

การวิเคราะห์การจำแนกข้อมูลที่ไม่สมดุลในแต่ละกลุ่ม กรณีศึกษาการจำแนกกลุ่มเสี่ยงโรคเบาหวาน

Discriminant Analysis for Classifying Imbalance Data: A Case Study of Diabetes Risk Group

ทัศนีย์ น้าเจริญ*
ลลิต อิงศรีสว่าง**

บทคัดย่อ

การศึกษานี้เป็นการเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์และการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ในการจำแนกข้อมูล 3 กลุ่ม โดยที่มีความไม่สมดุลของข้อมูลแต่ละกลุ่ม ข้อมูลที่นำมาศึกษาเป็นข้อมูลกลุ่มเสี่ยงโรคเบาหวานของเจ้าหน้าที่ที่ปฏิบัติงานในโรงพยาบาลรัฐแห่งหนึ่งจำนวน 599 คน มีการวินิจฉัยเป็น 3 กลุ่ม คือกลุ่มไม่เสี่ยงโรคเบาหวาน มีประมาณ 90% กลุ่มเสี่ยงโรคเบาหวาน มีประมาณ 5% และกลุ่มเป็นโรคเบาหวาน มีประมาณ 5% ข้อมูลดังกล่าวไม่มีการแจกแจงแบบปกติพหุและความแปรปรวนร่วมแต่ละกลุ่มไม่เท่ากัน ในการศึกษาได้ทำการบวตสเตรปข้อมูลกลุ่มเสี่ยงโรคเบาหวานจำนวน 50 ชุด แต่ละชุดจะทำการการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ ด้วยฟังก์ชันการจำแนกเชิงเส้นและกำลังสอง และวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธีเคเนียร์สแนเบอร์ เมื่อกำหนด k

เป็น 3, 4, 5 พร้อมทั้งกำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วนที่เท่ากันและสัดส่วนที่แตกต่างกันสำหรับสร้างกฎการจำแนกกลุ่มที่เหมาะสม และวัดประสิทธิภาพของวิธีการวิเคราะห์การจำแนกแต่ละวิธีด้วยค่าเฉลี่ยอัตราความผิดพลาด ผลการศึกษพบว่าจากข้อมูลบวตสเตรป 50 ชุด การวิเคราะห์การจำแนกด้วยวิธีเคเนียร์สแนเบอร์ เมื่อ $k=3$ หรือ 4 และเมื่อกำหนดค่าความน่าจะเป็นก่อนหน้าของกลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน ด้วยสัดส่วน 0.90 : 0.05 : 0.05 หรือ 0.80 : 0.10 : 0.10 พบว่าค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มผิดมีค่าต่ำที่สุด

คำสำคัญ: โรคเบาหวาน ค่าอัตราความผิดพลาด การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์

*นักศึกษาลัทธิสุตรมหาบัณฑิต สาขาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

**อาจารย์ที่ปรึกษา

Abstract

The purpose of this research is to compare the classification performance among parametric discriminant analysis and nonparametric discriminant analysis in three-group classification, imbalance in each group. The data is Diabetes Risk Group, nor multivariate normality and neither the group in the data nor the covariance matrices of the three groups were equal. The data have made 50 bootstrap samples, to analyze classification method that consists of parametric discriminant analysis, which linear discriminant function and quadratic discriminant function. And k - nearest neighbor when $k=3, 4, 5$ in the nonparametric discriminant analysis. The prior probability was proportional. Using the Error Rate performance in comparison to classify that 50 bootstrap samples, k - nearest neighbor will provide the error rate in the classification minimum when $k=3, 4$. And determining the prior probability $0.90 : 0.05 : 0.05$ หรือ $0.80 : 0.10 : 0.10$. To keep the error rate in the classification minimum.

Keywords: Parametric Discriminant Analysis, Nonparametric Discriminant Analysis, Diabetes, Error Rate

คำนำ

การวิเคราะห์การจำแนก (Discriminant Analysis) จัดเป็นเทคนิคหนึ่งของการวิเคราะห์พหุตัวแปร (Multivariate analysis) ใช้สำหรับแบ่งกลุ่มของสิ่งที่สนใจศึกษา (Objects) หรือค่าสังเกต (Observation) เป็น กลุ่มย่อย หรือเรียกว่า คลาส (Class) จำนวน k กลุ่ม ($k \geq 2$) โดยพิจารณาว่าตัวแปรใดบ้าง

ที่ทำให้กลุ่มหรือคลาสต่างกัน และตัวแปรใดบ้างที่มีความสำคัญต่อกลุ่มที่ทำการศึกษา โดยสร้างเป็นสมการการจำแนกที่แสดงความสัมพันธ์ระหว่างตัวแปรที่ถูกจัดเป็นกลุ่มย่อย (เรียกว่า ตัวแปรตาม ซึ่งเป็นตัวแปรกลุ่ม) กับกลุ่มของตัวแปรที่มีความสามารถในการแบ่งกลุ่ม หรือบอกความแตกต่างระหว่างกลุ่มได้ (เรียกว่า ตัวแปรอิสระ) นอกจากนี้ การวิเคราะห์การจำแนก ยังสามารถพยากรณ์กลุ่มให้กับหน่วยตัวอย่างใหม่หรือหน่วยสังเกตใหม่ โดยใช้สมการการจำแนกที่สร้างขึ้น

