การเปรียบเทียบวิธีประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติก เมื่อข้อมูลไม่สมดุล

สุดารัตน์ บุญธรรม

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ
ปีการศึกษา 2566

การเปรียบเทียบวิธีประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติก เมื่อข้อมูลไม่สมดุล

สุดารัตน์ บุญธรรม

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ
ปีการศึกษา 2566

วิทยานิพนธ์ เรื่อง "การเปรียบเทียบวิธีประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติก เมื่อข้อมูลไม่สมดุล"

ของ สุดารัตน์ บุญธรรม

ได้รับการพิจารณาให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติ

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการสอบวิทยานิพนธ์

(ศาสตราจารย์ ดร.เสาวณิต สุขภารังษี)

..... ประธานที่ปรึกษาวิทยานิพนธ์

(รองศาสตราจารย์ ดร.เกตุจันทร์ จำปาไชยศรี)

..... กรรมการผู้ทรงคุณวุฒิภายใน

(รองศาสตราจารย์ ดร.อนามัย นาอุดม)

อนุมัติ

.....

(รองศาสตราจารย์ ดร.กรองกาญจน์ ชูทิพย์)

คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การเปรียบเทียบวิธีประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติกเมื่อข้อมูลไม่สมดุล

ผู้วิจัย สุดารัตน์ บุญธรรม

ประธานที่ปรึกษา รองศาสตราจารย์ ดร.เกตุจันทร์ จำปาไชยศรี

ประเภทสารนิพนธ์ วิทยานิพนธ์ วท.ม. สถิติ, มหาวิทยาลัยนเรศวร, 2566

คำสำคัญ การถดถอยลอจิสติก ข้อมูลไม่สมดุล วิธีภาวะน่าจะเป็นสูงสุด วิธีฟังก์ชันสกอร์ที่ปรับปรุง วิธีเบส์เซียน

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบถดถอยลอจิสติก เมื่อประมาณค่าพารามิเตอร์ในตัวแบบ 3 วิธี ได้แก่ วิธีภาวะน่าจะเป็นสูงสุด (MLE) วิธีเบส์เซียน และวิธีฟังก์ชันสกอร์ที่ปรับปรุง (SCORE) ร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มลด (RUS) วิธีการสุ่มเกิน (ROS) และวิธีการสังเคราะห์ข้อมูลใหม่ (SMOTE) กำหนดขนาดตัวอย่างที่ใช้ในการศึกษาเท่ากับ 100 และ 500 จำนวนตัวแปรอิสระเท่ากับ 1 และ 3 ตัว อัตราส่วนความไม่สมดุลของข้อมูลในกลุ่ม 0 และ 1 เป็น 60:40, 70:30, 80:20 และ 90:10 ตามลำดับ และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30 และ 80:20 ทำการจำลองข้อมูลโดยกระทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์ที่กำหนด ใช้ค่าความแม่นยำ ความไว ความจำเพาะ และค่าความแม่นที่สมดุล เป็นเกณฑ์ในการเปรียบเทียบ

ผลการวิจัยพบว่า ในทุกระดับอัตราส่วนความไม่สมดุลของข้อมูล ขนาดตัวอย่าง และอัตราส่วนระหว่าง Training : Validation เมื่อพิจารณาความแม่นยำของการจำแนกกลุ่มถูกต้อง พบว่ากรณีที่มีตัวแปรอิสระ 1 ตัว วิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE และกรณีที่มีตัวแปรอิสระ 3 ตัว วิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE ให้ความแม่นยำสูงสุด เมื่อพิจารณาค่าความไว พบว่ากรณีที่มีตัวแปรอิสระ 1 ตัว วิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS หรือ ROS และกรณีที่มีตัวแปรอิสระ 3 ตัว วิธีเบส์เซียน กรณีทราบความรู้ก่อนหน้าร่วมกับ RUS หรือ ROS มีประสิทธิภาพสูงสุดเป็นส่วนใหญ่และมีค่าใกล้เคียงกัน เมื่อพิจารณาความจำเพาะ พบว่ากรณีที่มีตัวแปรอิสระ 1 ตัว วิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE และกรณีที่มีตัวแปรอิสระ 3 ตัว วิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุดเป็นส่วนใหญ่ และเมื่อพิจารณาค่าความแม่นที่สมดุล กรณีที่มีตัวแปรอิสระ 1 และ 3 ตัว พบว่าวิธีภาวะน่าจะเป็นสูงสุดและวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ RUS, ROS และ SMOTE ให้ความแม่นยำที่สมดุลสูงสุดใกล้เคียงกัน

นอกจากนี้ยังพบว่าเมื่ออัตราส่วนความไม่สมดุลของข้อมูลเพิ่มขึ้น ค่าความแม่นยำของการจำแนกกลุ่มถูกต้องด้วยวิธีภาวะน่าจะเป็นสูงสุดและวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ RUS, ROS และ SMOTE มีแนวโน้มลดลงเป็นส่วนใหญ่

Title A COMPARISON OF PARAMETER ESTIMATION METHODS IN LOGISTIC REGRESSION WITH UNBALANCED DATA

Author Sudarut Boontam

Advisor Associate Professor Katechan Jampachaisri, Ph.D.

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

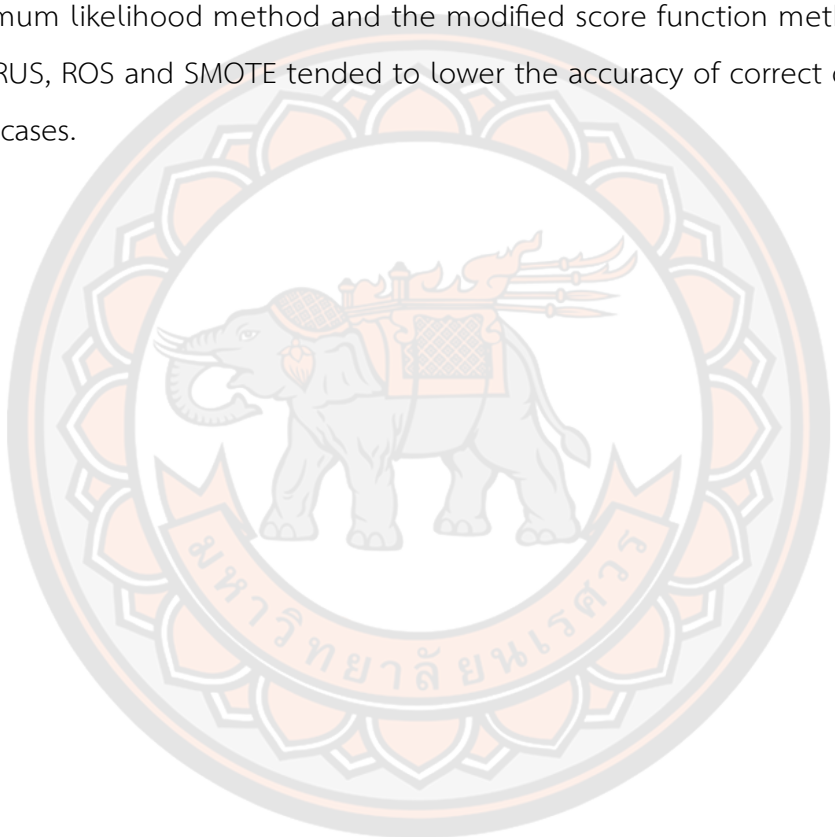
Keywords Logistic regression model Unbalanced data Maximum likelihood method Modified score function method Bayesian method

ABSTRACT

The objective of this research is to study and compare predictive efficiency of logistic regression using three parameter estimation methods: Maximum likelihood method (MLE), Bayesian method and Modified score function method (SCORE) in combination with unbalanced data handling using Random Under-Sampling (RUS), Random Over-Sampling (ROS), and Synthetic and Minority Over-sampling (SMOTE) techniques. The study is performed on 2 levels of sample size: 100 and 500, with one and three predictors. The ratios of unbalanced data for group 0 and 1 are 60:40, 70:30, 80:20 and 90:10 respectively. The ratios between Training : Validation are 70:30 and 80:20. In each situation, the simulation is conducted iteratively 1,000 times. The criteria for comparison are accuracy, sensitivity, specificity, and balanced accuracy.

The research results revealed that, for all ratios of unbalanced data, sample sizes and ratios between Training : Validation, when considering the accuracy of correct classification, it was found that, for one predictor, the Bayesian method with Informative prior in combination with SMOTE and, for three predictors, the modified score function method combined with SMOTE yielded the highest accuracy. When considering the sensitivity, for one predictor, the maximum likelihood method combined with RUS or ROS and, for three predictors, the Bayesian method with Informative prior combined with RUS or ROS are the most efficient in most cases and

provide similar values. When considering the specificity, for one predictor, the Bayesian method with Informative prior combined with SMOTE and, for three predictors, the modified score function method combined with SMOTE are the most efficient in most cases. When considering the balanced accuracy, for one and three predictors, the maximum likelihood method and the modified score function method combined with RUS, ROS and SMOTE yielded the highest balanced accuracy with similar values. In addition, as the ratio of unbalanced data increases, the maximum likelihood method and the modified score function method combined with RUS, ROS and SMOTE tended to lower the accuracy of correct classification in most cases.



ประกาศคุณูปการ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความกรุณาอย่างยิ่งจาก รองศาสตราจารย์ ดร. เกตุ จันทร์ จำปาไชยศรี ประธานที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำปรึกษา แนะนำ แก้ไขข้อบกพร่องตลอดระยะเวลาในการทำวิทยานิพนธ์ด้วยความเอาใจใส่อย่างยิ่ง จนวิทยานิพนธ์สำเร็จโดยสมบรูณ์ ผู้วิจัยรู้สึกทราบซึ้ง ตระหนักในพระคุณยิ่ง และขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.อนามัย นาอุดม และผู้ช่วยศาสตราจารย์ ดร. กัลยา บุญหล้า ที่ได้กรุณาให้คำแนะนำและข้อเสนอแนะเพิ่มเติม ซึ่งเป็นประโยชน์อย่างยิ่งต่อการทำวิทยานิพนธ์ และขอขอบพระคุณอาจารย์ทุกท่านที่ประสิทธิประสาทวิชาความรู้ แนวคิด และให้คำแนะนำต่างๆ เป็นอย่างดี ตลอดระยะเวลาที่ได้ศึกษาในมหาวิทยาลัยนเรศวร

ขอกราบขอบพระคุณ ดร.ดาริกา แย้มรับบุญ และ ดร.ทิพย์วัลย์ เกตุอินทร์ ที่ให้ความช่วยเหลือด้านการเขียนโปรแกรมคอมพิวเตอร์ (R Studio Program) และขอบคุณเพื่อนๆ และน้องๆ สาขาสถิติ ที่คอยให้คำแนะนำและเป็นกำลังใจให้ผู้วิจัยเสมอมา

ขอกราบขอบพระคุณ คุณพ่อ คุณแม่ และบูรพาจารย์ที่ให้การอบรมสั่งสอน ให้การศึกษา ให้ความรู้บังเกิดปัญญา ซึ่งมีส่วนทำให้เกิดความสำเร็จในครั้งนี้ ตลอดจนพี่ๆ น้องๆ และญาติๆ ของผู้วิจัยทุกท่าน ที่คอยสนับสนุน ส่งเสริม และเป็นกำลังใจที่ดีแก่ผู้วิจัยตลอดจนสำเร็จการศึกษา

สุดารัตน์ บุญธรรม

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์ถดถอยลอจิสติก (Logistic Regression Analysis) เป็นเทคนิคการวิเคราะห์ตัวแปรเชิงพหุ โดยมีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตาม (Dependent Variable) และตัวแปรอิสระ (Independent Variable) และนำสมการถดถอยลอจิสติกที่ได้ไปประมาณค่าหรือพยากรณ์ค่าตัวแปรตามเมื่อกำหนดค่าตัวแปรอิสระ ซึ่งตัวแบบถดถอยลอจิสติกประกอบไปด้วย ตัวแปรตาม (หรือตัวแปรเกณฑ์) เป็นตัวแปรเชิงกลุ่ม (Categorical Variable) หรือตัวแปรชนิดไม่ต่อเนื่อง และตัวแปรอิสระ (หรือตัวแปรทำนาย) สามารถเป็นตัวแปรเชิงกลุ่มหรือตัวแปรเชิงปริมาณ (Quantitative Variable) ก็ได้ (กาญจน์เขจร ชูชีพ, 2561) การวิเคราะห์ถดถอยลอจิสติกแบ่งออกเป็น 2 ประเภท คือ การถดถอยลอจิสติกแบบทวิภาค (Binary Logistic) เมื่อตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่าเพียง 2 ค่า (Dichotomous Variable) เช่น “ใช่” กับ “ไม่ใช่” หรือ “เสี่ยง” กับ “ไม่เสี่ยง” และการถดถอยลอจิสติกแบบพหุ (Multinomial Logistic) เมื่อตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่ามากกว่า 2 ค่า เช่น ทางการแพทย์ กำหนดให้ตัวแปรตามคือระยะของการเป็นโรคมะเร็ง มี 4 ระดับ คือ ระยะที่ 1 ระยะที่ 2 ระยะที่ 3 และระยะที่ 4 (กัลยา วานิชย์บัญชา, 2562) ในปัจจุบันการวิเคราะห์ถดถอยลอจิสติกเป็นวิธีการทางสถิติที่ได้รับความนิยมอย่างแพร่หลาย ใช้สำหรับปัญหาการจำแนก (Classification) แต่เมื่อตัวแปรตามเป็นตัวแปรแบบทวิภาค (Binary) มักจะเกิดปัญหาข้อมูลไม่สมดุล (Unbalanced Data) เช่น การศึกษาข้อมูลการผ่าตัดหัวใจที่ได้จากโรงพยาบาลท้องถิ่นในกรุงกัวลาลัมเปอร์ โดยตัวแปรตามเป็นแบบทวิภาค กำหนด 1 แทน เสียชีวิตหลังการผ่าตัด และ 0 แทน มีชีวิตหลังการผ่าตัด มีกลุ่มตัวอย่างผู้ป่วยทั้งหมด 4,976 ราย พบว่ามีผู้ป่วย 4,767 ราย (95.8%) มีชีวิตหลังการผ่าตัด และผู้ป่วย 209 ราย (4.2%) เสียชีวิตหลังการผ่าตัด (Wah, et al., 2016) จะเห็นได้ว่าข้อมูลมีความไม่สมดุลอย่างมาก ซึ่งข้อมูลไม่สมดุลหมายถึงข้อมูลกลุ่มหนึ่งมีจำนวนมากกว่าอีกกลุ่มหนึ่ง นั่นคือ กลุ่มส่วนใหญ่ (Majority Classes) จะมีจำนวนข้อมูลมากกว่า ในขณะที่กลุ่มส่วนน้อย (Minority Classes) จะมีจำนวนข้อมูลน้อยกว่า (อัจฉรา แผ้วบาง และ สายชล สินสมบูรณ์ทอง, 2562) อาจเกิดจากลักษณะของข้อมูลที่มีความแตกต่างกัน หรือเกิดจากการเก็บรวบรวมข้อมูล โดยเฉพาะข้อมูลทางการแพทย์ เช่น จำนวนผู้ติดเชื้อโควิด-19 จำนวน 4.52 ล้านคน จากประชากรไทยทั้งหมด 66 ล้านคน (ระบบสถิติทางการทะเบียน, 2565) หรือผู้ป่วยที่เข้ามารับการรักษาในโรงพยาบาลมีจำนวนมาก แต่ผู้ป่วยที่ถูกวินิจฉัยว่าเป็นโรคมะเร็งมีจำนวนน้อยเมื่อเทียบกับ

จำนวนผู้ป่วยทั้งหมด หรือบางโรคที่อาจพบได้ยากทางด้านการแพทย์อื่น ๆ (วิชญ์วิสิฐ เกสรสิทธิ์ และคณะ, 2561) เมื่อนำข้อมูลที่มีความไม่สมดุลไปประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติกจะทำให้ค่าพารามิเตอร์ที่ได้มีความเอนเอียง (Bias) ซึ่งค่าประมาณพารามิเตอร์ที่ได้จะไม่เข้าใกล้ค่าพารามิเตอร์ที่แท้จริง (Hezlin et.al, 2021) ดังนั้นการนำข้อมูลที่มีความไม่สมดุลไปวิเคราะห์หรือพยากรณ์จะส่งผลให้ประสิทธิภาพการจำแนกกลุ่มมีความเอนเอียง สามารถพยากรณ์กลุ่มส่วนใหญ่ได้อย่างถูกต้องและแม่นยำ แต่พยากรณ์กลุ่มส่วนน้อยได้ผิดพลาดและไม่แม่นยำ ซึ่งจะนำไปสู่ปัญหาที่เรียกว่า ปัญหาการแบ่งกลุ่มผิด (Misclassification) ตัวอย่างเช่น การศึกษาตรวจจับการฉ้อโกงบัตรเครดิต (Brandt & Lanzén, 2020) จากการทำธุรกรรมจำนวน 284,807 ครั้ง มีการฉ้อโกงเกิดขึ้นทั้งหมด 492 ครั้ง คิดเป็นร้อยละ 0.17 ของการทำธุรกรรมทั้งหมด ซึ่งทำให้ข้อมูลชุดนี้มีความไม่สมดุลอย่างมาก เมื่อทำการพยากรณ์ จะส่งผลให้ค่าพยากรณ์เอนเอียงไปยังกลุ่มส่วนใหญ่ คือไม่เกิดการฉ้อโกง และส่งผลเสียแก่บริษัทรับชำระเงินผ่านบัตรเครดิตที่ไม่สามารถตรวจสอบได้ว่าธุรกรรมแต่ละครั้งนั้นเป็นธุรกรรมที่ฉ้อโกงหรือไม่

จากปัญหาดังกล่าวได้มีการศึกษาคิดค้นวิธีในการแก้ปัญหาความไม่สมดุลของข้อมูลให้มีความสมดุลก่อนดำเนินการสร้างตัวแบบจำลอง (กีระชาติ สุขสุทธิ์, 2559) ซึ่งวิธีที่นิยมใช้คือการสุ่มข้อมูลของกลุ่มส่วนน้อยเพิ่มขึ้นให้มีจำนวนใกล้เคียงหรือเท่ากับข้อมูลในกลุ่มส่วนใหญ่ เรียกว่า วิธีการสุ่มเกิน (Random Over-Sampling) หรือสุ่มลดข้อมูลของกลุ่มส่วนใหญ่ให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มส่วนน้อย เรียกว่า วิธีการสุ่มลด (Random Under-Sampling) (เบญจภรณ์ จันทรกองกุล และคณะ, 2557) เช่น การศึกษาเทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่สมดุลในการรักษาภาวะข้อเข่าเสื่อมในผู้สูงอายุ (พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตกุล, 2561) และการแก้ปัญหาข้อมูลไม่สมดุลสำหรับจำแนกผู้ป่วยโรคเบาหวาน (วิชญ์วิสิฐ เกษรสิทธิ์, 2561) เป็นต้น

จากเหตุผลดังกล่าว ผู้วิจัยจึงสนใจศึกษา วิธีประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติกร่วมกับการจัดการข้อมูลไม่สมดุล เพื่อเปรียบเทียบประสิทธิภาพในการพยากรณ์หรือจำแนกข้อมูลได้ถูกต้องมากยิ่งขึ้น

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า สุภวรรณ มานะการ (2549) ได้ศึกษาการจำแนกกลุ่มโดยใช้วิธีถดถอยลอจิสติก ประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Method) และวิธีนาอีฟ เบส์ (Naïve Bayes) ซึ่งการวิเคราะห์ข้อมูลจะแบ่งเป็น 2 ชุด คือชุดตัวอย่างที่ใช้สร้างเกณฑ์การจำแนก และชุดตัวอย่างที่ใช้ตรวจสอบเกณฑ์การจำแนก โดยทำการเปรียบเทียบในอัตราส่วนระหว่างชุดตัวอย่างที่ใช้สร้างเกณฑ์การจำแนกต่อชุดตัวอย่างที่ใช้ตรวจสอบ

เกณฑ์การจำแนก เป็น 90:10, 80:20, 70:30, 60:40 และ 50:50 จำลองข้อมูลโดยใช้โปรแกรม MATLAB กระทำซ้ำ 1000 ครั้งในแต่ละสถานการณ์ โดยเกณฑ์การเปรียบเทียบคือ อัตราความผิดพลาดของการจำแนก (Error rate of misclassification) ผลการศึกษาพบว่า อัตราการจำแนกกลุ่มข้อมูลผิดพลาดทั้งสองวิธีมีแนวโน้มลดลง เมื่อขนาดตัวอย่างและอัตราส่วนของชุดตัวอย่างที่ใช้สร้างเกณฑ์การจำแนก (Training Data) เพิ่มขึ้น โดยอัตราการจำแนกกลุ่มข้อมูลผิดผลาดด้วยวิธีนาอีฟ เบส์ (Naïve Bayes) ส่วนใหญ่แล้วจะให้ค่าต่ำกว่าวิธีการถดถอยลอจิสติก Wah, et al. (2016) ได้ศึกษาเปรียบเทียบประสิทธิภาพของวิธีการจัดการข้อมูลไม่สมดุลด้วยเทคนิคการสุ่มตัวอย่างเกินและเทคนิคการสุ่มตัวอย่างลดร่วมกับวิธีเวกเตอร์ค้ำยัน (Support Vector Machine) วิธีเพื่อนบ้านที่ใกล้ที่สุด (k – Nearest Neighbours Algorithm) และวิธีการถดถอยลอจิสติก (Logistic Regression) ในการศึกษานี้ใช้ชุดข้อมูลการผ่าตัดหัวใจที่ได้จากโรงพยาบาลท้องถิ่นในกรุงกัวลาลัมเปอร์ โดยตัวแปรตามเป็นไบนารี่ กำหนด 1 แทน เสียชีวิตหลังการผ่าตัด และ 0 แทน มีชีวิตหลังการผ่าตัด จำนวนตัวแปรอิสระ 8 ตัว กลุ่มตัวอย่างผู้ป่วยทั้งหมด 4,976 ราย โดยผู้ป่วย 4,767 ราย (95.8%) มีชีวิตหลังการผ่าตัด และผู้ป่วยเพียง 209 ราย (4.2%) ที่เสียชีวิตหลังการผ่าตัด ซึ่งแสดงให้เห็นว่าข้อมูลการผ่าตัดหัวใจมีความไม่สมดุล ใช้เกณฑ์การวัดประสิทธิภาพคือ ค่าความแม่นยำ (Accuracy) ค่าความไว (Sensitivity) และ ค่าความจำเพาะ (Specificity) ผลการศึกษาพบว่าค่าความไวเพิ่มขึ้นเมื่อใช้วิธีจำแนกกลุ่มทั้ง 3 วิธี ร่วมกับเทคนิคการสุ่มตัวอย่างทั้งสองวิธี นอกจากนี้เทคนิคการสุ่มตัวอย่างเกิน มีประสิทธิภาพดี เมื่อใช้ร่วมกับวิธีเวกเตอร์ค้ำยันและวิธีเพื่อนบ้านที่ใกล้ที่สุด Febrianti, et al. (2018) ได้ศึกษาเปรียบเทียบประสิทธิภาพการประมาณค่าพารามิเตอร์ของตัวแบบถดถอยลอจิสติก โดยทำการเปรียบเทียบ 2 วิธี คือวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) และวิธีฟังก์ชันสกอร์ที่ปรับปรุง (Modified Score Function) ทำการศึกษาในชุดข้อมูลจริง เมื่อตัวอย่างมีขนาดเล็กและสัดส่วนของเหตุการณ์ที่สนใจน้อย ซึ่งพบปัญหาของวิธีภาวะน่าจะเป็นสูงสุดที่ไม่สามารถประมาณค่าพารามิเตอร์ได้ เนื่องจากกระบวนการวนซ้ำไม่ให้ผลลัพธ์ที่บรรจบกัน วิธีที่สามารถแก้ไขปัญหาดังกล่าวได้คือวิธีฟังก์ชันสกอร์ที่ปรับปรุง ทำให้กระบวนการวนซ้ำบรรจบกันและให้ค่าประมาณพารามิเตอร์ได้อย่างรวดเร็ว Hassan (2020) ได้ศึกษาเปรียบเทียบประสิทธิภาพการประมาณค่าพารามิเตอร์ของการถดถอยลอจิสติกด้วยวิธีแบบเบส์และวิธีอื่นอีก 5 วิธี ได้แก่ วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีเวกเตอร์ค้ำยัน (Support Vector Machine SVM) วิธีเพื่อนบ้านที่ใกล้ที่สุด (k – Nearest Neighbours Algorithm) วิธีการถดถอยลอจิสติก (Logistic Regression) และวิธีนาอีฟ เบส์ ในชุดข้อมูลผู้ป่วยเบาหวาน ในเมือง Zakho โดยกำหนดตัวแปรตามแบบไบนารี่ ให้ 1 แทน เป็นเบาหวาน และ 0 แทน สุขภาพดี มีตัวแปรอิสระ 7 ตัว การประมาณค่า

ด้วยวิธีแบบเบส์ กำหนดการแจกแจงความน่าจะเป็นก่อน คือ Gaussian, Laplace และ Cauchy ใช้การจำลอง Markov Chain Monte Carlo (MCMC) โดยการใช้ตัวสุ่มตัวอย่างด้วยวิธี Gibbs Sampling เพื่อประมาณการแจกแจงความน่าจะเป็นภายหลัง โดยเกณฑ์วัดประสิทธิภาพของแบบจำลองได้แก่ ค่าความแม่นยำ (Accuracy) ค่าความเที่ยง (Precision) ค่าการเรียกคืน (Recall) และค่าประสิทธิภาพ (F-measure) ผลการศึกษาพบว่า วิธีแบบเบส์กับการแจกแจงความน่าจะเป็นแบบเกาส์เซียนมีประสิทธิภาพสูงสุด และสามารถจำแนกผู้ป่วยได้ถูกต้องเกือบทั้งหมด Brandt and Lanzén (2020) ได้ศึกษาเปรียบเทียบประสิทธิภาพของการจำแนกกลุ่ม เมื่อข้อมูลไม่สมดุล โดยใช้เทคนิคการสุ่มตัวอย่าง 2 เทคนิค คือ Synthetic Minority Over – Sampling Technique (SMOTE) และ Adaptive Synthetic Sampling Approach (ADASYN) ใช้ตัวแบบที่แตกต่างกัน 3 แบบ คือ การถดถอยลอจิสติก (Logistic Regression) เทคนิคป่าสุ่ม (Random Forest Classifier) และวิธีเวกเตอร์ค้ำยัน (Support Vector Machines) เปรียบเทียบกับชุดข้อมูลไม่สมดุล 3 ชุด จากฐานข้อมูล Kaggle.com ที่มีระดับความไม่สมดุลที่แตกต่างกัน สำหรับเกณฑ์การเปรียบเทียบ ได้แก่ ค่าความไว (Sensitivity) ค่าประสิทธิภาพ (F-measure) และสัมประสิทธิ์สหสัมพันธ์ของแมทธิว (Matthews correlation coefficient : MCC) ผลการศึกษาพบว่า สำหรับเกณฑ์การเปรียบเทียบทั้ง 3 เกณฑ์ ไม่มีเทคนิคการสุ่มตัวอย่างร่วมกับตัวแบบที่สามารถปรับปรุงประสิทธิภาพของชุดข้อมูลทั้ง 3 ชุด ได้อย่างสม่ำเสมอ แต่อย่างไรก็ตาม ผลลัพธ์แสดงให้เห็นว่าการใช้เทคนิคการสุ่มตัวอย่าง SMOTE ช่วยปรับปรุงประสิทธิภาพของวิธีเวกเตอร์ค้ำยัน ได้เป็นส่วนใหญ่ โดยเฉพาะเมื่อระดับความไม่สมดุลเพิ่มขึ้น นอกจากนี้เทคนิคการสุ่มตัวอย่างทั้ง 2 เทคนิค สามารถปรับปรุงประสิทธิภาพของเทคนิคป่าสุ่มได้ เมื่อระดับความไม่สมดุลเพิ่มขึ้น Yilmaz and Celik (2021) ได้ศึกษาเปรียบเทียบประสิทธิภาพการประมาณค่าพารามิเตอร์ของแบบจำลองถดถอยลอจิสติกด้วยวิธีภาวะน่าจะเป็นสูงสุดและวิธีแบบเบส์ กับชุดข้อมูลองค์การเพื่อความร่วมมือทางเศรษฐกิจและการพัฒนา (OECD) ประกอบไปด้วยข้อมูลประชากรและเศรษฐกิจจาก 34 ประเทศที่เป็นสมาชิก OECD โดยตัวแปรตามคือ การเป็นสมาชิกสหภาพยุโรป (EU) กำหนด 1 แทน การเป็นสมาชิก และ 0 แทน การไม่เป็นสมาชิก มีตัวแปรอิสระ 9 ตัว เกณฑ์การวัดประสิทธิภาพของแบบจำลอง ได้แก่ AIC และ BIC ผลการศึกษาพบว่าวิธีแบบเบส์ มีสัดส่วนการจำแนกกลุ่มถูกต้องสูงกว่าวิธีภาวะน่าจะเป็นสูงสุด และเมื่อตัวอย่างมีขนาดเล็กการประมาณค่าพารามิเตอร์ด้วยวิธีแบบเบส์มีประสิทธิภาพดีกว่าวิธีภาวะน่าจะเป็นสูงสุด

จากการศึกษางานวิจัยที่กล่าวมาข้างต้นจะเห็นได้ว่า เมื่อข้อมูลไม่สมดุล วิธีจัดการความไม่สมดุลด้วยวิธีการสุ่มลด (Random Under – Sampling) วิธีการสุ่มเกิน (Random Over – Sampling) และวิธีสังเคราะห์ข้อมูลใหม่ (Synthetic Minority Over – Sampling Technique :

SMOTE) เป็นวิธีที่มีประสิทธิภาพในการจัดการชุดข้อมูลที่มีความไม่สมดุล ดังนั้นผู้วิจัยจึงสนใจนำวิธีจัดการข้อมูลไม่สมดุลทั้ง 3 วิธี ข้างต้น มาใช้ร่วมกับการประมาณค่าพารามิเตอร์ในการวิเคราะห์การถดถอยลอจิสติกด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Method) วิธีเบส์เซียน (Bayesian Method) และวิธีฟังก์ชันสกอร์ที่ปรับปรุง (Modified Score Function)

1.2 วัตุประสงค์ของการวิจัย

เพื่อศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของตัวแบบถดถอยลอจิสติกเมื่อประมาณค่าพารามิเตอร์ด้วย วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Method) วิธีเบส์เซียน (Bayesian method) และวิธีฟังก์ชันสกอร์ที่ปรับปรุง (Modified Score Function) ร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มลด (Random Under-Sampling) วิธีการสุ่มเกิน (Random Over-Sampling) และวิธีการสังเคราะห์ข้อมูลใหม่ (Synthetic Minority Over-sampling Technique)

1.3 ขอบเขตของงานวิจัย

ผู้วิจัยได้กำหนดขอบเขตของงานวิจัยดังนี้

1. กำหนดระดับความไม่สมดุลของข้อมูล 2 กลุ่ม ดังนี้ กลุ่ม 0 : กลุ่ม 1 เป็น 90:10, 80:20, 70:30 และ 60:40
2. กำหนดขนาดตัวอย่าง (n) ที่ใช้ในการศึกษาเท่ากับ คือ 100 และ 500
3. กำหนดจำนวนตัวแปรอิสระ (p) ที่ใช้ในการศึกษา คือ 1 และ 3 ตัว
4. กำหนดตัวแปรอิสระเป็นข้อมูลเชิงปริมาณที่มีการแจกแจงดังนี้
   - การแจกแจงแบบปรกติหลายตัวแปร (Multivariate Normal Distribution) โดยมีฟังก์ชันความหนาแน่นน่าจะเป็น ดังนี้

$$f(\mathbf{y};\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{\frac{p}{2}}\left|\Sigma\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})};-\infty<\mathbf{y}<\infty,-\infty<\boldsymbol{\mu}<\infty,\Sigma>0$$

เมื่อ เวกเตอร์ค่าเฉลี่ยของตัวแปรสุ่ม Y คือ $E(Y)=\mu=\begin{bmatrix}\mu_1\\ \vdots\\ \mu_p\end{bmatrix}$

เมทริกซ์ความแปรปรวนร่วมของตัวแปรสุ่ม Y คือ $Cov(Y)=\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{bmatrix}$

- ตัวแปรอิสระ 1 ตัวแปร

$\boldsymbol{\mu}=\mu_1=0$ และ $\Sigma=\sigma_{11}=1$

- ตัวแปรอิสระ 3 ตัวแปร

$$\boldsymbol{\mu}=\begin{bmatrix}0\\0\\0\end{bmatrix} \text{ และ } \Sigma=\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$$

5. กำหนดตัวแปรตามเป็นตัวแปรทวิภาคมีค่าเป็น 0 กับ 1 เมื่อกำหนดให้

   0 แทน เหตุการณ์ที่ไม่สนใจ

   1 แทน เหตุการณ์ที่สนใจ

   ซึ่งตัวแปรตามสร้างมาจากการแจกแจงเบอร์นูลลี

6. กำหนดความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจในประชากร (π) เป็น 0.1
7. กำหนดค่า β เริ่มต้นมีค่าเท่ากับ 1
   - ตัวแปรอิสระ 1 ตัวแปร

$$\beta=\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}$$

   - ตัวแปรอิสระ 3 ตัวแปร

$$\beta=\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\\\beta_3\end{bmatrix}=\begin{bmatrix}1\\1\\1\\1\end{bmatrix}$$

8. การวิเคราะห์จะแบ่งตัวอย่างออกเป็น 2 ชุด คือ ชุดตัวอย่างที่ใช้สร้างเกณฑ์การจำแนก และชุดตัวอย่างที่ใช้ตรวจสอบเกณฑ์การจำแนก คือ 70:30 และ 80:20
9. กำหนดฟังก์ชันความหนาแน่นน่าจะเป็นก่อน (Prior Probability Density Function) ในวิธีเบส์เซียน

