ง 3 เทคนิคเมื่อเปรียบเทียบกับชุดข้อมูลตั้งต้นของตัวแบบการจำแนกด้วยการถดถอยลอจิสติกทวิภาค เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย และเทคนิคนาอีฟเบย์ พบว่าการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่มมีประสิทธิภาพดีกว่าเทคนิคการสุ่มลด และเทคนิคการสุ่มผสมผสาน

บทที่ 5
สรุปผลการวิจัย

งานวิจัยเรื่องการเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกกับข้อมูลด้านการเงิน มีวัตถุประสงค์เพื่อศึกษากระบวนการทำงานและเปรียบเทียบประสิทธิภาพตัวแบบการจำแนกด้วยเทคนิคการจำแนก 3 เทคนิค ได้แก่ การถดถอยลอจิสติกทวิภาค เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย และเทคนิคนาอีฟเบย์ โดยใช้ชุดข้อมูลด้านการเงินที่มีจำนวนของตัวแปรอิสระเชิงคุณภาพและจำนวนของตัวแปรอิสระเชิงปริมาณแตกต่างกัน 3 แบบ ได้แก่ ชุดข้อมูลเครดิตเยอรมันที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ ชุดข้อมูลลูกค้าบัตรเครดิตที่มีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ และชุดข้อมูลการตลาดของธนาคารที่มีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ โดยศึกษาภายใต้ชุดข้อมูลตั้งต้นและชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค ได้แก่ การสุ่มเพิ่ม การสุ่มลด และการผสมผสาน จากนั้นทดสอบประสิทธิภาพด้วยหลักการ 5–Fold Cross-Validation โดยมีเกณฑ์ที่ใช้วัดประสิทธิภาพของตัวแบบการจำแนก คือ ค่าความแม่นยำ ค่าความเที่ยง ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม

ผลที่ได้จากการศึกษาสามารถสรุปได้ดังนี้

1. การเปรียบเทียบประสิทธิภาพบนชุดข้อมูลตั้งต้นภายใต้ชุดข้อมูลทดสอบ พบว่า การถดถอยลอจิสติกทวิภาคมีประสิทธิภาพดีที่สุดบนชุดข้อมูลเครดิตเยอรมัน ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพมากกว่าเชิงปริมาณ และชุดข้อมูลการตลาดของธนาคาร ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพเท่ากับเชิงปริมาณ ในขณะที่เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพดีที่สุดบนชุดข้อมูลลูกค้าบัตรเครดิต ซึ่งเป็นชุดข้อมูลที่มีจำนวนตัวแปรอิสระเชิงคุณภาพน้อยกว่าเชิงปริมาณ โดยแสดงผลลัพธ์ในตารางที่ 10 – 13

2. เมื่อพิจารณาการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลทางการเงินทั้ง 3 ชุด ภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มีการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่ม 3 เทคนิค พบว่า การนำเทคนิคการสุ่มทั้ง 3 เทคนิคมาประยุกต์ใช้นั้น ไม่ได้ทำให้ค่าความแม่นยำของตัวแบบการจำแนกด้วยเทคนิคใดมีค่าดีขึ้น ซึ่งแสดงผลลัพธ์ในตารางที่ 15 – 17

3. การนำเทคนิคการสุ่ม 3 เทคนิคมาใช้ จะทำให้มีการทำนายคำตอบในกลุ่ม Y=0 หรือกลุ่มส่วนน้อยมีค่าความถูกต้องที่มากขึ้น ซึ่งพิจารณาได้จากค่าเรียกคืน

4. จากผลการเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกบนชุดข้อมูลเครดิตเยอรมัน ชุดข้อมูลลูกค้าบัตรเครดิต และชุดข้อมูลการตลาดของธนาคารภายใต้ชุดข้อมูลตั้งต้นกับชุดข้อมูลที่มี

การปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มทั้ง 3 เทคนิค จากตารางที่ 18 เมื่อพิจารณาจากอัตราการเปลี่ยนแปลงของค่าความแม่นยำของตัวแบบการจำแนกด้วยการถดถอยลอจิสติกทวิภาค เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย และเทคนิคนาอีฟเบย์ พบว่าการปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่มมีประสิทธิภาพดีกว่าเทคนิคการสุ่มลด และเทคนิคการสุ่มผสมผสาน

5.1 ข้อเสนอแนะ

จากการวิจัยในครั้งนี้ เพื่อเป็นแนวทางในการพัฒนางานวิจัยให้มีประสิทธิภาพมากขึ้นผู้วิจัยมีข้อเสนอแนะดังนี้

1. ในขั้นตอนการคัดเลือกชุดข้อมูลที่นำมาวิเคราะห์ ควรเลือกใช้ชุดข้อมูลที่มีความหลากหลาย เช่น จำนวนของข้อมูล หรือคุณลักษณะของตัวแปรอิสระในชุดข้อมูล เพื่อที่จะสามารถเปรียบเทียบประสิทธิภาพได้ชัดเจนยิ่งขึ้น

2. ในการแบ่งข้อมูลเพื่อนำมาทดสอบประสิทธิภาพ ควรมีการประยุกต์ใช้วิธีอื่นเพิ่มเติม เช่น Split Test เป็นการแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน เช่น 70% : 30% หรือ 80% : 20% และ 10-Fold Cross-Validation

3. ควรมีการเปลี่ยนค่าพารามิเตอร์ที่สำคัญในการพัฒนาตัวแบบการจำแนกให้มีความหลากหลายมากยิ่งขึ้น เช่น จำนวนขั้นต่ำของค่าสังเกตที่มีในโหนด (minsplit) และจำนวนขั้นต่ำของค่าสังเกตที่โหนดสุดท้าย (minbucket) เป็นต้น

4. อาจมีการประยุกต์ใช้เทคนิคการจำแนกอื่น ๆ เช่น เทคนิคโครงข่ายประสาทเทียม เทคนิคเคเนียเรสเนเบอร์ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน และเทคนิคป่าสุ่ม

5. ประยุกต์ใช้เทคนิคการปรับปรุงชุดข้อมูลที่ไม่สมดุลด้วยเทคนิคอื่น ๆ เช่น เทคนิค SMOTE เทคนิค Tomek Link (T-Link) เทคนิค SMOTE Tomek Links เทคนิค DBSCAN (Density-Based Spatial Clustering of Applications with Noise) และ เทคนิค RUSBoost เป็นต้น