วิธีการวิเคราะห์การจำแนก มีหลายวิธีทั้งวิธีที่อิงพารามิเตอร์ (Parametric Method) และวิธีที่ไม่อิงพารามิเตอร์ (Nonparametric Method) ซึ่งวิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ มีข้อตกลงที่สำคัญคือ ตัวแปรอิสระจะต้องถูกสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติพหุ (Multivariate Normal Distribution) และเมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระแต่ละกลุ่มอาจจะเท่ากันหรือไม่เท่ากันก็ได้ กรณีความแปรปรวนร่วมเท่ากัน ใช้การวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกเชิงเส้น (Linear discriminant Function) ส่วนกรณีความแปรปรวนร่วมไม่เท่ากัน ใช้การวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสอง (Quadratic Discriminant Function) โดยฟังก์ชันการจำแนกทั้ง 2 ชนิดจะใช้ได้กับตัวแปรอิสระที่เป็นตัวแปรปริมาณชนิดต่อเนื่องเท่านั้น

ในทางปฏิบัติข้อมูลที่น่ามาศึกษาอาจมีคุณสมบัติไม่เป็นไปตามข้อตกลง อาจส่งผลให้การจำแนกกลุ่มเกิดความผิดพลาดได้ ในทางทฤษฎีมีทางเลือกอื่นสำหรับการวิเคราะห์การจำแนก เช่น 1) วิธีการวิเคราะห์การจำแนกด้วยตัวแบบการถดถอยโลจิสติก (Logistic Regression Model) สำหรับข้อมูลที่ไม่มีการแจกแจงปกติพหุและมีตัวแปรอิสระบางตัวเป็นตัวแปรกลุ่ม และ 2) วิธีการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ (Nonparametric

discriminant analysis) สำหรับข้อมูลไม่ได้มีการแจกแจงปกติพหุและความแปรปรวนร่วมไม่เท่ากัน แต่ตัวแปรอิสระทุกตัวเป็นตัวแปรเชิงปริมาณ เช่น วิธีเคอร์เนล (kernel method) หรือวิธี k-Nearest Neighbour: kNN แต่วิธี kNN เป็นที่นิยมมากกว่าวิธีเคอร์เนล (Khattree and Naik, 2000)

อย่างไรก็ตามในการวิเคราะห์การจำแนกข้อมูลออกเป็น k กลุ่มหรือ k คลาส บางครั้งผู้วิจัยอาจจะเผชิญปัญหาว่ามีข้อมูลในกลุ่มหนึ่งมากกว่ากลุ่มอื่นเป็นจำนวนมาก สมมติว่า $k = 3$ และกลุ่มที่ 1 มีข้อมูล 90% ส่วนกลุ่มที่ 2 และกลุ่มที่ 3 มีข้อมูลกลุ่มละ 5% จึงเป็นที่สงสัยว่าในการวิเคราะห์การจำแนก หน่วยตัวอย่างหรือหน่วยสังเกตส่วนใหญ่จะถูกพยากรณ์ไปอยู่ในกลุ่มหรือคลาสที่มีข้อมูลมากกว่า ซึ่งจะมีผลต่ออัตราความผิดพลาด (Error rate) ของการจำแนกกลุ่ม

การศึกษานี้จึงสนใจที่จะเปรียบเทียบประสิทธิภาพของการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ และการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ กับปัญหาข้อมูลกลุ่มเสี่ยงโรคเบาหวานของเจ้าหน้าที่ปฏิบัติงานในโรงพยาบาลรัฐแห่งหนึ่ง ซึ่งมีการวินิจฉัยเป็น 3 กลุ่ม คือ กลุ่มไม่เสี่ยงโรคเบาหวาน กลุ่มเสี่ยงโรคเบาหวาน และกลุ่มที่เป็นโรคเบาหวาน ซึ่งข้อมูลแต่ละกลุ่มมีความไม่สมดุลกัน คือมีจำนวนข้อมูลในแต่ละกลุ่มแตกต่างกันมาก คือ 90% จัดอยู่ในกลุ่มไม่เสี่ยงโรคเบาหวาน 5% จัดอยู่ในกลุ่มเสี่ยงโรคเบาหวาน และอีก 5% จัดอยู่ในกลุ่มที่เป็นโรคเบาหวาน สำหรับตัวแปรอิสระที่นำมาพิจารณาในการจำแนกกลุ่มเสี่ยงประกอบด้วย อายุ เพศ ระดับน้ำตาลในเลือดและค่าดัชนีมวลกาย โดยประมาณค่าอัตราความผิดพลาดของการจำแนกด้วยวิธีบูตสเตรป (การสุ่มตัวอย่างซ้ำแบบใส่คืน จากข้อมูลกลุ่มเสี่ยงโรคเบาหวาน) จำนวน 50 ชุดแต่ละชุดมีขนาดตัวอย่าง 599 ค่าสังเกต และเปรียบเทียบอัตราความผิดพลาดเฉลี่ยที่ได้จากข้อมูล 50 ชุด ระหว่างวิธีวิเคราะห์การจำแนกที่อิงพารามิเตอร์ และการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์พร้อมทั้ง

เปรียบเทียบอัตราความผิดพลาดที่วิเคราะห์ได้จากข้อมูลจริง

วัตถุประสงค์

1. เพื่อเปรียบเทียบประสิทธิภาพของความต้องการในการจำแนกข้อมูล 3 กลุ่มระหว่างการวิเคราะห์การจำแนกที่อิงพารามิเตอร์และที่ไม่อิงพารามิเตอร์เมื่อมีความไม่สมดุลของข้อมูลในแต่ละกลุ่ม
2. เพื่อประยุกต์วิธีการวิเคราะห์การจำแนกที่เหมาะสมกับการจำแนกกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่มกรณีที่มีความไม่สมดุลของข้อมูลในแต่ละกลุ่ม