   9.1 กรณีไม่ทราบความรู้เดิมเกี่ยวกับพารามิเตอร์ β ดังนี้

$$f(\beta)\propto c \quad ;-\infty<\beta<\infty$$ เมื่อ c เป็นค่าคงที่

   9.2 กรณีทราบความรู้เดิมเกี่ยวกับพารามิเตอร์ β ดังนี้

   - สำหรับตัวแปรอิสระ 1 ตัว

$$\beta\sim N_{n\times(p+1)}(\mu,\Sigma)$$ เมื่อ $\mu=\begin{bmatrix}0\\0\end{bmatrix}$ และ $\Sigma=\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$

- สำหรับตัวแปรอิสระ 3 ตัว

$$\beta \sim N_{n\times(p+1)}(\mu,\Sigma) \quad \text{เมื่อ} \quad \mu=\begin{bmatrix}0\\0\\0\\0\end{bmatrix} \quad \text{และ} \quad \Sigma=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

10. จำลองข้อมูลโดยใช้โปรแกรม R ทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ทราบประสิทธิภาพการจำแนกกลุ่มของตัวแบบถดถอยลอจิสติกเมื่อประมาณพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด วิธีเบส์เซียน และวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับการจัดการข้อมูลไม่สมดุล
2. เป็นแนวทางในการนำตัวแบบไปประยุกต์ใช้กับข้อมูลจริง

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของการพยากรณ์ในตัวแบบถดถอยลอจิสติก เมื่อประมาณค่าพารามิเตอร์ ด้วยวิธีภาวะน่าจะเป็นสูงสุด วิธีเบส์เซียน และวิธีฟังก์ชันสกอร์ที่ปรับปรุง ร่วมกับการจัดการข้อมูลไม่สมดุล โดยมีเอกสารและงานวิจัยที่เกี่ยวข้องแบ่งหัวข้อตามลำดับการนำเสนอดังนี้

1. ฟังก์ชันความหนาแน่นน่าจะเป็น
2. ตัวแบบถดถอยลอจิสติก
3. ข้อมูลไม่สมดุล และวิธีจัดการข้อมูลไม่สมดุล
4. ระเบียบวิธีของนิวตัน-ราฟสัน
5. ฟังก์ชันภาวะน่าจะเป็น
6. วิธีภาวะน่าจะเป็นสูงสุด
7. วิธีปรับเปลี่ยนฟังก์ชันคะแนน
8. แนวคิดของการประมาณค่าพารามิเตอร์แบบเบส์
9. การสุ่มตัวอย่างแบบกิบส์
10. การประมาณค่าด้วยวิธีเบส์เซียน
11. เกณฑ์การตัดสินใจ
12. งานวิจัยที่เกี่ยวข้อง

2.1 ฟังก์ชันความหนาแน่นน่าจะเป็น (Probability density function)

ฟังก์ชันความหนาแน่นน่าจะเป็นของตัวแปรสุ่มที่เกี่ยวข้องมี ดังนี้

1. การแจกแจงแบบเบอร์นูลลี (Bernoulli distribution)

ในการทดลองสุ่มเมื่อเกิดเหตุการณ์ที่สนใจจะเรียกว่า ประสบผลสำเร็จ (Success) แต่ถ้าเกิดเหตุการณ์ที่ไม่สนใจจะเรียกว่า ล้มเหลว (Failure)

ถ้าให้ Y เป็นตัวแปรสุ่มที่สัมพันธ์กับการทดลองเบอร์นูลลี ดังนั้นเรนจ์สเปสของ Y จะมีค่าที่เป็นไปได้ 2 ค่า เท่านั้น ในการทดลองหนึ่งๆ ซึ่งมีความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจเท่ากับ π และความน่าจะเป็นของการเกิดเหตุการณ์ที่ไม่สนใจเท่ากับ $1-\pi$ เช่น การเสี่ยงเป็นโรคมะเร็งลำไส้ใหญ่ (ถ้าเสี่ยงเป็นโรคมะเร็งลำไส้ใหญ่ กำหนดให้ $Y=1$ และไม่เสี่ยงเป็นโรคมะเร็งลำไส้ใหญ่ กำหนดให้ $Y=0$) หรือเพศของทารกที่กำลังจะคลอด (ถ้าเป็นเพศหญิง กำหนดให้ $Y=1$

และเพศชาย กำหนดให้ $Y=0$) เป็นต้น ให้ Y เป็นตัวแปรสุ่มที่มีการแจกแจงแบบเบอร์นูลลี โดยมีพารามิเตอร์ π เขียนแทนด้วย $Y \sim Ber(\pi)$ ซึ่งมีฟังก์ชันความหนาแน่นน่าจะเป็น ดังนี้

$$f(y;\pi)=\pi^{y}(1-\pi)^{1-y}\ ;\ y=0,1,\ 0\le\pi\le 1$$

โดย ค่าเฉลี่ยของตัวแปรสุ่ม Y คือ $E(Y)=\pi$

และความแปรปรวนของตัวแปรสุ่ม Y คือ $Var(Y)=\pi(1-\pi)$

2. การแจกแจงเอกรูป (Uniform distribution)

ให้ Y เป็นตัวแปรสุ่มที่มีการแจกแจงแบบเอกรูป เขียนแทนด้วย $Y \sim U(a,b)$ มีฟังก์ชันความหนาแน่นน่าจะเป็น ดังนี้

$$f(y;a,b)=\frac{1}{b-a}\ ;a<y<b$$

โดย ค่าเฉลี่ยของตัวแปรสุ่ม Y คือ $\mathrm{E}(Y)=\frac{a+b}{2}$

และ ความแปรปรวนของตัวแปรสุ่ม Y คือ $Var(Y)=\frac{(b-a)^2}{12}$

3. การแจกแจงปรกติ (Normal distribution)

ให้ Y เป็นตัวแปรสุ่มชนิดต่อเนื่อง มีการแจกแจงแบบปรกติ ด้วยค่าเฉลี่ย μ และความแปรปรวน σ^2 เขียนแทนด้วย $Y \sim N(\mu,\sigma^2)$ ซึ่งมีฟังก์ชันความหนาแน่นน่าจะเป็น ดังนี้

$$f(y;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y-\mu)^2}\ ;-\infty<y<\infty\ ;-\infty<\mu<\infty\ ;\sigma^2>0$$

โดย ค่าเฉลี่ยของตัวแปรสุ่ม Y คือ $E(Y)=\mu$

และ ความแปรปรวนของตัวแปรสุ่ม Y คือ $Var(Y)=\sigma^2$

4. การแจกแจงปรกติหลายตัวแปร (Multivariate Normal distribution)

เวกเตอร์ของตัวแปรสุ่มชนิดต่อเนื่อง $\boldsymbol{Y}$ มีขนาด $n\times 1$ มีการแจกแจงแบบปรกติหลายตัวแปร โดยที่ $\boldsymbol{\mu}$ แทน เวกเตอร์ค่าเฉลี่ยขนาด $n\times 1$ และ $\sum$ แทน เมทริกซ์ความแปรปรวนร่วม (Covariance matrix) ขนาด $n\times n$ เขียนแทนด้วย $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu},\sum)$ ซึ่งมีฟังก์ชันความหนาแน่นน่าจะเป็น ดังนี้

$$f(\boldsymbol{y};\boldsymbol{\mu},\sum)=\frac{1}{(2\pi)^{\frac{n}{2}}\left|\sum\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})};-\infty<\boldsymbol{y}<\infty,-\infty<\boldsymbol{\mu}<\infty,\sum>0$$

โดย เวกเตอร์ค่าเฉลี่ยของตัวแปรสุ่ม Y คือ $E(Y)=\boldsymbol{\mu}$

และ เมทริกซ์ความแปรปรวนร่วมของตัวแปรสุ่ม Y คือ $Var(Y)=\sum$

2.2 ตัวแบบถดถอยลอจิสติก (Logistic regression model)

การวิเคราะห์การถดถอยลอจิสติก เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามที่มีลักษณะเป็นตัวแปรเชิงกลุ่ม สามารถแบ่งออกเป็น 2 ประเภท คือ ตัวแปรตามเป็นตัวแปรทวิภาค (Binary Logistic Regression) และตัวแปรตามที่มีค่ามากกว่า 2 ค่า (Multinomial Logistic Regression) (กัลยา วานิชย์บัญชา, 2544) ซึ่งในที่นี้จะกล่าวถึงการวิเคราะห์การถดถอยลอจิสติกที่ตัวแปรตามมีเพียง 2 ค่าเท่านั้น โดยตัวแปรตามมีการแจกแจงแบบเบอร์นูลลี เขียนแทนด้วย $Y \sim Ber\left(\pi\left(x_{ij}\right)\right)$ และ x_{ij} แทนเวกเตอร์ของตัวแปรอิสระเมื่อ $i = 1,2,3,...,n$ และ $j = 1,2,3,...,p$ ซึ่งความน่าเป็นที่จะเกิดเหตุการณ์ที่สนใจ คือ $\pi\left(x_{ij}\right) = P\left(Y = 1\right)$ นั่นคือ

$$Y = \begin{cases} 1 \text{ ; ด้วยความน่าจะเป็น } \pi(x_{ij}) \\ 0 \text{ ; ด้วยความน่าจะเป็น } 1-\pi(x_{ij}) \end{cases}$$

นั่นคือ

$$P(Y=y) = \pi(x_{ij})^{y}(1-\pi(x_{ij}))^{1-y} \text{ ; } y = 0,1 \tag{1}$$

ค่าเฉลี่ยของ Y คือ

$$\begin{aligned} \mathrm{E}(Y) &= \sum_{y_i=0}^{1} y_i P(Y=y_i) \\ &= 0 \cdot P(Y=0) + 1 \cdot P(Y=1) \\ &= 0 \cdot (1-\pi(x_{ij})) + 1 \cdot (\pi(x_{ij})) \\ &= \pi(x_{ij}) \end{aligned} \tag{2}$$

จะได้

$$\begin{aligned} \mathrm{E}(Y^2) &= \sum_{y_i=0}^{1} (y_i)^2 P(Y=y_i) \\ &= 0^2 \cdot P(Y=0) + 1^2 \cdot P(Y=1) \\ &= 0^2 \cdot (1-\pi(x_{ij})) + 1^2 \cdot (\pi(x_{ij})) \\ &= \pi(x_{ij}) \end{aligned} \tag{3}$$

ดังนั้น จะได้ความแปรปรวนของ Y คือ

$$\begin{aligned} Var(Y) &= \pi(x_{ij}) - [\pi(x_{ij})]^2 \\ &= \pi(x_{ij})[1-\pi(x_{ij})] \end{aligned} \tag{4}$$

เนื่องจากตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มี 2 ค่า คือ 0 และ 1 จึงทำให้ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระไม่ได้อยู่ในรูปเชิงเส้น แต่จะอยู่ในรูปแบบดังนี้

$$Y = \frac{e^{\boldsymbol{X}_{ij}^T \boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_{ij}^T \boldsymbol{\beta}}} \tag{5}$$

เมื่อ $\boldsymbol{\beta} = \left[\beta_0, \beta_1, ..., \beta_p\right]'$ แทน สัมประสิทธิ์ถดถอยของตัวแบบถดถอยลอจิสติก ซึ่งเป็นพารามิเตอร์ไม่ทราบค่า

เมื่อตัวแปรตามเป็นตัวแปรทวิภาค พบว่าความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระไม่เป็นเชิงเส้น จึงทำการแปลงความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจจากช่วง (0,1) ให้อยู่ในช่วง $(-\infty, \infty)$ โดยใช้การแปลงลอจิท (Logit Transformation) จะได้ตัวแบบถดถอยลอจิสติกเชิงเส้น ดังนี้

$$logit(\pi(x_{ij})) = \boldsymbol{X}_{ij}^T \boldsymbol{\beta} \tag{6}$$

โดยที่

$$logit(\pi(x_{ij})) = ln\left(\frac{\pi(x_{ij})}{1-\pi(x_{ij})}\right) \tag{7}$$

ดังนั้น

$$\frac{\pi(x_{ij})}{1-\pi(x_{ij})} = e^{X_{ij}^T \beta} \tag{8}$$

และได้ว่า

$$\pi(x_{ij}) = E(Y) = \frac{e^{X_{ij}^T \beta}}{1+e^{X_{ij}^T \beta}} \tag{9}$$

โดยเรียกตัวแบบที่ได้ว่า ตัวแบบถดถอยลอจิสติก (Logistic Regression Model) จะได้ว่า

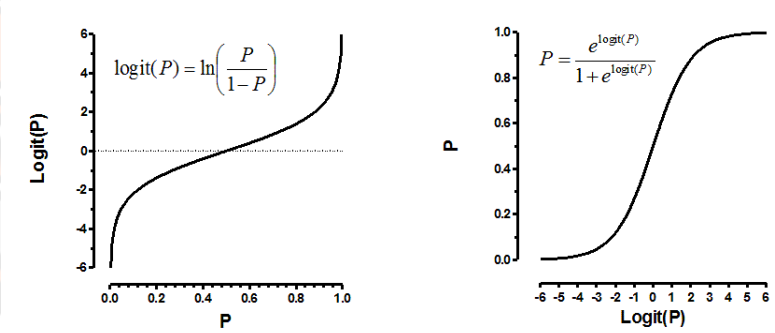
$$P(Y=1 \mid x_{ij}) = \frac{e^{X_{ij}^T \beta}}{1+e^{X_{ij}^T \beta}} \tag{10}$$

และ

$$P(Y=0 \mid x_{ij}) = \frac{1}{1+e^{X_{ij}^T \beta}} \tag{11}$$

ตัวแบบถดถอยลอจิสติกกรณีตัวแปรตามเป็นทวิภาค มีลักษณะที่สำคัญ (วีรานันท์ พงศาภักดี, 2541) ดังนี้

1. ค่าเฉลี่ยแบบมีเงื่อนไขของตัวแบบถดถอยลอจิสติก ต้องมีค่าอยู่ระหว่าง 0 และ 1
2. ค่าความคลาดเคลื่อนสุ่มมีการแจกแจงแบบเบอร์นูลลี ถ้าตัวอย่างมีขนาดใหญ่การแจกแจงดังกล่าวจะใกล้เคียงกับการแจกแจงแบบปรกติ
3. การวิเคราะห์ตัวแบบถดถอยลอจิสติกสามารถใช้หลักเกณฑ์เดียวกับการวิเคราะห์ตัวแบบถดถอยเชิงเส้นอย่างง่ายและแบบพหุคูณได้ เนื่องจากฟังก์ชันถดถอยลอจิสติกมีลักษณะเป็นเส้นโค้งรูปตัว S (S-Shape) หรือที่เรียกว่าโค้งซิกมอยด์ (Sigmoid Curve) ที่เป็นฟังก์ชันเพิ่มทางเดียว (Monotonic Increasing Function) และฟังก์ชันลดทางเดียว (Monotonic Decreasing Function) ขึ้นอยู่กับเครื่องหมายของสัมประสิทธิ์ถดถอย ดังแสดงในภาพ 1 ซึ่งจะเห็นได้ว่าเส้นโค้งมีลักษณะสมมาตรที่ $E(Y)=\pi(x_{ij})=0.5$ โดยที่ความสัมพันธ์มีลักษณะเหมือนการถดถอยเชิงเส้นและใกล้เคียงเส้นตรงเมื่อ $E(Y)$ อยู่ระหว่าง 0.2 ถึง 0.8 ซึ่ง $E(Y)$ มีค่าอยู่ภายในช่วง 0 และ 1 สำหรับทุกค่าของตัวแปรอิสระ



ภาพ 1 เส้นโค้งของฟังก์ชันถดถอยลอจิสติก (Logit Regression Function)

ที่มา : Building a Logistic Regression model from scratch, by Srivastava, T. (2013)

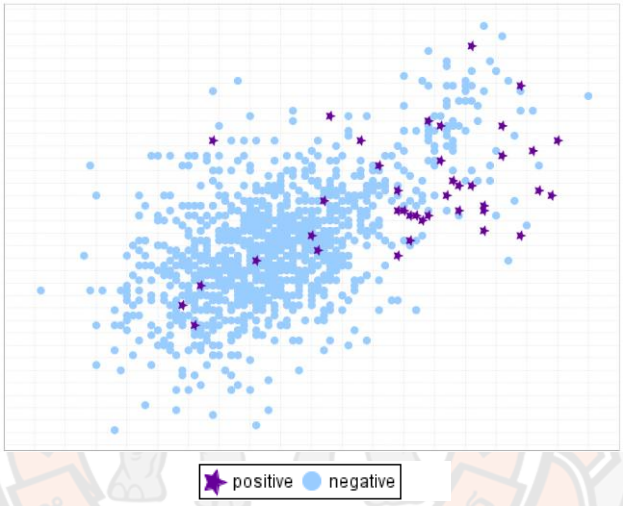
4. ตัวแปรตาม Y มีการแจกแจงแบบเบอร์นูลลี ที่มี $\pi(x_{ij})=P(y=1)$ โดยที่ $0\le\pi(x_{ij})\le1$ แต่ $X_{ij}^T\beta$ ไม่จำเป็นต้องอยู่ในช่วง 0 กับ 1 โดยใช้ฟังก์ชันการเชื่อมโยง (Link Function) มาสร้างตัวแบบแทนความสัมพันธ์เชิงเส้นระหว่าง $\pi(x_{ij})$ กับตัวแปรอิสระ x_{ij} ในการวิเคราะห์การถดถอยลอจิสติก ส่วนใหญ่จะทำการแปลงความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจให้อยู่ในรูปลอจิท (Logit) ของ $\pi(x_{ij})$ ดังนี้

$$logit(\pi(x_{ij}))=ln\left(\frac{\pi(x_{ij})}{1-\pi(x_{ij})}\right)=X_{ij}^T\beta$$

ซึ่งลอจิทของ $\pi(x_{ij})$ ไม่จำเป็นต้องอยู่ในช่วง $(0,1)$

2.3 ข้อมูลไม่สมดุล และวิธีจัดการข้อมูลไม่สมดุล

ข้อมูลไม่สมดุล หมายถึง จำนวนข้อมูลในกลุ่มหนึ่งมากกว่าจำนวนข้อมูลในอีกกลุ่มเป็นจำนวนมาก (Chawla et al., 2002) สาเหตุของความไม่สมดุลของข้อมูลเกิดได้จากหลายปัจจัย เช่น การเก็บข้อมูลผิดผลาด หรือเก็บข้อมูลไม่ครบจำนวนที่ได้กำหนดไว้ หรือมีข้อจำกัดในการเก็บรวบรวมข้อมูล เนื่องจากมีค่าใช้จ่ายในการเก็บข้อมูลที่สูงมาก และใช้ระยะเวลาในการเก็บข้อมูลนาน นอกจากนี้ ข้อมูลไม่สมดุลอาจเกิดจากลักษณะทางธรรมชาติของข้อมูล (กีระชาติ สุขสุทธิ์, 2559) ส่วนใหญ่มักจะเกิดกับข้อมูลทางการแพทย์ เช่น ผู้ป่วยที่เป็นโรคมะเร็งลำไส้ระยะสุดท้ายมีจำนวนน้อยกว่าผู้ป่วยที่เป็นโรคมะเร็งลำไส้ระยะเริ่มต้น เป็นต้น



ภาพ 2 ตัวอย่างข้อมูลไม่สมดุล

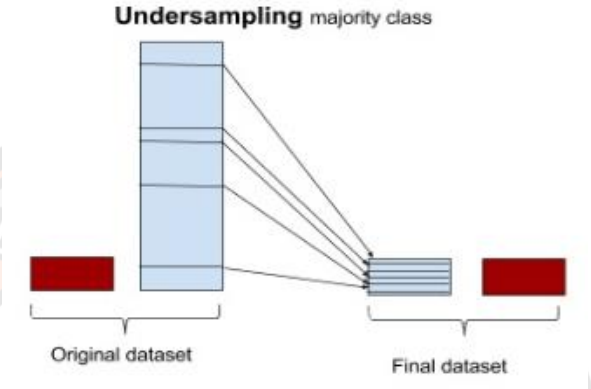
ที่มา : López, v. et al. (2017, p.130)

จากภาพ 2 ข้อมูลกลุ่มส่วนน้อยแทนกลุ่ม Positive และกลุ่มส่วนใหญ่แทนกลุ่ม Negative โดยทั่วไปจะเรียกกลุ่มข้อมูลที่มีจำนวนมากว่า กลุ่มส่วนมาก (Majority Class) และเรียกกลุ่มข้อมูลที่มีจำนวนน้อยว่า กลุ่มส่วนน้อย (Minority Class) เมื่อข้อมูลมีความไม่สมดุลจะส่งผลกระทบต่อการจำแนกข้อมูลและการพยากรณ์ของข้อมูล เนื่องจากอัลกอริทึมทั่วไปจะทำงานได้อย่างมีประสิทธิภาพสูงสุดก็ต่อเมื่อจำนวนข้อมูลในแต่ละกลุ่มมีจำนวนใกล้เคียงหรือข้อมูลมีความสมดุล แต่เมื่อข้อมูลไม่มีความสมดุล อัลกอริทึมทั่วไปจะเกิดการเอนเอียงในการจำแนกของคำตอบ ซึ่งจะเอนเอียงไปทางกลุ่มส่วนมาก ส่งผลให้การจำแนกกลุ่มส่วนน้อยเกิดความผิดพลาด จึงจำเป็นต้องจัดการกับข้อมูลไม่สมดุลก่อน

การจัดการข้อมูลไม่สมดุล เป็นการแก้ไขปัญหากับชุดข้อมูล โดยจะทำการปรับปรุงข้อมูลที่มีความไม่สมดุลให้เป็นชุดข้อมูลที่มีความสมดุลด้วยเทคนิคการสุ่มเลือกข้อมูล (Data Sampling Technique) หรือเทคนิคการเลือกข้อมูล (Data Selection Technique) ในงานวิจัยนี้จะใช้ 3 วิธี ดังนี้

วิธีการสุ่มลด (Random Under-sampling : RUS)

วิธีการสุ่มลด เป็นเทคนิคหรือวิธีที่ใช้ในการลดข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มส่วนน้อย (กิตติพงษ์ ชมบุญ, 2559) ข้อดีของวิธีการสุ่มลดคือ สามารถช่วยลดขนาดของข้อมูลในกลุ่มส่วนมาก ซึ่งเป็นการทำงานที่ดีหากข้อมูลมีปริมาณมากและลักษณะคล้ายกัน แต่ข้อเสียคือ การลดข้อมูลด้วยการสุ่ม อาจจะทำให้ข้อมูลที่สำคัญหายไป

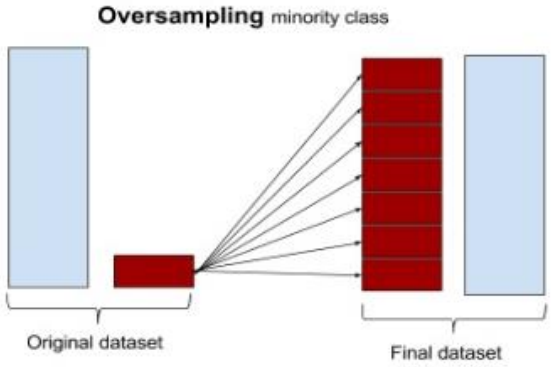


ภาพ 3 วิธีการสุ่มลด (Random Under-sampling)

ที่มา : Learning from Imbalanced Classes by Fawcett, T. (2016)

วิธีการสุ่มเกิน (Random Over-sampling : ROS)

วิธีการสุ่มเกิน เป็นเทคนิคหรือวิธีที่ใช้ในการเพิ่มข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มส่วนมาก โดยการสุ่มเลือกข้อมูลจากข้อมูลเดิม หรือสร้างข้อมูลขึ้นมาใหม่จากตัวอย่างของข้อมูลเดิม (กิตติพงษ์ ชมบุญ, 2559) ข้อดีของวิธีการสุ่มเพิ่มคือ สามารถเพิ่มปริมาณข้อมูลในกลุ่มส่วนน้อยให้มีจำนวนมากขึ้น แต่ข้อเสียคือ การเพิ่มข้อมูลลักษณะนี้อาจจะทำให้เกิดปัญหาข้อมูลซ้ำซ้อนและถูกรบกวนได้ง่าย



ภาพ 4 วิธีการสุ่มเกิน (Random Over-sampling)

ที่มา : Learning from Imbalanced Classes by Fawcett, T. (2016)

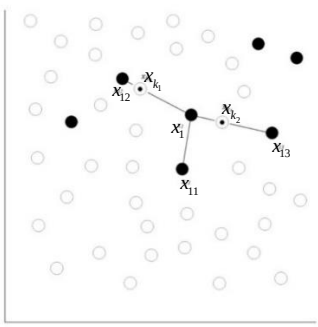
วิธีการสุ่มเกินด้วย SMOTE Technique (Synthetic Minority Over-sampling Technique : SMOTE)

วิธีสุ่มเกินด้วย SMOTE Technique (Cgawla et al., 2002) เป็นการสุ่มตัวอย่างแบบพิเศษของวิธีสุ่มเกิน แทนที่จะสุ่มเพิ่มข้อมูลเดิม แต่จะทำการสังเคราะห์ข้อมูลใหม่จากข้อมูลเดิม ใช้หลักการเพื่อนบ้านที่ใกล้ที่สุด โดยการสุ่มข้อมูลจากกลุ่มส่วนน้อยตามจำนวนที่กำหนด แล้วสร้างข้อมูลสังเคราะห์จากข้อมูลตัวอย่างด้วยการวัดระยะห่างจากจุดข้อมูลตัวอย่างไปยังจุดข้อมูลใกล้เคียง ทำการสุ่มสร้างข้อมูลสังเคราะห์ขึ้น โดยข้อมูลสังเคราะห์ที่สร้างขึ้นจะอยู่ภายในระยะห่างจากจุดข้อมูลตัวอย่างไปยังจุดข้อมูลเพื่อนบ้านที่ใกล้ที่สุด ซึ่งขั้นตอนการสังเคราะห์ข้อมูลใหม่มีดังนี้

SMOTE Algorithm

กำหนด k และ N แทน จำนวนเพื่อนบ้านที่ใกล้ที่สุดและจำนวนตัวอย่างข้อมูลสังเคราะห์ตามลำดับ มีขั้นตอน ดังนี้

1. กำหนด $x_i, i = 1,...,n_s$ แทนค่าสังเกตของกลุ่มน้อย และให้ A แทนเซตทั้งหมดของ x_i ดังนั้น $x_i \in A$ สำหรับทุกๆค่าของ x_i
2. คำนวณระยะทางยูคลิด (Euclidean distance) ระหว่าง x_i และทุกค่าของค่าสังเกตในกลุ่มส่วนน้อย เพื่อได้ค่า k – nearest neighbors ของ x_i
3. ให้ S_{ik} แทนเซตของ k – nearest neighbors ของ x_i
4. สุ่มตัวอย่างข้อมูลสังเคราะห์ N จำนวน แทนด้วย x_{ij} , $(j = 1,...,N)$
5. กำหนด λ แทนตัวเลขในช่วง $(0,1)$ และสามารถสร้างข้อมูลสังเคราะห์ได้ดังสมการนี้ $x_k = x_i + \lambda(x_i - x_{ij})$
6. ทำซ้ำในขั้นตอนที่ 5 สำหรับทุกๆ ค่าของ x_{ij}
7. หยุดอัลกอริทึม

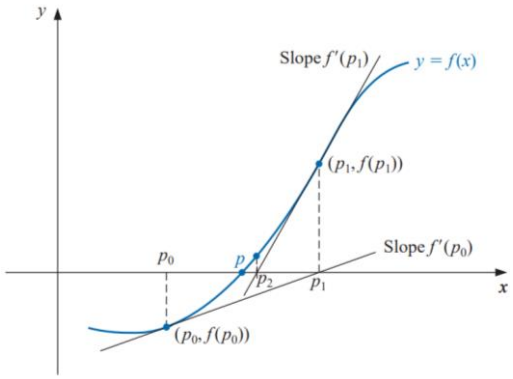


ภาพ 5 วิธีการสุ่มเกินด้วย SMOTE Technique บนชุดข้อมูลที่ไม่สมดุลในพื้นที่สองมิติ (Two - dimensional) ที่มา : Brandt, J. & Lanzén, E. (2020, p.13)

จากภาพ 5 พิจารณาค่าสังเกตในกลุ่มส่วนน้อย x_1 เมื่อกำหนดค่า $k = 3$ และ $N = 2$ จะได้ค่าสังเกตที่สังเคราะห์ใหม่คือ x_{k_1} และ x_{k_2} อยู่ในระยะสุ่มตามแนวเส้นตรงระหว่างเพื่อนบ้านที่ใกล้ที่สุด

2.4 ระเบียบวิธีของนิวตัน-ราฟสัน

ระเบียบวิธีของนิวตัน-ราฟสัน เป็นวิธีการหาค่ารากของสมการ โดยใช้ความชันของเส้นตรงเข้ามาช่วยหาการลู่เข้าสู่คำตอบของสมการ (Faires and Burden, 2010) และกำหนดให้การหาอนุพันธ์ของสมการ $f(x)$ จะต้องหาค่าได้บนช่วงของ p เมื่อ p จุดหรือตำแหน่งของสมการบนแกน x



ภาพ 6 ความหมายทางเรขาคณิตของวิธีการนิวตัน-ราฟสัน

ที่มา : Dey, S. (2023, p.6)

จากภาพ 6 จะได้ ความชันของเส้นสัมผัสกราฟ $f(x)$ ณ จุด คือ $(p_0, f(p_0))$ ดังนั้น $f'(x)$ สมการเส้นสัมผัส คือ

$$y - f(p_0) = f'(p_0)(x - p_0)$$

โดยที่ค่าของ y ณ เส้นสัมผัส บนแกน x มีค่าเท่ากับ 0 ดังนั้น จุด p_1 จะคำนวณได้จาก

$$0 - f(p_0) = f'(p_0)(p_1 - p_0)$$

เมื่อแก้สมการเพื่อหาค่า p_1 จะได้ว่า

$$p_1 = p_0 - \frac{f(p_0)}{f'(p_0)},\ f'(p_0) \neq 0$$

ดังนั้น ในการหาค่า p_{r+1} ของสมการ $f(x) = 0$ จะใช้ค่า p_r มาช่วยในการคำนวณ ดังนี้

$$p_{r+1} = p_r - \frac{f(p_r)}{f'(p_r)},\ r \geq 0$$

โดยผลต่างของรากของสมการที่ $r+1$ และ r ควรมีค่าน้อย หรือมีค่าน้อยกว่าความคลาดเคลื่อนที่กำหนด ในงานวิจัยนี้กำหนดให้เป็น 0.0001 ซึ่งในการหาคำตอบของระบบสมการด้วยวิธีนิวตัน-ราฟสัน จะหาอนุพันธ์ โดยอาศัยฟังก์ชันภาวะน่าจะเป็น ดังนี้

2.5 ฟังก์ชันภาวะน่าจะเป็น (Likelihood function)

กำหนดให้ $Y^T=\left[y_1,y_2,...,y_i,...,y_n\right]$ เป็นเวกเตอร์สุ่มขนาด $1\times n$ และมีฟังก์ชันความหนาแน่นร่วม $f(Y,\beta)$ จะได้ฟังก์ชันภาวะน่าจะเป็นดังนี้

$$\begin{aligned} L(\beta) &= f(Y,\beta) \\ &= \prod_{i=1}^{n} f(y_i,\beta) \end{aligned} \tag{12}$$

การประมาณค่า β ด้วยฟังก์ชันภาวะน่าจะเป็น เป็นการหาค่า β ที่ทำให้ $L(\beta)$ มีค่าสูงสุด พิจารณาฟังก์ชันภาวะน่าจะเป็นของตัวแบบถดถอยลอจิสติก ดังนี้

$$L(\beta)=\prod_{i=1}^{n}\pi(x_{ij})^{y_i}\left(1-\pi(x_{ij})\right)^{1-y_i} \tag{13}$$

คำนวณลอการึทึมของฟังก์ชันภาวะน่าจะเป็น (Log Likelihood Function) ในตัวแบบถดถอยลอจิสติกได้ดังนี้

$$\begin{aligned} \ln l(\beta) &= \ln L(\beta) \\ &= \ln\left(\prod_{i=1}^{n}\pi(x_{ij})^{y_i}\left(1-\pi(x_{ij})\right)^{1-y_i}\right) \\ &= \sum_{i=1}^{n}\left(y_i\ln\pi(x_{ij})+(1-y_i)\ln\left(1-\pi(x_{ij})\right)\right) \\ &= \sum_{i=1}^{n}\left(y_i\ln\pi(x_{ij})+\ln(1-\pi(x_{ij}))-y_i\ln\left(1-\pi(x_{ij})\right)\right) \\ &= \sum_{i=1}^{n}\left(y_i\ln\left(\frac{\pi(x_{ij})}{1-\pi(x_{ij})}\right)+\ln\left(1-\pi(x_{ij})\right)\right) \\ &= \sum_{i=1}^{n}\left(y_i(X_{ij}^T\beta)+\ln\left(1-\frac{e^{X_{ij}^T\beta}}{1+e^{X_{ij}^T\beta}}\right)\right) \\ &= \sum_{i=1}^{n}\left(y_i(X_{ij}^T\beta)+\ln\left(\frac{1}{1+e^{X_{ij}^T\beta}}\right)\right) \\ &= \sum_{i=1}^{n}\left(y_i(X_{ij}^T\beta)-\ln\left(1+e^{X_{ij}^T\beta}\right)\right) \\ &= \sum_{i=1}^{n}y_i(X_{ij}^T\beta)-\sum_{i=1}^{n}\ln\left(1+e^{X_{ij}^T\beta}\right) \end{aligned} \tag{14}$$

สำหรับงานวิจัยนี้ การประมาณค่าพารามิเตอร์ที่ทำให้ $\ln l(\beta)$ มีค่ามากที่สุด ทำได้โดยหาอนุพันธ์ย่อย (Partial Differentiate) อันดับที่หนึ่งของ $\ln l(\beta)$ เทียบกับ $\beta_0,\beta_1,...,\beta_i,...,\beta_p$