ภาคผนวก ก

ชุดข้อมูลที่ 1 ชุดข้อมูลเครดิตเยอรมัน

1. ชุดข้อมูลตั้งต้นของชุดข้อมูลเครดิตเยอรมันภายใต้ชุดข้อมูลเรียนรู้

ตารางผนวกที่ 1 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลตั้งต้นของชุดข้อมูลเครดิตเยอรมัน (ชุดข้อมูลเรียนรู้)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	502	117	123	58	67.96	51.25	58.43	81.10	89.64	85.16	78.13
2	499	117	123	61	66.85	51.25	58.02	81.01	89.11	84.86	77.75
3	498	125	115	62	64.97	47.92	55.16	79.94	88.93	84.19	76.63
4	499	117	123	61	66.85	51.25	58.02	81.01	89.11	84.86	77.75
5	500	128	112	60	65.12	46.67	54.37	79.62	89.29	84.18	76.50
		เฉลี่ย			66.35	49.67	56.80	80.53	89.21	84.65	77.35

ตารางผนวกที่ 2 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลตั้งต้นของชุดข้อมูลเครดิตเยอรมัน (ชุดข้อมูลเรียนรู้)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	517	101	139	43	76.37	57.92	65.88	83.66	92.32	87.78	82.00
2	528	109	131	32	80.37	54.58	65.01	82.89	94.29	88.22	82.38
3	532	122	118	28	80.82	49.17	61.14	81.35	95.00	87.64	81.25
4	518	98	142	42	77.17	59.17	66.98	84.09	92.50	88.10	82.50
5	516	94	146	44	76.84	60.83	67.91	84.59	92.14	88.21	82.75
		เฉลี่ย			78.32	56.33	65.38	83.31	93.25	87.99	82.18

ตารางผนวกที่ 3 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลตั้งต้นของชุดข้อมูลเครดิตเยอรมัน (ชุดข้อมูลเรียนรู้)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	476	123	137	84	61.99	52.69	56.96	79.47	85.00	82.14	74.76
2	475	109	131	85	60.65	54.58	57.46	81.34	84.82	83.04	75.75
3	463	110	130	97	57.27	54.17	55.67	80.80	82.68	81.73	74.13
4	477	107	133	83	61.57	55.42	58.33	81.68	85.18	83.39	76.25

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
5	468	114	126	92	57.80	52.50	55.02	80.41	83.57	81.96	74.25
		เฉลี่ย			59.86	53.87	56.69	80.74	84.25	82.45	75.03

2. ชุดข้อมูลตั้งต้นของชุดข้อมูลเครดิตเยอรมันภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 4 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลตั้งต้นของชุดข้อมูลเครดิตเยอรมัน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	125	35	25	15	62.50	41.67	50.00	78.13	89.29	83.33	75.00
2	124	30	30	16	65.22	50.00	56.60	80.52	88.57	84.35	77.00
3	129	35	25	11	69.44	41.67	52.08	78.66	92.14	84.87	77.00
4	119	33	27	21	56.25	45.00	50.00	78.29	85.00	81.51	73.00
5	121	25	35	19	64.81	58.33	61.40	82.88	86.43	84.62	78.00
		เฉลี่ย			63.65	47.33	54.02	79.69	88.29	83.74	76.00

ตารางผนวกที่ 5 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลตั้งต้นของชุดข้อมูลเครดิตเยอรมัน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	117	31	29	23	55.77	48.33	51.79	79.05	83.57	81.25	73.00
2	125	36	24	15	61.54	40.00	48.48	77.64	89.29	83.06	74.50
3	127	41	19	13	59.38	31.67	41.30	75.60	90.71	82.47	73.00
4	119	35	25	21	54.35	41.67	47.17	77.27	85.00	80.95	72.00
5	116	31	29	24	54.72	48.33	51.33	78.91	82.86	80.84	72.50
		เฉลี่ย			57.15	42.00	48.01	77.69	86.29	81.71	73.00

ตารางผนวกที่ 6 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลตั้งต้นของชุดข้อมูลเครดิตเยอรมัน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	106	30	30	34	46.88	50.00	48.39	77.94	75.71	76.81	68.00
2	119	31	29	21	58.00	48.33	52.73	79.33	85.00	82.07	74.00
3	127	30	30	13	69.77	50.00	58.25	80.89	90.71	85.52	78.50
4	113	31	29	27	51.79	48.33	50.00	78.47	80.71	79.58	71.00
5	118	23	37	22	62.71	61.67	62.18	83.69	84.29	83.99	77.50
		เฉลี่ย			57.83	51.67	54.31	80.07	83.29	81.59	73.80

3. ชุดข้อมูลเครดิตเยอรมันด้วยเทคนิคการสุ่มเพิ่มภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 7 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลเครดิตเยอรมันที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	96	16	44	44	50.00	73.33	59.46	85.71	68.57	76.19	70.00
2	102	21	39	38	50.65	65.00	56.93	82.93	72.86	77.57	70.50
3	101	20	40	39	50.63	66.67	57.55	83.47	72.14	77.39	70.50
4	99	22	38	41	48.10	63.33	54.68	81.82	70.71	75.86	68.50
5	107	15	45	33	57.69	75.00	65.22	87.70	76.43	81.68	76.00
		เฉลี่ย			51.42	68.67	58.77	84.33	72.14	77.74	71.10

ตารางผนวกที่ 8 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลเครดิตเยอรมันที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	91	21	39	49	44.32	65.00	52.70	81.25	65.00	72.22	65.00
2	112	22	38	28	57.58	63.33	60.32	83.58	80.00	81.75	75.00
3	113	30	30	27	52.63	50.00	51.28	79.02	80.71	79.86	71.50
4	115	34	26	25	50.98	43.33	46.85	77.18	82.14	79.58	70.50

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
5	101	21	39	39	50.00	65.00	56.52	82.79	72.14	77.10	70.00
		เฉลี่ย			51.10	57.33	53.53	80.76	76.00	78.10	70.40

ตารางผนวกที่ 9 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลเครดิตเยอรมันที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	91	16	44	49	47.31	73.33	57.52	85.05	65.00	73.68	67.50
2	102	20	40	38	51.28	66.67	57.97	83.61	72.86	77.86	71.00
3	111	23	37	29	56.06	61.67	58.73	82.84	79.29	81.02	74.00
4	99	22	38	41	48.10	63.33	54.68	81.82	70.71	75.86	68.50
5	106	15	45	34	56.96	75.00	64.75	87.60	75.71	81.23	75.50
		เฉลี่ย			51.94	68.00	58.73	84.18	72.71	77.93	71.30

4. ชุดข้อมูลเครดิตเยอรมันด้วยเทคนิคการสุ่มลดภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 10 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลเครดิตเยอรมันที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลด (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	92	15	45	48	48.39	75.00	58.82	85.98	65.71	74.49	68.50
2	100	18	42	40	51.22	70.00	59.15	84.75	71.43	77.52	71.00
3	98	17	43	42	50.59	71.67	59.31	85.22	70.00	76.86	70.50
4	93	19	41	47	46.59	68.33	55.41	83.04	66.43	73.81	67.00
5	96	13	47	44	51.65	78.33	62.25	88.07	68.57	77.11	71.50
		เฉลี่ย			49.69	72.67	58.99	85.41	68.43	75.96	69.70