วิธีดำเนินการวิจัย

งานวิจัยนี้มีวัตถุประสงค์ เพื่อศึกษาการวิเคราะห์การจำแนกแบบพารามิเตอร์ โดยใช้สมการการจำแนกแบบเส้นตรง แบบกำลังสอง และวิธีการจำแนกแบบนอนพารามิเตอร์ โดยวิธีเคเนียร์สเนเบอร์ ได้กำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อน (Priors Probability) จากค่าสัดส่วนที่แตกต่างกันในการจำแนกกลุ่มผู้ป่วยโรคเบาหวาน โดยใช้ค่าอัตราความผิดพลาด (Error Rate) ในการเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่ม

1. ข้อมูลที่ใช้ในการศึกษา

ข้อมูลที่นำมาศึกษาครั้งนี้เป็นข้อมูลการตรวจสุขภาพประจำปีของเจ้าหน้าที่ที่ปฏิบัติงานในโรงพยาบาลนพรัตน์ราชธานี ระหว่างเดือนสิงหาคม 2551 - เมษายน 2552 จำนวน 599 คน โดยแพทย์ได้จัดกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่มคือ

- 1) กลุ่มไม่เสี่ยงโรคเบาหวาน จำนวน 561 คน
- 2) กลุ่มเสี่ยงโรคเบาหวาน จำนวน 17 คน
- 3) กลุ่มที่เป็นโรคเบาหวาน จำนวน 21 คน

2. ตัวแปรที่ใช้ในการศึกษา

2.1 ตัวแปรตาม (y) คือ ตัวแปรกลุ่มแบ่งเป็น 3 กลุ่มคือ กลุ่มไม่เสี่ยงโรคเบาหวาน

กลุ่มที่เสี่ยงโรคเบาหวาน และกลุ่มเป็นโรคเบาหวาน ดังนี้

กลุ่มไม่เสี่ยงโรคเบาหวาน กำหนดให้ $y = 0$

กลุ่มเสี่ยงโรคเบาหวาน กำหนดให้ $y = 1$

กลุ่มเป็นโรคเบาหวาน กำหนดให้ $y = 2$

2.2 ตัวแปรอิสระที่นำมาศึกษาประกอบด้วย

1. อายุ (AGE) หน่วยเป็น ปี
2. เพศ (SEX) (ให้ 0 = เพศชาย, 1 = เพศหญิง)
3. ระดับน้ำตาลในเลือด หน่วยเป็น (mg/dl)
4. ค่าดัชนีมวลกาย หน่วยเป็น (kg/m^2)

3. ขั้นตอนการศึกษาวิเคราะห์

3.1 สุ่มตัวอย่างซ้ำทั้งหมดจากข้อมูลจริงที่เป็นข้อมูลการตรวจสุขภาพประจำปีของเจ้าหน้าที่ที่ปฏิบัติงานในโรงพยาบาลนพรัตนราชธานี ระหว่างเดือนสิงหาคม 2551 - เมษายน 2552 ที่เป็นกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่ม จำนวน 50 ชุดข้อมูล โดยข้อมูลในแต่ละชุดมีจำนวน 599 ค่าสังเกต (แบ่งข้อมูลเป็นชุดข้อมูลสำหรับฝึกสอนจำนวน 2/3 ของข้อมูลทั้งหมด และเป็นชุดข้อมูลสำหรับทดสอบจำนวน 1/3 ของข้อมูลทั้งหมด โดยใช้วิธีการสุ่ม)

3.2 นำข้อมูลชุดฝึกสอนแต่ละชุดที่ได้จากข้อมูลจริงและข้อมูลบุตสแตรปแต่ละชุด มาตรวจสอบข้อตกลง การแจกแจงปกติพหุด้วยวิธีการ Mardia's multivariate skewness and kurtosis ตรวจสอบการเท่ากันของความแปรปรวนร่วมด้วยสถิติทดสอบ Box's M และตรวจสอบการเท่ากันของค่าเฉลี่ยของแต่ละประชากรด้วยสถิติ Hotelling T²

3.3 ผลจากการตรวจสอบข้อตกลง

1. หากข้อมูลมีการแจกแจงแบบปกติพหุ มีความแปรปรวนร่วมเท่ากัน วิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกเชิงเส้น

2. หากข้อมูลมีการแจกแจงแบบปกติพหุ แต่ความแปรปรวนร่วมไม่เท่ากัน วิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง

3. หากข้อมูลไม่มีการแจกแจงแบบปกติพหุ และความแปรปรวนร่วมไม่เท่ากัน วิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN โดยที่ k คือจำนวนค่าสังเกตที่อยู่ล้อมรอบค่าสังเกต x ที่ต้องการถูกจัดเข้ากลุ่มใดกลุ่มหนึ่งใน 3 กลุ่มเสี่ยง ในที่นี้กำหนด k เป็น 3, 4, 5

3.4 สร้างกฎการจำแนกที่เหมาะสมจากชุดข้อมูลฝึกสอนแต่ละชุด โดยกำหนดค่าความน่าจะเป็นก่อนหน้าในการจัดจำแนกหน่วยตัวอย่างหรือค่าสังเกตเข้ากลุ่มที่ 1 (กลุ่มไม่เสี่ยงโรคเบาหวาน) : กลุ่มที่ 2 (กลุ่มเสี่ยงโรคเบาหวาน) : กลุ่มที่ 3 (กลุ่มที่เป็นโรคเบาหวาน) ตามลำดับ ดังนี้