เรียกว่า ฟังก์ชันสกอร์ (Score Function) แล้วนำมาเป็นสมาชิกเวกเตอร์ $U(\beta)$ ขนาด $(p+1)\times 1$ ซึ่งมีรูปแบบดังนี้

$$U(\beta)=\begin{bmatrix}\frac{\partial \ln(\beta)}{\partial \beta_0}\\ \frac{\partial \ln(\beta)}{\partial \beta_1}\\ \vdots \\ \frac{\partial \ln(\beta)}{\partial \beta_p}\end{bmatrix}_{(p+1)\times 1}$$

กำหนดเมทริกซ์ $H(\beta)$ เรียกว่าเมทริกซ์เฮสเซียน (Hessian matrix) มีขนาด $(p+1)\times(p+1)$ โดยมีสมาชิกเป็นอนุพันธ์ย่อยอันดับที่สอง (Second Partial Differentiate) ของ $\ln l(\beta)$

โดยที่สมาชิกตัวที่ $(j,k)=\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k}\ ;\ j,k=0,1,2,...,p$

ซึ่งสามารถคำนวณหาค่าเวกเตอร์ $U(\hat{\beta})$ เป็นเวกเตอร์ของค่าประมาณพารามิเตอร์ที่ได้จากวิธีภาวะน่าจะเป็นสูงสุด โดยใช้อนุกรมเทเลอร์ (Taylor Series) กระจาย $U(\beta)$ รอบ $\beta^{(r)}$ จะได้ว่า

$$U(\hat{\beta})\approx U(\beta^{(r)})+H(\beta^{(r)})(\hat{\beta}-\beta^{(r)})$$

จากนิยามฟังก์ชันภาวะน่าจะเป็นสูงสุดของ β จะได้

$$\left.\frac{\partial \ln L(\beta)}{\partial \beta_j}\right|_{\hat{\beta}}=0\ ;\ j=0,1,2,...,p$$

และ $U(\hat{\beta})=0$

ดังนั้น $\hat{\beta}\approx \beta^{(r)}-H^{-1}(\beta^{(r)})U\beta^{(r)})$

ค่าประมาณ $\hat{\beta}$ ณ รอบที่ $r+1$ คือ

$$\hat{\beta}^{(r+1)}=\hat{\beta}^{(r)}-H^{-1}(\beta^{(r)})U(\beta^{(r)}) \qquad (15)$$

เมื่อ $\hat{\beta}^{(r+1)}$ แทน ค่าประมาณของพารามิเตอร์ β รอบที่ $r+1$

$\hat{\beta}^{(r)}$ แทน ค่าประมาณของพารามิเตอร์ β รอบที่ r

$H\left(\beta^{(r)}\right)$ แทน เมทริกซ์เฮสเซี่ยนรอบที่ r

$U(\beta^{(r)})$ แทน ฟังก์ชันสกอร์รอบที่ r

โดยทำซ้ำจนกว่าจะได้ค่าประมาณที่ลู่เข้าสู่ค่าใดค่าหนึ่ง ซึ่งมีเกณฑ์การหยุดกระบวนการทำซ้ำ ดังนี้

$$\left|\hat{\beta}^{(r+1)} - \hat{\beta}^{(r)}\right| < 0.0001$$

2.6 วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation)

วิธีภาวะน่าจะเป็นสูงสุดเป็นวิธีหาค่าประมาณของพารามิเตอร์ β ในเทอมของค่าสังเกตที่ทำให้ฟังก์ชันภาวะน่าจะเป็น $L(\beta)$ มีค่าสูงสุด โดยหาจากอนุพันธ์ของฟังก์ชันภาวะน่าจะเป็น $L(\beta)$ เทียบกับพารามิเตอร์ β เพื่อความสะดวกในทางปฏิบัติการประมาณค่ามักใช้ $\ln L(\beta)$ ในการหาตัวประมาณแบบภาวะน่าจะเป็นสูงสุด

จากรูปแบบเชิงเส้น $\beta_0 + \beta_1 x_1 + ... + \beta_p x_p$ สามารถเขียนในอีกรูปแบบหนึ่งได้เป็น $X_{ij}^T \beta$ ดังนั้น ตัวแบบถดถอยลอจิสติก คือ

$$P(y_i = 1) = \pi(x_{ij}) = \frac{e^{x_{ij}^T \beta}}{1 + e^{x_{ij}^T \beta}}$$

$$P(y_i = 0) = 1 - \pi(x_{ij}) = \frac{1}{1 + e^{x_{ij}^T \beta}}$$

การแจกแจงความน่าจะเป็นของการเกิดเหตุการณ์เป็นดังนี้

$$P(Y = y) = \pi(x_{ij})^y (1 - \pi(x_{ij}))^{1-y} \; ; y = (0,1)$$

โดยมีฟังก์ชันภาวะน่าจะเป็น ซึ่งเป็นพารามิเตอร์ β ดังนี้

$$L(\beta) = \prod_{i=1}^{n} \pi(x_{ij})^{y_i} \left(1 - \pi(x_{ij})\right)^{1-y_i}$$

จะได้

$$\begin{aligned} l(\beta) &= \ln L(\beta) \\ &= \ln \prod_{i=1}^{n} \pi(x_{ij})^{y_i} \left(1 - \pi(x_{ij})\right)^{1-y_i} \\ &= \sum_{i=1}^{n} \left(y_i (X_{ij}^T \beta) - \ln\left(1 + e^{X_{ij}^T \beta}\right)\right) \end{aligned} \tag{16}$$

ได้สมการปกติ ดังนี้

$$\begin{aligned}\frac{\partial l(\beta)}{\partial \beta_k} &= \sum_{i=1}^{n}\left\{ y_i x_k - \frac{x_k e^{x_{ij}^T\beta}}{1+e^{x_{ij}^T\beta}} \right\} = 0 \\ &= \sum_{i=1}^{n}\left\{ y_i x_k - x_k \pi\left(x_{ij}\right) \right\} = 0 \\ &= \sum_{i=1}^{n} x_k \left\{ y_i - \pi\left(x_{ij}\right) \right\} = 0 \qquad \text{เมื่อ } k = 0,1,...,\mathrm{p}\end{aligned} \tag{17}$$

โดยสามารถเขียนให้อยู่ในรูปของเมทริกซ์ (Matrix) ได้ดังนี้

$$\frac{\partial l(\beta)}{\partial \beta_k} = \begin{bmatrix} \sum_{i=1}^{n}\left(y_i - \pi\left(x_{ij}\right)\right)x_0 \\ \sum_{i=1}^{n}\left(y_i - \pi\left(x_{ij}\right)\right)x_1 \\ \vdots \\ \sum_{i=1}^{n}\left(y_i - \pi\left(x_{ij}\right)\right)x_p \end{bmatrix}$$

จากการหาค่าอนุพันธ์ย่อยในลำดับที่ 1 โดยจะเรียกฟังก์ชันที่ได้ว่านี้ว่า ฟังก์ชันสกอร์ (Scoring function) เมื่อแทนค่า $\pi\left(x_{ij}\right)$ ลงในสมการที่ (17) จะได้ว่า

$$\begin{aligned}\frac{\partial l(\beta)}{\partial \beta_k} &= \sum_{i=1}^{n} x_k \left[y_i - \left(\frac{x_k e^{x_{ij}^T\beta}}{1+e^{x_{ij}^T\beta}} \right) \right] \\ &= \sum_{i=1}^{n} \left[x_k y_i - x_k \left(\frac{x_k e^{x_{ij}^T\beta}}{1+e^{x_{ij}^T\beta}} \right) \right]\end{aligned}$$

การประมาณค่าพารามิเตอร์จะใช้หลักกการทำซ้ำ (Iterative procedures) โดยใช้วิธีการของนิวตัน-ราฟสัน (Newton-Raphson) เข้ามาช่วยในการประมาณค่า ซึ่งวิธีนิวตัน-ราฟสันเป็นเทคนิคการวิเคราะห์เชิงตัวเลขที่นิยมใช้กันอย่างแพร่หลาย โดยจะใช้สมการ p สมการ เมื่อต้องการหาค่าพารามิเตอร์ที่ไม่ทราบค่า p พารามิเตอร์ ซึ่งได้จากการหาค่าอนุพันธ์ย่อยในอันดับที่ 2 และเรียกว่า เฮสเซียนเมทริกซ์ (Hessian Matrix)

$$\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_l} = -\sum_{i=1}^{n}\frac{\left(1+e^{x_{ij}^T\beta}\right)e^{x_{ij}^T\beta}x_k x_l - \left(e^{x_{ij}^T\beta}\right)^2 x_k x_l}{\left(1+e^{x_{ij}^T\beta}\right)^2}$$

$$= -\sum_{i=1}^{n}\left[x_k x_l y_i \pi\left(x_{ij}\right) - x_k x_l \left(\pi\left(x_{ij}\right)\right)^2 \right]$$

$$= -\sum_{i=1}^{n}\left[x_k x_l y_i \pi\left(x_{ij}\right)\left(1-\pi\left(x_{ij}\right)\right)' \right]$$

ซึ่งสามารถเขียนให้อยู่ในรูปของเมทริกซ์ได้ดังนี้

$$\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_l} = \begin{bmatrix} \sum_{i=1}^{n}\pi(x_{ij})(1-\pi(x_i))x_0^2 & \sum_{i=1}^{n}\pi(x_{ij})(1-\pi(x_i))x_0x_1 & \cdots & \sum_{i=1}^{n}\pi(x_{ij})(1-\pi(x_i))x_0x_p \\ \sum_{i=1}^{n}\pi(x_{ij})(1-\pi(x_i))x_1x_0 & \sum_{i=1}^{n}\pi(x_{ij})(1-\pi(x_i))x_1^2 & \cdots & \sum_{i=1}^{n}\pi(x_{ij})(1-\pi(x_i))x_1x_p \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}\pi(x_{ij})(1-\pi(x_i))x_px_0 & \sum_{i=1}^{n}\pi(x_{ij})(1-\pi(x_i))x_px_1 & \cdots & \sum_{i=1}^{n}\pi(x_{ij})(1-\pi(x_i))x_p^2 \end{bmatrix}$$

หรือสามารถเขียนให้อยู่ในอีกรูปแบบได้เป็น

$$\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_l} = -\begin{bmatrix} \frac{\partial^2 \ln L(\beta)}{\partial\beta_0^2} & \frac{\partial^2 \ln L(\beta)}{\partial\beta_0\partial\beta_1} & \cdots & \frac{\partial^2 \ln L(\beta)}{\partial\beta_0\partial\beta_p} \\ \frac{\partial^2 \ln L(\beta)}{\partial\beta_1\partial\beta_0} & \frac{\partial^2 \ln L(\beta)}{\partial\beta_1^2} & \cdots & \frac{\partial^2 \ln L(\beta)}{\partial\beta_1\partial\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln L(\beta)}{\partial\beta_p\partial\beta_0} & \frac{\partial^2 \ln L(\beta)}{\partial\beta_p\partial\beta_1} & \cdots & \frac{\partial^2 \ln L(\beta)}{\partial\beta_p^2} \end{bmatrix}$$

การประมาณพารามิเตอร์ด้วยวิธีนิวตัน-ราฟสัน สามารถทำได้ดังนี้

$$\beta^{i+1} = \beta^i - \left(\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_l}\right)^{-1}\left(\frac{\partial l\left(\beta^i\right)}{\partial\beta_k}\right), \qquad i = 1,...,1000$$

เมื่อ β^{i+1} แทน ค่าประมาณของสัมประสิทธิ์ถดถอยที่ได้จากรอบที่ $i+1$

β^i แทน ค่าประมาณของสัมประสิทธิ์ถดถอยที่ได้จากรอบที่ i

2.7 วิธีฟังก์ชันสกอร์ที่ปรับปรุง (Modified Score Function)

จากตัวแบบถดถอยลอจิสติกที่ใช้วิธีภาวะน่าจะเป็นสูงสุดในการประมาณค่าพารามิเตอร์ ด้วยกระบวนการ Newton-Raphson พบว่าเกิดการวนซ้ำที่ไม่บรรจบกัน เมื่อตัวอย่างมีขนาดเล็ก และสัดส่วนของเหตุการณ์ที่สนใจมีน้อย ซึ่งวิธีแก้ไขปัญหาทำได้โดยปรับฟังก์ชันสกอร์ (Score Function) (Firth, 1993)

การปรับเปลี่ยนฟังก์ชันสกอร์ ใช้เวกเตอร์ความเอนเอียง (Bias vector) และเมทริกซ์สารสนเทศ (Information matrix) มาประกอบการประมาณค่าพารามิเตอร์ในแบบจำลองถดถอยลอจิสติก และปรับฟังก์ชันสกอร์ จาก $U(\beta)$ เป็น $U^*(\beta)$ ดังนี้

$$U^*(\beta) = U(\beta) - J(\beta)b(\beta) = 0 \tag{18}$$

โดย $J(\beta)$ แทนเมทริกซ์สารสนเทศ (Information matrix) ดังนี้

$$J(\beta) = \begin{bmatrix} -E\left(\frac{\partial^2 l(\beta_0,\beta_1,...,\beta_p)}{\partial\beta_0\partial\beta_0}\right) & -E\left(\frac{\partial^2 l(\beta_0,\beta_1,...,\beta_p)}{\partial\beta_0\partial\beta_1}\right) & \cdots & -E\left(\frac{\partial^2 l(\beta_0,\beta_1,...,\beta_p)}{\partial\beta_0\partial\beta_p}\right) \\ -E\left(\frac{\partial^2 l(\beta_0,\beta_1,...,\beta_p)}{\partial\beta_1\partial\beta_0}\right) & -E\left(\frac{\partial^2 l(\beta_0,\beta_1,...,\beta_p)}{\partial\beta_1\partial\beta_1}\right) & \cdots & -E\left(\frac{\partial^2 l(\beta_0,\beta_1,...,\beta_p)}{\partial\beta_1\partial\beta_p}\right) \\ \vdots & \vdots & \ddots & \vdots \\ -E\left(\frac{\partial^2 l(\beta_0,\beta_1,...,\beta_p)}{\partial\beta_p\partial\beta_0}\right) & -E\left(\frac{\partial^2 l(\beta_0,\beta_1,...,\beta_p)}{\partial\beta_p\partial\beta_1}\right) & \cdots & -E\left(\frac{\partial^2 l(\beta_0,\beta_1,...,\beta_p)}{\partial\beta_p\partial\beta_p}\right) \end{bmatrix}$$

$$J(\beta) = \begin{bmatrix} \sum_{i=1}^{n}\pi(x_i)(1-\pi(x_i))x_0^2 & \sum_{i=1}^{n}\pi(x_i)(1-\pi(x_i))x_0x_1 & \cdots & \sum_{i=1}^{n}\pi(x_i)(1-\pi(x_i))x_0x_p \\ \sum_{i=1}^{n}\pi(x_i)(1-\pi(x_i))x_1x_0 & \sum_{i=1}^{n}\pi(x_i)(1-\pi(x_i))x_1^2 & \cdots & \sum_{i=1}^{n}\pi(x_i)(1-\pi(x_i))x_1x_p \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}\pi(x_i)(1-\pi(x_i))x_px_0 & \sum_{i=1}^{n}\pi(x_i)(1-\pi(x_i))x_px_1 & \cdots & \sum_{i=1}^{n}\pi(x_i)(1-\pi(x_i))x_p^2 \end{bmatrix}$$

$$= X^TWX$$

เมื่อ $W = diag\left\{\pi(x_i)(1-\pi(x_i))\right\}$ และ x แทน เมทริกซ์ข้อมูล (Design matrix)
$b(\beta)$ แทนเวกเตอร์ความเอนเอียง (Bias vector) เขียนได้ดังนี้

$$b(\beta) = (X^TWX)^{-1}X^TW\xi \tag{19}$$

เมื่อ $\xi = \begin{bmatrix} \frac{h_{11}}{\pi(x_1)(1-\pi(x_1))}\left(\pi(x_1)-\frac{1}{2}\right) \\ \frac{h_{22}}{\pi(x_2)(1-\pi(x_2))}\left(\pi(x_2)-\frac{1}{2}\right) \\ \vdots \\ \frac{h_{ii}}{\pi(x_i)(1-\pi(x_i))}\left(\pi(x_i)-\frac{1}{2}\right) \end{bmatrix}$

โดย h_{ii} เป็นองค์ประกอบของเส้นทแยงมุมของ Hat matrix

สามารถคำนวน Hat matrix ได้ดังนี้

$$H = W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}} \tag{20}$$

ดังนั้น

$$\begin{aligned} U^*(\beta) &= U(\beta) - J(\beta)b(\beta) = 0 \\ 0 &= U(\beta) - (X^T W X)(X^T W X)^{-1} X^T W \xi \\ 0 &= U(\beta) - I X^T W \xi \\ 0 &= U(\beta) - X^T W \xi \end{aligned}$$

จะได้ผลลัพธ์ของการประมาณค่าพารามิเตอร์ ดังนี้

$$\beta^*_{r+1} = \beta^*_r - b(\beta^*_r) + I_r(\beta)^{-1} U^*(\beta^*_r) \tag{21}$$

เมื่อ β^*_{r+1} แทน ค่าประมาณของพารามิเตอร์ β รอบที่ $r+1$

β^*_r แทน ค่าประมาณของพารามิเตอร์ β รอบที่ r

$b(\beta^*_r)$ แทน เวกเตอร์ความเอนเอียงรอบที่ r

$I_r(\beta)^{-1}$ แทน เมทริกซ์สารสนเทศผกผันรอบที่ r

$U^*(\beta^*_r)$ แทน ฟังก์ชันสกอร์ที่ปรับปรุงรอบที่ r

โดยทำซ้ำจนกว่าจะได้ค่าประมาณที่ลู่เข้าสู่ค่าใดค่าหนึ่ง ซึ่งมีเกณฑ์การหยุดกระบวนการทำซ้ำ ดังนี้

$$\left|\beta^*_{r+1} - \beta^*_r\right| < 0.0001$$

2.8 แนวคิดของการประมาณค่าพารามิเตอร์แบบเบส์

สถิติตามแนวของเบส์ (Bayesian approach) แตกต่างจากสถิติตามแนวเดิม (Classical approach) ในแนวเดิม การประมาณพารามิเตอร์ θ เริ่มจากการสุ่มตัวอย่างจากประชากรที่มีฟังก์ชันความหนาแน่น $f(y;\theta)$ และถือว่าพารามิเตอร์ θ เป็นค่าคงที่แต่ไม่ทราบค่า แต่ในแนวคิดของเบส์จะพยายามใช้ความรู้เดิมหรือข้อมูลเกี่ยวกับ θ ให้เป็นประโยชน์ในการประมาณ θ ให้ได้ดียิ่งขึ้น ดังนั้นจึงให้ θ เป็นค่าของตัวแปรสุ่ม Θ มีการแจกแจงความน่าจะเป็นรูปใดรูปหนึ่ง

ให้ $Y_1,...,Y_n$ เป็นตัวอย่างสุ่มจากประชากรที่มีฟังก์ชันความหนาแน่น $f(y;\theta)=f(y|\theta)$ โดยที่ θ เป็นค่าของตัวแปรสุ่ม Θ ในที่นี้ $f(y|\theta)$ เป็นฟังก์ชันความหนาแน่นน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability Density Function) ของตัวแปรสุ่ม Y เมื่อกำหนดให้ $\Theta=\theta$ ให้ตัวแปรสุ่ม Θ มีฟังก์ชันความหนาแน่น $g(\theta)$ เรียกว่า ฟังก์ชันความหนาแน่นน่าจะเป็นก่อน (Prior or Initial Probability Density Function) ของ Θ

ให้ $h(\theta|Y_1,...,Y_n)$ เป็นฟังก์ชันความหนาแน่นแบบมีเงื่อนไขของตัวแปรสุ่ม Θ เมื่อกำหนด $Y_1=y_1,...,Y_n=y_n$ เรียกว่า ฟังก์ชันความหนาแน่นน่าจะเป็นภายหลัง (Posterior Probability Density Function) ของ Θ

ในที่นี้ ฟังก์ชันความหนาแน่น $Y_1,...,Y_n$ เมื่อกำหนดให้ $\Theta=\theta$ ได้แก่

$$f(Y_1,...,Y_n;\theta)=f(Y_1,...,Y_n|\theta)=\prod_{i=1}^{n}f(y_i|\theta)$$

การหาค่าฟังก์ชันความหนาแน่นภายหลังของ Θ ได้จากฟังก์ชันความหนาแน่นน่าจะเป็นก่อน ฟังก์ชันความหนาแน่นแบบมีเงื่อนไขของตัวอย่างสุ่ม ให้ Θ มีฟังก์ชันความหนาแน่นน่าจะเป็นก่อนเป็น $g(\theta)$ ดังนั้นฟังก์ชันความหนาแน่นร่วมของ $Y_1,...,Y_n$ และ Θ ได้แก่

$$f(Y_1,...,Y_n|\theta)\mathrm{g}(\theta)=\prod_{i=1}^{n}f(y_i|\theta)\mathrm{g}(\theta)$$

ฟังก์ชันความหนาแน่นร่วมของ $(y_1,...,y_n)$ คือ $\int_{\theta}f(Y_1,...,Y_n|\theta)\mathrm{g}(\theta)d\theta$ ดังนั้นสามารถคำนวณฟังก์ชันความหนาแน่นน่าจะเป็นภายหลังของ Θ (ประชุม สุวัตถี, 2545) ได้ดังนี้

กรณี Θ เป็นตัวแปรสุ่มชนิดต่อเนื่อง

$$h(\theta|Y_1,...,Y_n)=\frac{\prod_{i=1}^{n}f(y_i|\theta)\mathrm{g}(\theta)}{\int_{\theta}\prod_{i=1}^{n}f(y_i|\theta)\mathrm{g}(\theta)d\theta}=\frac{L(\theta)\mathrm{g}(\theta)}{\int_{\theta}L(\theta)\mathrm{g}(\theta)d\theta}\propto L(\theta)\mathrm{g}(\theta)$$

กรณี Θ เป็นตัวแปรสุ่มชนิดไม่ต่อเนื่อง

$$h(\theta \mid Y_1,...,Y_n) = \frac{\prod_{i=1}^{n} f(y_i \mid \theta)\,\mathrm{g}(\theta)}{\sum_{\theta}\prod_{i=1}^{n} f(y_i \mid \theta)\,\mathrm{g}(\theta)} = \frac{L(\theta)\,\mathrm{g}(\theta)}{\sum_{\theta} L(\theta)\,\mathrm{g}(\theta)} \propto L(\theta)\,\mathrm{g}(\theta)$$

โดยที่ $L(\theta)$ แทน ฟังก์ชันภาวะน่าจะเป็น

$\mathrm{g}(\theta)$ แทน ฟังก์ชันความหนาแน่นน่าจะเป็นก่อน

2.9 การสุ่มตัวอย่างแบบกิบส์ (Gibbs Sampling)

การสุ่มตัวอย่างแบบกิบส์ ถูกพัฒนาขึ้นครั้งแรกโดย Geman and Gemen ในปี ค.ศ 1984 เพื่อใช้ในการจำลองชุดข้อมูลหรือสถานการ์ที่ศึกษา ต่อมาได้นำมาใช้ในงานที่เกี่ยวข้องกับสาขาสถิติและสาขาอื่นอย่างแพร่หลาย โดยเฉพาะในสาขาสถิติถูกนำมาใช้ในงานที่เกี่ยวข้องกับเบส์เซียน (Carlin and Louis, 2009)

ถ้าให้ θ เป็นพารามิเตอร์ที่สนใจ และ $y_1,...,y_n$ เป็นค่าสังเกต ซึ่งแนวคิดแบบเบส์มีจุดมุ่งหมายเพื่อจะประมาณฟังก์ชันความน่าแน่นน่าจะเป็นภายหลัง $h(\theta \mid y_1,...,y_n)$ ที่เกิดจากผลคูณของฟังก์ชันภาวะน่าจะเป็นและฟังก์ชันความหนาแน่นก่อน เพื่อนำไปใช้ประโยชน์ในการประมาณค่าพารามิเตอร์ แต่บางครั้งไม่สามารถหาฟังก์ชันความน่าแน่นน่าจะเป็นภายหลังที่มีรูปแบบชัดเจนได้ ในกรณีนี้อาจแก้ปัญหาโดยใช้การจำลองมอนติคาร์โล (Monte Carlo Simulation) และนำวิธีการสุ่มตัวอย่างแบบกิบส์มาประยุกต์ใช้เพื่อช่วยในการหาฟังก์ชันความน่าแน่นน่าจะเป็นภายหลัง (มานพ วราภักดิ์, 2551)

1. **หลักการของการสุ่มตัวอย่างแบบกิบส์**

ให้ $\boldsymbol{\theta}$ เป็นเวกเตอร์ที่มี p องค์ประกอบ (Components) นั่นคือ $\theta^T = \left[\theta_1, \theta_2, ..., \theta_p\right]$ การสุ่ม $\boldsymbol{\theta}$ ในรอบที่ $t+1$ ทำได้โดยการสุ่มแต่ละองค์ประกอบของ θ_i ในรอบที่ t จากการแจกแจงในรูปแบบ $\pi(\theta_i \mid \theta_{-i})$ เมื่อ $\theta_{-i} = \left[\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p\right]$ ซึ่งเป็นการแจกแจงแบบมีเงื่อนไขเต็มรูป (Full Conditional Distribution) ที่ได้มาจาก $\pi(\theta)$ ซึ่งเป็นการแจกแจงร่วม (Joint Distribution) ของทุกองค์ประกอบของ $\boldsymbol{\theta}$ และ

$$\pi(\theta_i \mid \theta_{-i}) = \frac{\pi(\theta)}{\int_{\theta_i} \pi(\theta) d\theta_i}$$

โดยที่

$$\begin{aligned}\theta_{-i} &= \{\theta_k\};\ i \neq k;\ i,k = 1,..., p \\ &= \{\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p\}\end{aligned}$$

เมื่อสุ่ม θ ในรอบที่ t ครบทุกองค์ประกอบแล้ว จะได้ θ ใหม่ ในรอบที่ $t+1$ และเมื่อจำนวนรอบของการสุ่มมากขึ้น นั่นคือ $t \to \infty$ จะได้ $\theta^{(t)}$ มีการแจกแจงลู่เข้าสู่การแจกแจงที่สเตชันนารี (Stationary Distribution)

2. **ขั้นตอนในการสุ่มตัวอย่าง**

หลังจากที่มีการสร้างตัวแบบของปัญหาและกำหนดพารามิเตอร์ที่สนใจ และเกี่ยวข้องแล้ว การสุ่มตัวอย่างแบบกิบส์เพื่อประมาณฟังก์ชันความหนาแน่นน่าจะเป็นภายหลัง มีขั้นตอนดังนี้

ขั้นตอนที่ 1 กำหนดค่าเริ่มต้นให้กับ $\theta^{(0)} = \left[\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)}\right]$ โดยที่ $\theta^{(0)}$ เป็นพารามิเตอร์ที่สนใจและพารามิเตอร์ที่เกี่ยวข้อง

ขั้นตอนที่ 2 สุ่มตัวอย่าง $\theta^{(1)} = \left[\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_p^{(1)}\right]$ จากการแจกแจงแบบมีเงื่อนไขเต็มรูปดังนี้

สุ่ม $\theta_1^{(1)}$ จาก $\pi\left(\theta_1 \mid \theta_2^{(0)}, ..., \theta_p^{(0)}\right)$

สุ่ม $\theta_2^{(1)}$ จาก $\pi\left(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_p^{(0)}\right)$

$\vdots \qquad \vdots$

สุ่ม $\theta_p^{(1)}$ จาก $\pi\left(\theta_p \mid \theta_1^{(1)}, ..., \theta_{p-1}^{(1)}\right)$

จากนั้นใช้ $\theta^{(1)}$ ในการสุ่มตัวอย่างขั้นต่อไป

ขั้นตอนที่ 3 สุ่มตัวอย่าง $\theta^{(t+1)}$ เมื่อ $t = 2, 3, ...$ จากการแจกแจงแบบมีเงื่อนไขเต็มรูป ดังนี้

สุ่ม $\theta_1^{(t+1)}$ จาก $\pi\left(\theta_1 \mid \theta_2^{(t)}, ..., \theta_p^{(t)}\right)$

สุ่ม $\theta_2^{(t+1)}$ จาก $\pi\left(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_p^{(t)}\right)$

$\vdots \qquad \vdots$

สุ่ม $\theta_p^{(t+1)}$ จาก $\pi\left(\theta_p \mid \theta_1^{(t+1)}, ..., \theta_{p-1}^{(t+1)}\right)$

จากนั้นใช้ $\theta^{(t+1)}$ ในการสุ่มตัวอย่างขั้นต่อไป ทำซ้ำในขั้นตอนที่ 3 จนกระทั่ง Θ มีขนาดใหญ่พอที่จะทำให้ θ มีการลู่เข้า (Convergence)

2.10 การประมาณค่าพารามิเตอร์ด้วยวิธีเบส์เซียน

กำหนดให้ Y เป็นเวกเตอร์ของตัวแปรตามมีการแจกแจงแบบเบอร์นูลลี เขียนแทนด้วย $y_i \sim Ber\left(\pi\left(x_{ij}\right)\right)$ ซึ่งความน่าจะเป็นที่เกิดเหตุการณ์ที่สนใจคือ $\pi\left(x_{ij}\right) = P\left(y_i = 1\right)$ เพื่อให้สะดวกและง่ายต่อการวิเคราะห์ข้อมูล Albert and Chib (1993) ได้เสนอการแปลงค่าของตัวแปรตามแบบไบนารี่ ให้เป็นตัวแปรสุ่มชนิดต่อเนื่องที่มีค่าอยู่ในช่วง $\left(-\infty, \infty\right)$ ดังนี้

$$Logit\left(\pi\left(x_{ij}\right)\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = x_{ij}^T \beta$$

โดยที่

$$Logit\left(\pi\left(x_{ij}\right)\right)=\log\left(\frac{\pi\left(x_{ij}\right)}{1-\pi\left(x_{ij}\right)}\right)=x_{ij}^{T}\beta$$

ดังนั้น

$$\pi\left(x_{ij}\right)=\frac{e^{x_{ij}^{T}\beta}}{1+e^{x_{ij}^{T}\beta}}$$

พิจารณาฟังก์ชันภาวะน่าจะเป็น ดังนี้

$$\begin{aligned}
f\left(Y\mid x\beta\right)&=\prod_{i=1}^{n}f\left(y_i\mid x_{ij}\beta\right)\\
&=\prod_{i=1}^{n}\pi\left(x_{ij}\right)^{y_i}\left(1-\pi\left(x_{ij}\right)\right)^{1-y_i}\\
&=\pi\left(x_{ij}\right)^{\Sigma y_i}\left(1-\pi\left(x_{ij}\right)\right)^{n-\Sigma y_i}\\
&=\left(\frac{e^{x_{ij}^{T}\beta}}{1+e^{x_{ij}^{T}\beta}}\right)^{\Sigma y_i}\left(1-\frac{e^{x_{ij}^{T}\beta}}{1+e^{x_{ij}^{T}\beta}}\right)^{n-\Sigma y_i}\\
&=\left(\frac{e^{x_{ij}^{T}\beta}}{1+e^{x_{ij}^{T}\beta}}\right)^{\Sigma y_i}\left(\frac{1+e^{x_{ij}^{T}\beta}-e^{x_{ij}^{T}\beta}}{1+e^{x_{ij}^{T}\beta}}\right)^{n-\Sigma y_i}\\
&=\left(\frac{e^{x_{ij}^{T}\beta}}{1+e^{x_{ij}^{T}\beta}}\right)^{\Sigma y_i}\left(\frac{1}{1+e^{x_{ij}^{T}\beta}}\right)^{n-\Sigma y_i}
\end{aligned}$$

1. วิธีเบส์เซียนกรณีไม่ทราบความรู้เดิมเกี่ยวกับพารามิเตอร์

กำหนดฟังก์ชันความหนาแน่นจะเป็นก่อนของพารามิเตอร์ β ดังนี้

$$f\left(\beta\right)\propto C\quad ;-\infty<\beta<\infty$$

พิจารณาฟังก์ชันความหนาแน่นจะเป็นภายหลังของพารามิเตอร์ β ดังนี้

$$\begin{aligned}
h\left(\beta\mid y,x\right)&=\frac{f\left(Y\mid x,\beta\right)\cdot f\left(\beta\right)}{\int_{-\infty}^{\infty}f\left(Y\mid x,\beta\right)\cdot f\left(\beta\right)d\beta}\\
&\propto f\left(Y\mid x,\beta\right)\cdot f\left(\beta\right)
\end{aligned}$$

จะได้ฟังก์ชันความหนาแน่นจะเป็นภายหลังของพารามิเตอร์ β คือ

$$\begin{aligned} h(\beta \mid y,x) &\propto \prod_{i=1}^{n} f\left(y_i \mid x_{ij}\beta\right)\cdot f\left(\beta\right) \\ &\propto \left(\frac{e^{x_{ij}^{T}\beta}}{1+e^{x_{ij}^{T}\beta}}\right)^{\Sigma y_i}\left(\frac{1}{1+e^{x_{ij}^{T}\beta}}\right)^{n-\Sigma y_i} \end{aligned} \tag{22}$$

เนื่องจาก ฟังก์ชันความหนาแน่นน่าจะเป็นภายหลังของพารามิเตอร์ β ไม่ได้อยู่ในรูปของการแจกแจงเฉพาะ (Non-close form) ดังนั้นจะนำเทคนิคการจำลอง Markov chain Monte Carlo (MCMC) ด้วยการสุ่มตัวอย่างแบบกิบส์ (Gibbs Sampling) มาใช้เพื่อหาค่าประมาณพารามิเตอร์แบบเบส์

2. วิธีเบส์เซียนกรณีทราบความรู้เดิมเกี่ยวกับพารามิเตอร์

กำหนดฟังก์ชันความหนาแน่นจะเป็นก่อนของพารามิเตอร์ β ดังนี้

$$f\left(\beta_j\right)=\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2} \qquad ;\mathrm{j}=0,\ldots,\mathrm{k}$$