ตารางผนวกที่ 11 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลเครดิตเยอรมันที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลด (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	85	18	42	55	43.30	70.00	53.50	82.52	60.71	69.96	63.50
2	81	16	44	59	42.72	73.33	53.99	83.51	57.86	68.35	62.50
3	95	23	37	45	45.12	61.67	52.11	80.51	67.86	73.64	66.00
4	81	16	44	59	42.72	73.33	53.99	83.51	57.86	68.35	62.50
5	93	18	42	17	71.19	70.00	70.59	83.78	84.55	84.16	79.41
		เฉลี่ย			49.01	69.67	56.84	82.77	65.77	72.89	66.78

ตารางผนวกที่ 12 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลเครดิตเยอรมันที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลด (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	89	21	39	51	43.33	65.00	52.00	80.91	63.57	71.20	64.00
2	106	19	41	34	54.67	68.33	60.74	84.80	75.71	80.00	73.50
3	109	23	37	31	54.41	61.67	57.81	82.58	77.86	80.15	73.00
4	98	22	38	42	47.50	63.33	54.29	81.67	70.00	75.38	68.00
5	103	15	45	37	54.88	75.00	63.38	87.29	73.57	79.84	74.00
		เฉลี่ย			50.96	66.67	57.64	83.45	72.14	77.32	70.50

5. ชุดข้อมูลเครดิตเยอรมันด้วยเทคนิคการสุ่มผสมผสานภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 13 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลเครดิตเยอรมันที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสาน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	91	16	44	49	47.31	73.33	57.52	85.05	65.00	73.68	67.50
2	106	19	41	34	54.67	68.33	60.74	84.80	75.71	80.00	73.50
3	97	20	40	43	48.19	66.67	55.94	82.91	69.29	75.49	68.50
4	98	20	40	42	48.78	66.67	56.34	83.05	70.00	75.97	69.00

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
5	100	17	43	40	51.81	71.67	60.14	85.47	71.43	77.82	71.50
		เฉลี่ย			50.15	69.33	58.14	84.25	70.29	76.59	70.00

ตารางผนวกที่ 14 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลเครดิตเยอรมันที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสาน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	77	21	39	63	38.24	65.00	48.15	78.57	55.00	64.71	58.00
2	94	18	42	46	47.73	70.00	56.76	83.93	67.14	74.60	68.00
3	92	18	42	48	46.67	70.00	56.00	83.64	65.71	73.60	67.00
4	99	17	43	41	51.19	71.67	59.72	85.34	70.71	77.34	71.00
5	87	20	40	53	43.01	66.67	52.29	81.31	62.14	70.45	63.50
		เฉลี่ย			45.37	68.67	54.58	82.56	64.14	72.14	65.50

ตารางผนวกที่ 15 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลเครดิตเยอรมันที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสาน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	88	21	39	52	42.86	65.00	51.66	80.73	62.86	70.68	63.50
2	102	18	42	38	52.50	70.00	60.00	85.00	72.86	78.46	72.00
3	108	19	41	32	56.16	68.33	61.65	85.04	77.14	80.90	74.50
4	99	23	37	41	47.44	61.67	53.62	81.15	70.71	75.57	68.00
5	104	18	42	36	53.85	70.00	60.87	85.25	74.29	79.39	73.00
		เฉลี่ย			50.56	67.00	57.56	83.43	71.57	77.00	70.20

ชุดข้อมูลที่ 2 ชุดข้อมูลลูกค้าบัตรเครดิต

1. ชุดข้อมูลตั้งต้นของชุดข้อมูลลูกค้าบัตรเครดิตภายใต้ชุดข้อมูลเรียนรู้

ตารางผนวกที่ 16 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลตั้งต้นของชุดข้อมูลลูกค้าบัตรเครดิต (ชุดข้อมูลเรียนรู้)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	17783	3780	1555	882	63.81	29.15	40.02	82.47	95.27	88.41	80.58
2	17804	3770	1531	895	63.11	28.88	39.63	82.53	95.21	88.42	80.56
3	17802	3734	1581	883	64.16	29.75	40.65	82.66	95.27	88.52	80.76
4	17984	3803	1393	820	62.95	26.81	37.60	82.54	95.64	88.61	80.74
5	17722	3868	1529	881	63.44	28.33	39.17	82.08	95.26	88.18	80.21
		เฉลี่ย			63.49	28.58	39.41	82.46	95.33	88.43	80.57

ตารางผนวกที่ 17 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลตั้งต้นของชุดข้อมูลลูกค้าบัตรเครดิต (ชุดข้อมูลเรียนรู้)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	17919	3534	1801	746	70.71	33.76	45.70	83.53	96.00	89.33	82.17
2	17939	3563	1738	760	69.58	32.79	44.57	83.43	95.94	89.25	81.99
3	17935	3546	1769	750	70.23	33.28	45.16	83.49	95.99	89.30	82.10
4	18042	3575	1621	762	68.02	31.20	42.78	83.46	95.95	89.27	81.93
5	17809	3618	1779	794	69.14	32.96	44.64	83.11	95.73	88.98	81.62
		เฉลี่ย			69.54	32.80	44.57	83.41	95.92	89.23	81.96

ตารางผนวกที่ 18 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลตั้งต้นของชุดข้อมูลลูกค้าบัตรเครดิต (ชุดข้อมูลเรียนรู้)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	14795	2199	3136	3870	44.76	58.78	50.82	87.06	79.27	82.98	74.71
2	14706	2176	3125	3993	43.90	58.95	50.33	87.11	78.65	82.66	74.30
3	14663	2095	3220	4022	44.46	60.58	51.29	87.50	78.47	82.74	74.51
4	14797	2155	3041	4007	43.15	58.53	49.67	87.29	78.69	82.77	74.33

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
5	14619	2194	3203	3984	44.57	59.35	50.91	86.95	78.58	82.56	74.26
		เฉลี่ย			44.17	59.24	50.60	87.18	78.73	82.74	74.42

2. ชุดข้อมูลตั้งต้นของชุดข้อมูลลูกค้าบัตรเครดิตภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 19 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลตั้งต้นของชุดข้อมูลลูกค้าบัตรเครดิต (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	4467	949	352	232	60.27	27.06	37.35	82.48	95.06	88.32	80.32
2	4460	962	373	205	64.53	27.94	39.00	82.26	95.61	88.43	80.55
3	4434	959	362	245	59.64	27.40	37.55	82.22	94.76	88.05	79.93
4	4325	971	469	235	66.62	32.57	43.75	81.67	94.85	87.76	79.90
5	4572	904	335	189	63.93	27.04	38.00	83.49	96.03	89.32	81.78
		เฉลี่ย			63.00	28.40	39.13	82.42	95.26	88.38	80.50