1. จัดสรรด้วยสัดส่วนที่เท่ากัน คือ 0.3333 : 0.3333 : 0.3333
2. จัดสรรตามสัดส่วน 0.90 : 0.05 : 0.05
3. จัดสรรตามสัดส่วน 0.80 : 0.10 : 0.10
4. จัดสรรตามสัดส่วน 0.70 : 0.15 : 0.15

3.5 นำกฎการจำแนกที่ได้จากข้อมูลชุดฝึกสอนจากข้อ 3.4 มาใช้กับข้อมูลชุดทดสอบ มาคำนวณค่าอัตราความผิดพลาด

3.6 เปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มด้วยวิธีการวิเคราะห์การจำแนกที่ได้จากข้อ 3.3 โดยวัดจากอัตราความผิดพลาดเฉลี่ย วิธีการวิเคราะห์การจำแนกที่ให้ค่าอัตราความผิดพลาดเฉลี่ยต่ำสุด จะเป็นวิธีการที่เหมาะสมที่สุดกับข้อมูลกลุ่มเสี่ยงโรคเบาหวาน

3.7 ทำซ้ำข้อ 3.2 - 3.6 กับข้อมูลจริง แล้วนำอัตราความผิดพลาดเฉลี่ยจากข้อมูลบุตสแตรปเปรียบเทียบกับอัตราความผิดพลาดที่ได้จากข้อมูลจริง

วิธีการทางสถิติที่ใช้ในการศึกษาวิจัย

1. การวิเคราะห์การจำแนก (Discriminant Analysis)

วัตถุประสงค์ของการวิเคราะห์การจำแนก เพื่อศึกษาว่าตัวแปรอิสระตัวใดบ้างเป็นตัวแปรที่ทำให้กลุ่มต่างกัน โดยการสร้างสมการการจำแนกกลุ่มและเพื่อพยากรณ์หน่วยใหม่ที่ยังไม่ทราบกลุ่มว่าควรอยู่กลุ่มใด โดยที่ตัวแปรตามเป็นตัวแปรเชิงกลุ่มตั้งแต่ 2 กลุ่มขึ้นไป และตัวแปรอิสระเป็นตัวแปรเชิงปริมาณหรือเป็นตัวแปรเชิงกลุ่มรวมอยู่ด้วย

ข้อตกลงเบื้องต้นของการวิเคราะห์จำแนกแบบอิงพารามิเตอร์ (Hair et al., 2010) มีดังนี้

1. ตัวแปรอิสระมีการแจกแจงแบบปกติพหุ (Multivariate Normality)
2. เมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระแต่ละกลุ่มเท่ากัน
3. ตัวแปรอิสระมีความสัมพันธ์เชิงเส้น
4. ตัวแปรอิสระไม่มีความสัมพันธ์กันในระดับสูง (Multicollinearity)

2. วิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์

2.1 วิธีการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกเชิงเส้น

มีประชากร k กลุ่ม มีค่าความแปรปรวนร่วมเท่ากัน คือ Σ และค่าใช้จ่าย (cost) ในการจำแนกผิดในแต่ละกลุ่มสมมติว่ามีค่าเท่ากัน ค่าระยะห่างกำลังสองคำนวณได้จาก

$$D_t^2(x) = (x - \mu_t)' \Sigma^{-1} (x - \mu_t) - 2 \ln(\pi_t)$$

ค่าสังเกต x จะถูกจัดให้อยู่ในประชากรกลุ่มที่ t ถ้า $D_t^2(x) = \min_{j=1, \dots, k} (D_j^2(x))$

2.2 วิธีการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสอง

วิธีการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสอง ดังนี้ เมื่อเมทริกซ์ความแปรปรวนร่วมของประชากรไม่เท่ากัน กฎการจำแนก สำหรับประชากร k กลุ่ม จะมีความซับซ้อนมากขึ้น ถึงแม้ว่าข้อมูล

จะเป็นไปตามข้อตกลงของการแจกแจงแบบปกติพหุ กฎการจำแนกจะขึ้นอยู่กับเมทริกซ์ความแปรปรวนร่วมของประชากรแต่ละกลุ่ม $\Sigma_1, \Sigma_2, \dots, \Sigma_k$, การจัดกลุ่มให้กับค่าสังเกต x เข้าประชากรกลุ่มที่ t ; $t = 1, \dots, k$ จะพิจารณาจากค่า

$$D_t^2(x) = \min_{j=1, \dots, k} (D_j^2(x))$$

3. วิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ : วิธี เคเนียร์เซนเบอร์ (kNN)

วิธีการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ว่าในการจัดกลุ่มให้กับหน่วยตัวอย่างใหม่ให้อยู่ในกลุ่มเดียวกับหน่วยตัวอย่างที่อยู่รอบข้างที่ใกล้ที่สุด การพิจารณาความใกล้เคียงโดยใช้ค่าระยะห่างกำลังสองระหว่าง 2 หน่วยในประชากรกลุ่มที่ t ที่มีเวกเตอร์ของค่าสังเกต x_1 และ x_2 ตามลำดับ ดังสมการข้างล่างนี้

$$d_t^2(x_1, x_2) = (x_1 - x_2)' V_t^{-1} (x_1 - x_2)$$

โดยที่ V_t คือ เมทริกซ์ความแปรปรวนร่วมภายในกลุ่มที่ t
 x_1, x_2 คือ เวกเตอร์ของค่าสังเกตของหน่วยตัวอย่างที่ 1 และ 2