พิจารณาฟังก์ชันความหนาแน่นจะเป็นภายหลังของพารามิเตอร์ β ดังนี้

$$\begin{aligned} h(\beta \mid y,x) &= \frac{f\left(Y \mid x,\beta\right)\cdot f\left(\beta_j\right)}{\int_{-\infty}^{\infty} f\left(Y \mid x,\beta\right)\cdot f\left(\beta_j\right)d\beta} \\ &\propto f\left(Y \mid x,\beta\right)\cdot f\left(\beta_j\right) \end{aligned}$$

จะได้ฟังก์ชันความหนาแน่นจะเป็นภายหลังของพารามิเตอร์ β คือ

$$\begin{aligned} h(\beta \mid y,x) &\propto \prod_{i=1}^{n} f\left(y_i \mid x_{ij},\beta\right)\cdot\prod_{j=0}^{k} f\left(\beta_j\right) \\ &\propto \left(\frac{e^{x_{ij}^{T}\beta}}{1+e^{x_{ij}^{T}\beta}}\right)^{\Sigma y_i}\left(\frac{1}{1+e^{x_{ij}^{T}\beta}}\right)^{n-\Sigma y_i}\prod_{j=1}^{k}\left[\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2}\right] \end{aligned} \tag{23}$$

เนื่องจาก ฟังก์ชันความหนาแน่นน่าจะเป็นภายหลังของพารามิเตอร์ β ไม่ได้อยู่ในรูปของการแจกแจงเฉพาะ (Non-close form) ดังนั้นจะนำเทคนิคการจำลอง Markov chain Monte Carlo (MCMC) ด้วยการสุ่มตัวอย่างแบบกิบส์ (Gibbs Sampling) มาใช้เพื่อหาค่าประมาณพารามิเตอร์แบบเบส์

2.11 เกณฑ์การตัดสินใจ

การเปรียบเทียบประสิทธิภาพของการพยากรณ์ในตัวแบบถดถอยลอจิสติก เมื่อประมาณค่าพารามิเตอร์ ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Method) วิธีเบส์เซียน (Bayesian Method) และวิธีฟังก์ชันสกอร์ที่ปรับปรุง (Modified Score Function) ร่วมกับการจัดการข้อมูลไม่สมดุล จะพิจารณาจากค่าความแม่นยำ (Accuracy) อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยหรือความไว (Sensitivity) อัตราความถูกต้องในการทำนายกลุ่มส่วนมากหรือความจำเพาะ (Specificity) และค่าความแม่นที่สมดุล (Balanced Accuracy) ซึ่งโดยมีรายละเอียดดังนี้

ตาราง 1 เมทริกซ์ความสับสน (Confusion Matrix) แสดงผลของค่าจริงและผลการพยากรณ์

ค่าพยากรณ์	ค่าจริง	
	Positive (กลุ่มส่วนน้อย)	Negative (กลุ่มส่วนมาก)
Positive (กลุ่มส่วนน้อย)	True positive (TP)	False positive (FP)
Negative (กลุ่มส่วนมาก)	False negative (FN)	True negative (TN)

โดยที่ True positive (TP) แสดงถึงจำนวนข้อมูลที่อยู่ในกลุ่ม Positive และทำนายว่าอยู่ในกลุ่ม Positive

False negative (FN) แสดงถึงจำนวนข้อมูลที่อยู่ในกลุ่ม Positive และทำนายว่าอยู่ในกลุ่ม Negative

False positive (FP) แสดงถึงจำนวนข้อมูลที่อยู่ในกลุ่ม Negative และทำนายว่าอยู่ในกลุ่ม Positive

True negative (TN) แสดงถึงจำนวนข้อมูลที่อยู่ในกลุ่ม Negative และทำนายว่าอยู่ในกลุ่ม Negative

1) ค่าความแม่นยำ (Accuracy) แสดงถึงประสิทธิภาพความถูกต้องในการทำนายในภาพรวม ซึ่งคำนวณได้ดังนี้

$$Accurancy = \frac{TP + TN}{TP + TN + FP + FN}$$

2) อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยหรือความไว (True positive rate or Sensitivity) แสดงถึงความถูกต้องในการทำนายกลุ่มส่วนน้อยว่าอยู่กลุ่มส่วนน้อย ซึ่งคำนวณได้ดังนี้

$$Sensitivity = \frac{TP}{TP + FN}$$

3) อัตราความถูกต้องในการทำนายกลุ่มส่วนมากหรือความจำเพาะ (True negative rate or Specificity) แสดงถึงความถูกต้องในการทำนายกลุ่มส่วนมากว่าอยู่กลุ่มส่วนมาก ซึ่งคำนวณได้ดังนี้

$$Specificity = \frac{TN}{FP + TN}$$

4) ค่าความแม่นที่สมดุล (Balanced Accuracy) คือ ค่าที่ใช้วัดความแม่นยำในการจำแนกประเภทตัวแปรหรือข้อมูลที่ถ่วงน้ำหนักระหว่างค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity) ซึ่งคำนวณได้ดังนี้

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

โดยเกณฑ์การวัดแต่ละเกณฑ์ หากมีค่าสูงแสดงถึงประสิทธิภาพในการจำแนกที่ดี

2.12 งานวิจัยที่เกี่ยวข้อง

สุภวรรณ มานะการ (2549) ได้ศึกษาการจำแนกกลุ่มโดยใช้วิธีถดถอยลอจิสติกประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด และวิธีนาอีฟ เบส์ ซึ่งการวิเคราะห์ข้อมูลจะแบ่งเป็น 2 ชุด คือชุดตัวอย่างที่ใช้สร้างเกณฑ์การจำแนก และชุดตัวอย่างที่ใช้ตรวจสอบเกณฑ์การจำแนก โดยมีขนาดตัวอย่างที่ใช้ในการศึกษา เป็น 30, 50, 70, 100 และ 500 จำนวนตัวแปรอิสระเป็น 4, 6 และ 10 ตามลำดับ ทำการเปรียบเทียบในอัตราส่วน Training : Validation เป็น 90:10, 80:20, 70:30, 60:40 และ 50:50 จำลองข้อมูลโดยใช้โปรแกรม MATLAB กระทำซ้ำ 1,000 ครั้งในแต่ละสถานการณ์ เกณฑ์การเปรียบเทียบคือ อัตราความผิดพลาดของการจำแนก (Error Rate of Misclassification) ผลการศึกษาพบว่า อัตราการจำแนกกลุ่มข้อมูลผิดพลาดทั้งสองวิธีมีแนวโน้มลดลง เมื่อขนาดตัวอย่างและอัตราส่วนของข้อมูล Training เพิ่มขึ้น และกรณีที่สัดส่วนค่าของตัวแปรอิสระของข้อมูลทั้งสองกลุ่มมีค่าใกล้เคียงกัน พบว่าอัตราการจำแนกกลุ่มข้อมูลผิดผลาดของทั้งสองวิธีมีแนวโน้มน้อยกว่ากรณีที่สัดส่วนค่าของตัวแปรอิสระของข้อมูลทั้งสองกลุ่มมีค่าแตกต่างกัน และอัตราการจำแนกกลุ่มข้อมูลผิดผลาดด้วยวิธีนาอีฟ เบส์ ส่วนใหญ่แล้วจะให้ค่าต่ำกว่าวิธีการถดถอยลอจิสติกเล็กน้อย

Wah, et al. (2016) ได้ศึกษาเปรียบเทียบประสิทธิภาพของ วิธีการจัดการข้อมูลไม่สมดุลด้วยเทคนิคการสุ่มตัวอย่างเกินและเทคนิคการสุ่มตัวอย่างลดร่วมกับวิธีเวกเตอร์ค้ำยัน (Support Vector Machine) วิธีเพื่อนบ้านที่ใกล้ที่สุด (k - Nearest Neighbour Algorithm) และวิธีการถดถอยลอจิสติก (Logistic Regression) ในการศึกษานี้ใช้ชุดข้อมูลการผ่าตัดหัวใจที่ได้จาก

โรงพยาบาลท้องถิ่นในกรุงกัวลาลัมเปอร์ โดยตัวแปรตามเป็นไบนารี่ กำหนด 1 แทน เสียชีวิตหลังการผ่าตัด และ 0 แทน มีชีวิตหลังการผ่าตัด ตัวแปรอิสระ 8 ตัว กลุ่มตัวอย่างทั้งหมด 4,976 ราย โดยผู้ป่วย 4,767 ราย (95.8%) มีชีวิตหลังการผ่าตัด และผู้ป่วยเพียง 209 ราย (4.2%) ที่เสียชีวิตหลังการผ่าตัด ซึ่งแสดงให้เห็นว่าข้อมูลการผ่าตัดหัวใจมีความไม่สมดุล ใช้เกณฑ์การวัดประสิทธิภาพคือ Accuracy, Sensitivity และ Specificity ผลการศึกษาพบว่า ค่า Sensitivity เพิ่มขึ้นเมื่อใช้วิธีจำแนกกลุ่มทั้ง 3 วิธี ร่วมกับเทคนิคการสุ่มตัวอย่างทั้งสองวิธี นอกจากนี้เทคนิคการสุ่มตัวอย่างเกิน มีประสิทธิภาพดี เมื่อใช้ร่วมกับวิธีเวกเตอร์ค้ำยันและวิธีเพื่อนบ้านที่ใกล้ที่สุด

Febrianti, et al. (2018) ได้ศึกษาเปรียบเทียบประสิทธิภาพการประมาณค่าพารามิเตอร์ของตัวแบบถดถอยลอจิสติก โดยทำการเปรียบเทียบ 2 วิธี คือวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) และวิธีฟังก์ชันสกอร์ที่ปรับปรุง (Modified Score Function) ทำการศึกษากับชุดข้อมูลมะเร็งเยื่อบุโพรงมดลูก กำหนดตัวแปรตามแบบไบนารี่ โดย 1 แทนอาการรุนแรง และ 0 แทนอาการไม่รุนแรง ข้อมูลทั้งหมด 79 ค่าสังเกต ประกอบไปด้วย y = 1 จำนวน 30 ค่าสังเกต และ y = 0 จำนวน 49 ค่าสังเกต สุ่มตัวอย่างขนาดเท่ากับ 10, 20 และ 30 ตัว มีสัดส่วนของเหตุการณ์ที่สนใจเท่ากับ 0.1 และกำหนดเกณฑ์การวนซ้ำสูงสุดที่ 10,000 รอบ จากผลการศึกษาพบว่า เมื่อขนาดตัวอย่างและสัดส่วนของเหตุการณ์ที่สนใจมีขนาดเล็ก จะเกิดปัญหาการวนซ้ำไม่ให้ผลลัพธ์ที่บรรจบกัน เมื่อใช้วิธีภาวะน่าจะเป็นสูงสุดร่วมกับการวนซ้ำของ Newton-Raphson ทำให้ไม่สามารถประมาณค่าพารามิเตอร์ได้ ซึ่งแก้ไขได้โดยใช้วิธีฟังก์ชันสกอร์ที่ปรับปรุง ทำให้กระบวการวนซ้ำบรรจบกันและให้ค่าประมาณพารามิเตอร์ได้อย่างรวดเร็ว

Hassan (2020) ได้ศึกษาการประมาณค่าพารามิเตอร์ของตัวแบบถดถอยลอจิสติก ด้วยวิธีแบบเบส์ และวิธีอื่นอีก 5 วิธี ได้แก่ วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีเวกเตอร์ค้ำยัน (Support Vector Machine SVM) วิธีเพื่อนบ้านที่ใกล้ที่สุด (k – Nearest Neighbours Algorithm) วิธีการถดถอยลอจิสติก (Logistic Regression) และวิธีนาอีฟ เบส์ (Naïve Bayes) ในชุดข้อมูลผู้ป่วยเบาหวาน จำนวน 909 ราย จากเมือง Zakho โดยกำหนดตัวแปรตามแบบไบนารี่ให้ 1 แทน มีอาการเบาหวาน และ 0 แทน สุขภาพดี ซึ่งมีตัวแปรอิระ 7 ตัว ซึ่งการประมาณค่าด้วยวิธีแบบเบส์ ได้กำหนดการแจกแจงความน่าจะเป็นก่อน แบบ Gaussian, Laplace และ Cauchy หาการแจกแจงภายหลังโดยใช้การจำลองแบบ Markov Chain Monte Carlo (MCMC) สุ่มตัวอย่างด้วยวิธี Gibbs Sampling เพื่อประมาณการแจกแจงความน่าจะเป็นภายหลัง และเกณฑ์วัดประสิทธิภาพของแบบจำลองได้แก่ ค่าความแม่นยำ (Accuracy) ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) และค่าประสิทธิภาพ (F-measure) ผลการศึกษาพบว่าวิธีแบบเบส์กับการแจกแจงความน่าจะเป็นก่อนแบบเกาส์เซียนมีประสิทธิภาพสูงสุด สามารถจำแนกผู้ป่วยได้ถูกต้องเกือบทั้งหมด

Brandt and Lanzén (2020) ได้ศึกษาเปรียบเทียบประสิทธิภาพของการจำแนกกลุ่ม เมื่อข้อมูลไม่สมดุล โดยใช้เทคนิคการสุ่มตัวอย่าง 2 เทคนิค คือ Synthetic Minority Over - sampling Technique (SMOTE) และ Adaptive Synthetic sampling approach (ADASYN) ใช้แบบจำลองที่แตกต่างกัน 3 วิธี คือ วิธีการถดถอยลอจิสติก (Logistic Regression) เทคนิคป่าสุ่ม (Random Forest Classifier) และวิธีเวกเตอร์ค้ำยัน (Support Vector Machines) เปรียบเทียบกับชุดข้อมูลไม่สมดุล 3 ชุด จากฐานข้อมูล Kaggle.com โดยมีระดับความไม่สมดุลที่แตกต่างกัน ข้อมูลชุดแรกเรียกว่า Predicting Churn for Bank Customers สำหรับลูกค้าธนาคาร โดยมีตัวแปรตามคือ ลูกค้ายกเลิกการใช้บริการจากธนาคารหรือไม่ กลุ่มส่วนน้อยคิดเป็นร้อยละ 20.37 ของการสังเกตทั้งหมด ชุดข้อมูลที่สองเรียกว่า Home Credit Default Risk โดยมีตัวแปรตามคือลูกค้าผิดนัดชำระหนี้หรือไม่ โดยกลุ่มส่วนน้อยคิดเป็นร้อยละ 8.07 ของค่าสังเกตทั้งหมด และข้อมูลชุดที่สามเรียกว่า Credit Card Fraud Detection เรียกว่าการตรวจจับการฉ้อโกงของบัตร มีตัวแปรตาม คือ ธุรกรรมบัตรเครดิตเป็นการฉ้อโกงหรือไม่ โดยกลุ่มส่วนน้อยคิดเป็นร้อยละ 0.17 ของค่าสังเกตทั้งหมด เกณฑ์การเปรียบเทียบได้แก่ ค่าความไว(Sensitivity) ค่าประสิทธิภาพ (F-measure) และ Matthews correlation coefficient (MCC) ผลการศึกษาพบว่า สำหรับเกณฑ์การเปรียบเทียบทั้ง 3 เกณฑ์ ไม่มีเทคนิคการสุ่มตัวอย่างร่วมกับตัวแบบที่สามารถปรับปรุงประสิทธิภาพของชุดข้อมูลทั้ง 3 ชุด ได้อย่างสม่ำเสมอ แต่อย่างไรก็ตาม ผลลัพธ์แสดงให้เห็นว่าการใช้เทคนิคการสุ่มตัวอย่าง SMOTE ช่วยปรับปรุงประสิทธิภาพของวิธีเวกเตอร์ค้ำยัน ได้เป็นส่วนใหญ่ โดยเฉพาะเมื่อระดับความไม่สมดุลเพิ่มขึ้น นอกจากนี้เทคนิคการสุ่มตัวอย่างทั้ง 2 เทคนิค สามารถปรับปรุงประสิทธิภาพของเทคนิคป่าสุ่มได้เมื่อระดับความไม่สมดุลเพิ่มขึ้น

Hezlin, et al. (2021) ได้ศึกษาการประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติก เมื่อข้อมูลไม่สมดุล ด้วยวิธีภาวะน่าจะเป็นสูงสุด กำหนดขนาดตัวอย่างเป็น 100, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500 และ 5000 ตามลำดับ และอัตราส่วนของความไม่สมดุล (Imbalanced Ratio : IR) ประกอบไปด้วย 8 อัตราส่วน คือ 1%, 2%, 5%, 10%, 20%, 30%, 40% และ 50% และใช้เกณฑ์การเปรียบเทียบ คือ ความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error) ผลการศึกษาพบว่า ข้อมูลที่มีความไม่สมดุลมีผลกระทบต่อการประมาณค่าพารามิเตอร์และประสิทธิภาพในการจัดกลุ่มของแบบจำลองถดถอยลอจิสติกแบบทวิ ซึ่งสรุปได้ว่าผลกระทบของอัตราส่วนความไม่สมดุลต่อการประมาณค่าพารามิเตอร์ลดลง เมื่อขนาดตัวอย่างเพิ่มขึ้น นอกจากนี้อัตราส่วนที่ไม่สมดุลในตัวแปรตามไม่เพียงส่งผลต่อการประมาณค่าพารามิเตอร์เท่านั้น แต่ยังส่งผลต่อค่า p - value และ ค่า odds- ratio ดังนั้นเมื่อข้อมูลไม่สมดุลจะนำไปสู่ปัญหาการจำแนกผิดกลุ่ม โดยแนวทางสำหรับการจัดการปัญหาข้อมูลไม่สมดุล เช่น การสุ่มตัวอย่างด้วยวิธีการสุ่มลด วิธีการสุ่มเกิน และวิธีการสังเคราะห์ข้อมูลใหม่ เป็นต้น

Yilmaz and Celik (2021) ได้ศึกษาเปรียบเทียบประสิทธิภาพการประมาณค่าพารามิเตอร์ของแบบจำลองถดถอยลอจิสติกทวิ ด้วยวิธีภาวะน่าจะเป็นสูงสุด และวิธีแบบเบส์ เมื่อตัวอย่างมีขนาดเล็ก โดยใช้ชุดข้อมูลองค์การเพื่อความร่วมมือทางเศรษฐกิจและการพัฒนา (OECD) ประกอบไปด้วยข้อมูลประชากรและเศรษฐกิจต่างๆ จาก 34 ประเทศที่เป็นสมาชิก OECD โดยตัวแปรตามคือ การเป็นสมาชิกสหภาพยุโรป (EU) กำหนด 1 แทนการเป็นสมาชิก และ 0 แทน การไม่เป็นสมาชิก มีตัวแปรอิสระ 9 ตัว เกณฑ์การวัดประสิทธิภาพของแบบจำลองได้แก่ AIC และ BIC ผลการศึกษาพบว่าวิธีแบบเบส์ ที่กำหนดการแจกแจงความน่าจะเป็นก่อน แบบ Gaussian มีอัตราส่วนการจำแนกกลุ่มที่ถูกต้องสูงกว่า และมีค่า AIC และ BIC ต่ำกว่าวิธีภาวะน่าจะเป็นสูงสุด นั่นคือเมื่อตัวอย่างมีขนาดเล็กวิธีการประมาณค่าพารามิเตอร์แบบเบส์มีประสิทธิภาพที่ดีกว่าวิธีภาวะน่าจะเป็นสูงสุด

จากการศึกษางานวิจัยที่กล่าวมาข้างต้นจะเห็นได้ว่า เมื่อข้อมูลมีความไม่สมดุล นำไปวิเคราะห์จะส่งผลให้การประมาณค่าพารามิเตอร์ หรือการพยากรณ์ไม่ถูกต้อง จำเป็นต้องจัดการความไม่สมดุลของข้อมูลก่อนเป็นดันดับแรก ซึ่งวิธีที่นิยมได้แก่ วิธีการการสุ่มลด วิธีการสุ่มเกิน และวิธีการสังเคราะห์ข้อมูลใหม่ โดยการปรับข้อมูลทั้งสองกลุ่มคือ ในกลุ่มส่วนใหญ่และกลุ่มส่วนน้อยมีค่าเท่ากันหรือใกล้เคียงกัน เมื่อปรับปรุงข้อมูลให้มีความสมดุลแล้วนำไปประมาณค่าพารามิเตอร์ในการวิเคราะห์ถดถอยลอจิสติก จากงานวิจัยข้างต้นพบว่า วิธีภาวะน่าจะเป็นสูงสุด เป็นวิธีที่นิยมใช้อย่างแพร่หลายในการวิเคราะห์การถดถอยลอจิสติก แต่เมื่อขนาดตัวอย่างเล็ก วิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีเบส์เซียน ให้ประสิทธิภาพการวิเคราะห์ที่ดีเป็นส่วนใหญ่ ดังนั้นผู้วิจัยจึงสนใจวิธีจัดการข้อมูลไม่สมดุลทั้ง 3 วิธี ข้างต้น มาใช้ร่วมกับการประมาณค่าพารามิเตอร์ในการวิเคราะห์ถดถอยลอจิสติกด้วยวิธีภาวะน่าจะเป็นสูงสุด วิธีฟังก์ชันสกอร์ที่ปรับปรุง และวิธีเบส์เซียน โดยวัดประสิทธิภาพด้วยค่าความแม่นยำ ค่าความไว ค่าความจำเพาะ และค่าความแม่นที่สมดุล

บทที่ 3
วิธีดำเนินงานวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการพยากรณ์ในตัวแบบถดถอยลอจิสติก เมื่อประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด วิธีเบส์เซียน และวิธีฟังก์ชันสกอร์ที่ปรับปรุง ร่วมกับการจัดการข้อมูลไม่สมดุล โดยใช้ค่าความแม่นยำ อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยหรือความไว อัตราความถูกต้องในการทำนายกลุ่มส่วนมากหรือความจำเพาะ และค่าความแม่นที่สมดุล เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพ

ขอบเขตของการวิจัย

ผู้วิจัยได้กำหนดขอบเขตของงานวิจัยดังนี้

1. กำหนดระดับความไม่สมดุลของข้อมูล 2 กลุ่ม ดังนี้ กลุ่ม 0 : กลุ่ม 1 เป็น 90:10, 80:20, 70:30 และ 60:40
2. กำหนดขนาดตัวอย่าง (n) ที่ใช้ในการศึกษาเท่ากับ คือ 100 และ 500
3. กำหนดจำนวนตัวแปรอิสระ (p) ที่ใช้ในการศึกษา คือ 1 และ 3 ตัว
4. กำหนดตัวแปรอิสระเป็นข้อมูลเชิงปริมาณที่มีการแจกแจงดังนี้
   - การแจกแจงแบบปรกติหลายตัวแปร (Multivariate Normal Distribution) โดยมีฟังก์ชันความหนาแน่นน่าจะเป็น ดังนี้

$$f(\boldsymbol{y};\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})};-\infty<\boldsymbol{y}<\infty,-\infty<\boldsymbol{\mu}<\infty,\Sigma>0$$

เมื่อ เวกเตอร์ค่าเฉลี่ยของตัวแปรสุ่ม Y คือ $E(Y)=\mu=\begin{bmatrix}\mu_1\\ \vdots\\ \mu_p\end{bmatrix}$

เมทริกซ์ความแปรปรวนร่วมของตัวแปรสุ่ม Y คือ $Cov(Y)=\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{bmatrix}$

   - ตัวแปรอิสระ 1 ตัวแปร

$$\boldsymbol{\mu}=\mu_1=0 \text{ และ } \Sigma=\sigma_{11}=1$$

- ตัวแปรอิสระ 3 ตัวแปร

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{ และ } \Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

5. กำหนดตัวแปรตามเป็นตัวแปรทวิภาคมีค่าเป็น 0 กับ 1 เมื่อกำหนดให้

   0 แทน เหตุการณ์ที่ไม่สนใจ

   1 แทน เหตุการณ์ที่สนใจ

   ซึ่งตัวแปรตามสร้างมาจากการแจกแจงเบอร์นูลลี

6. กำหนดความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจในประชากร (π) เป็น 0.1

7. กำหนดค่า β เริ่มต้นมีค่าเท่ากับ 1

   - ตัวแปรอิสระ 1 ตัวแปร

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

   - ตัวแปรอิสระ 3 ตัวแปร

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

8. การวิเคราะห์จะแบ่งตัวอย่างออกเป็น 2 ชุด คือ ชุดตัวอย่างที่ใช้สร้างเกณฑ์การจำแนก และชุดตัวอย่างที่ใช้ตรวจสอบเกณฑ์การจำแนก คือ 70:30 และ 80:20

9. กำหนดฟังก์ชันความหนาแน่นน่าจะเป็นก่อน (Prior Probability Density Function) ในวิธีเบส์เซียน

   9.1 กรณีไม่ทราบความรู้เดิมเกี่ยวกับพารามิเตอร์ β ดังนี้

$$f(\beta) \propto c \quad ; -\infty < \beta < \infty \text{ เมื่อ } c \text{ เป็นค่าคงที่}$$

   9.2 กรณีทราบความรู้เดิมเกี่ยวกับพารามิเตอร์ β ดังนี้

   - สำหรับตัวแปรอิสระ 1 ตัว

$$\beta \sim N_{n\times(p+1)}(\mu, \Sigma) \text{ เมื่อ } \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ และ } \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

- สำหรับตัวแปรอิสระ 3 ตัว

$$\beta \sim N_{n\times(p+1)}(\mu,\Sigma) \quad \text{เมื่อ} \quad \mu=\begin{bmatrix}0\\0\\0\\0\end{bmatrix} \quad \text{และ} \quad \Sigma=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

10. จำลองข้อมูลโดยใช้โปรแกรม R ทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์

ขั้นตอนการวิจัย

การดำเนินการวิจัยมีขั้นตอนดังนี้

1. กำหนดขนาดประชากร (N)
2. กำหนดค่าพารามิเตอร์เริ่มต้น
   - กรณีตัวแปรอิสระ 1 ตัว

$$\beta=\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}$$

   - กรณีตัวแปรอิสระ 3 ตัว

$$\beta=\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\\\beta_3\end{bmatrix}=\begin{bmatrix}1\\1\\1\\1\end{bmatrix}$$

3. กำหนดจำนวนตัวแปรอิสระ (p) ที่ใช้ในการศึกษา เป็น 1 และ 3 ตัว
4. กำหนดขนาดตัวอย่าง (n) เป็น 100 และ 500
5. สร้างตัวแปรอิสระจากการแจกแจงปรกติหลายตัวแปร โดยกำหนดเวกเตอร์ค่าเฉลี่ยของตัวแปรสุ่ม Y และเมทริกซ์ความแปรปรวนร่วมของตัวแปรสุ่ม Y ดังนี้
   - กรณีตัวแปรอิสระ 1 ตัว

     $\mu=\mu_1=0$ และ $\Sigma=\sigma_{11}=1$

   - กรณีตัวแปรอิสระ 3 ตัว

$$\boldsymbol{\mu}=\begin{bmatrix}0\\0\\0\end{bmatrix} \text{ และ } \boldsymbol{\Sigma}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$$

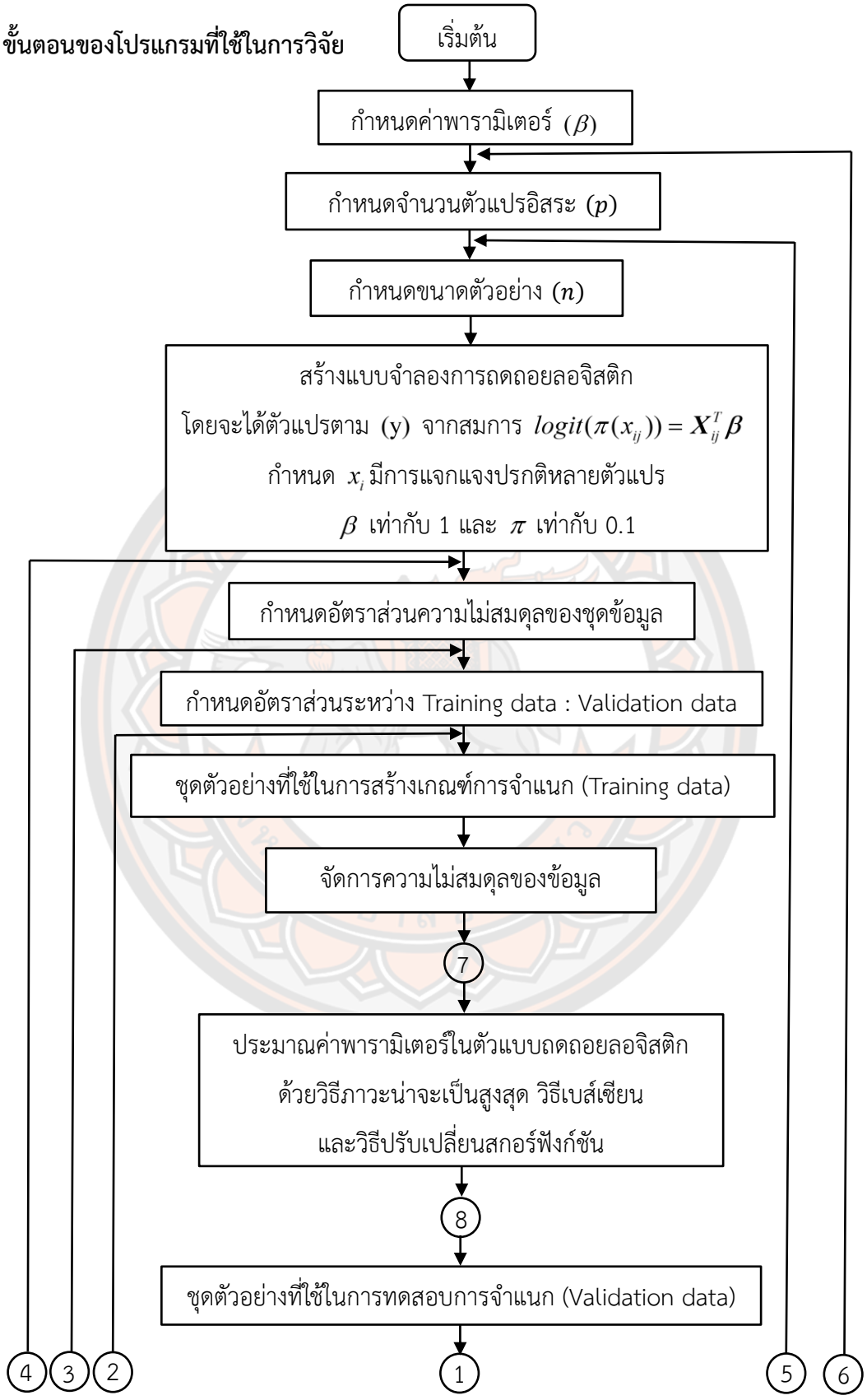
6. สร้างตัวแบบถดถอยลอจิสติก โดยนำตัวแปรอิสระแทนค่าลงในสมการถดถอยลอจิสติก
7. กำหนดความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจในประชากร (π) เท่ากับ 0.1

   ถ้า $\pi>0.1$ ตัวแปรตามอยู่ในกลุ่มส่วนมาก

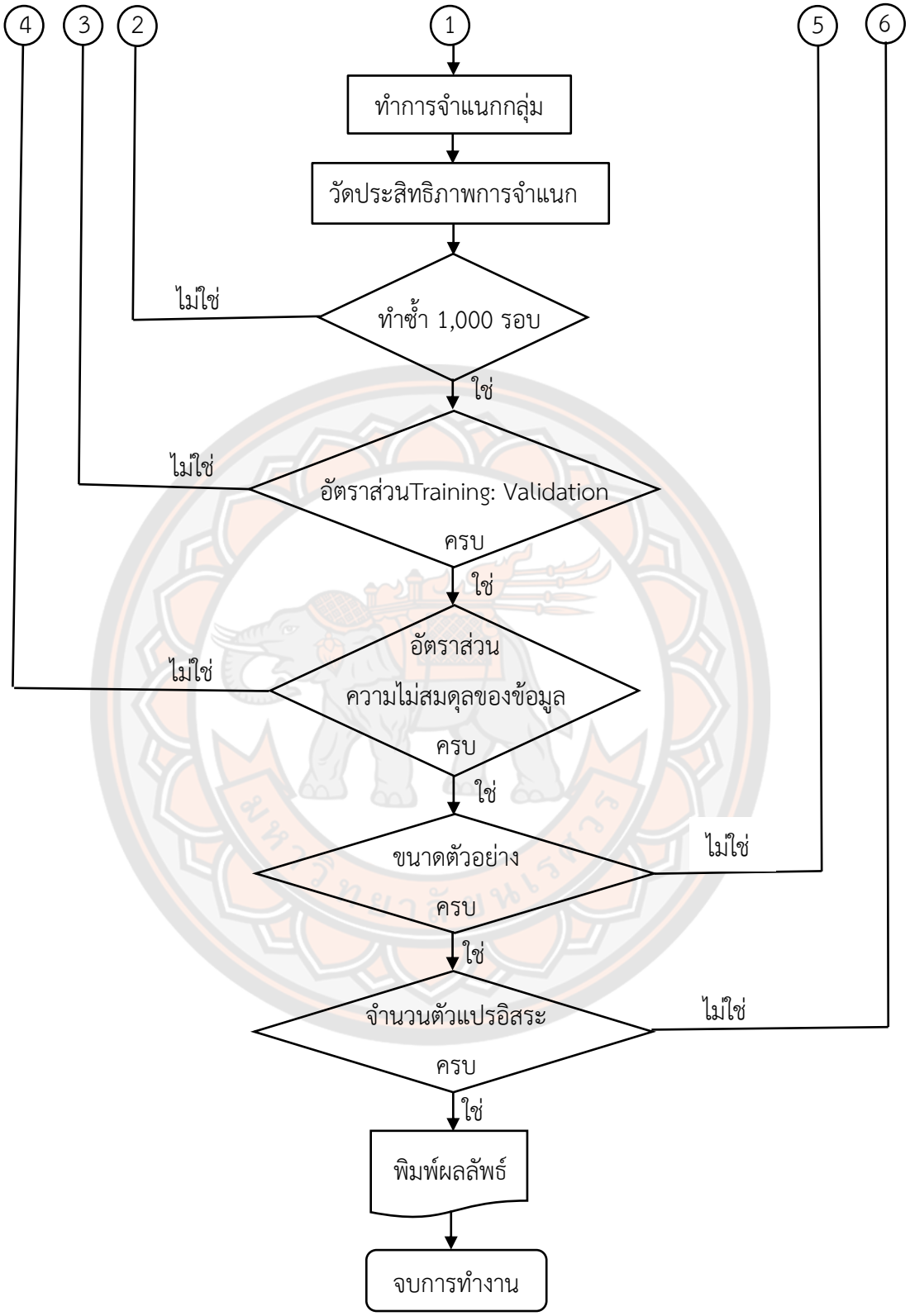
   ถ้า $\pi\le 0.1$ ตัวแปรตามอยู่ในกลุ่มส่วนน้อย