ตารางผนวกที่ 20 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลตั้งต้นของชุดข้อมูลลูกค้าบัตรเครดิต (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	4492	925	376	207	64.49	28.90	39.92	82.92	95.59	88.81	81.13
2	4472	896	493	193	71.87	35.49	47.52	83.31	95.86	89.15	82.01
3	4476	913	408	203	66.78	30.89	42.24	83.06	95.66	88.92	81.40
4	4369	884	556	191	74.43	38.61	50.85	83.17	95.81	89.05	82.08
5	4602	841	398	159	71.45	32.12	44.32	84.55	96.66	90.20	83.33
		เฉลี่ย			69.80	33.20	44.97	83.40	95.92	89.22	81.99

ตารางผนวกที่ 21 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลตั้งต้นของชุดข้อมูลลูกค้าบัตรเครดิต (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3669	561	740	1030	41.81	56.88	48.19	86.74	78.08	82.18	73.48
2	3711	543	792	954	45.36	59.33	51.41	87.24	79.55	83.22	75.05
3	3675	550	771	1004	43.44	58.36	49.81	86.98	78.54	82.55	74.10
4	3500	490	950	1060	47.26	65.97	55.07	87.72	76.75	81.87	74.17
5	3839	553	686	922	42.66	55.37	48.19	87.41	80.63	83.89	75.42
		เฉลี่ย			44.11	59.18	50.53	87.22	78.71	82.74	74.44

3. ชุดข้อมูลลูกค้าบัตรเครดิตด้วยเทคนิคการสุ่มเพิ่มภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 22 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลลูกค้าบัตรเครดิตที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3802	567	734	897	45.00	56.42	50.07	87.02	80.91	83.86	75.60
2	3855	572	763	810	48.51	57.15	52.48	87.08	82.64	84.80	76.97
3	3790	563	758	889	46.02	57.38	51.08	87.07	81.00	83.92	75.80
4	3691	536	904	869	50.99	62.78	56.27	87.32	80.94	84.01	76.58
5	4057	558	681	704	49.17	54.96	51.91	87.91	85.21	86.54	78.97
		เฉลี่ย			47.94	57.74	52.36	87.28	82.14	84.63	76.78

ตารางผนวกที่ 23 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลลูกค้าบัตรเครดิตที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3785	553	748	914	45.01	57.49	50.49	87.25	80.55	83.77	75.55
2	3800	566	769	865	47.06	57.60	51.80	87.04	81.46	84.15	76.15
3	3772	579	742	907	45.00	56.17	49.97	86.69	80.62	83.54	75.23
4	3673	537	903	887	50.45	62.71	55.91	87.24	80.55	83.76	76.27

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
5	3990	552	687	771	47.12	55.45	50.95	87.85	83.81	85.78	77.95
		เฉลี่ย			46.93	57.88	51.82	87.21	81.40	84.20	76.23

ตารางผนวกที่ 24 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลลูกค้าบัตรเครดิตที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3568	503	798	1131	41.37	61.34	49.41	87.64	75.93	81.37	72.77
2	3556	496	839	1109	43.07	62.85	51.11	87.76	76.23	81.59	73.25
3	3525	515	806	1154	41.12	61.01	49.13	87.25	75.34	80.86	72.18
4	3359	449	991	1201	45.21	68.82	54.57	88.21	73.66	80.28	72.50
5	3730	480	759	1031	42.40	61.26	50.12	88.60	78.34	83.16	74.82
		เฉลี่ย			42.63	63.06	50.87	87.89	75.90	81.45	73.10

4. ชุดข้อมูลลูกค้าบัตรเครดิตด้วยเทคนิคการสุ่มลดภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 25 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลลูกค้าบัตรเครดิตที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลด (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3790	571	730	909	44.54	56.11	49.66	86.91	80.66	83.66	75.33
2	3846	566	769	819	48.43	57.60	52.62	87.17	82.44	84.74	76.92
3	3792	566	755	887	45.98	57.15	50.96	87.01	81.04	83.92	75.78
4	3695	535	905	865	51.13	62.85	56.39	87.35	81.03	84.07	76.67
5	4050	555	684	711	49.03	55.21	51.94	87.95	85.07	86.48	78.90
		เฉลี่ย			47.82	57.78	52.31	87.28	82.05	84.58	76.72

ตารางผนวกที่ 26 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลลูกค้าบัตรเครดิตที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลด (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3785	553	748	914	45.01	57.49	50.49	87.25	80.55	83.77	75.55
2	3800	566	769	865	47.06	57.60	51.80	87.04	81.46	84.15	76.15
3	3772	579	742	907	45.00	56.17	49.97	86.69	80.62	83.54	75.23
4	3673	537	903	887	50.45	62.71	55.91	87.24	80.55	83.76	76.27
5	3990	552	687	771	47.12	55.45	50.95	87.85	83.81	85.78	77.95
		เฉลี่ย			46.93	57.88	51.82	87.21	81.40	84.20	76.23

ตารางผนวกที่ 27 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลลูกค้าบัตรเครดิตที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลดภายใต้ชุดข้อมูลทดสอบ

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3614	531	770	1085	41.51	59.19	48.80	87.19	76.91	81.73	73.07
2	3469	475	860	1196	41.83	64.42	50.72	87.96	74.36	80.59	72.15
3	3627	535	786	1052	42.76	59.50	49.76	87.15	77.52	82.05	73.55
4	3456	467	973	1104	46.85	67.57	55.33	88.10	75.79	81.48	73.82
5	3794	510	729	967	42.98	58.84	49.68	88.15	79.69	83.71	75.38
		เฉลี่ย			43.19	61.90	50.86	87.71	76.85	81.91	73.59

5. ชุดข้อมูลลูกค้าบัตรเครดิตด้วยเทคนิคการสุ่มผสมผสานภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 28 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลลูกค้าบัตรเครดิตที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสาน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3795	565	736	904	44.88	56.57	50.05	87.04	80.76	83.78	75.52
2	3847	572	763	818	48.26	57.15	52.33	87.06	82.47	84.70	76.83
3	3772	565	756	907	45.46	57.23	50.67	86.97	80.62	83.67	75.47
4	3679	532	908	881	50.75	63.06	56.24	87.37	80.68	83.89	76.45

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
5	4050	558	681	711	48.92	54.96	51.77	87.89	85.07	86.46	78.85
		เฉลี่ย			47.66	57.79	52.21	87.27	81.92	84.50	76.62

ตารางผนวกที่ 29 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลลูกค้าบัตรเครดิตที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสาน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3785	553	748	914	45.01	57.49	50.49	87.25	80.55	83.77	75.55
2	3800	566	769	865	47.06	57.60	51.80	87.04	81.46	84.15	76.15
3	3772	579	742	907	45.00	56.17	49.97	86.69	80.62	83.54	75.23
4	3673	537	903	887	50.45	62.71	55.91	87.24	80.55	83.76	76.27
5	3990	552	687	771	47.12	55.45	50.95	87.85	83.81	85.78	77.95
		เฉลี่ย			46.93	57.88	51.82	87.21	81.40	84.20	76.23