4. การประมาณค่าอัตราความผิดพลาด

4.1 วิธีบูตสเตรป (Bootstrap)

ชุดข้อมูลบูตสเตรป ($x_1^*, x_2^*, \dots, x_n^*$) ประกอบด้วยสมาชิกของข้อมูลชุดเดิม ($x_1, x_2, \dots, x_n$) บางค่าสังเกตอาจจะไม่ถูกสุ่มเลยก็ได้ บางค่าสังเกตอาจถูกสุ่มเพียงครั้งเดียว หรือบางค่าสังเกตอาจถูกสุ่มมากกว่าหนึ่งครั้งก็ได้

ขั้นตอนของบูตสเตรป สุ่มตัวอย่างบูตสเตรปที่เป็นอิสระต่อกันหลายๆ ตัวอย่าง ประเมินการสุ่มชุดตัวอย่างบูตสเตรปด้วยการคำนวณค่าสถิติ $\hat{\theta}$ ประมาณค่าความคลาดเคลื่อนมาตรฐาน (Standard Error: SE) ของค่าสถิติ $\hat{\theta}$ จากค่าเบี่ยงเบนมาตรฐานของตัวอย่าง

บูตสเตรปที่สุ่มซ้ำ เรียกเป็น ค่าประมาณบูตสเตรปของ SE เขียนแทนด้วย se_B เมื่อ B คือจำนวนตัวอย่างบูตสเตรปที่ได้ (Efron and Tibshirani, 1993)

4.2 การประมาณค่าอัตราความผิดพลาด

(Error Rate Estimation)

การประมาณค่าอัตราความผิดพลาดของวิธีการวิเคราะห์การจำแนกมีหลายวิธี ในที่นี้จะกล่าวถึง วิธีการประมาณค่าจากข้อมูล เรียกว่าเป็นการประมาณจำนวนความผิดพลาด (Error count estimates) สมมติว่ามีชุดข้อมูลทดสอบชุดหนึ่ง ทราบว่า ค่าสังเกตใดเป็นสมาชิกของประชากรใด ในการประเมินความสามารถ

ของฟังก์ชันการจำแนก โดยการหาจำนวนและร้อยละของค่าสังเกตที่มาจากประชากรกลุ่มที่ t แต่ถูกจัดให้อยู่ในประชากรกลุ่มที่ s ด้วยค่าความน่าจะเป็นแบบมีเงื่อนไข

ดังนั้น สัดส่วนของการจำแนกผิดของประชากรที่มาจากประชากรกลุ่มที่ t คือ

$$ER(t) = \sum_{s=1, s \neq t}^k \hat{P}(s/t)$$

ค่า apparent error rate คือ ค่าที่ได้จากการประมาณ $\sum_{t=1}^k \pi_t \hat{E}(R(t))$

ตารางที่ 1 จำนวนของการจัดกลุ่มที่ได้จากฟังก์ชันการจำแนกของชุดข้อมูล

กลุ่มจริง	กลุ่มที่ทำนาย			รวม
	1	2	3	
1	n_{11}	n_{12}	n_{13}	$n_{1\bullet}$
2	n_{21}	n_{22}	n_{23}	$n_{2\bullet}$
3	n_{31}	n_{32}	n_{33}	$n_{3\bullet}$
รวม	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	n_i

โดยที่ n_i คือ ขนาดตัวอย่างของชุดข้อมูลที่ $i : i = 1, \dots, 50$

ค่าอัตราความผิดพลาดของแต่ละชุดข้อมูลทดสอบ $(ER)_i$ คำนวณจาก

$$(ER)_i = \frac{n_{12} + n_{13} + n_{21} + n_{23} + n_{31} + n_{32}}{n_i}$$

อัตราความผิดพลาดเฉลี่ย $E(ER)_i$ คำนวณจากผลรวมของค่าอัตราความผิดพลาดทั้ง 50 ชุดข้อมูลหารด้วยจำนวนชุดข้อมูลทั้งหมด

$$E(ER)_i = \frac{\sum_{i=1}^{50} (ER)_i}{50}$$

5. โรคเบาหวาน

เบาหวาน คือ ภาวะร่างกายมีระดับน้ำตาลในเลือดสูงกว่าปกติ ซึ่งเป็นความผิดปกติของร่างกายที่มีการผลิตฮอร์โมนอินซูลินไม่เพียงพอ อินซูลินจะเป็นตัวพาน้ำตาลกลูโคสในเลือดเข้าไปเลี้ยงเนื้อเยื่อต่างๆ หากร่างกายขาดฮอร์โมนอินซูลินหรือผลิตฮอร์โมนไม่เพียงพอจะทำให้น้ำตาลไม่สามารถเข้าไปในเนื้อเยื่อได้จึงมีน้ำตาลในเลือดเหลือค้างอยู่มากและมีระดับที่สูงกว่าปกติ เมื่อในเลือดมีน้ำตาลสูง ไตจะทำหน้าที่กรองน้ำตาลออกมากับน้ำปัสสาวะ ทำให้ปัสสาวะมีรสหวาน จึงเรียกว่า “เบาหวาน”

ประเภทของโรคเบาหวาน สมาคมโรคเบาหวานแห่งสหรัฐอเมริกา (American Diabetes Association) ปี ค.ศ.1997 ได้จำแนกโรคเบาหวานออกเป็น 4 ประเภทดังนี้