8. กำหนดอัตราส่วนความไม่สมดุลของกลุ่ม 0 และกลุ่ม 1 เป็น 90:10, 80:20, 70:30 และ 60:40
9. กำหนดอัตราส่วนระหว่าง Training : Validation เป็น 70:30 และ 80:20
10. จัดการความไม่สมดุลของข้อมูลในชุดตัวอย่างที่ใช้สร้างเกณฑ์การจำแนก (Training Data) ด้วยวิธีสุ่มลด สุ่มเกิน และวิธีสังเคราะห์ข้อมูลใหม่
11. ประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติก ด้วยวิธีภาวะน่าจะเป็นสูงสุด วิธีเบส์เซียน และวิธีฟังก์ชันสกอร์ที่ปรับปรุง
12. คำนวณค่าความแม่นยำ อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยหรือความไว อัตราความถูกต้องในการทำนายกลุ่มส่วนใหญ่หรือความจำเพาะ และค่าความแม่นที่สมดุล ของแต่ละวิธีในชุดตัวอย่างที่ใช้ตรวจสอบเกณฑ์การจำแนก (Validation Data)
13. สรุปผลการวิจัยและอภิปรายผล

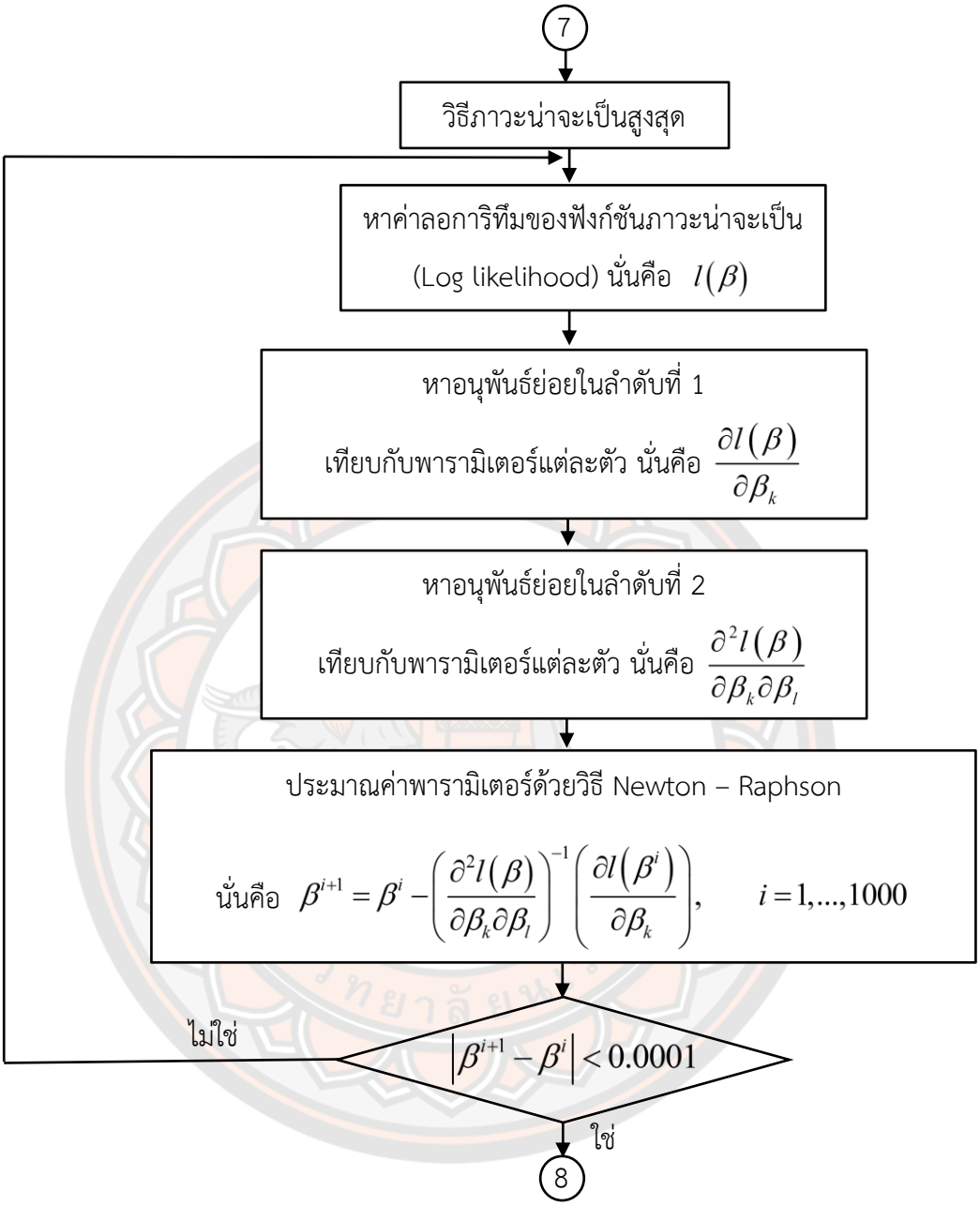
ขั้นตอนของโปรแกรมที่ใช้ในการวิจัย



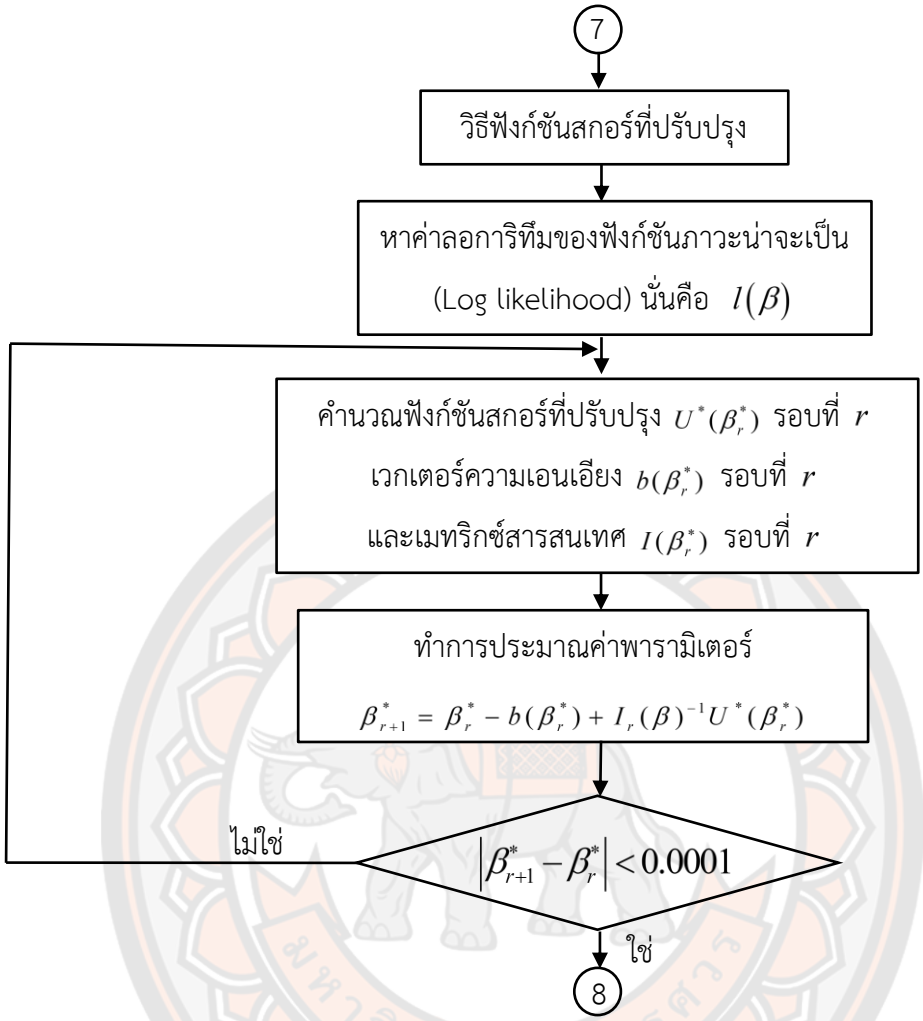
ภาพ 7 แผนผังแสดงขั้นตอนการดำเนินการวิจัย


4
3
2
1
5
6
ทำการจำแนกกลุ่ม
วัดประสิทธิภาพการจำแนก
ทำซ้ำ 1,000 รอบ
ไม่ใช่
ใช่
อัตราส่วนTraining: Validation
ไม่ใช่
ครบ
ใช่
อัตราส่วน
ความไม่สมดุลของข้อมูล
ไม่ใช่
ครบ
ใช่
ขนาดตัวอย่าง
ไม่ใช่
ครบ
ใช่
จำนวนตัวแปรอิสระ
ไม่ใช่
ครบ
ใช่
พิมพ์ผลลัพธ์
จบการทำงาน

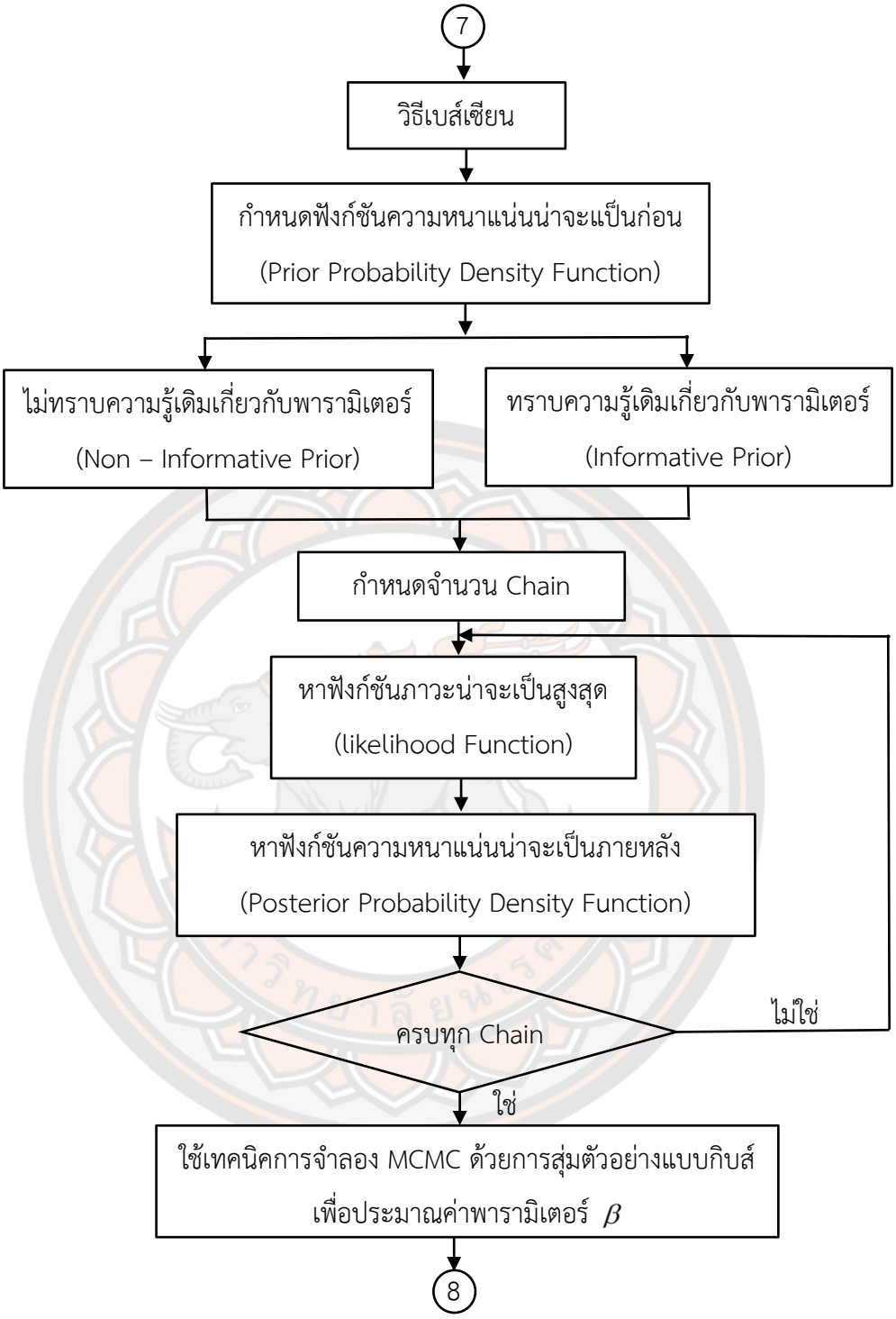
ภาพ 7 (ต่อ)



ภาพ 8 แผนผังแสดงขั้นตอนการทำงานด้วยวิธีภาวะน่าจะเป็นสูงสุด



ภาพ 9 แผนผังแสดงขั้นตอนการทำงานด้วยวิธีฟังก์ชันสกอร์ที่ปรับปรุง



ภาพ 10 แผนผังแสดงขั้นตอนการทำงานด้วยวิธีเบส์เซียนกรณีไม่ทราบและทราบความรู้เดิมเกี่ยวกับพารามิเตอร์

บทที่ 4
ผลการวิจัย

การวิจัยนี้มีจุดมุ่งหมายเพื่อศึกษาและเปรียบเทียบประสิทธิภาพของการพยากรณ์ในตัวแบบถดถอยลอจิสติก เมื่อประมาณค่าพารามิเตอร์ ด้วยวิธีภาวะน่าจะเป็นสูงสุด วิธีเบส์เซียน และวิธีฟังก์ชันสกอร์ที่ปรับปรุง ร่วมกับการจัดการข้อมูลไม่สมดุล เกณฑ์การเปรียบเทียบพิจารณาจาก ค่าความแม่นยำ อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยหรือความไว อัตราความถูกต้องในการทำนายกลุ่มส่วนมากหรือความจำเพาะ และค่าความแม่นที่สมดุล ในที่นี้กำหนดสัญลักษณ์ที่ใช้ในการวิจัย ดังต่อไปนี้

n	แทน	ขนาดตัวอย่าง
MLE	แทน	วิธีภาวะน่าจะเป็นสูงสุด
Score	แทน	วิธีฟังก์ชันสกอร์ที่ปรับปรุง
Baye non	แทน	วิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อน
Baye	แทน	วิธีเบส์เซียน กรณีทราบความรู้ก่อน
IR	แทน	อัตราส่วนความไม่สมดุล
RUS	แทน	วิธีการสุ่มลด
ROS	แทน	วิธีการสุ่มเกิน
SMOTE	แทน	วิธีการสังเคราะห์ข้อมูลใหม่
Acc	แทน	ความแม่นยำ
Sen	แทน	ความไว
Spec	แทน	ความจำเพาะ
Balanced Acc	แทน	ความแม่นที่สมดุล

ผลการวิเคราะห์ข้อมูล

จากการจำแนกกลุ่มข้อมูลที่มีความไม่สมดุลเมื่อประมาณพารามิเตอร์ในตัวแบบถดถอยลอจิสติกด้วยวิธีภาวะน่าจะเป็นสูงสุด วิธีเบส์เซียน และวิธีฟังก์ชันสกอร์ที่ปรับปรุง ตามสถานการณ์ที่กำหนด ได้ผลการวิเคราะห์ดังต่อไปนี้

ตาราง 2 ผลการวัดประสิทธิภาพของตัวแบบถดถอยลอจิสติก เมื่อกำหนดตัวแปรอิสระเท่ากับ 1 ตัว ขนาดตัวอย่างเท่ากับ 100 และกำหนดอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30

IR	เกณฑ์การวัดประสิทธิภาพ	วิธีการประมาณค่าพารามิเตอร์											
		MLE			SCORE			Bayesians					
								Non-informative			Informative		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	Acc.	***0.8521***	***0.8526***	***0.8584***	***0.8567***	***0.8562***	***0.8575***	0.8097	0.8004	0.7687	0.8033	0.7878	0.7468
	Sen.	***0.8931***	***0.8925***	0.8408	0.8739	0.8677	0.8177	0.5804	0.5522	0.4552	0.5622	0.5149	0.3926
	Spec.	0.8247	0.8259	0.8701	0.8453	0.8486	0.8840	0.9626	0.9659	0.9777	0.9641	0.9697	***0.9829***
	Balanced Acc.	***0.8589***	***0.8592***	***0.8554***	***0.8596***	***0.8581***	***0.8508***	0.7715	0.7591	0.7164	0.7631	0.7423	0.6877
70:30	Acc.	0.8428	0.8441	0.8498	0.8516	0.8544	0.8597	***0.8642***	0.8580	0.8533	***0.8619***	0.8505	0.8426
	Sen.	***0.8942***	***0.8929***	0.8758	0.8748	0.8677	0.8501	0.6978	0.6471	0.6107	0.69611	0.6074	0.5598
	Spec.	0.8208	0.8232	0.8386	0.8417	0.8487	0.8639	0.9355	0.9484	0.9573	0.9330	0.9547	***0.9638***
	Balanced Acc.	***0.8575***	***0.8580***	***0.8572***	***0.8582***	***0.8582***	***0.8570***	0.8167	0.7978	0.7840	0.8146	0.7811	0.7618
80:20	Acc.	0.8366	0.8373	0.8374	0.8475	0.8515	0.8574	0.8712	***0.8863***	***0.8873***	0.8642	***0.8862***	***0.8877***
	Sen.	***0.8930***	***0.8977***	***0.8965***	0.878	0.8698	0.8615	0.8255	0.7657	0.7640	0.8458	0.7243	0.7143
	Spec.	0.8225	0.8223	0.8226	0.8399	0.8470	0.8564	0.8826	0.9165	0.9181	0.8688	0.9267	***0.9310***
	Balanced Acc.	0.8578	***0.8600***	0.8596	0.8590	0.8584	0.8590	0.8540	0.8411	0.8411	0.8573	0.8255	0.8227
90:10	Acc.	0.8303	0.8331	0.8420	0.8419	0.8558	0.8742	0.7951	0.8686	0.8775	0.7263	0.8806	***0.8926***
	Sen.	0.9047	0.9103	0.9037	0.8937	0.8763	0.8440	0.9423	0.8703	0.8550	***0.9682***	0.8403	0.8100
	Spec.	0.8220	0.8245	0.8351	0.8361	0.8535	0.8776	0.7787	0.8684	0.8800	0.7002	0.8850	***0.9017***
	Balanced Acc.	***0.8634***	***0.8674***	***0.8694***	***0.8649***	***0.8649***	***0.8608***	***0.8605***	***0.8694***	***0.8675***	0.8338	***0.8627***	0.8559

หมายเหตุ : ตัวหนา และเอียง แทน วิธีประมาณที่ดีที่สุดในแต่ละสถานการณ์ (ค่าที่มีความแตกต่างกันในทศนิยมที่ 3 ถือว่าไม่แตกต่างกัน)

จากตาราง 2 เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 60:40 จาก**ค่าความแม่นยำ**และ**ค่าความแม่นที่สมดุล** พบว่าวิธีภาวะน่าจะเป็นสูงสุดและวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS และ ROS **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงที่สุด

เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 70:30 จาก**ค่าความแม่นยำ** วิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มลดมีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงที่สุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีภาวะน่าจะเป็นสูงสุดและวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 80:20 จาก**ค่าความแม่นยำ** วิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อน ROS และ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงที่สุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีภาวะน่าจะเป็นสูงสุดและวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 90:10 จาก**ค่าความแม่นยำ** วิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ RUS ด้วยวิธีการสุ่มลดมีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีการประมาณค่าพารามิเตอร์ทั้ง 4 วิธี ร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุดเป็นส่วนใหญ่

ตาราง 3 ผลการวัดประสิทธิภาพของตัวแบบถดถอยลอจิสติก เมื่อกำหนดตัวแปรอิสระเท่ากับ 1 ตัว ขนาดตัวอย่างเท่ากับ 100 และกำหนดอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20

IR	เกณฑ์การวัดประสิทธิภาพ	วิธีการประมาณค่าพารามิเตอร์											
		MLE			SCORE			Bayesians					
								Non-informative			Informative		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	Acc.	***0.8532***	***0.8537***	***0.8587***	***0.8565***	***0.8567***	***0.8575***	0.8031	0.7962	0.7673	0.7966	0.7827	0.743
	Sen.	***0.8964***	***0.8968***	0.8444	0.8765	0.8723	0.8221	0.5615	0.5391	0.4500	0.5423	0.5001	0.3801
	Spec.	0.8244	0.8250	0.8682	0.8432	0.8463	0.8811	0.9642	0.9675	0.9788	0.9662	0.971	***0.9833***
	Balanced Acc.	***0.8604***	***0.8609***	0.8563	0.8598	0.8593	0.8516	0.7628	0.7533	0.7144	0.7542	0.7356	0.6000
70:30	Acc.	0.8487	0.8499	0.8559	0.8570	0.8592	***0.8648***	***0.8684***	***0.8615***	0.8554	***0.8655***	0.8538	0.8457
	Sen.	***0.9007***	0.8980	0.8818	0.8818	0.8737	0.8550	0.7018	0.6522	0.6133	0.6905	0.6100	0.5626
	Spec.	0.8264	0.8293	0.8448	0.8464	0.8530	0.8690	0.9397	0.9511	0.9591	0.9406	0.9582	***0.9670***
	Balanced Acc.	***0.8635***	***0.8636***	***0.8633***	***0.8641***	***0.8633***	***0.8620***	0.8208	0.8017	0.7862	0.8155	0.7841	0.7648
80:20	Acc.	0.8434	0.8433	0.8433	0.8538	0.8593	0.8633	0.8826	***0.8935***	***0.895***	0.8783	***0.8934***	***0.8947***
	Sen.	***0.8970***	***0.8973***	***0.8985***	0.8833	0.8798	0.8700	0.8273	0.7680	0.7685	0.8365	0.7253	0.7173
	Spec.	0.8299	0.8299	0.8294	0.8464	0.8543	0.8616	0.8964	0.9248	0.9266	0.8887	***0.9354***	***0.9391***
	Balanced Acc.	***0.8635***	***0.8636***	***0.8640***	***0.8648***	***0.8670***	***0.8658***	***0.8618***	0.8464	0.8476	***0.8626***	0.8393	0.8282
90:10	Acc.	0.8292	0.8289	0.8302	0.8402	0.8546	0.8673	0.8043	0.8681	0.8699	0.7533	***0.8800***	***0.8882***
	Sen.	0.9045	0.9020	0.9040	0.8940	0.8715	0.8534	0.9430	0.8545	0.8545	***0.9674***	0.8225	0.8105
	Spec.	0.8208	0.8208	0.8220	0.8343	0.8527	0.8688	0.7889	0.8696	0.8716	0.7299	0.8863	***0.8968***
	Balanced Acc.	***0.8626***	***0.8614***	***0.8630***	***0.8641***	***0.8621***	***0.8611***	***0.8659***	***0.8620***	***0.8631***	0.8485	0.8544	0.8536

หมายเหตุ : ตัวหนา และเอียง แทน วิธีประมาณที่ดีที่สุดในแต่ละสถานการณ์ (ค่าที่มีความแตกต่างกันในทศนิยมที่ 3 ถือว่าไม่แตกต่างกัน)

จากตาราง 3 เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 60:40 จาก**ค่าความแม่นยำ** พบว่าวิธีภาวะน่าจะเป็นสูงสุดและวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด **ค่าความไว** และ**ค่าความแม่นที่สมดุล** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด และ**ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด

เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 70:30 จาก**ค่าความแม่นยำ** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ RUS มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีภาวะน่าจะเป็นสูงสุดและวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 80:20 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพราะ** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีภาวะน่าจะเป็นสูงสุด วิธีฟังก์ชันสกอร์ที่ปรับปรุง และวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 90:10 จาก**ค่าความแม่นยำ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ RUS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีภาวะน่าจะเป็นสูงสุด วิธีฟังก์ชันสกอร์ที่ปรับปรุง และวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

ตาราง 4 ผลการวัดประสิทธิภาพของตัวแบบถดถอยลอจิสติก เมื่อกำหนดตัวแปรอิสระเท่ากับ 3 ตัว ขนาดตัวอย่างเท่ากับ 100 และกำหนดอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30

IR	เกณฑ์การวัดประสิทธิภาพ	วิธีการประมาณค่าพารามิเตอร์											
		MLE			SCORE			Bayesians					
								Non-informative			Informative		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	Acc.	0.8817	0.8819	***0.8924***	***0.8928***	***0.8925***	***0.8942***	0.8692	0.8705	***0.8940***	0.8610	0.8665	***0.8928***
	Sen.	0.9428	0.9388	0.9091	0.9111	0.8989	0.8752	***0.9658***	***0.9650***	0.8728	***0.9694***	***0.9627***	0.9013
	Spec.	0.8409	0.8439	0.8813	0.8806	0.8883	***0.9069***	0.8047	0.8075	***0.9081***	0.7888	0.8024	0.8872
	Balanced Acc.	***0.8918***	***0.8913***	***0.8952***	***0.8958***	***0.8936***	***0.8910***	0.8853	0.8863	0.8904	0.8791	0.8825	***0.8943***
70:30	Acc.	0.8681	0.8745	0.8794	0.8873	***0.8938***	***0.8994***	0.8084	0.8130	0.8502	0.7899	0.8126	0.8436
	Sen.	0.9398	0.9417	0.9331	0.9157	0.9022	0.8893	***0.9863***	***0.9853***	0.9690	***0.9881***	***0.9836***	0.9729
	Spec.	0.8374	0.8458	0.8563	0.8753	0.8902	***0.9037***	0.7321	0.7392	0.7992	0.7050	0.7393	0.7882
	Balanced Acc.	0.8886	***0.8937***	***0.8947***	***0.8955***	***0.8962***	***0.8965***	0.8592	0.8623	0.8841	0.8466	0.8614	0.8806
80:20	Acc.	0.8549	0.8650	0.8658	0.8740	0.8946	***0.9001***	0.7303	0.7407	0.7468	0.6885	0.7436	0.7546
	Sen.	0.9437	0.9482	0.9500	0.9272	0.9110	0.8993	***0.9957***	***0.9955***	***0.9948***	***0.9948***	***0.9942***	***0.9927***
	Spec.	0.8328	0.8442	0.8448	0.8608	0.8905	***0.9003***	0.6640	0.6771	0.6848	0.6119	0.6809	0.6950
	Balanced Acc.	0.8882	0.8962	0.8974	0.894	***0.9007***	0.8998	0.8298	0.8363	0.8398	0.8034	0.8375	0.8439
90:10	Acc.	0.8630	0.8870	0.8870	0.8832	0.8928	***0.9022***	0.7728	0.7799	0.8673	0.7529	0.7828	0.8566
	Sen.	0.9431	0.9490	0.9176	0.9198	0.9127	0.8743	***0.9902***	***0.9903***	0.9499	***0.9908***	0.9877	0.9603
	Spec.	0.8363	0.8433	0.8769	0.8710	0.8861	***0.9115***	0.7003	0.7097	0.8398	0.6736	0.7145	0.8220
	Balanced Acc.	0.8897	***0.8961***	***0.8972***	***0.8954***	***0.8994***	***0.8929***	0.8453	0.8500	***0.8948***	0.8322	0.8511	***0.8912***

หมายเหตุ : ตัวหนา และเอียง แทน วิธีประมาณที่ดีที่สุดในแต่ละสถานการณ์ (ค่าที่มีความแตกต่างกันในทศนิยมที่ 3 ถือว่าไม่แตกต่างกัน)

จากตาราง 4 เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 60:40 จาก**ค่าความแม่นยำ** พบว่าวิธีการประมาณค่าพารามิเตอร์ทั้ง 4 วิธี ร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมและวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีการประมาณค่าพารามิเตอร์ทั้ง 4 วิธี ร่วมกับ SMOTE มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 70:30 จาก**ค่าความแม่นยำ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด สำหรับ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีภาวะน่าจะเป็นสูงสุด วิธีฟังก์ชันสกอร์ที่ปรับปรุง และวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 80:20 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ ROS มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 90:10 จาก**ค่าความแม่นยำ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** และ**ค่าความจำเพาะ สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีการประมาณค่าพารามิเตอร์ทั้ง 4 วิธี ร่วมกับ SMOTE มีประสิทธิภาพสูงสุดเป็นส่วนใหญ่

ตาราง 5 ผลการวัดประสิทธิภาพของตัวแบบถดถอยลอจิสติก เมื่อกำหนดตัวแปรอิสระเท่ากับ 3 ตัว ขนาดตัวอย่างเท่ากับ 100 และกำหนดอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20

IR	เกณฑ์การวัดประสิทธิภาพ	วิธีการประมาณค่าพารามิเตอร์											
		MLE			SCORE			Bayesians					
								Non-informative			Informative		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	Acc.	0.8829	0.8840	***0.8931***	***0.8940***	***0.8955***	***0.8956***	0.8708	0.8725	***0.8983***	0.8638	0.8717	***0.8983***
	Sen.	0.9440	0.9432	0.9109	0.9140	0.9080	0.8799	***0.9678***	***0.9664***	0.8770	***0.9683***	***0.9624***	0.9040
	Spec.	0.8421	0.8445	0.8812	0.8807	0.8872	0.9060	0.8062	0.8098	***0.9125***	0.7941	0.8112	0.8944
	Balanced Acc.	***0.8930***	***0.8939***	***0.8960***	***0.8973***	***0.8976***	***0.8929***	0.8870	0.8881	***0.8948***	0.8812	0.8868	***0.8992***
70:30	Acc.	0.8737	0.8774	0.8837	0.8920	0.8972	***0.9034***	0.8133	0.8176	0.8561	0.7991	0.8212	0.8499
	Sen.	0.9455	0.9457	0.9363	0.9212	0.9113	0.8982	***0.9845***	***0.9843***	0.9658	***0.9858***	0.9793	0.9693
	Spec.	0.8429	0.8481	0.8611	0.8794	0.8911	***0.9056***	0.7399	0.7461	0.8090	0.7191	0.7534	0.7986
	Balanced Acc.	0.8942	0.8969	0.8987	***0.9003***	***0.9012***	***0.9019***	0.8622	0.8652	0.8874	0.8525	0.8663	0.8840
80:20	Acc.	0.8557	0.8664	0.8670	0.879	0.8972	***0.9031***	0.7294	0.7387	0.7454	0.6979	0.7465	0.7568
	Sen.	0.9485	0.9483	0.9473	0.9305	0.9173	0.9033	***0.9945***	***0.9945***	***0.9943***	***0.9960***	***0.9935***	***0.9918***
	Spec.	0.8324	0.8459	0.8469	0.8661	0.8921	***0.9031***	0.6631	0.6748	0.6831	0.6234	0.6847	0.6980
	Balanced Acc.	0.8905	0.8971	0.8971	0.8983	***0.9047***	***0.9032***	0.8288	0.8346	0.8387	0.8097	0.8391	0.8449
90:10	Acc.	0.8280	0.8548	0.8563	0.8470	***0.9001***	***0.9090***	0.6180	0.6393	0.6500	0.5323	0.6535	0.6705
	Sen.	0.9399	0.9530	0.9510	0.9345	0.9185	0.8924	***0.9945***	***0.9955***	***0.9945***	***0.9940***	***0.9950***	***0.9930***
	Spec.	0.8156	0.8438	0.8457	0.8372	0.8980	***0.9108***	0.5761	0.5997	0.6117	0.4809	0.6156	0.6346
	Balanced Acc.	0.8777	0.8984	0.8984	0.8859	***0.9083***	***0.9016***	0.7853	0.7976	0.8031	0.7375	0.8053	0.8138

หมายเหตุ : ตัวหนา และเอียง แทน วิธีประมาณที่ดีที่สุดในแต่ละสถานการณ์ (ค่าที่มีความแตกต่างกันในทศนิยมที่ 3 ถือว่าไม่แตกต่างกัน)

จากตาราง 5 เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 60:40 จาก**ค่าความแม่นยำ** และ**ค่าความแม่นที่สมดุล** พบว่าวิธีการประมาณค่าพารามิเตอร์ทั้ง 4 วิธี ร่วมกับ SMOTE มีประสิทธิภาพสูงสุดเป็นส่วนใหญ่ **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด และ**ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 70:30 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ RUS มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** วิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3วิธี

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 80:20 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 90:10 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลทั้ง 3 วิธี มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด

ตาราง 6 ผลการวัดประสิทธิภาพของตัวแบบถดถอยลอจิสติก เมื่อกำหนดตัวแปรอิสระเท่ากับ 1 ตัว ขนาดตัวอย่างเท่ากับ 500 และกำหนดอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30