ตารางผนวกที่ 30 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลลูกค้าบัตรเครดิตที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสาน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	3588	523	778	1111	41.19	59.80	48.78	87.28	76.36	81.45	72.77
2	3614	503	832	1051	44.18	62.32	51.71	87.78	77.47	82.30	74.10
3	3540	516	805	1139	41.41	60.94	49.31	87.28	75.66	81.05	72.42
4	3369	456	984	1191	45.24	68.33	54.44	88.08	73.88	80.36	72.55
5	3621	472	767	1140	40.22	61.90	48.76	88.47	76.06	81.79	73.13
		เฉลี่ย			42.45	62.66	50.60	87.78	75.88	81.39	72.99

ชุดข้อมูลที่ 3 ชุดข้อมูลการตลาดของธนาคาร

1. ชุดข้อมูลตั้งต้นของชุดข้อมูลการตลาดของธนาคารภายใต้ชุดข้อมูลเรียนรู้

ตารางผนวกที่ 31 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลตั้งต้นของชุดข้อมูลการตลาดของธนาคาร (ชุดข้อมูลเรียนรู้)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	28468	2206	1556	771	66.87	41.36	51.11	92.81	97.36	95.03	90.98
2	28423	1933	1729	816	67.94	47.21	55.71	93.63	97.21	95.39	91.64
3	28520	2176	1536	718	68.15	41.38	51.49	92.91	97.54	95.17	91.22
4	28490	2125	1578	748	67.84	42.61	52.35	93.06	97.44	95.20	91.28
5	28730	1541	2171	508	81.04	58.49	67.94	94.91	98.26	96.56	93.78
		เฉลี่ย			70.37	46.21	55.72	93.46	97.56	95.47	91.78

ตารางผนวกที่ 32 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลตั้งต้นของชุดข้อมูลการตลาดของธนาคาร (ชุดข้อมูลเรียนรู้)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	28279	1846	1916	960	66.62	50.93	57.73	93.87	96.72	95.27	91.50
2	28170	1365	2297	1069	68.24	62.73	65.37	95.38	96.34	95.86	92.60
3	28479	1861	1851	759	70.92	49.87	58.56	93.87	97.40	95.60	92.05
4	28118	1202	2510	1120	69.15	67.62	68.37	95.90	96.17	96.03	92.95
5	28757	1394	2318	481	82.82	62.45	71.20	95.38	98.35	96.84	94.31
		เฉลี่ย			71.55	58.72	64.25	94.88	97.00	95.92	92.68

ตารางผนวกที่ 33 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลตั้งต้นของชุดข้อมูลการตลาดของธนาคาร (ชุดข้อมูลเรียนรู้)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	24987	1164	2598	4252	37.93	69.06	48.96	95.55	85.46	90.22	83.59
2	26593	1494	2168	2646	45.04	59.20	51.16	94.68	90.95	92.78	87.42
3	26771	1680	2032	2467	45.17	54.74	49.49	94.10	91.56	92.81	87.41
4	25609	1610	2102	3629	36.68	56.63	44.52	94.09	87.59	90.72	84.10

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
5	25551	1152	2560	3687	40.98	68.97	51.41	95.69	87.39	91.35	85.31
		เฉลี่ย			41.16	61.72	49.11	94.82	88.59	91.58	85.57

2. ชุดข้อมูลตั้งต้นของชุดข้อมูลการตลาดของธนาคารภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 34 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลตั้งต้นของชุดข้อมูลการตลาดของธนาคาร (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	7258	763	115	51	69.28	13.10	22.03	90.49	99.30	94.69	90.06
2	7053	856	122	256	32.28	12.47	17.99	89.18	96.50	92.69	86.58
3	7209	770	158	101	61.00	17.03	26.62	90.35	98.62	94.30	89.43
4	6992	814	114	318	26.39	12.28	16.76	89.57	95.65	92.51	86.26
5	4773	153	775	2537	23.40	83.51	36.56	96.89	65.29	78.02	67.35
		เฉลี่ย			42.47	27.68	23.99	91.30	91.07	90.44	83.93

ตารางผนวกที่ 35 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลตั้งต้นของชุดข้อมูลการตลาดของธนาคาร (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	7309	856	22	0	100.0	2.51	4.89	89.52	100.0	94.47	89.54
2	2963	978	0	4346	0.00	0.00	N/A*	75.18	40.54	52.68	35.75
3	6990	918	10	320	3.03	1.08	1.59	88.39	95.62	91.86	84.97
4	1637	805	123	5673	2.12	13.25	3.66	67.04	22.39	33.57	21.36
5	753	0	928	6557	12.40	100.0	22.06	100.0	10.30	18.68	20.41
		เฉลี่ย			23.51	23.37	N/A*	84.03	53.77	58.25	50.41

* มี 1 ชุดทดสอบที่ TN (True Negative) เท่ากับ 0

ตารางผนวกที่ 36 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลตั้งต้นของชุดข้อมูลการตลาดของธนาคาร (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	7291	821	57	17	77.03	6.49	11.97	89.88	99.77	94.57	89.76
2	7264	774	204	45	81.93	20.86	33.25	90.37	99.38	94.66	90.12
3	7253	584	344	57	85.79	37.07	51.77	92.55	99.22	95.77	92.22
4	5429	78	850	1881	31.12	91.59	46.46	98.58	74.27	84.72	76.22
5	67	0	928	7243	11.36	100.0	20.40	100.0	0.92	1.82	12.08
		เฉลี่ย			57.44	51.20	32.77	94.28	74.71	74.31	72.08

3. ชุดข้อมูลการตลาดของธนาคารด้วยเทคนิคการสุ่มเพิ่มภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 37 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลการตลาดของธนาคารที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	6098	341	537	1211	30.72	61.16	40.90	94.70	83.43	88.71	81.04
2	5971	553	425	1338	24.11	43.46	31.01	91.52	81.69	86.33	77.18
3	6800	165	763	510	59.94	82.22	69.33	97.63	93.02	95.27	91.81
4	5304	279	649	2006	24.44	69.94	36.23	95.00	72.56	82.28	72.26
5	2962	85	843	4348	16.24	90.84	27.55	97.21	40.52	57.20	46.19
		เฉลี่ย			31.09	69.52	41.00	95.21	74.25	81.96	73.70

ตารางผนวกที่ 38 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลการตลาดของธนาคารที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	7309	840	38	0	100.0	4.33	8.30	89.69	100.0	94.57	89.74
2	6628	631	347	681	33.75	35.48	34.60	91.31	90.68	90.99	84.17
3	6672	819	109	638	14.59	11.75	13.01	89.07	91.27	90.16	82.31
4	1549	532	396	5761	6.43	42.67	11.18	74.44	21.19	32.99	23.61