1. โรคเบาหวานชนิดที่ 1 (Type 1 Diabetes) หรือเบาหวานชนิดพึ่งอินซูลิน โรคเบาหวานชนิดนี้เกิดจากเบต้าเซลล์ในตับอ่อนมีจำนวนน้อยหรือไม่มีเลย ทำให้ตับอ่อนไม่สามารถผลิตอินซูลินได้ ส่งผลให้ร่างกายไม่สามารถนำน้ำตาลเข้าไปเลี้ยงเนื้อเยื่อได้ ระดับน้ำตาลในเลือดจึงสูง ส่วนใหญ่จะพบในวัยเด็กหรือคนที่มีอายุไม่เกิน 40 ปี

2. โรคเบาหวานชนิดที่ 2 (Type 2 Diabetes) หรือเบาหวานชนิดไม่พึ่งอินซูลิน พบในคนที่มีอายุมากกว่า 40 ปีขึ้นไป เกิดจากการที่ตับอ่อนสร้างอินซูลินได้น้อยลง มีอาการเบาหวานเกิดขึ้นอย่างช้าๆ ไม่รุนแรง อาจรักษาด้วยการควบคุมอาหารหรือใช้ยาลดระดับน้ำตาล

3. โรคเบาหวานขณะตั้งครรภ์ โรคเบาหวานชนิดนี้ผู้ป่วยไม่เคยมีประวัติเป็นโรคเบาหวานมาก่อน โดยทั่วไปจะเกิดหลังตั้งครรภ์ได้ 20-24 สัปดาห์ หลังคลอดอาการโรคเบาหวานจะหายไปแต่มีความเสี่ยงในการเกิดโรคเบาหวานได้มาก จึงควรตรวจเช็คโรคเป็นระยะเพื่อป้องกันการเกิดโรคเบาหวาน

4. โรคเบาหวานชนิดอื่น ได้แก่ โรคเบาหวาน

ที่เกิดจากความผิดปกติทางพันธุกรรมที่ทราบชนิดชัดเจน โรคของตับอ่อน ความผิดปกติของฮอร์โมนหรือสารเคมี และอื่น ๆ

อาการของโรคเบาหวาน อาการที่พบบ่อยในผู้ป่วยเบาหวาน คือ คอแห้ง กระหายน้ำ ดื่มน้ำมาก ปัสสาวะบ่อยและมีจำนวนมาก รับประทานอาหารจุ หิวบ่อยแต่น้ำหนักตัวลดลง ตาพร่ามัว ซาปลายมือปลายเท้า คันตามผิวหนัง ถ้าเป็นผลจะหายช้ากว่าปกติ มีการติดเชื้อตามผิวหนัง

ผลการวิจัย

1. การตรวจสอบข้อตกลง

1.1 การแจกแจงแบบปกติพหุ

เมื่อนำข้อมูลตัวแปรเชิงปริมาณ 3 ตัวมา คือ อายุ ระดับน้ำตาลในเลือด และดัชนีมวลกาย มาทดสอบการแจกแจงแบบปกติพหุ โดยวิธี Mardia's Multivariate Kurtosis พบว่ามีกลุ่มไม่เสี่ยงโรคเบาหวานที่มีการแจกแจงปกติพหุ ส่วนกลุ่มเสี่ยงโรคเบาหวานและกลุ่มเป็นโรคเบาหวาน ไม่มีการแจกแจงปกติพหุ

1.2 การเท่ากันของค่าความแปรปรวนร่วมของกลุ่มเสี่ยงแต่ละกลุ่ม

การเท่ากันของค่าความแปรปรวนร่วมของกลุ่มเสี่ยงแต่ละกลุ่ม ผลการวิเคราะห์โดยสถิติ Box'M สรุปได้ว่าข้อมูลทั้ง 3 กลุ่มมีเมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระไม่เท่ากันอย่างมีนัยสำคัญทางสถิติ

ผลการตรวจสอบข้อตกลงการแจกแจงปกติพหุ และการเท่ากันของความแปรปรวนร่วมสำหรับการวิเคราะห์ที่อิงพารามิเตอร์ ปรากฏว่า ข้อมูลที่นำมาศึกษาไม่คุณสมบัติไม่เป็นไปตามข้อตกลง ดังกล่าว

2. ผลการสุ่มตัวอย่างซ้ำทั้งหมดจากข้อมูลจริง (ข้อมูลบุตสเตรป)

จากการสุ่มตัวอย่างซ้ำแบบใส่คืนจากข้อมูลจริง ขึ้นมาจำนวน 50 ชุดข้อมูลและแต่ละชุดมีจำนวน

ข้อมูล 599 ค่าสังเกตพบว่า ข้อมูลบัพเตสเตรปแต่ละชุดที่สุ่มได้มีข้อมูลในกลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน ในอัตราส่วน 0.90 : 0.05 : 0.05

3. ค่าอัตราความผิดพลาดในการจำแนกกลุ่มผู้ป่วยโรคเบาหวาน

ในการจำแนกผู้ป่วยโรคเบาหวาน ใช้การสุ่มตัวอย่างซ้ำทั้งหมดโดยวิธีบัพเตสเตรป จำนวน 50 ชุด ข้อมูล โดยข้อมูลในแต่ละชุดมีจำนวน 599 ค่าสังเกต โดยแบ่งเป็นชุดข้อมูลสำหรับฝึกสอนจำนวน 2/3 ของข้อมูลทั้งหมดและเป็นชุดข้อมูลสำหรับทดสอบจำนวน 1/3 ของข้อมูลทั้งหมด โดยการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ ใช้สมการการจำแนกด้วยฟังก์ชันเส้นและฟังก์ชันการจำแนกกำลังสอง และการวิเคราะห์