IR	เกณฑ์การวัดประสิทธิภาพ	วิธีการประมาณค่าพารามิเตอร์											
		MLE			SCORE			Bayesians					
								Non-informative			Informative		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	Acc.	0.8555	0.8556	***0.8614***	0.8599	0.8597	***0.8606***	0.7830	0.7798	0.7591	0.7678	0.7641	0.72745
	Sen.	***0.8963***	***0.8972***	0.8461	0.8751	0.8747	0.8232	0.4980	0.4881	0.4255	0.4521	0.4412	0.3364
	Spec.	0.8282	0.8279	0.8716	0.8497	0.8497	0.8855	0.9731	0.9744	***0.9815***	0.9783	0.9793	***0.9883***
	Balanced Acc.	***0.8623***	***0.8625***	0.8588	***0.8624***	***0.8622***	0.8543	0.7355	0.7312	0.7035	0.7152	0.7103	0.6623
70:30	Acc.	0.8470	0.8472	0.8537	0.8553	0.8559	***0.8607***	0.8545	0.8505	0.8444	0.8465	0.8412	0.8331
	Sen.	***0.8964***	***0.8957***	0.8801	0.8745	0.8732	0.8575	0.6232	0.6000	0.5663	0.5756	0.5510	0.5110
	Spec.	0.8258	0.8264	0.8423	0.8471	0.8484	0.8621	0.9536	0.9578	0.9635	0.9613	0.9655	***0.9712***
	Balanced Acc.	***0.8611***	***0.8611***	***0.8612***	***0.8608***	***0.8608***	0.8598	0.7884	0.7789	0.7649	0.7700	0.7583	0.7411
80:20	Acc.	0.8408	0.8412	0.8414	0.8528	0.8534	0.8544	***0.8849***	***0.8868***	***0.8869***	***0.8863***	***0.8870***	***0.8870***
	Sen.	***0.8957***	***0.8962***	***0.8964***	0.8720	0.8706	0.8701	0.7448	0.7104	0.7094	0.7098	0.6676	0.6639
	Spec.	0.8271	0.8275	0.8276	0.8479	0.8491	0.8505	0.9199	0.9309	0.9312	0.9305	***0.9419***	***0.9428***
	Balanced Acc.	***0.8614***	***0.8618***	***0.8620***	***0.8600***	0.8598	***0.8603***	0.8324	0.8207	0.8203	0.8201	0.8047	0.8033
90:10	Acc.	0.8323	0.8332	0.8332	0.8489	0.8514	0.8550	0.8587	0.8797	0.8807	0.8639	***0.8904***	***0.8928***
	Sen.	***0.8966***	***0.8985***	***0.8985***	0.8764	0.8755	0.8712	0.8648	0.8209	0.8204	0.8499	0.7897	0.7842
	Spec.	0.8252	0.8259	0.8259	0.8458	0.8488	0.8532	0.8580	0.8862	0.8874	0.8655	***0.9016***	***0.9049***
	Balanced Acc.	***0.8609***	***0.8622***	***0.8622***	***0.8611***	***0.8621***	***0.8622***	***0.8614***	0.8535	0.8539	0.8577	0.8456	0.8445

หมายเหตุ : ตัวหนา และเอียง แทน วิธีประมาณที่ดีที่สุดในแต่ละสถานการณ์ (ค่าที่มีความแตกต่างกันในทศนิยมที่ 3 ถือว่าไม่แตกต่างกัน)

จากตาราง 6 เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 60:40 จาก**ค่าความแม่นยำ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 70:30 จาก**ค่าความแม่นยำ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 80:20 จาก**ค่าความแม่นยำ** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด และ**ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 90:10 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

ตาราง 7 ผลการวัดประสิทธิภาพของตัวแบบถดถอยลอจิสติก เมื่อกำหนดตัวแปรอิสระเท่ากับ 1 ตัว ขนาดตัวอย่างเท่ากับ 500 และกำหนดอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20

IR	เกณฑ์การวัดประสิทธิภาพ	วิธีการประมาณค่าพารามิเตอร์											
		MLE			SCORE			Bayesians					
								Non-informative			Informative		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	Acc.	0.8569	0.8568	***0.8622***	***0.8610***	***0.8606***	***0.8619***	0.7828	0.7797	0.7592	0.7673	0.7633	0.7255
	Sen.	***0.8984***	***0.8982***	0.8484	0.8774	0.8763	0.8268	0.4954	0.4854	0.4240	0.4485	0.4369	0.3293
	Spec.	0.8292	0.8292	0.8714	0.8502	0.8502	0.8853	0.9745	0.9758	***0.9826***	0.9799	***0.9810***	***0.9896***
	Balanced Acc.	***0.8638***	***0.8637***	0.8599	***0.8638***	***0.8632***	0.8560	0.7349	0.7306	0.7033	0.7142	0.7089	0.6595
70:30	Acc.	0.8491	0.8495	0.8553	0.8572	0.8579	***0.8623***	0.8548	0.8507	0.8448	0.8462	0.8418	0.8339
	Sen.	***0.8961***	***0.8968***	0.8804	0.8750	0.8746	0.8574	0.6211	0.5978	0.5652	0.5751	0.5511	0.5107
	Spec.	0.8289	0.8293	0.8446	0.8495	0.8508	0.8645	0.9549	0.9591	0.9646	0.9624	0.9664	***0.9725***
	Balanced Acc.	***0.8625***	***0.8630***	***0.8625***	***0.8623***	***0.8627***	***0.8609***	0.7880	0.7785	0.7649	0.7687	0.7587	0.7416
80:20	Acc.	0.8406	0.8408	0.8410	0.8537	0.8547	0.8552	***0.8848***	***0.8873***	***0.8873***	***0.8866***	***0.8876***	***0.8877***
	Sen.	***0.8969***	***0.8971***	***0.8977***	0.8766	0.8748	0.8744	0.7404	0.7092	0.7087	0.7069	0.6684	0.6667
	Spec.	0.8265	0.8268	0.8269	0.8480	0.8497	0.8504	0.9209	0.9318	0.9320	0.9315	***0.9425***	***0.9430***
	Balanced Acc.	***0.8617***	***0.8619***	***0.8623***	***0.8623***	***0.8622***	***0.8624***	0.8306	0.8205	0.8204	0.8192	0.8054	0.8048
90:10	Acc.	0.8336	0.8340	0.8334	0.8494	0.8517	0.8540	0.8611	0.8811	0.8818	0.8677	***0.8923***	***0.8938***
	Sen.	0.8985	***0.9005***	***0.9001***	0.8796	0.8776	0.8742	0.8629	0.8259	0.8252	0.8484	0.7938	0.7884
	Spec.	0.8263	0.8267	0.8260	0.8460	0.8488	0.8518	0.8609	0.8873	0.8880	0.8699	***0.9032***	***0.9055***
	Balanced Acc.	***0.8624***	***0.8636***	***0.8631***	***0.8628***	***0.8632***	***0.8630***	***0.8619***	0.8566	0.8566	0.8591	0.8485	0.8469

หมายเหตุ : ตัวหนา และเอียง แทน วิธีประมาณที่ดีที่สุดในแต่ละสถานการณ์ (ค่าที่มีความแตกต่างกันในทศนิยมที่ 3 ถือว่าไม่แตกต่างกัน)

จากตาราง 7 เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 60:40 จาก**ค่าความแม่นยำ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** วิธีภาวะน่าจะเป็นสูงสุดและวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 70:30 จาก**ค่าความแม่นยำ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 80:20 จาก**ค่าความแม่นยำ** วิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุด ร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 90:10 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุด ร่วมกับ ROS และ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

ตาราง 8 ผลการวัดประสิทธิภาพของตัวแบบถดถอยลอจิสติก เมื่อกำหนดตัวแปรอิสระเท่ากับ 3 ตัว ขนาดตัวอย่างเท่ากับ 500 และกำหนดอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30

IR	เกณฑ์การวัดประสิทธิภาพ	วิธีการประมาณค่าพารามิเตอร์											
		MLE			SCORE			Bayesians					
								Non-informative			Informative		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	Acc.	0.8816	0.8823	***0.8905***	***0.8905***	***0.8902***	***0.8908***	0.8690	0.8695	***0.8900***	0.8740	0.8760	***0.8915***
	Sen.	0.9466	0.9462	0.9112	0.9112	0.9074	0.8732	***0.9671***	***0.9667***	***0.8638***	***0.9611***	0.9585	0.8824
	Spec.	0.8383	0.8396	0.8767	0.8768	0.8788	***0.9025***	0.8036	0.8047	***0.9074***	0.8160	0.8211	0.8975
	Balanced Acc.	***0.8925***	***0.8929***	***0.8940***	***0.8940***	***0.8931***	0.8878	0.8853	0.8857	0.8856	0.8885	0.8898	0.8899
70:30	Acc.	0.8583	0.8600	0.8598	0.8824	0.8858	***0.8906***	0.7322	0.7351	0.7403	0.7355	0.7514	0.7615
	Sen.	0.9450	0.9470	0.9478	0.9124	0.9101	0.9002	***0.9945***	***0.9940***	***0.9936***	***0.9936***	***0.9920***	***0.9913***
	Spec.	0.8366	0.8382	0.8378	0.8749	0.8797	***0.8882***	0.6666	0.6704	0.6769	0.6710	0.6913	0.7040
	Balanced Acc.	***0.8908***	***0.8926***	***0.8928***	***0.8936***	***0.8949***	***0.8942***	0.8305	0.8322	0.8353	0.8323	0.8416	0.8477
80:20	Acc.	0.8408	0.8412	0.8414	0.8528	0.8534	0.8544	***0.8849***	***0.8868***	***0.8869***	***0.8863***	***0.8870***	***0.8870***
	Sen.	***0.8957***	***0.8962***	***0.8964***	0.8720	0.8706	0.8701	0.7448	0.7104	0.7094	0.7098	0.6676	0.6639
	Spec.	0.8271	0.8275	0.8276	0.8479	0.8491	0.8505	0.9199	0.9309	0.9312	0.9305	***0.9419***	***0.9428***
	Balanced Acc.	***0.8614***	***0.8618***	***0.8620***	***0.8600***	0.8598	***0.8603***	0.8324	0.8207	0.8203	0.8201	0.8047	0.8033
90:10	Acc.	0.8655	0.8667	0.8666	0.8863	0.8883	***0.8911***	0.7769	0.7786	0.7819	0.7848	0.7942	0.8011
	Sen.	0.9466	0.9476	0.9478	0.9125	0.9109	0.9044	***0.9903***	0.9898	0.9893	0.9889	0.9872	0.9857
	Spec.	0.8385	0.8397	0.8395	0.8775	***0.8808***	***0.8866***	0.7058	0.7081	0.7128	0.7168	0.7298	0.7396
	Balanced Acc.	***0.8925***	***0.8937***	***0.8937***	***0.8950***	***0.8959***	***0.8955***	0.8481	0.8490	0.8510	0.8528	0.8585	0.8627

หมายเหตุ : ตัวหนา และเอียง แทน วิธีประมาณที่ดีที่สุดในแต่ละสถานการณ์ (ค่าที่มีความแตกต่างกันในทศนิยมที่ 3 ถือว่าไม่แตกต่างกัน)

จากตาราง 8 เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 60:40 จาก**ค่าความแม่นยำ** พบว่าวิธีการประมาณค่าพารามิเตอร์ที่ง 4 วิธี ร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ RUS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมและวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 70:30 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธีมีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 80:20 จาก**ค่าความแม่นยำ** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธีมีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด และ**ค่าความจำเพาะ** พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 90:10 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธีมีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

ตาราง 9 ผลการวัดประสิทธิภาพของตัวแบบถดถอยลอจิสติก เมื่อกำหนดตัวแปรอิสระเท่ากับ 3 ตัว ขนาดตัวอย่างเท่ากับ 500 และกำหนดอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20

IR	เกณฑ์การวัดประสิทธิภาพ	วิธีการประมาณค่าพารามิเตอร์											
		MLE			SCORE			Bayesians					
								Non-informative			Informative		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	Acc.	0.8819	0.8824	***0.8916***	***0.8920***	***0.8919***	***0.8924***	0.8689	0.8690	***0.8915***	0.8747	0.8756	***0.8924***
	Sen.	0.9470	0.9465	0.9118	0.9127	0.9095	0.8739	***0.9655***	***0.9650***	0.8650	0.9594	0.9575	0.8816
	Spec.	0.8385	0.8398	0.8782	0.8782	0.8802	***0.9047***	0.8045	0.8049	***0.9092***	0.8183	0.8210	0.8997
	Balanced Acc.	***0.8927***	***0.8931***	***0.8950***	***0.8954***	***0.8948***	0.8893	0.8850	0.8850	0.8871	0.8889	0.8893	***0.8906***
70:30	Acc.	0.8720	0.8727	0.8794	0.8888	0.8898	***0.8940***	0.8163	0.8176	0.8509	0.8237	0.8299	0.8555
	Sen.	0.9470	0.9460	0.9375	0.9135	0.9114	0.8942	**0.9854**	***0.9852***	0.9696	***0.9823***	***0.9805***	0.9666
	Spec.	0.8398	0.8413	0.8545	0.8782	0.8806	***0.8939***	0.7439	0.7457	0.8000	0.7558	0.7653	0.8079
	Balanced Acc.	***0.8934***	***0.8936***	***0.8960***	***0.8958***	***0.8960***	***0.8941***	0.8647	0.8655	0.8848	0.8690	0.8729	0.8872
80:20	Acc.	0.8585	0.8609	0.8605	0.8833	0.8863	***0.8907***	0.7341	0.7366	0.7407	0.7380	0.7545	0.7637
	Sen.	0.9456	0.9470	0.9478	0.9123	0.9118	0.9021	***0.9947***	***0.9944***	***0.9940***	***0.9935***	***0.9920***	***0.9904***
	Spec.	0.8368	0.8393	0.8387	0.8761	***0.8800***	***0.8879***	0.6689	0.6722	0.6774	0.6741	0.6951	0.7070
	Balanced Acc.	***0.8912***	***0.8931***	***0.8932***	***0.8942***	***0.8959***	***0.8950***	0.8318	0.8333	0.8357	0.8338	0.8435	0.8487
90:10	Acc.	0.8457	0.8514	0.8503	0.8768	0.8852	***0.8937***	0.6305	0.6365	0.6451	0.6240	0.6598	0.6783
	Sen.	0.9492	0.9518	0.9528	0.9161	0.9118	0.8979	***0.9980***	***0.9979***	***0.9979***	***0.9980***	***0.9973***	***0.9966***
	Spec.	0.8342	0.8403	0.8389	0.8724	0.8822	***0.8933***	0.5897	0.5964	0.6059	0.5824	0.6223	0.6429
	Balanced Acc.	***0.8917***	***0.8960***	***0.8958***	***0.8943***	***0.8970***	***0.8956***	0.7938	0.7971	0.8019	0.7902	0.8098	0.8198

หมายเหตุ : ตัวหนา และเอียง แทน วิธีประมาณที่ดีที่สุดในแต่ละสถานการณ์ (ค่าที่มีความแตกต่างกันในทศนิยมที่ 3 ถือว่าไม่แตกต่างกัน)

จากตาราง 9 เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 60:40 จาก**ค่าความแม่นยำ** พบว่าวิธีการประมาณค่าพารามิเตอร์ที่ง 4 วิธี ร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด **ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมและวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 70:30 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับ RUS และ ROS มีประสิทธิภาพสูงสุด และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 80:20 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

อัตราส่วนความไม่สมดุลของข้อมูลเท่ากับ 90:10 จาก**ค่าความแม่นยำ** และ**ค่าความจำเพาะ** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด **ค่าความไว** พบว่าวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนและกรณีทราบความรู้ก่อนร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี และ**ค่าความแม่นที่สมดุล** พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยทั้ง 3 วิธี มีประสิทธิภาพสูงสุด

บทที่ 5
บทสรุป

งานวิจัยนี้มีจุดมุ่งหมายเพื่อศึกษาและเปรียบเทียบประสิทธิภาพการพยากรณ์ในตัวแบบถดถอยลอจิสติก เมื่อประมาณพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด วิธีเบส์เซียน และวิธีฟังก์ชันสกอร์ที่ปรับปรุง ร่วมกับการจัดการความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มลด วิธีการสุ่มเกิน และวิธีการสังเคราะห์ข้อมูลใหม่ เมื่อกำหนดจำนวนตัวแปรอิสระเท่ากับ 1 และ 3 ตัว ขนาดตัวอย่างเท่ากับ 100 และ 500 อัตราส่วนความไม่สมดุลของข้อมูลในกลุ่ม 0 และ 1 เป็น 60:40, 70:30, 80:20 และ 90:10 การวิเคราะห์ข้อมูลจะแบ่งตัวอย่างออกเป็น 2 ชุด คือ ชุดตัวอย่างที่ใช้สร้างเกณฑ์ในการจำแนก (Training data) และชุดตัวอย่างที่ใช้ตรวจสอบเกณฑ์การจำแนก (Validation data) ในอัตราส่วน Training : Validation เป็น 70:30 และ 80:20 ทำการจำลองข้อมูลในแต่ละสถานการณ์และทำซ้ำ 1,000 ครั้ง โดยใช้ค่าความแม่นยำ อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยหรือความไว อัตราความถูกต้องในการทำนายกลุ่มส่วนมากหรือความจำเพาะ และค่าความแม่นที่สมดุล เป็นเกณฑ์ในการเปรียบเทียบ

5.1 สรุปผลการวิจัย

ผลการเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบถดถอยลอจิสติก สามารถสรุปได้ดังนี้

ตาราง 10 สรุปวิธีการประมาณค่าพารามิเตอร์ที่ให้ค่าความแม่นยำสูงสุดในแต่ละสถานการณ์ที่ศึกษา

IR	ขนาดตัวอย่าง (n)	$p=1$						$p=3$					
		70 : 30			80 : 20			70:30			80:20		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	100	Score/ MLE	Score/ MLE	Score/ MLE	Score/ MLE	Score/ MLE	Score/ MLE	Score	Score	ทั้ง 4 วิธี	Score	Score	ทั้ง 4 วิธี
	500			Score/ MLE	Score	Score	Score/ MLE	Score	Score	ทั้ง 4 วิธี	Score	Score	ทั้ง 4 วิธี
70:30	100	Baye non/ Baye			Baye non/Baye	Baye non	Score		Score	Score			Score
	500			Score			Score			Score			Score
80:20	100		Baye non/ Baye	Baye non/ Baye		Baye non/ Baye	Baye non/ Baye			Score			Score
	500	Baye non/ Baye	Baye non/ Baye	Baye non/ Baye	Baye non/ Baye	Baye non/ Baye	Baye non/ Baye	Baye non/ Baye	Baye non/ Baye	Baye non/ Baye			Score
90:10	100			Baye		Baye	Baye			Score		Score	Score
	500		Baye	Baye		Baye	Baye			Score			Score

จากตาราง 10 กรณีตัวแปรอิสระ 1 ตัว เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเป็น 60:40 พบว่าวิธีภาวะน่าจะเป็นสูงสุดและวิธีภาวะน่าจะเป็นสูงสุดเมื่อจัดการข้อมูลไม่สมดุลทั้ง 3 วิธี มีค่าความแม่นยำมากที่สุดใกล้เคียงกัน แต่เมื่ออัตราส่วนความไม่สมดุลเพิ่มขึ้น พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อนร่วมกับ SMOTE มีค่าความแม่นยำมากที่สุดเป็นส่วนใหญ่ และกรณีตัวแปรอิสระ 3 ตัว ในทุกระดับของอัตรส่วนความไม่สมดุลของข้อมูล พบว่าวิธีภาวะน่าจะเป็นสงสดร่วมกับ SMOTE มีค่าความแม่นยำมากที่สดเป็นส่วนใหญ่

ตาราง 11 สรุปวิธีการประมาณค่าพารามิเตอร์ที่ให้ค่าความไวสูงสุดในแต่ละสถานการณ์ที่ศึกษา

IR	ขนาดตัวอย่าง (n)	$p=1$						$p=3$					
		70 : 30			80 : 20			70:30			80:20		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	100	MLE	MLE		MLE	MLE		Baye non / Baye	Baye non / Baye		Baye non / Baye	Baye non / Baye	
	500	MLE	MLE		MLE	MLE		Baye non / Baye	Baye non		Baye non	Baye non	
70:30	100	MLE	MLE		MLE			Baye non / Baye	Baye non / Baye		Baye non / Baye	Baye non	
	500	MLE	MLE		MLE	MLE		Baye non / Baye	Baye non / Baye	Baye non / Baye	Baye non / Baye	Baye non / Baye	
80:20	100	MLE	MLE	MLE	MLE	MLE	MLE	Baye non / Baye	Baye non / Baye	Baye non / Baye	Baye non / Baye	Baye non / Baye	Baye non / Baye
	500	MLE	MLE	MLE	MLE	MLE	MLE	MLE	MLE	MLE	Baye non / Baye	Baye non / Baye	Baye non / Baye
90:10	100	Baye			Baye			Baye non / Baye	Baye non		Baye non / Baye	Baye non / Baye	Baye non / Baye
	500	MLE	MLE	MLE		MLE	MLE	Baye non			Baye non / Baye	Baye non / Baye	Baye non / Baye

จากตาราง 11 กรณีตัวแปรอิสระ 1 ตัว ในทุกระดับของอัตรส่วนความไม่สมดุลของข้อมูล พบว่าวิธีภาวะน่าจะเป็นสูงสุดเมื่อจัดการข้อมูลไม่สมดุลทั้ง 3 แบบ มีค่าความไวมากที่สุดเป็นส่วนใหญ่ และกรณีตัวแปรอิสระ 3 ตัว ในทุกระดับของอัตรส่วนความไม่สมดุลของข้อมูล พบว่าวิธีเบส์เซียนกรณีไม่ทราบความรู้ก่อน และทราบความรู้ก่อน ร่วมกับ RUS และ ROS มีค่าความไวมากที่สุดเป็นส่วนใหญ่

ตาราง 12 สรุปวิธีการประมาณค่าพารามิเตอร์ที่ให้ค่าความจำเพาะสูงสุดในแต่ละสถานการณ์ที่ศึกษา

IR	ขนาดตัวอย่าง (n)	$p=1$						$p=3$					
		70 : 30			80 : 20			70:30			80:20		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	100			Baye			Baye			Score /Baye non			Baye non
	500			Baye		Baye	Baye non /Baye			Score /Baye non			Score /Baye non
70:30	100			Baye			Baye			Score			Score
	500			Baye			Baye			Score			Score
80:20	100			Baye		Baye	Baye			Score			Score
	500		Baye	Baye		Baye	Baye		Baye	Baye		Score	Score
90:10	100			Baye			Baye			Score			Score
	500		Baye	Baye		Baye	Baye		Score	Score			Score

จากตาราง 12 กรณีตัวแปรอิสระ 1 ตัว ในทุกระดับของอัตรส่วนความไม่สมดุลของข้อมูล พบว่าวิธีเบส์เซียน กรณีทราบความรู้ก่อน เมื่อจัดการความไม่สมดุลแบบ SMOTE มีค่าความจำเพาะมากที่สุด

กรณีตัวแปรอิสระ 3 ตัว เมื่ออัตราส่วนความไม่สมดุลของข้อมูลเป็น 60:40 วิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนร่วมกับ SMOTE มีค่าความจำเพาะมากที่สุดเป็นส่วนใหญ่ แต่เมื่ออัตราส่วนความไม่สมดุลเพิ่มขึ้น พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีค่าความจำเพาะมากที่สุดเป็นส่วนใหญ่

ตาราง 13 สรุปวิธีการประมาณค่าพารามิเตอร์ที่ให้ค่าความแม่นที่สมดุลสูงสุดในแต่ละสถานการณ์ที่ศึกษา

IR	ขนาดตัวอย่าง (n)	$p=1$						$p=3$					
		70 : 30			80 : 20			70:30			80:20		
		RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE	RUS	ROS	SMOTE
60:40	100	MLE / Score	MLE / Score	MLE / Score	MLE	MLE		MLE / Score	MLE / Score	ทั้ง 4 วิธี	MLE / Score	MLE / Score	ทั้ง 4 วิธี
	500	MLE / Score	MLE / Score		MLE / Score	MLE / Score		MLE / Score	MLE / Score	MLE	MLE / Score	MLE / Score	MLE / Baye
70:30	100	MLE / Score	MLE / Score	MLE / Score	MLE / Score	MLE / Score	MLE / Score	Score	MLE / Score	MLE / Score	Score	Score	Score
	500	MLE / Score	MLE / Score	MLE	MLE / Score	MLE / Score	MLE / Score	MLE / Score	MLE / Score	MLE / Score	MLE / Score	MLE / Score	MLE / Score
80:20	100	MLE / Score	MLE / Score	MLE / Score	MLE / Score/ Baye non	MLE / Score	MLE / Score		Score			Score	Score
	500	MLE / Score	MLE	MLE	MLE / Score	MLE / Score	MLE / Score	MLE / Score	MLE	MLE / Score	MLE / Score	MLE / Score	MLE / Score
90:10	100		Baye non		MLE / Score/ Baye non	MLE / Score/ Baye non	MLE / Score/ Baye non	Score	MLE / Score	ทั้ง 4 วิธี		Score	Score
	500	MLE / Score/ Baye non	MLE / Score	MLE / Score	MLE / Score/ Baye non	MLE / Score	MLE / Score	MLE / Score	MLE	MLE / Score	MLE / Score	MLE	MLE / Score

จากตาราง 13 กรณีตัวแปรอิสระ 1 และ 3 ในทุกระดับของอัตรส่วนความไม่สมดุลของข้อมูล พบว่าวิธีภาวะน่าจะเป็นสูงสุดและวิธีฟังก์ชันสกอร์ที่ปรับปรุงเมื่อจัดการข้อมูลไม่สมดุลทั้ง 3 แบบ มีค่าความแม่นที่สมดุลมากที่สุดเป็นส่วน

5.2 อภิปรายผลการวิจัย

จากการศึกษาการประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยลอจิสติก 3 วิธี ได้แก่ วิธีภาวะน่าจะเป็นสูงสุด วิธีเบส์เซียน และวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับจัดการความไม่สมดุลของข้อมูล ด้วยวิธีการสุ่มลด วิธีการสุ่มเกิน และวิธีการสังเคราะห์ข้อมูลใหม่ เพื่อหาวิธีที่มีประสิทธิภาพในการพยากรณ์ที่ดีที่สุดสำหรับสถานการณ์ที่ศึกษา เมื่อพิจารณาค่าความแม่นยำ เมื่อกำหนดตัวแปรอิสระ 1 ตัว เมื่อข้อมูลอยู่ในระดับมีความไม่สมดุลเล็กน้อย พบว่าวิธีภาวะน่าจะเป็นสูงสุด และวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงสุด ในทุกขนาดของตัวอย่างและอัตราส่วนระหว่าง Training : Validation แต่เมื่อระดับอัตราส่วนความไม่สมดุลของข้อมูลเพิ่มขึ้น พบว่าส่วนใหญ่วิธีเบส์เซียน ในกรณีทราบความรู้ก่อน ร่วมกับ SMOTE มีประสิทธิภาพสูงสุด ในทุกขนาดของตัวอย่างและอัตราส่วนระหว่าง Training : Validation เมื่อกำหนดตัวแปรอิสระ 3 ตัว พบว่าในทุกระดับของอัตราส่วนความไม่สมดุลของข้อมูล พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงที่สุดเป็นส่วนใหญ่ ในทุกขนาดของตัวอย่างและอัตราส่วนระหว่าง Training : Validation

เมื่อพิจารณาในส่วนของค่าความไว หรืออัตราการจำแนกกลุ่มถูกในกลุ่มส่วนน้อย กรณีมีตัวแปรอิสระ 1 ตัว ในทุกระดับของอัตราส่วนความไม่สมดุลของข้อมูล พบวิธีภาวะน่าจะเป็นสูงสุดร่วมกับการจัดการความไม่สมดุลทั้ง 3 แบบมีประสิทธิภาพที่ดีที่สุดเป็นส่วนใหญ่ ในทุกขนาดของตัวอย่างและอัตราส่วนระหว่าง Training : Validation กรณีมีตัวแปรอิสระ 3 ตัว ในทุกระดับของอัตราส่วนความไม่สมดุลของข้อมูล พบว่าวิธีเบส์เซียน ในกรณีไม่ทราบความรู้ก่อน และกรณีทราบความรู้ก่อนร่วมกับ RUS และ ROS มีประสิทธิภาพสูงที่สุดเป็นส่วนใหญ่ ในทุกขนาดของตัวอย่างและอัตราส่วนระหว่าง Training : Validation

เมื่อพิจารณาในส่วนของค่าความจำเพาะ หรืออัตราการจำแนกกลุ่มถูกในกลุ่มส่วนใหญ่ กรณีมีตัวแปรอิสระ 1 ตัว ในทุกระดับของอัตราส่วนความไม่สมดุลของข้อมูล พบว่าวิธีเบส์เซียน ในกรณีทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงที่สุด ในทุกขนาดของตัวอย่างและอัตราส่วนระหว่าง Training : Validation กรณีมีตัวแปรอิสระ 3 ตัว ข้อมูลอยู่ในระดับมีความไม่สมดุลเล็กน้อย พบว่าวิธีฟังก์ชันสกอร์ที่ปรับปรุงและวิธีเบส์เซียน ในกรณีไม่ทราบความรู้ก่อนร่วมกับ SMOTE มีประสิทธิภาพสูงที่สุด เมื่อระดับอัตราส่วนความไม่สมดุลของข้อมูลเพิ่มขึ้น พบว่าส่วนใหญ่วิธีฟังก์ชัน

สกอร์ที่ปรับปรุงร่วมกับ SMOTE มีประสิทธิภาพสูงที่สุด ในทุกขนาดของตัวอย่างและอัตราส่วนระหว่าง Training : Validation

เมื่อดูภาพรวมของอัตราการจำแนกถูกในกลุ่มส่วนใหญ่และกลุ่มส่วนน้อย จะพิจารณาค่าความแม่นที่สมดุล กรณีมีตัวแปรอิสระ 1 และ 3 ตัว ในทุกระดับของอัตราส่วนความไม่สมดุลของข้อมูล พบว่าส่วนใหญ่วิธีภาวะน่าจะเป็นสูงสุด และวิธีฟังก์ชันสกอร์ที่ปรับปรุง ร่วมกับการจัดการความไม่สมดุลทั้ง 3 แบบมีประสิทธิภาพที่ดีที่สุดเป็นส่วนใหญ่ ในทุกขนาดของตัวอย่างและอัตราส่วนระหว่าง Training : Validation

จากผลการวิจัยพบว่าการสุ่มตัวอย่างซ้ำโดยเพิ่มข้อมูลในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงกับกลุ่มส่วนมากด้วยวิธีการสังเคราะห์ข้อมูลใหม่ (SMOTE) มีประสิทธิภาพที่ดีที่สุดเป็นส่วนใหญ่ ซึ่งสอดคล้องกับงานวิจัยของ วิชญ์วิสิฐ เกษรสิทธิ์ และคณะ (2561) และกิตติภพ แซ่เตีย และจิรภัทร์ หยกรัตนศักดิ์ (2564) ที่พบว่าวิธีการสังเคราะห์ข้อมูลใหม่ มีประสิทธิภาพในการแก้ปัญหาข้อมูลไม่สมดุล ในส่วนของวิธีการสุ่มเกิน (ROS) ซึ่งเป็นการสุ่มตัวอย่างเพิ่ม โดยการทำซ้ำเพื่อเพิ่มจำนวนตัวอย่างในกลุ่มส่วนน้อยจากข้อมูลเดิมซึ่งไม่ได้ให้สารสนเทศของข้อมูลเพิ่มขึ้นเนื่องจากไม่ใช้ข้อมูลใหม่ ซึ่งแตกต่างจากวิธีการสังเคราะห์ข้อมูลใหม่ (SMOTE) ที่เป็นการสร้างข้อมูลใหม่ จึงทำให้ได้ผลลัพธ์ที่ดีกว่า (นพมาศ อัครจันทโชติ และคณะ, 2562) และวิธีการสุ่มลด (RUS) เป็นการลดจำนวนข้อมูลในกลุ่มส่วนใหญ่ให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มส่วนน้อย แต่วิธีนี้จะทำให้สูญเสียข้อมูลในการวิเคราะห์ จนทำให้ตัวแบบที่ได้สูญเสียสารสนเทศในการอธิบายตัวแปรตาม ซึ่งจะส่งผลให้การพยากรณ์การจำแนกกลุ่มมีความผิดพลาดมากขึ้น

เมื่อพิจารณาการกำหนดขนาดตัวอย่างเท่ากับ 100 และ 500 และการกำหนดอัตราส่วนระหว่าง Training : Validation เท่ากับ 70:30 และ 80:20 พบว่าผลการวิเคราะห์ที่ได้มีค่าใกล้เคียงกัน อาจเนื่องมาจากขนาดตัวอย่างที่กำหนดมีจำนวนมากและใกล้เคียงกัน เช่นเดียวกับการกำหนด Training : Validation มีค่าใกล้เคียงกัน

5.3 ข้อเสนอแนะ

1. **ด้านการนำไปใช้ประโยชน์**

จากการศึกษาเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติก เมื่อข้อมูลไม่สมดุล ร่วมกับการจัดการความไม่สมดุลด้วยวิธีการสุ่มลด วิธีการสุ่มเกิน และวิธีการสังเคราะห์ข้อมูลใหม่ เพื่อหาวิธีพยากรณ์ของตัวแบบถดถอยลอจิสติกที่ดีที่สุดสำหรับสถานการณ์ที่ศึกษา เมื่อข้อมูลมีความไม่สมดุล ควรจัดการความไม่สมดุลด้วยวิธีการสังเคราะห์ข้อมูลใหม่ ก่อนนำเข้า

แบบจำลอง ดังนั้นการจัดการความไม่สมดุลของข้อมูลมีประโยชน์ทำให้การจำแนกกลุ่มข้อมูลเกิดความผิดพลาดน้อยลง กล่าวคือ ข้อมูลที่อยู่ในกลุ่มส่วนน้อยจะไม่ถูกจัดให้ไปอยู่ในกลุ่มส่วนใหญ่ทั้งหมด ส่วนวิธีการประมาณค่าพารามิเตอร์ เมื่อกำหนดจำนวนตัวแปรอิสระเท่ากับ 1 ตัว ควรเลือกใช้วิธีภาวะน่าจะเป็นสูงสุด ในกรณีที่ระดับความไม่สมดุลของข้อมูลไม่แตกต่างกันมาก แต่เมื่อระดับความไม่สมดุลเพิ่มขึ้น ควรเลือกใช้วิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อน และเมื่อกำหนดจำนวนตัวแปรอิสระเท่ากับ 3 ตัว พบว่าควรเลือกใช้วิธีฟังก์ชันสกอร์ที่ปรับปรุง ในทุกระดับของความไม่สมดุลของข้อมูล โดยพิจารณาภาพรวมของค่าความแม่นยำ