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
5	30	0	928	7258	11.34	100.0	20.36	100.0	0.41	0.82	11.66
		เฉลี่ย			33.22	38.85	17.49	88.90	60.71	61.90	58.30

ตารางผนวกที่ 39 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลการตลาดของธนาคารที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มเพิ่ม (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	7289	802	76	20	79.17	8.66	15.61	90.09	99.73	94.66	89.96
2	7224	351	577	86	87.03	62.18	72.53	95.37	98.82	97.06	94.70
3	7239	668	310	70	81.58	31.70	45.66	91.55	99.04	95.15	91.09
4	3730	5	923	3580	20.50	99.46	33.99	99.87	51.03	67.54	56.48
5	30	0	928	7280	11.31	100.0	20.32	100.0	0.41	0.82	11.63
		เฉลี่ย			55.92	60.40	37.62	95.37	69.81	71.05	68.77

4. ชุดข้อมูลการตลาดของธนาคารด้วยเทคนิคการสุ่มลดภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 40 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลการตลาดของธนาคารที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลด (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	5903	333	545	1379	28.33	62.07	38.90	94.66	81.06	87.34	79.02
2	6071	555	423	1238	25.47	43.25	32.06	91.62	83.06	87.13	78.36
3	6803	156	772	507	60.36	83.19	69.96	97.76	93.06	95.35	91.95
4	5290	308	620	2020	23.48	66.81	34.75	94.50	72.37	81.96	71.74
5	2913	65	863	4397	16.41	93.00	27.89	97.82	39.85	56.63	45.84
		เฉลี่ย			30.81	69.66	40.71	95.27	73.88	81.68	73.38

ตารางผนวกที่ 41 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลการตลาดของธนาคารที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลด (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	7309	840	38	0	100.0	4.33	8.30	89.69	100.0	94.57	89.74
2	6628	829	149	681	17.95	15.24	16.48	88.88	90.68	89.77	81.78
3	6672	819	109	638	14.59	11.75	13.01	89.07	91.27	90.16	82.31
4	1549	532	396	5761	6.43	42.67	11.18	74.44	21.19	32.99	23.61
5	36	0	928	7274	11.31	100.0	20.33	100.0	0.49	0.98	11.70
		เฉลี่ย			30.06	34.80	13.86	88.42	60.73	61.69	57.83

ตารางผนวกที่ 42 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลการตลาดของธนาคารที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มลด (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	7286	784	94	23	80.34	10.71	18.89	90.29	99.69	94.75	90.14
2	7218	331	597	92	86.65	64.33	73.84	95.62	98.74	97.15	94.87
3	7238	664	314	71	81.56	32.11	46.07	91.60	99.03	95.17	91.13
4	3619	5	923	3691	20.00	99.46	33.31	99.86	49.51	66.20	55.13
5	36	0	928	7274	11.31	100.0	20.33	100.0	0.49	0.98	11.70
		เฉลี่ย			55.97	61.32	38.49	95.47	69.49	70.85	68.60

5. ชุดข้อมูลการตลาดของธนาคารด้วยเทคนิคการสุ่มผสมผสานภายใต้ชุดข้อมูลทดสอบ

ตารางผนวกที่ 43 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากการถดถอยลอจิสติกทวิภาคบนชุดข้อมูลการตลาดของธนาคารที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสาน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	6205	343	535	1104	32.64	60.93	42.51	94.76	84.90	89.56	82.33
2	5799	551	427	1510	22.04	43.66	29.30	91.32	79.34	84.91	75.13
3	6805	177	751	505	59.79	80.93	68.77	97.46	93.09	95.23	91.72

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
4	5235	289	639	2057	23.70	68.86	35.26	94.77	71.79	81.69	71.46
5	2779	45	883	4531	16.31	95.15	27.85	98.41	38.02	54.85	44.45
		เฉลี่ย			30.90	69.91	40.74	95.34	73.43	81.25	73.02

ตารางผนวกที่ 44 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยบนชุดข้อมูลการตลาดของธนาคารที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสาน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	7309	838	40	0	100.0	4.56	8.71	89.71	100.0	94.58	89.76
2	6628	472	506	681	42.63	51.74	46.74	93.35	90.68	92.00	86.09
3	6805	819	109	505	17.75	11.75	14.14	89.26	93.09	91.13	83.93
4	1549	532	396	5761	6.43	42.67	11.18	74.44	21.19	32.99	23.61
5	36	7274	928	0	100.0	11.31	20.33	0.49	100.0	0.98	11.70
		เฉลี่ย			53.36	24.41	20.22	69.45	80.99	62.34	59.02

ตารางผนวกที่ 45 เมทริกซ์ความสับสนของตัวแบบการจำแนกจากเทคนิคนาอีฟเบย์บนชุดข้อมูลการตลาดของธนาคารที่ปรับปรุงความไม่สมดุลด้วยเทคนิคการสุ่มผสมผสาน (ชุดข้อมูลทดสอบ)

	TP	FP	TN	FN	Y = 0			Y = 1			Acc
					Pre	Re	F1	Pre	Re	F1	
1	76	802	76	20	79.17	8.66	15.61	8.66	79.17	15.61	15.61
2	7238	667	311	71	81.41	31.80	45.74	91.56	99.03	95.15	91.09
3	7224	344	584	86	87.16	62.93	73.09	95.45	98.82	97.11	94.78
4	3719	5	923	3591	20.45	99.46	33.92	99.87	50.88	67.41	56.35
5	27	0	928	7283	11.30	100.0	20.31	100.0	0.37	0.74	11.59
		เฉลี่ย			55.90	60.57	37.73	79.11	65.65	55.20	53.88

ภาคผนวก ข

โปรแกรมอาร์

1. การสร้างตัวแบบการจำแนกด้วยการถดถอยลอจิสติกทวิภาค

```
library(caret)
library(MASS)
library(lmtest)
library(blorr)
library(fmsb)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
  fit.LR <- glm(Y ~ .,data = train, family = binomial("logit"))
  table(train$Y)
  table(test$Y)
  Train.Result.LR <- blr_confusion_matrix(fit.LR, data = train)
  Test.Result.LR <- blr_confusion_matrix(fit.LR, data = test)
  print(i)
  print(Train.Result.LR$conf_matrix)
  print(c(Train.Result.LR$conf_matrix[1],Train.Result.LR$conf_matrix[3],Train.Result
        .LR$conf_matrix[4],Train.Result.LR$conf_matrix[2]))
```

```
  print(Test.Result.LR$conf_matrix)
  print(c(Test.Result.LR$conf_matrix[1],Test.Result.LR$conf_matrix[3],Test.Result
        .LR$conf_matrix[4],Test.Result.LR$conf_matrix[2]))
}
```