การจำแนกที่อิงพารามิเตอร์ ด้วยวิธี kNN และกำหนดค่าความน่าจะเป็นก่อนหน้าจากค่าสัดส่วนแบบต่าง ๆ

เป็นการวิเคราะห์การจำแนกของกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่ม ซึ่งข้อมูลกลุ่มเสี่ยงโรคเบาหวานทั้งสามกลุ่มพบว่า กลุ่มไม่เสี่ยงโรคเบาหวานเพียงกลุ่มเดียวที่ไม่มีการแจกแจงปกติพิพ และการเท่ากันของความแปรปรวนร่วมทั้ง 3 กลุ่มไม่เท่ากัน ข้อมูลมีสัดส่วนของกลุ่มไม่เสี่ยง

โรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวาน เท่ากับ 0.90 : 0.05 : 0.05

การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ ด้วยวิธี kNN จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มที่ต่ำที่สุดเมื่อเปรียบเทียบกับการวิเคราะห์การจำแนกแบบอื่น ส่วนการวิเคราะห์การจำแนกที่อิง

ตารางที่ 2 ค่าเฉลี่ยและความคลาดเคลื่อนมาตรฐานของอัตราความผิดพลาด (Error Rate) ของกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่ม จากชุดข้อมูลทดสอบของตัวอย่างบัพเตสเตรป 50 ชุด

ค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน)				
วิธีการวิเคราะห์การจำแนก	สัดส่วน (0.33 : 0.33 : 0.33)	สัดส่วน (0.90 : 0.05 : 0.05)	สัดส่วน (0.80 : 0.10 : 0.10)	สัดส่วน (0.70 : 0.15 : 0.15)
ที่อิงพารามิเตอร์				
ฟังก์ชันการจำแนกเชิงเส้น	0.15 (0.15)	0.13 (0.11)	0.14 (0.12)	0.14 (0.16)
ฟังก์ชันการจำแนกกำลังสอง	0.06 (0.08)	0.03 (0.10)	0.04 (0.10)	0.05 (0.11)
ที่ไม่อิงพารามิเตอร์				
3 NN	0.05 (0.07)	0.02 (0.02)	0.02 (0.02)	0.03 (0.03)
4 NN	0.06 (0.08)	0.02 (0.02)	0.02 (0.03)	0.03 (0.04)
5 NN	0.07 (0.08)	0.03 (0.06)	0.05 (0.04)	0.05 (0.04)

พารามิเตอร์ด้วยฟังก์ชันการจำแนกเชิงเส้นจะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มมากที่สุด การกำหนดค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวาน) จะถูกกำหนดเป็นแบบสัดส่วนต่างกัน พบว่า การกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนด้วยสัดส่วน (0.90 : 0.05 : 0.05) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกต่ำที่สุด

สำหรับ $k = 3$ หรือ 4 และการกำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วน ที่เท่ากับ (0.90 : 0.05 : 0.05) ซึ่งใกล้เคียงกับสัดส่วนของข้อมูล จะให้ค่าเฉลี่ยอัตราความผิดพลาดที่มีการกระจายน้อยที่สุด

4. ค่าอัตราความผิดพลาดในการจำแนกกลุ่มเสี่ยงโรคเบาหวานที่ได้จากข้อมูลจริง

ตารางที่ 3 เป็นการวิเคราะห์การจำแนกของ

ข้อมูลจริงในกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่มพบว่า กลุ่มไม่เสี่ยงโรคเบาหวานเพียงกลุ่มเดียวที่ไม่มีการแจกแจงปกติพหุ และการเท่ากันของความแปรปรวนร่วมทั้ง 3 กลุ่มไม่เท่ากัน ข้อมูลมีสัดส่วนของกลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวานเท่ากับ 0.90 : 0.05 : 0.05

การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกาลังสองจะให้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด การกำหนดค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวาน) จะถูกกำหนดเป็นแบบสัดส่วนต่างกัน พบว่า การกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนด้วยสัดส่วน (0.90 : 0.05 : 0.05) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกต่ำที่สุด

ตารางที่ 3 แสดงค่าอัตราความผิดพลาด (Error Rate) ของข้อมูลจริงกับข้อมูลที่ได้จากการสุ่ม

	ค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน)			
	สัดส่วน	สัดส่วน	สัดส่วน	สัดส่วน
วิธีการวิเคราะห์การจำแนก	(0.33 : 0.33 : 0.33)	(0.90 : 0.05 : 0.05)	(0.80 : 0.10 : 0.10)	(0.70 : 0.15 : 0.15)
ที่อิงพารามิเตอร์				
ฟังก์ชันการจำแนกเชิงเส้น	0.203	0.061	0.107	0.164
ฟังก์ชันการจำแนกกาลังสอง	0.117	0.029	0.043	0.061
ที่ไม่อิงพารามิเตอร์				
3 NN	0.172	0.041	0.062	0.086
4 NN	0.231	0.038	0.088	0.118
5 NN	0.280	0.040	0.089	0.143