2. **ด้านการศึกษาวิจัย**

2.1 เป็นแนวทางในการจัดการความไม่สมดุลของข้อมูลด้วยวิธีการอื่นๆ เช่น วิธี ผสมผสาน (Hybrid Method) และวิธีสังเคราะห์ข้อมูลเพิ่ม ADASYN (Adaptive Synthetic Sampling Approach) ซึ่งเป็นวิธีที่ปรับปรุงการทำงานของวิธี SMOTE ให้ดีขึ้น

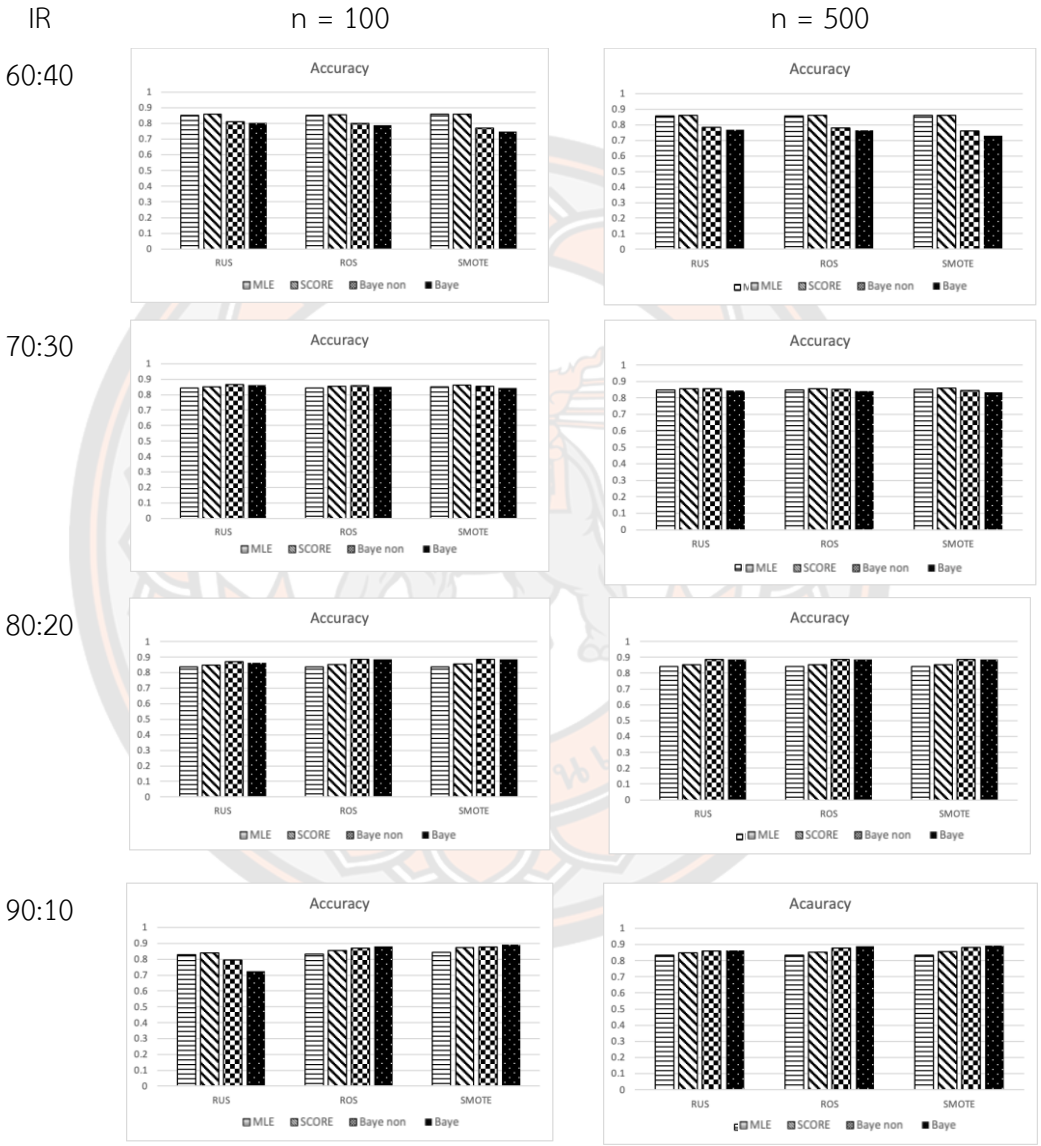
2.2 เป็นแนวทางในการศึกษาวิธีการประมาณค่าพารามิเตอร์ในตัวแบบถดถอยลอจิสติก เมื่อข้อมูลไม่สมดุลด้วยวิธีอื่น นอกเหนือจากที่ใช้ในการศึกษาวิจัย เช่น วิธีริดจ์ (Ridge Regression)

2.3 เป็นแนวทางในการศึกษาและเปรียบเทียบวิธีการจำแนกกลุ่มของข้อมูลมากกว่า 2 กลุ่มขึ้นไป

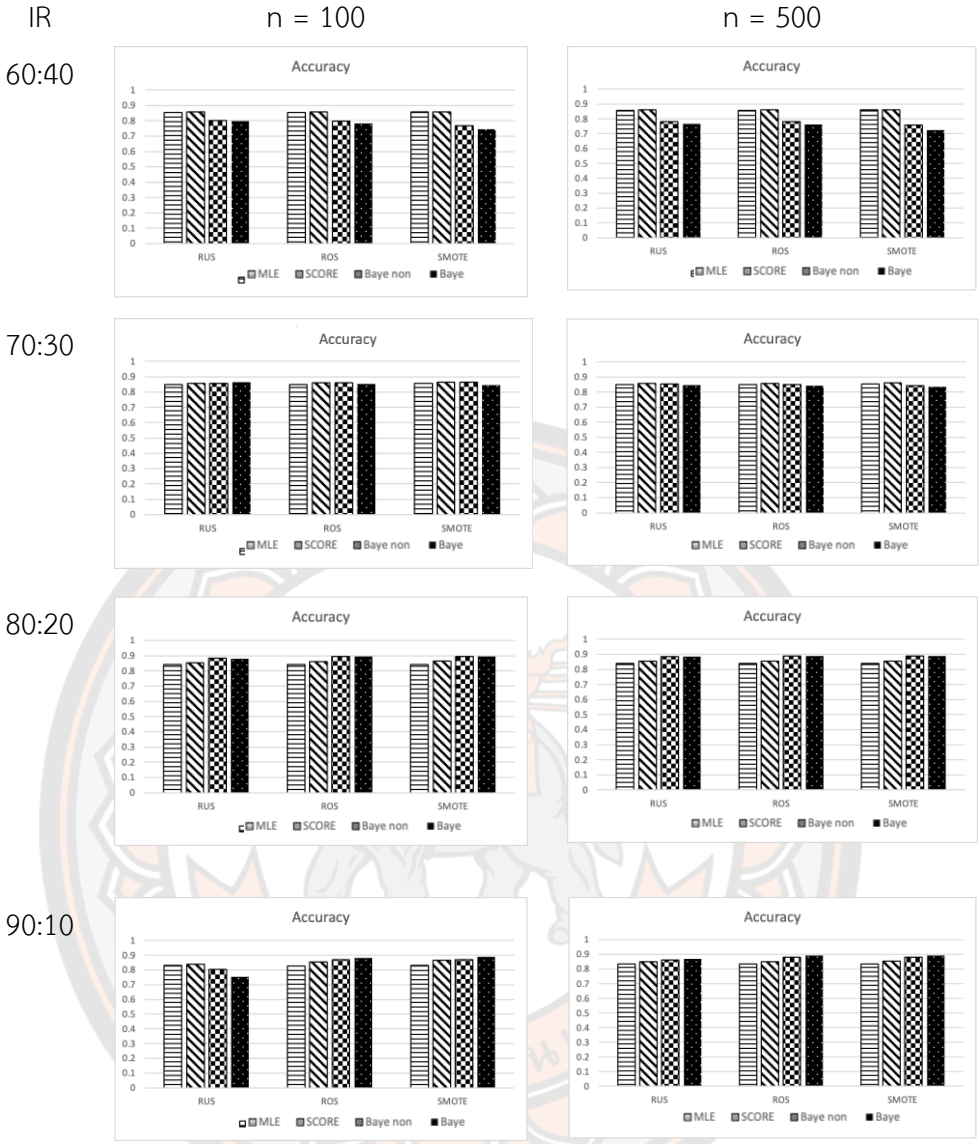
ภาคผนวก

แผนภูมิแท่งแสดงการวัดประสิทธิภาพ

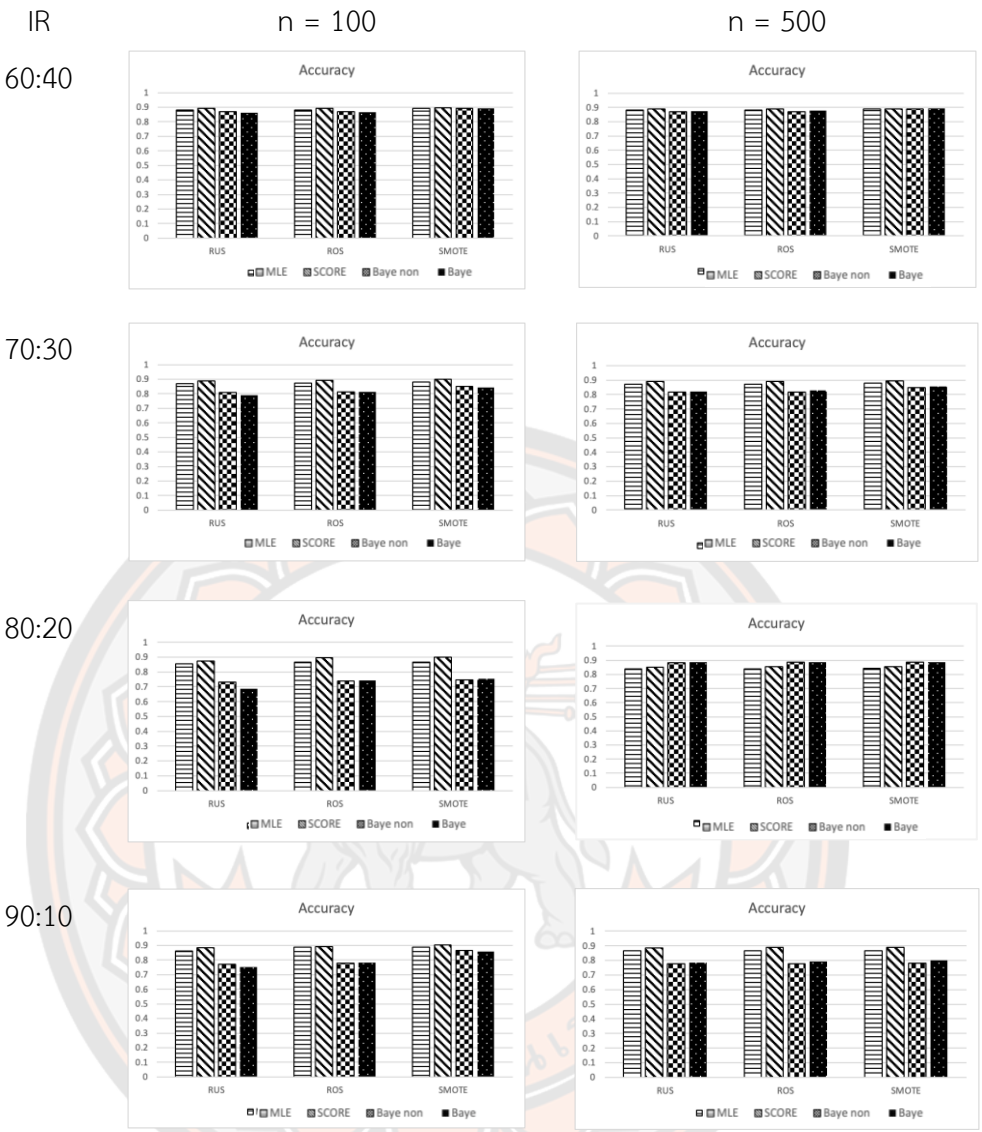
แผนภูมิแท่งแสดงวิธีการประมาณค่าพารามิเตอร์ ประกอบไปด้วย วิธีภาวะน่าจะเป็นสูงสุด (MLE) วิธีฟังก์ชันสกอร์ที่ปรับปรุง (SCORE) วิธีเบส์เซียน กรณีไม่ทราบความรู้ก่อนหน้า (Baye non) และวิธีเบส์เซียน กรณีทราบความรู้ก่อนหน้า (Baye) จัดการความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มลด (RUS) วิธีการสุ่มเกิน (ROS) และวิธีการสังเคราะห์ข้อมูลใหม่ (SMOTE) โดยวัดประสิทธิภาพของค่าความแม่นยำ ความไว ความจำเพาะ และความแม่นที่สมดุล ดังนี้



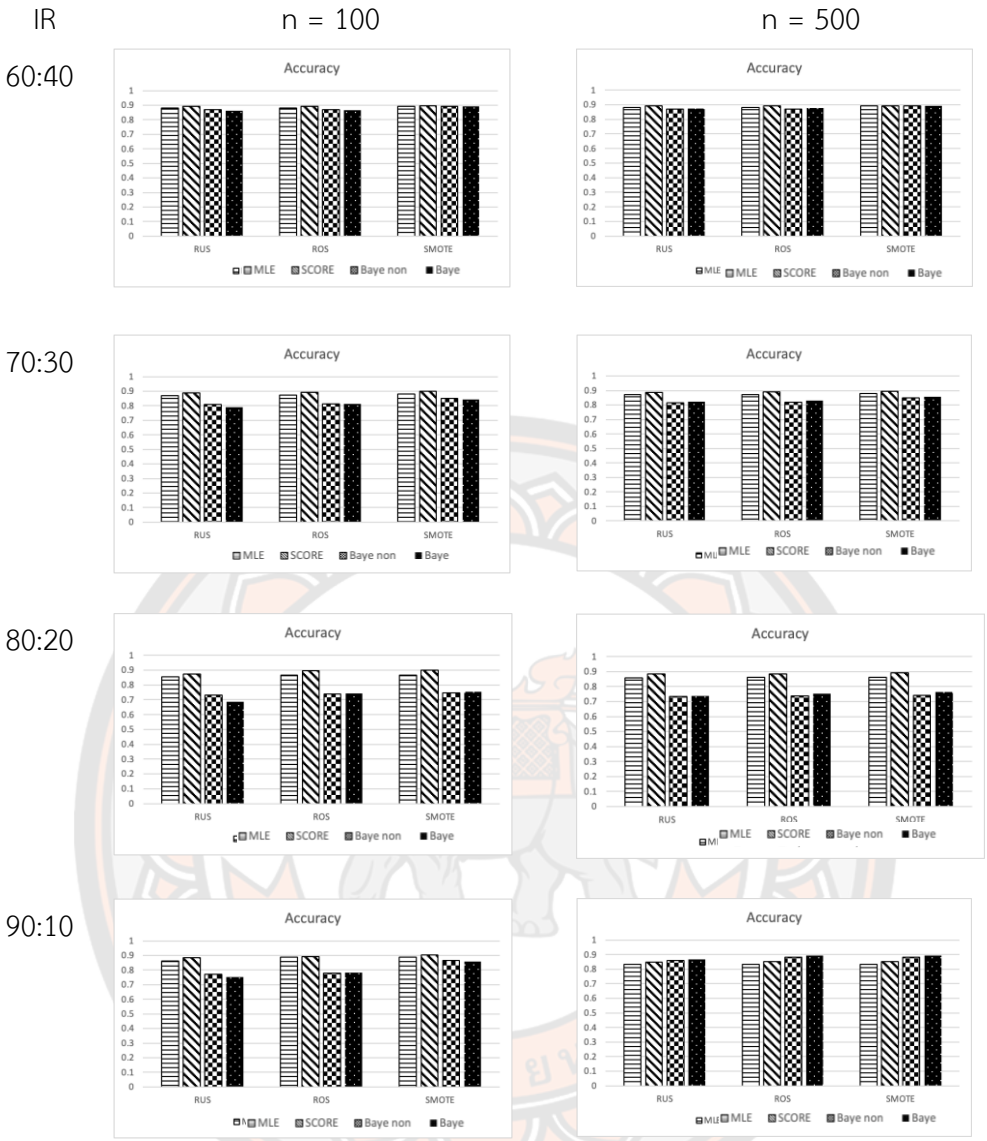
ภาคผนวก 1 แสดงแผนภูมิแท่งค่าความแม่นยำ โดยกำหนดตัวแปรอิสระเท่ากับ 1 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30



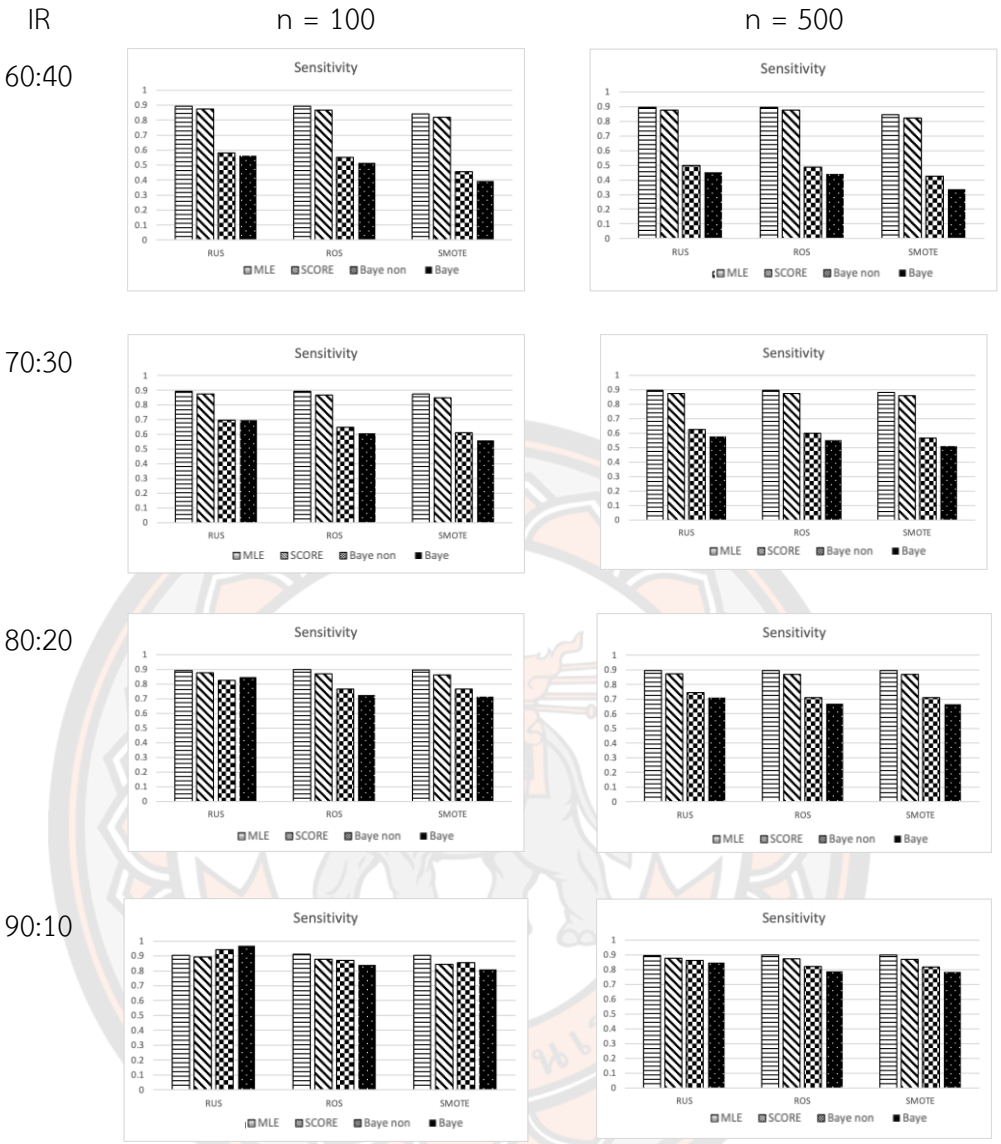
ภาพผนวก 12 แสดงแผนภูมิแท่งค่าความแม่นยำ โดยกำหนดตัวแปรอิสระเท่ากับ 1 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20



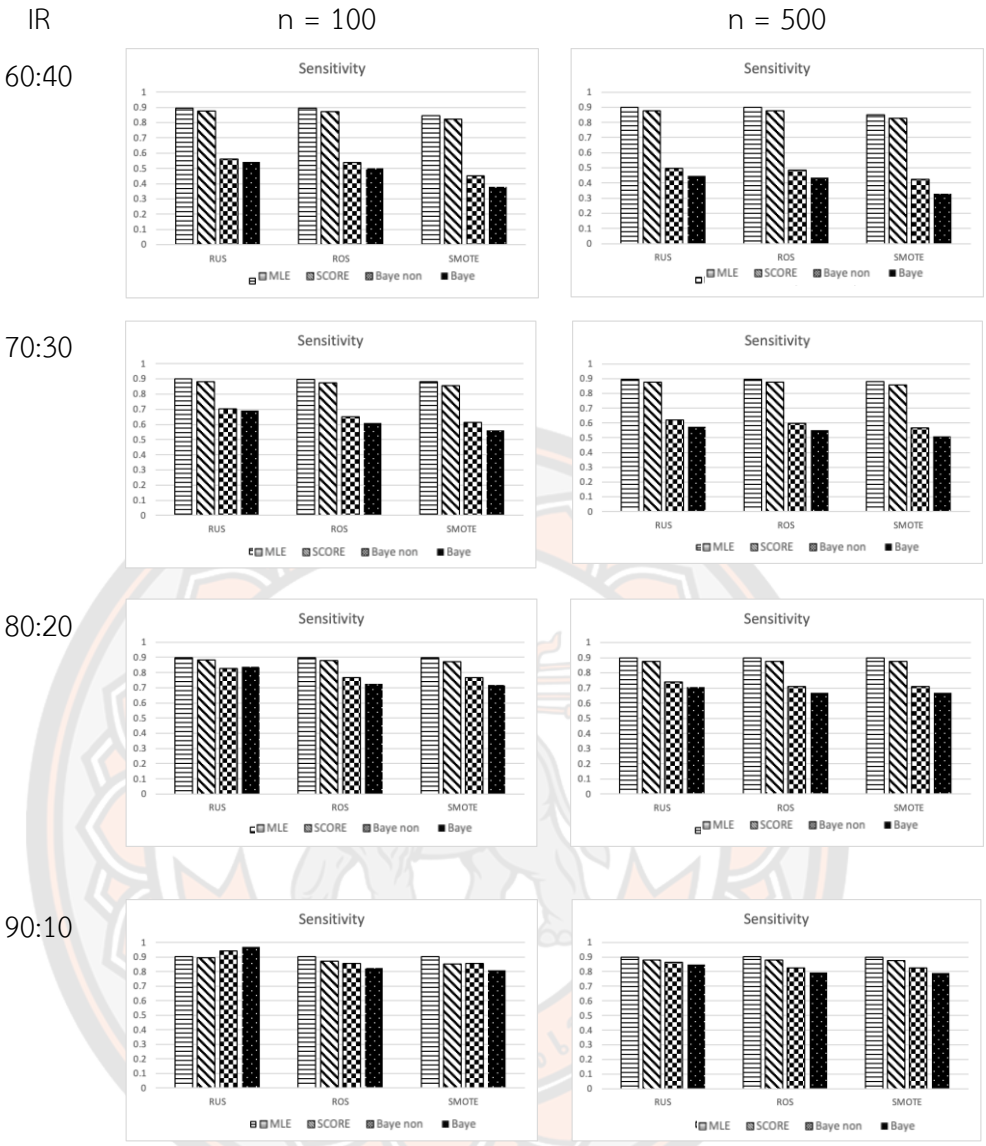
ภาพผนวก 3 แสดงแผนภูมิแท่งค่าความแม่นยำ โดยกำหนดตัวแปรอิสระเท่ากับ 3 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30



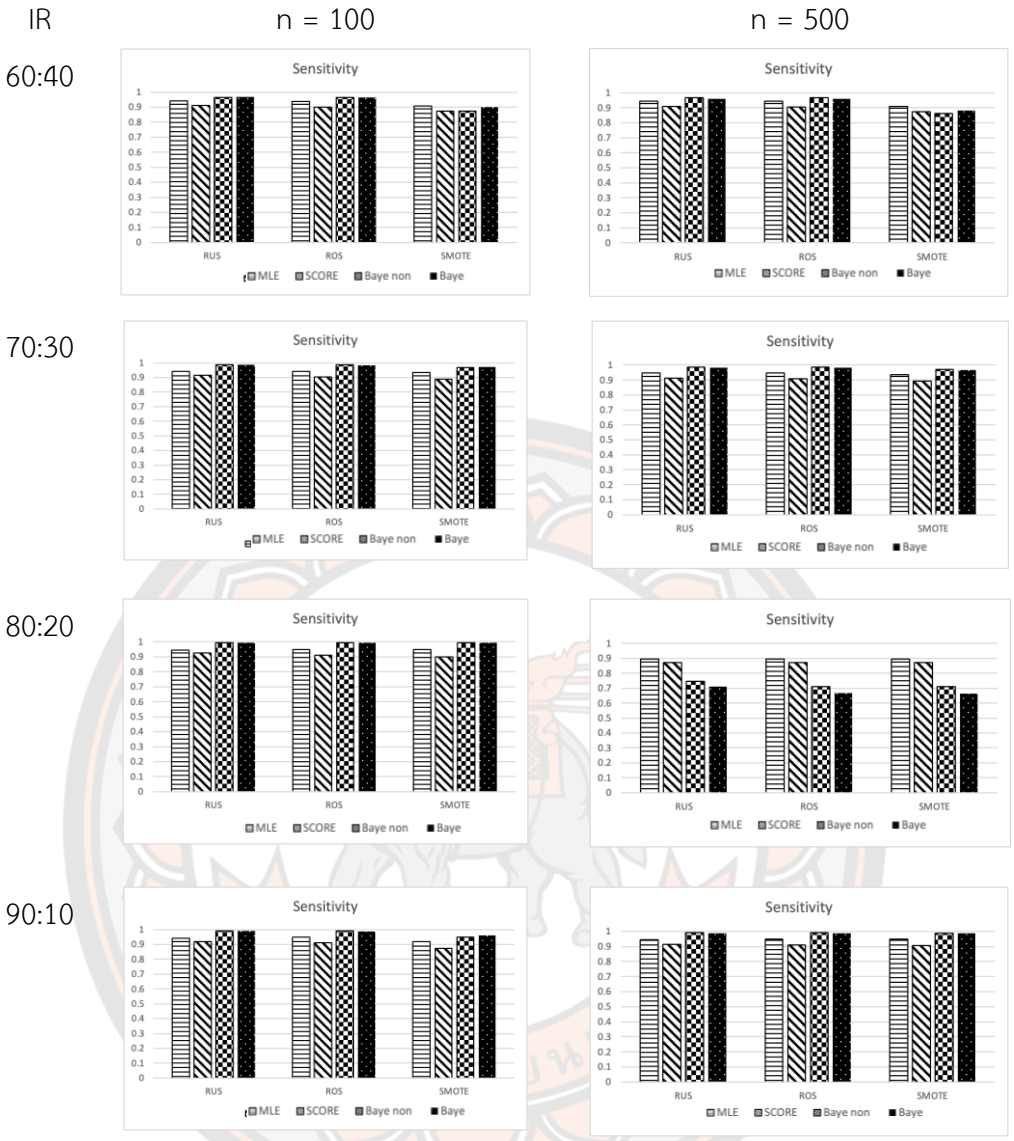
ภาพผนวก 4 แสดงแผนภูมิแท่งค่าความแม่นยำ โดยกำหนดตัวแปรอิสระเท่ากับ 3 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20



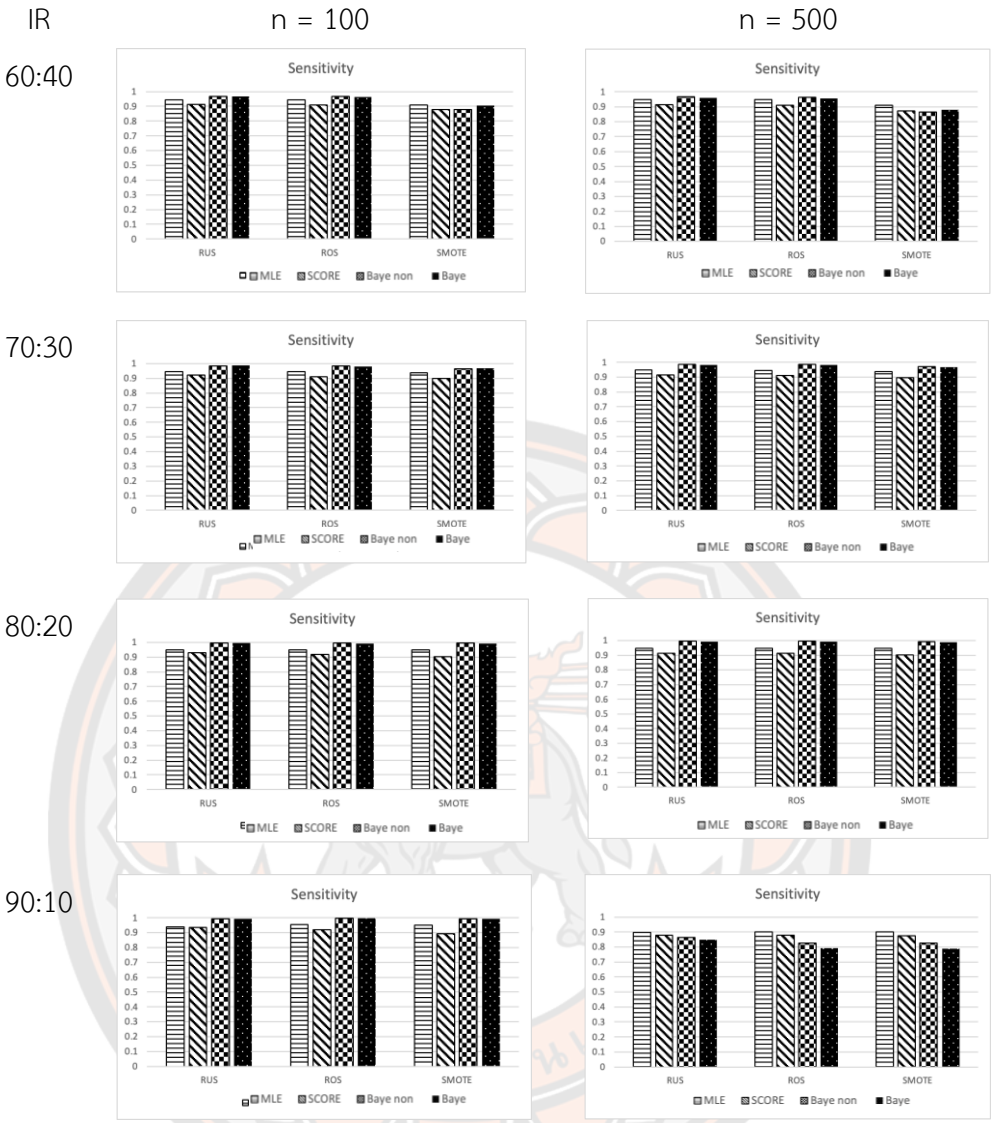
ภาพผนวก 5 แสดงแผนภูมิแท่งค่าความไว โดยกำหนดตัวแปรอิสระเท่ากับ 1 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30



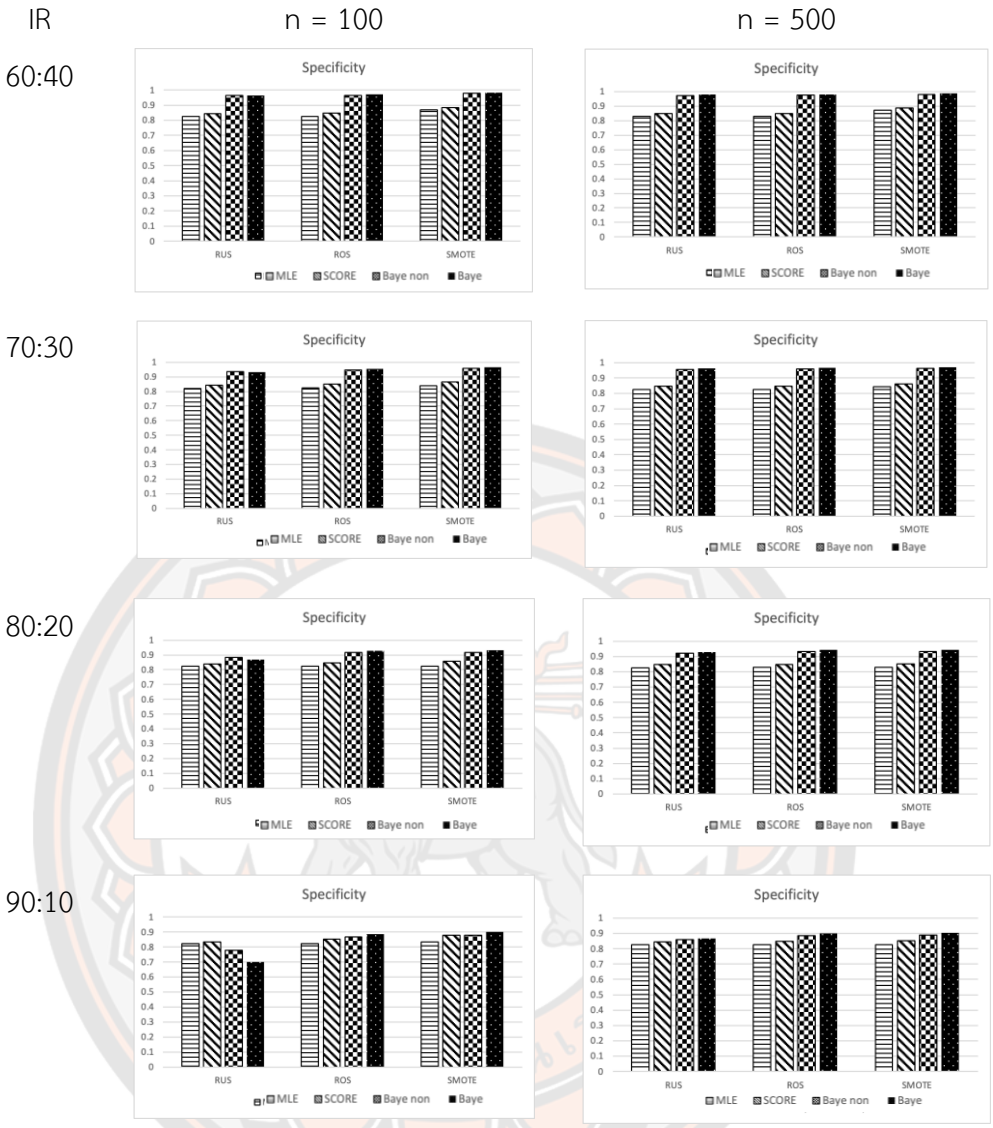
ภาพผนวก 6 แสดงแผนภูมิแท่งค่าความไว โดยกำหนดตัวแปรอิสระเท่ากับ 1 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20