2. การสร้างตัวแบบการจำแนกด้วยเทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

```
library(rpart)
library(rpart.plot)
library(mlbench)
library(caret)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
  minsplit = 10
  tree.model <- rpart(Y ~ ., data = train, method = "class", minsplit = 10, minbucket =
                        round(minsplit/3))
  ans.train <- predict(tree.model, newdata = train, type = "class")
  ans.test <- predict(tree.model, newdata = test, type = "class")
```

```
  train.ans <- table(ans.train, train$Y)
  test.ans <- table(ans.test, test$Y)
  print(i)
  print(train.ans)
  print(test.ans)
}
```

3. การสร้างตัวแบบการจำแนกด้วยเทคนิคนาอีฟเบย์

```
library(naivebayes)
library(caret)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
  fit.NB = naive_bayes(Y ~ ., data = train, laplace = 1)
  ans.train <- predict(fit.NB, newdata = train[,1:20])
  ans.test <- predict(fit.NB, newdata = test[,1:20])
  train.ans <- table(ans.train, train$Y)
  test.ans <- table(ans.test, test$Y)
```

```
  print(i)
  print(train.ans)
  print(test.ans)
}
```

4. การสร้างตัวแบบการจำแนกภายใต้เทคนิคการสุ่มเพิ่ม

4.1 การถดถอยลอจิสติกทวิภาค

```
library(caret)
library(MASS)
library(lmtest)
library(blorr)
library(fmsb)
library(ROSE)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
  train <- ovun.sample(Y~., data=train, p=0.5, seed=1, method="over")$data
  fit.LR <- glm(Y ~ .,data = train, family = binomial("logit"))
```

```
  Test.Result.LR <- blr_confusion_matrix(fit.LR, data = test)
  print(i)
  print(Test.Result.LR$conf_matrix)
  print(c(Test.Result.LR$conf_matrix[1],Test.Result.LR$conf_matrix[3],Test.Result
        .LR$conf_matrix[4],Test.Result.LR$conf_matrix[2]))
}
```

4.2 เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

```
library(rpart)
library(rpart.plot)
library(mlbench)
library(caret)
library(ROSE)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
  train <- ovun.sample(Y~., data=train, p=0.5, seed=1, method="over")$data
  minsplit = 10
```

```
  tree.model <- rpart(Y ~ ., data = train, method = "class", minsplit = 10, minbucket =
                        round(minsplit/3))
  ans.test <- predict(tree.model, newdata = test, type = "class")
  test.ans <- table(ans.test, test$Y)
  print(i)
  print(test.ans)
  print(c(test.ans[1],test.ans[2],test.ans[4],test.ans[3]))
}
```

4.3 เทคนิคนาอีฟเบย์

```
library(ROSE)
library(naivebayes)
library(caret)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
  train <- ovun.sample(Y~., data=train, p=0.5, seed=1, method="over")$data
  fit.NB = naive_bayes(Y ~ ., data = train, laplace = 1)
```

```
  ans.test <- predict(fit.NB, newdata = test[,1:20])
  test.ans <- table(ans.test, test$Y)
  print(i)
  print(test.ans)
}
```

5. การสร้างตัวแบบการจำแนกภายใต้เทคนิคการสุ่มลด

5.1 การถดถอยลอจิสติกทวิภาค

```
library(caret)
library(MASS)
library(lmtest)
library(blorr)
library(fmsb)
library(ROSE)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
  train <- ovun.sample(Y~., data=train, p=0.5, seed=1, method="under")$data
```

```
  fit.LR <- glm(Y ~ .,data = train, family = binomial("logit"))
  Test.Result.LR <- blr_confusion_matrix(fit.LR, data = test)
  print(i)
  print(Test.Result.LR$conf_matrix)
  print(c(Test.Result.LR$conf_matrix[1],Test.Result.LR$conf_matrix[3],Test.Result
        .LR$conf_matrix[4],Test.Result.LR$conf_matrix[2]))
}
```

5.2 เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

```
library(rpart)
library(rpart.plot)
library(mlbench)
library(caret)
library(ROSE)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
  train <- ovun.sample(Y~., data=train, p=0.5, seed=1, method="under")$data
```

```
  minsplit = 10
  tree.model <- rpart(Y ~ ., data = train, method = "class", minsplit = 10, minbucket =
                      round(minsplit/3))
  ans.test <- predict(tree.model, newdata = test, type = "class")
  test.ans <- table(ans.test, test$Y)
  print(i)
  print(test.ans)
  print(c(test.ans[1],test.ans[2],test.ans[4],test.ans[3]))
}
```

5.3 เทคนิคนาอีฟเบย์

```
library(ROSE)
library(naivebayes)
library(caret)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
  train <- ovun.sample(Y~., data=train, p=0.5, seed=1, method="under")$data
```

```
  fit.NB = naive_bayes(Y ~ ., data = train, laplace = 1)
  ans.test <- predict(fit.NB, newdata = test[,1:20])
  test.ans <- table(ans.test, test$Y)
  print(i)
  print(test.ans)
}
```

6. การสร้างตัวแบบการจำแนกภายใต้เทคนิคการสุ่มผสมผสาน

6.1 การถดถอยลอจิสติกทวิภาค

```
library(caret)
library(MASS)
library(lmtest)
library(blorr)
library(fmsb)
library(ROSE)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
```

```
  train <- ovun.sample(Y~., data=train, p=0.5,  seed=1, method="both")$data
  fit.LR <- glm(Y ~ .,data = train, family = binomial("logit"))
  Test.Result.LR <- blr_confusion_matrix(fit.LR, data = test)
  print(i)
  print(Test.Result.LR$conf_matrix)
  print(c(Test.Result.LR$conf_matrix[1],Test.Result.LR$conf_matrix[3],Test.Result
        .LR$conf_matrix[4],Test.Result.LR$conf_matrix[2]))
}
```

6.2 เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

```
library(rpart)
library(rpart.plot)
library(mlbench)
library(caret)
library(ROSE)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
```

```
  train <- ovun.sample(Y~., data=train, p=0.5,  seed=1, method=" both")$data
  minsplit = 10
  tree.model <- rpart(Y ~ ., data = train, method = "class", minsplit = 10, minbucket =
                        round(minsplit/3))
  ans.test <- predict(tree.model, newdata = test, type = "class")
  test.ans <- table(ans.test, test$Y)
  print(i)
  print(test.ans)
  print(c(test.ans[1],test.ans[2],test.ans[4],test.ans[3]))
}
```