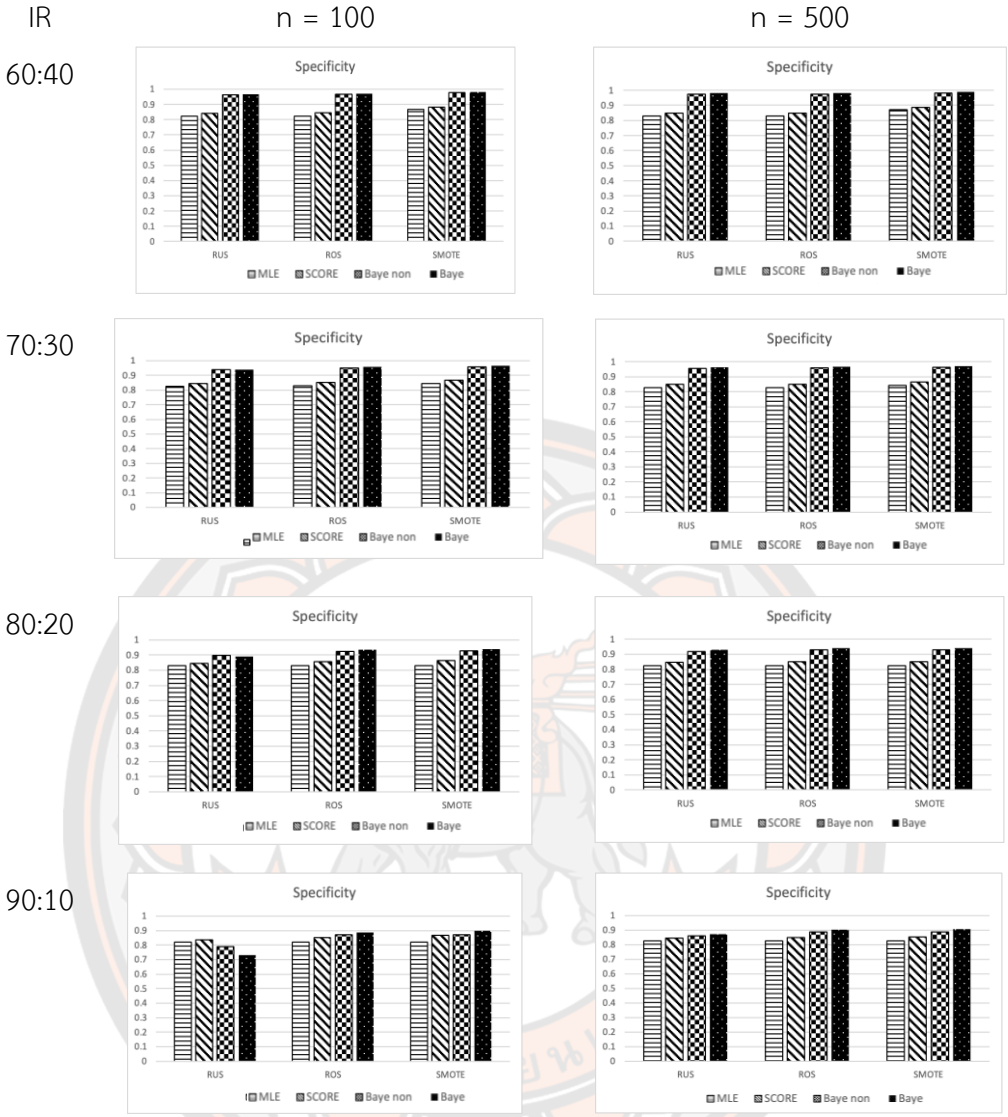
ภาพผนวก 7 แสดงแผนภูมิแท่งค่าความไว โดยกำหนดตัวแปรอิสระเท่ากับ 3 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30



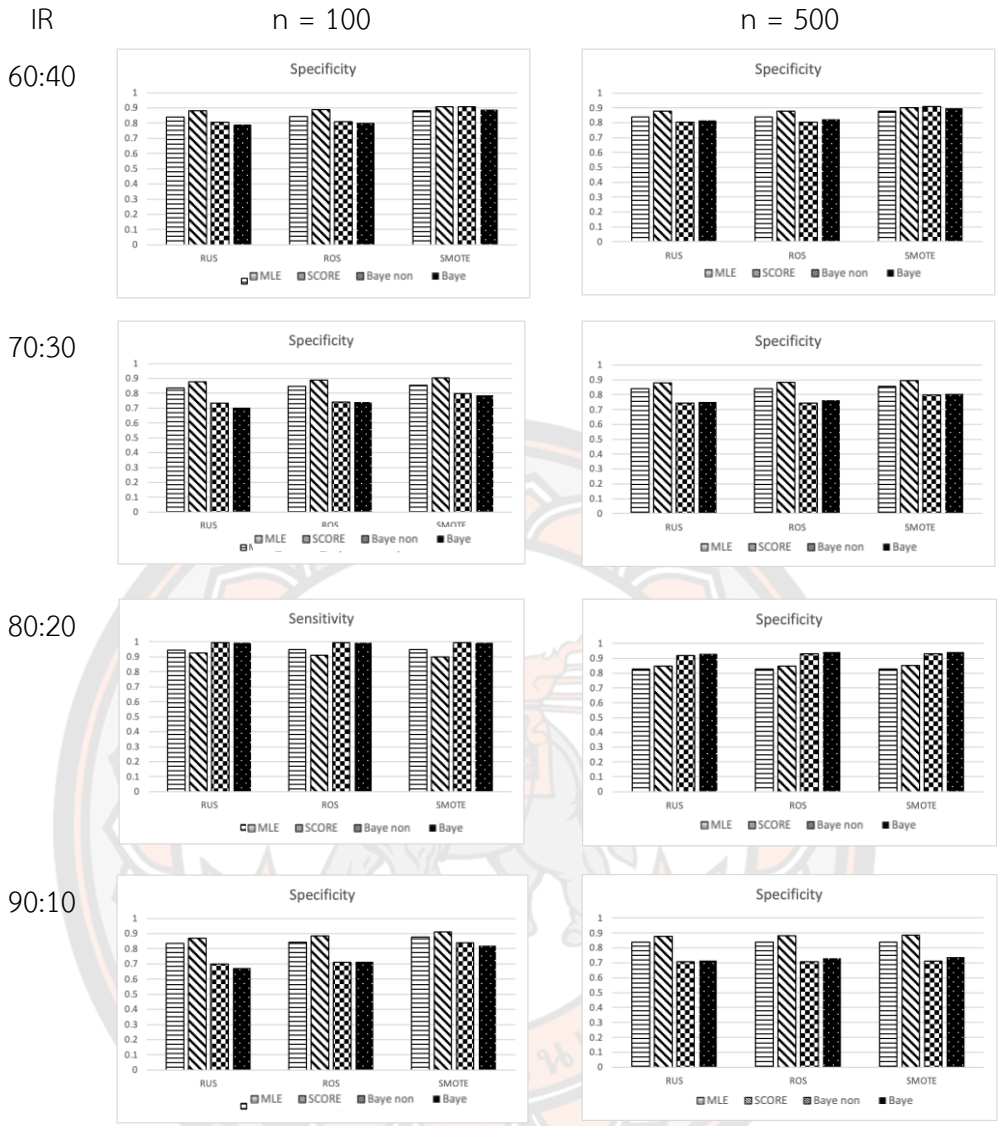
ภาพผนวก 8 แสดงแผนภูมิแท่งค่าความไว โดยกำหนดตัวแปรอิสระเท่ากับ 3 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20



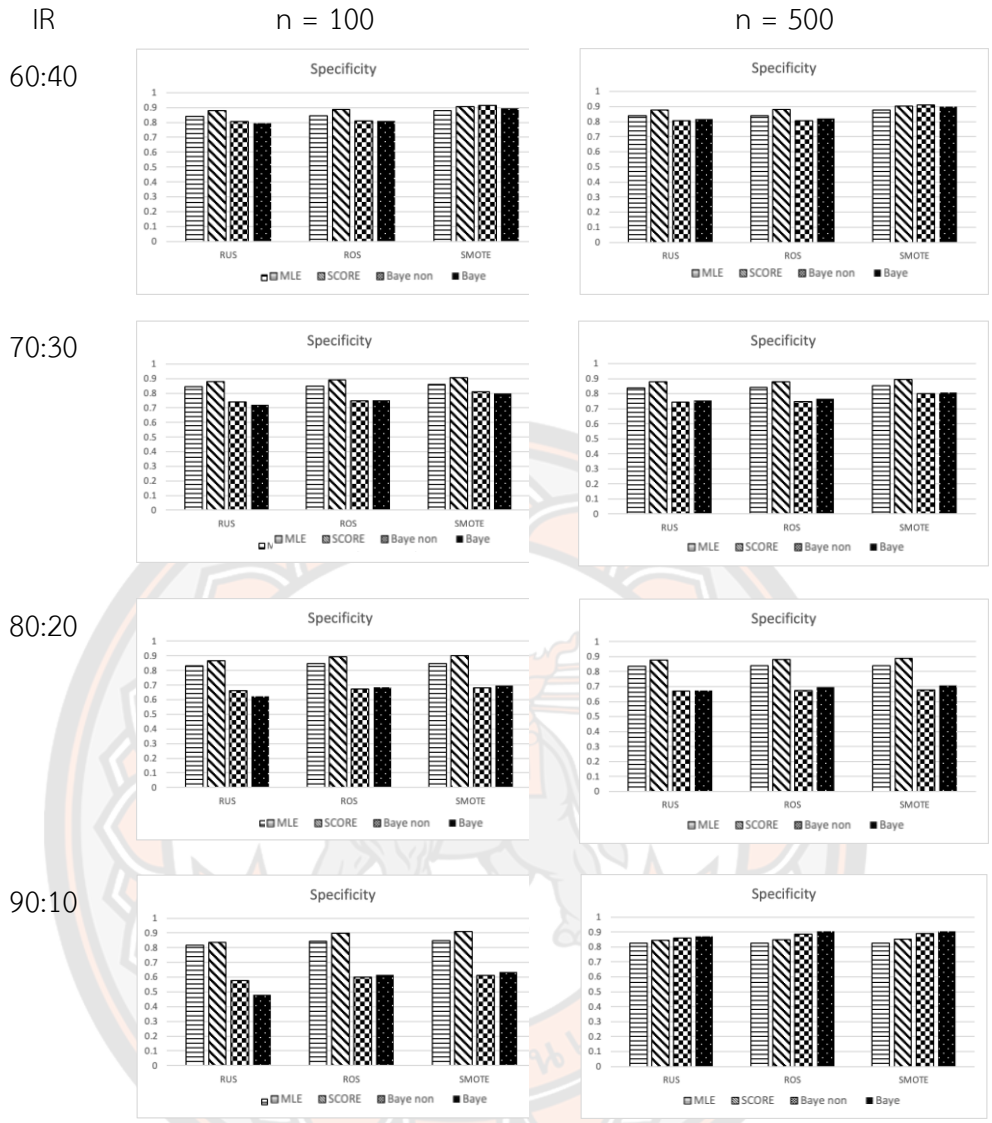
ภาพผนวก 9 แสดงแผนภูมิแท่งค่าความจำเพาะ โดยกำหนดตัวแปรอิสระเท่ากับ 1 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30



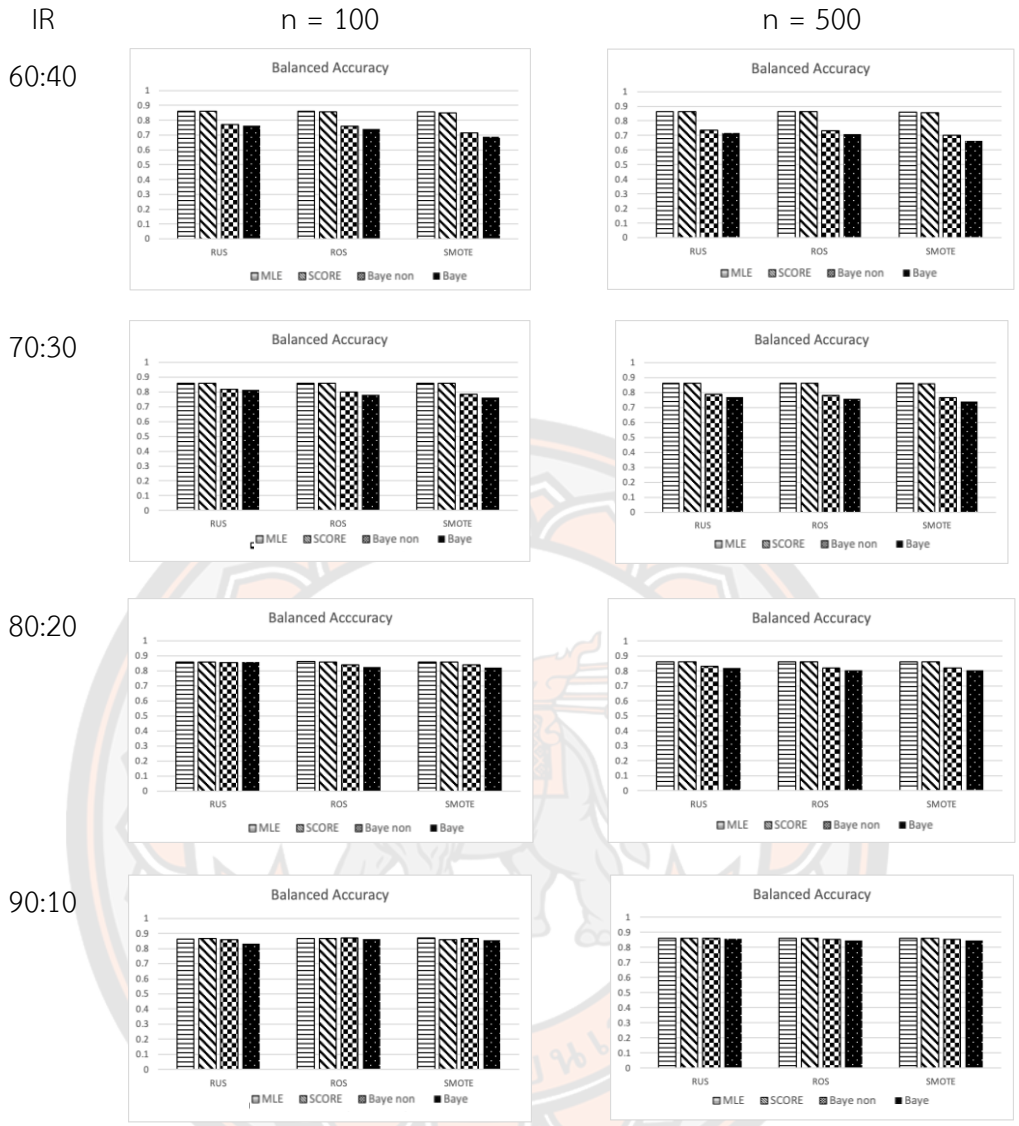
ภาพผนวก 10 แสดงแผนภูมิแท่งค่าความจำเพาะ โดยกำหนดตัวแปรอิสระเท่ากับ 1 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20



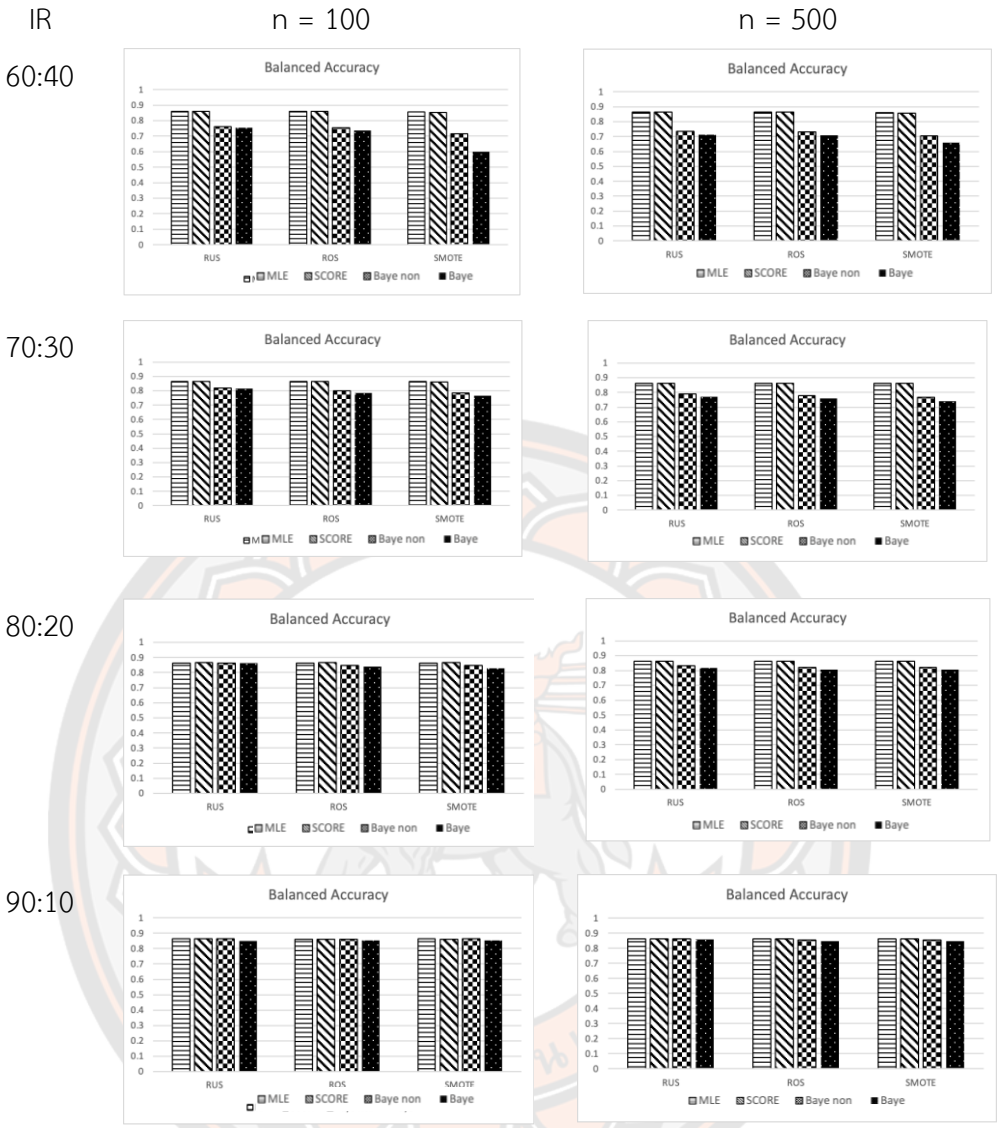
ภาพผนวก 11 แสดงแผนภูมิแท่งค่าความจำเพาะ โดยกำหนดตัวแปรอิสระเท่ากับ 3 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30



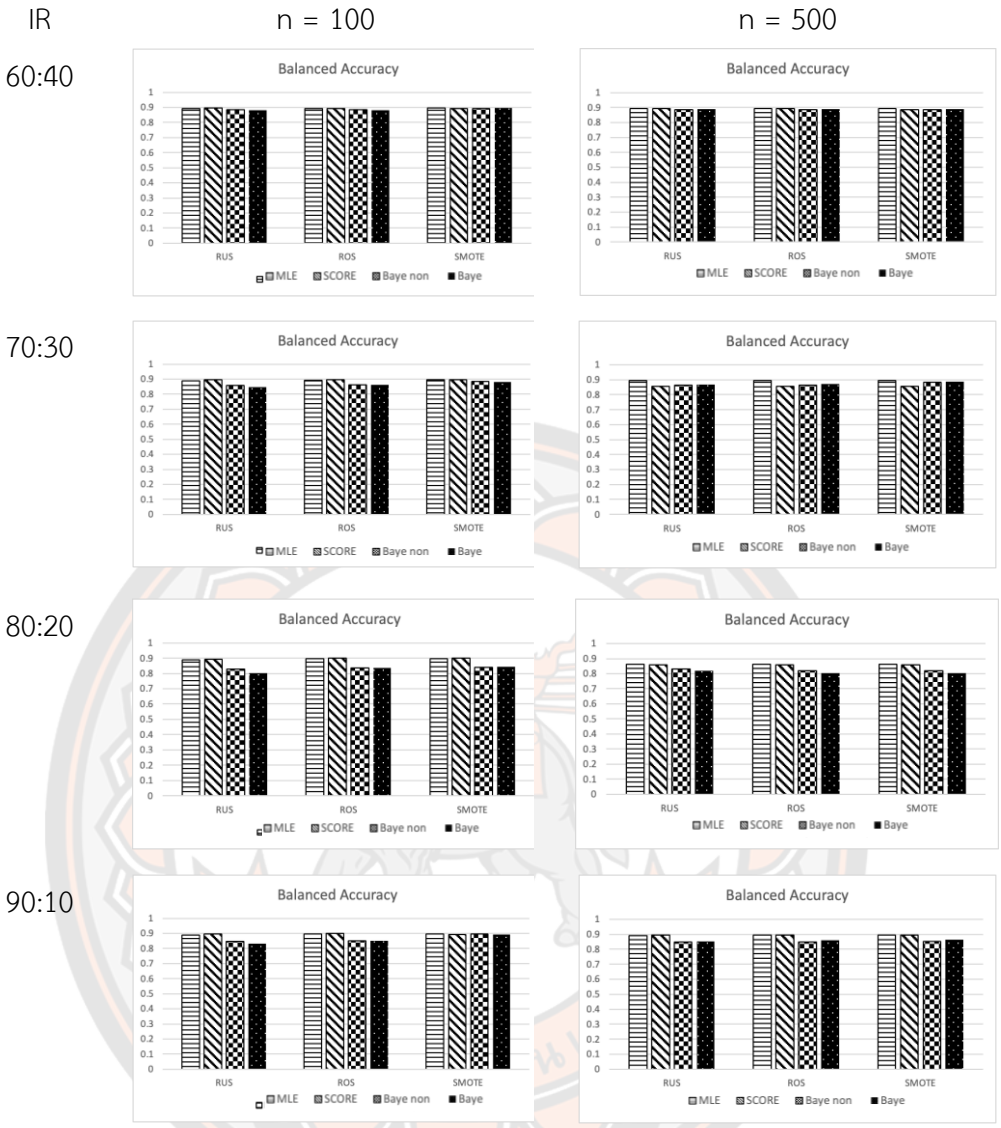
ภาพผนวก 12 แสดงแผนภูมิแท่งค่าความจำเพาะ โดยกำหนดตัวแปรอิสระเท่ากับ 3 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20



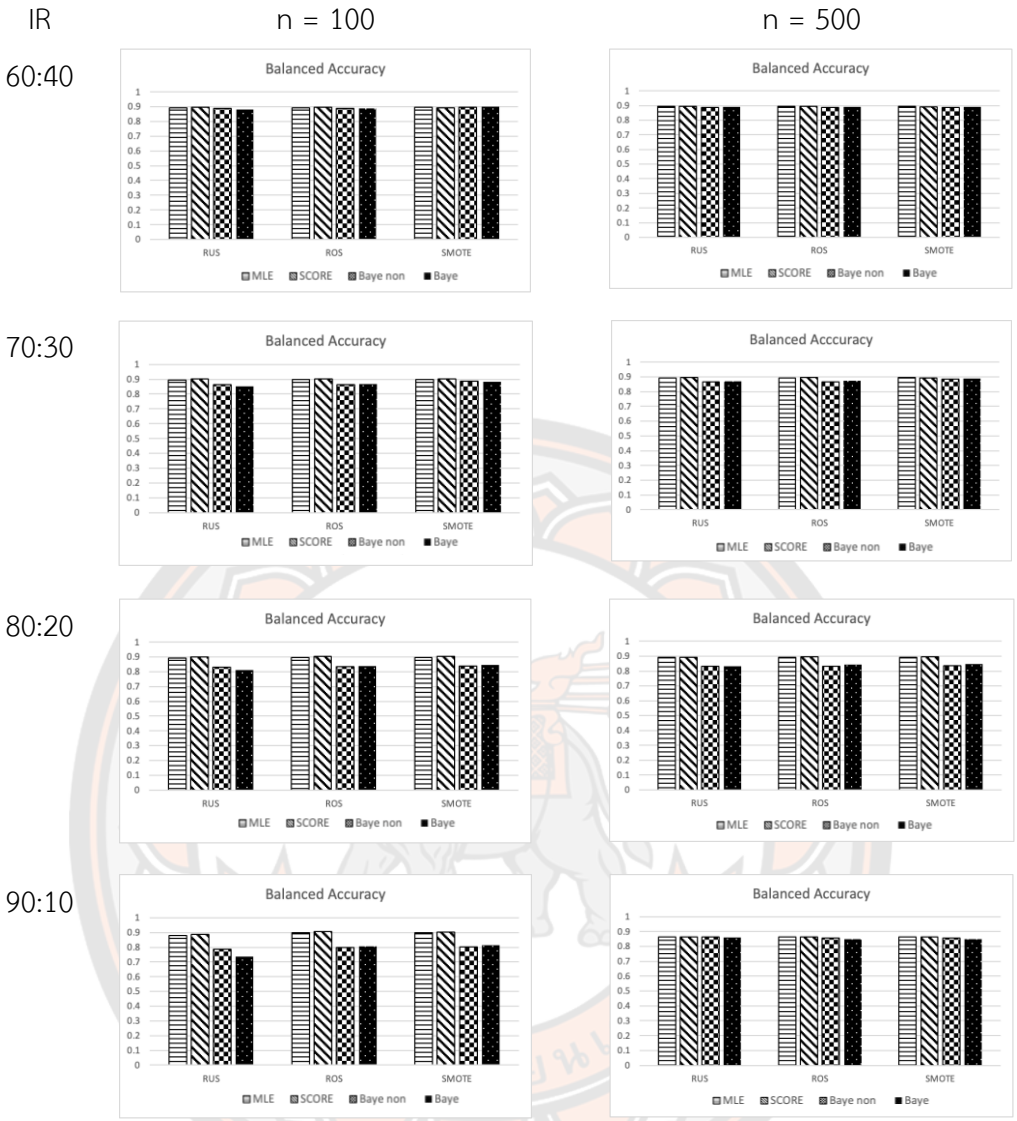
ภาพผนวก 13 แสดงแผนภูมิแท่งค่าความแม่นที่สมดุล โดยกำหนดตัวแปรอิสระเท่ากับ 1 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30



ภาพผนวก 14 แสดงแผนภูมิแท่งค่าความแม่นที่สมดุล โดยกำหนดตัวแปรอิสระเท่ากับ 1 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20



ภาพผนวก 15 แสดงแผนภูมิแท่งค่าความแม่นที่สมดุล โดยกำหนดตัวแปรอิสระเท่ากับ 3 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 70:30



ภาพผนวก 16 แสดงแผนภูมิแท่งค่าความแม่นที่สมดุล โดยกำหนดตัวแปรอิสระเท่ากับ 3 ตัว และอัตราส่วนของข้อมูลระหว่าง Training : Validation เป็น 80:20

บรรณานุกรม

บรรณานุกรม

บรรณานุกรม

กิตติพงษ์ ชมบุญ. (2559). **เทคนิคการค้นหคลาสที่ค้นพบได้ยากสำหรับข้อมูลที่ขนาดแตกต่างกันมาก.** วิทยานิพนธ์ วิศวกรรมศาสตร์ ดุษฎีบัณฑิต. สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

กิตติภพ แซ่เตีย และจิรภัทร์ หยกรัตนศักดิ์. (2564). **การจัดการข้อมูลไม่สมดุลของการทำกลยุทธ์เสนอขายประกันต่อยอดสำหรับผู้ถือบัตรเครดิต.** ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง กรุงเทพฯ.

กาญจน์เขจร ชูชีพ. (2561). **การถดถอยโลจิสติก (Logistic Regression)**. คณะวนศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.

กีระชาติ สุขสุทธิ์. (2559). **การจำแนกข้อมูลไม่สมดุลโดยใช้การปรับปรุงข้อมูลร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยขั้นตอนวิธีเชิงพันธุกรรมที่มีการเริ่มต้นใหม่**. วิทยานิพนธ์ วิศวกรรมศาสตร์ดุษฎีบัณฑิต. สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

กัลยา วานิชย์บัญชา. (2544). **การสิเคราะห์ตัวแปรหลายตัวด้วย SPSS for Windows** (พิมพ์ครั้งที่ 2). กรุงเทพฯ: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

กัลยา วานิชย์บัญชา. (2548). **การสิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows** (พิมพ์ครั้งที่ 4). กรุงเทพฯ: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

นพมาศ อัครจันทโชติ และคณะ. (2562). **การเปรียบเทียบวิธีการแก้ปัญหาข้อมูลไม่สมดุล สำหรับการจำแนกกลุ่มรายได้ของผู้ประกอบการร้านยาประเภท ข.ย.1.** ภาควิชาคณิตศาสตร์และสถิติ มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ.

เบญจภรณ์ จันทรกองกุล และคณะ. (2557). **วิธีการที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุลสูง.** ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยรามคำแหง.

พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตย์กุล. (2562). **เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ**. หลักสูตรวิศวกรรมซอฟต์แวร์ สำนักวิชาสารสนเทศศาสตร์ มหาวิทยาลัยวลัยลักษณ์.

มานพ วราภักดิ์. (2551) **ตัวประมาณแบบเบส์และการสุ่มตัวอย่างแบบกิบส์**. จุฬาลงกรณ์ธุรกิจปริทัศน์, 30 (115-116), 65-80.

วิชญ์วิสิฐ เกษรสิทธิ์ และคณะ. (2562). **การแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน**. วิทยาศาสตรมหาบัณฑิต สาขาสถิติ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์.

วีรานันท์ พงศาภักดี. (2541). **การวิเคราะห์ข้อมูลเชิงกลุ่ม: ทฤษฎีและการประยุกต์ กับ GLM และ SPSS/FW** (พิมพ์ครั้งที่ 2). ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์.

สุภวรรณ มานะการ. (2549). **การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มเมื่อตัวแปรอิสระมีลักษณะแบบไบนารี่**. วิทยานิพจธ์ วท.ม., มหาวิทยาลัยนเรศวร, พิษณุโลก

อัจฉรา แผ้วบาง และสายชล สินสมบรูณ์ทอง. (2562). **การปรับความไม่สมดุลของข้อมูลด้วยการจำแนก 5 วิธี**. ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

Albert, J. H. & Chib, S. (1993). **Bayesian analysis of binary and polychotomous response data**. Journey of the american statistics association, 88 (422), 669-679.

Brandt, J. & Lanzén, E. (2020). **A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification**. Department of Statistics, University of Uppsala.

Carlin, B. P. & Louis, T. A. (2009). **Bayesian methods for data analysis** $\left(3^{rd}\, ed.\right)$. United States of America: Chapman and Hall/CRC

Chawla, N. V. (2002). **SMOTE: Synthetic Minority Over-Sampling Technique**. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Firth, D. (1993). **Bias Reduction of Maximum Likelihood Estimates**. Biometrika, 80, 27-38. https://doi.org/10.1093/biomet/80.1.27

Faires, J. D. and Burden, R. (2010). **Numerical Analysis** $\left(3^{rd}\, ed.\right)$**.** United States of America: Cole-Thomson Learning.

Fawcett, T. (2016). **Learning from Imbalanced Classes.** https://www.svds.com/learning-imbalanced-classes/.

Febrianti, R., Widyaningsih, Y. & Soemartojo, S. (2018). **The parameter estimation of logistic regression with maximum likelihood method and score function modification**. Journal of Physics : Conference Series 1725 (2021) 012014 doi:10.1088/1742-6596/1725/1/012014.

Hassan, M. M., (2020). **A Fully Bayesian Logistic Regression Model for Classification of ZADA Diabetes Data**. Journal of JUOZ, pp. 95-111.

López, v. et al. (2017). **An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics**. Information Science 250 (2013) 113-141. journal homepage: www.elsevier.com/locate/ins.

Rahman, H. A., Wah, Y. B. & Huat, O. S. (2021). **Predictive Performance of Logistic Regression for Imbalanced Data with Categorical Covariate.** Pertanika J. Sci. & Technol. 29 (1): 181 - 197 (2021).

Dey, S. (2023). **Some naive and computationally polished techniques to approximate roots of equations**. Indian Institute of Science Education and Research - Bhopal Madhya Pradesh, India.

Srivastava, T. (2013). **Building a Logistic Regression model from scratch**. https://www.analyticsvidhya.com/blog/2015/10/basics-logistic-regression/.

Wah, Y. B., Rahman, H. A. & Haibo, H. (2016). **Handling imbalanced dataset using SVM and K-NN approach**. AIP conference Proceedings 1750, 020023 (2016)

YILMAZ, A. & CULIK, E. (2021). **A Bayesian Approach to Binary Logistics Regression Model with Application to OECD Data**. 94-101, 2021.