6.3 เทคนิคนาอีฟเบย์

```
library(ROSE)
library(naivebayes)
library(caret)
Train.File <- c("Train1.csv","Train2.csv","Train3.csv","Train4.csv","Train5.csv")
Test.File <- c("Test1.csv","Test2.csv","Test3.csv","Test4.csv","Test5.csv")
for (i in 1:5) {
  print(i)
  train <- read.csv(Train.File[i])
  test <- read.csv(Test.File[i])
  set.seed(1)
  train$Y <- factor(train$Y, levels = c(1,0))
  train$X19 <- factor(train$X19, levels = c("A191","A192"))
  train$X20 <- factor(train$X20, levels = c("A201","A202"))
  test$Y <- factor(test$Y, levels = c(1,0))
  test$X19 <- factor(test$X19, levels = c("A191","A192"))
  test$X20 <- factor(test$X20, levels = c("A201","A202"))
```

```
  train <- ovun.sample(Y~., data=train, p=0.5,  seed=1, method=" both")$data
  fit.NB = naive_bayes(Y ~ ., data = train, laplace = 1)
  ans.test <- predict(fit.NB, newdata = test[,1:20])
  test.ans <- table(ans.test, test$Y)
  print(i)
  print(test.ans)
}
```

บรรณานุกรม

บรรณานุกรม

Aida, K. (2017). Using a naive Bayesian classifier methodology for loan risk assessment (Evidence from a Tunisian commercial bank). *Journal of Economics, Finance and Administrative Science,* 42(22), 3-24.

Barış, A., & Derviş, B. (2020). Comparison of Machine Learning Methods in Prediction of Financial Failure of Businesses in The Manufacturing Industry: Evidence from Borsa Istanbul. *Anadolu University Journal of Social Sciences,* 20(4), 237-268.

Begüm, Ç., & Deniz, U. (2019). Comparison of Data Mining Classification Algorithms Determining the Default Risk. *Hindawi Scientific Programming,* 2019, 1-8.

Breiman L., Friedman J.H., Olshen R. & Stone, C.J. (1984). Classification and Regression Tree. *Wadsworth & Brooks/Cole Advanced Books & Software*, Pacific California.

Chawla, N. V., Bowyer, K. W., Hall L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research,* 16(1), 321-357.

Chawla, N. V., Japkowicz, N., & Kołcz, A. (2014). Editorial: Special Issue on Learning from Imbalanced Dat Sets. *ACM SIGKDD Explorations*, 6(1), 1-6.

Deepika, S. (2020). *Explore R Libraries: Rpart*. Pluralsight. https://www.pluralsight.com /guides/explore-r-libraries:-rpart?fbclid=IwAR2HQqxHA-RH7BzsdfKf0UZt0UG8uenbGFtjtKqznwZtUwBZ_-mb7ryU7d0

Farquad, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support - vectormachine. *Decision Support Systems*, 53(1), 226-233.

Fayaz, I., Meenakshi., & Satwinder, S. (2020). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*

Han, J., Kamber, M. & Pei, J. (2012). *Data mining: Concepts and techniques*. (3rd ed.). San Francisco, CA: Morgan Kaufmann, 327-344.

Hong, C., Songhua, H., Rui, H. & Xiuju, Z. (2021). Improved naïve Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing,* 30, 1-12.

Inthasone, S., Pasquier, N., Tettamanzi, A. G. B., & Pereira, C. C. (2014). The BioKET Biodiversity Data Warehouse: Data and Knowledge Integration and Extraction. *Advances in Intelligent Data Analysis XIII*, 131-142.

Iqbal, H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3).

Irimia Dieguez, A., Blanco Oliver, A., & Vazquez Cueto, M.J. (2015). A Comparison of Classification/Regression Trees and Logistic Regression in Failure Models. *Procedia Economics and Finance,* 23(2015), 9–14.

Ishwarappa & Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop technology. *Procedia Computer Science,* 48, 319–324.

Jie, H., & Jennifer, L. (2016). A Comparison of Machine Learning Techniques and Logistic Regression Method for the Prediction of Past-Due Amount. Kennesaw State University DigitalCommons@Kennesaw State University.

Nasritha, K., Kerdprasop, K., & Kerdprasop, N. (2017). Comparison of sampling techniques for imbalanced data classification. *Journal of Applied Informatics and Technology*, 1(1), 20-37.

Tiplawan, K. (2021). Big Data Innovation and Factors Influencing Thai Bank's Performance: Invariance Testing between Different Types of Banks. *JOURNAL OF YALA RAJABHAT UNIVERSITY,* 17(2), 126-135.

Usama, F., Gregory, P., & Padhraic, S. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3). 37-54.

Yuzhen, W., & Jingqiao, Q. (2020). Analysis of financial product purchases based on logistic regression. *Journal of Physics: Conference Series,* 1-6

กีระชาติ สุขสุทธิ์. (2559). *การจำแนกข้อมูลไม่สมดุล โดยใช้การปรับปรุงข้อมูลร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยขั้นตอนวิธีทางพันธุกรรมที่มีการเริ่มต้นใหม่*. วิทยานิพนธ์ปริญญาเอก มหาวิทยาลัยเทคโนโลยีสุรนารี.

กัลยา วานิชย์บัญชา. (2555). การวิเคราะห์ข้อมูลหลายตัวแปร (พิมพ์ครั้งที่ 3 ed.). ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

เกรียงไกร ชัยมินทร์. (2557). *ระบบตรวจจับการบุกรุกเครือข่ายสำหรับสำนักหอสมุด มหาวิทยาลัยเชียงใหม่ โดยการใช้ตัวจำแนกข้อมูลนาอีฟเบส์*. วิทยานิพนธ์ วท.ม สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเชียงใหม่.

เบญจภรณ์ จันทรกองกุล, สุวรรณา รัศมีขวัญ, สุนิสา ริมเจริญ, ภูสิต กุลเกษม, กฤษณะ ชินสาร, อัณณ์นุ พันธ์ รอดทุกข์, ปิยนุช วรบุตร และจรรยา อ้นปันส์. (2557). *วิธีการที่เหมาะสมสำหรับการ แบ่งกลุ่มข้อมูลที่ไม่สมดุลสูง*.

ภัคสุภางค์ มาปรีดา. (2560). ตัวแบบการถดถอยลอจิสติกในการพยากรณ์ความน่าจะเป็นของการชำระหนี้ได้ของครัวเรือน กรณีศึกษาจังหวัดปทุมธานี มหาวิทยาลัยธรรมศาสตร์. ปทุมธานี.

ยุทธ ไกยวรรณ์. (2555). หลักการและการใช้การวิเคราะห์การถดถอยลอจิสติกสาหรับการวิจัย. *วารสารวิจัยมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย*, 4(1), 1-12.

ศิริชัย พงษ์วิชัย. (2549). การวิเคราะห์ข้อมูลทางสถิติด้วยคอมพิวเตอร์ (พิมพ์ครั้งที่ 16 ed.). จุฬาลงกรณ์มหาวิทยาลัย.

สาครรัตน์ นักปราชญ์ และคัดนางค์ จามะริก. (2016). การเปิดเผยข้อมูลภาครัฐในรูปแบบ Business Intelligence (BI) ในยุค Big Data. *วารสารวิชาการ กสทช.*, 1(1), 553-583.