อภิปรายและสรุปผลการวิจัย

1. ผลจากตัวอย่างชุดสแตมป์

1.1 ค่าเฉลี่ยอัตราความผิดพลาดของกลุ่มที่ไม่เสี่ยงโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN และกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนเป็นสัดส่วน (0.90 : 0.05 : 0.05) ที่ใกล้เคียงกับข้อมูลจริงจะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด

1.2 ค่าเฉลี่ยอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN และกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนเป็นสัดส่วนที่เท่ากัน คือ (0.3333 : 0.3333 : 0.3333) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด

1.3 ค่าเฉลี่ยอัตราความผิดพลาดของกลุ่มที่เป็นโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN และกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนเป็นสัดส่วนที่เท่ากัน คือ (0.3333 : 0.3333 : 0.3333) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด

1.4 ค่าอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่ม

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN และกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนเป็นสัดส่วน (0.90 : 0.05 : 0.05) ที่ใกล้เคียงกับข้อมูลจริงจะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด

2. ผลจากการศึกษาด้วยข้อมูลจริง

2.1 ค่าอัตราความผิดพลาดของกลุ่มไม่เสี่ยงโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN และกำหนดค่าความน่าจะเป็นที่

เกิดขึ้นก่อนเป็นสัดส่วน (0.90 : 0.05 : 0.05) ที่ใกล้เคียงกับข้อมูลจริงจะให้อัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด

2.2 ค่าอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง จะให้อัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด แต่การกำหนดค่าความน่าจะเป็นก่อนหน้าไม่มีผลต่ออัตราความผิดพลาดในการจำแนกกลุ่ม

2.3 ค่าอัตราความผิดพลาดของกลุ่มเป็นโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN จะให้อัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด และการกำหนดค่าความน่าจะเป็นก่อนหน้าไม่มีผลต่ออัตราความผิดพลาดในการจำแนกกลุ่ม

2.4 ค่าอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่ม

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง และกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนด้วยสัดส่วน (0.90 : 0.05 : 0.05) ที่ใกล้เคียงกับข้อมูลจริงจะให้อัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด

ในการเปรียบเทียบประสิทธิภาพของการวิเคราะห์การจำแนกระหว่างวิธีที่อิงพารามิเตอร์ โดยใช้ฟังก์ชันการจำแนกเชิงเส้น และฟังก์ชันการจำแนกกำลังสอง กับวิธีที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อมีความไม่สมดุลของข้อมูลในแต่ละกลุ่ม ในการจำแนกกลุ่มเสี่ยงโรคเบาหวาน โดยพิจารณาจากค่าอัตราความผิดพลาดที่ต่ำที่สุดพบว่า การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN โดยกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนด้วยสัดส่วนที่ใกล้เคียงกับสัดส่วนจริงของข้อมูลจะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด

การศึกษาจากข้อมูลจริงเพียงชุดเดียว ไม่สามารถบอกได้ถึงระดับความน่าเชื่อถือของอัตราความผิดพลาดในการจำแนกกลุ่ม แต่การศึกษาอัตราความผิดพลาดในการจำแนกกลุ่มด้วยตัวอย่างบูตสเตรป ซึ่งเป็นการสุ่มตัวอย่างซ้ำจากข้อมูลชุดเดิมมา 50 ชุด จะให้ผลที่ได้จากการจำแนกมีความน่าเชื่อถือ

ข้อเสนอแนะ

1. ด้านการนำไปใช้ประโยชน์

ผลที่ได้จากการศึกษานี้ มีข้อเสนอแนะว่า เมื่อข้อมูลที่นำมาวิเคราะห์การจำแนกมีความไม่สมดุลในแต่ละกลุ่มและข้อมูลไม่เป็นไปตามข้อตกลงของการแจกแจงแบบปกติพหุและความแปรปรวนร่วมไม่เท่ากัน ควรใช้การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ จะให้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มที่มีค่าต่ำกว่าเมื่อเปรียบเทียบกับวิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์

และในการกำหนดค่าความน่าจะเป็นก่อนหน้า (Priors Probability) สำหรับการสร้างกฎการจำแนกกลุ่ม ควรกำหนดให้มีอัตราส่วนที่ใกล้เคียงกับอัตราส่วนของข้อมูลจริง จะได้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มน้อยที่สุด ซึ่งผู้ที่สนใจสามารถประยุกต์ใช้กับข้อมูลลักษณะอื่น ที่มี ความแตกต่างระหว่างกลุ่ม อาจจะมี 3 กลุ่มหรือมากกว่า 3 กลุ่มก็ได้ ที่มีข้อมูลแต่ละกลุ่มแตกต่างกันมาก ทำการวิเคราะห์การจำแนกตั้งแต่

ตรวจสอบข้อตกลงและวิเคราะห์การจำแนกเปรียบเทียบระหว่างวิธีอิงพารามิเตอร์และวิธีที่ไม่อิงพารามิเตอร์ โดยเปรียบเทียบอัตราความผิดพลาดของการจำแนกผิด ซึ่งที่กล่าวมาทั้งหมดนี้ ผู้สนใจสามารถใช้โปรแกรมสำเร็จรูปทางสถิติทำได้สะดวก

2. ด้านการศึกษาและวิจัยต่อไป

2.1 ในการเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การจำแนกระหว่างวิธีที่อิงพารามิเตอร์และวิธีที่ไม่อิงพารามิเตอร์ กรณีที่ข้อมูลแต่ละกลุ่มมีความไม่สมดุลกัน ผู้สนใจศึกษาทำวิจัยต่อ สามารถใช้วิธีการจำลองข้อมูล โดยคำนึงถึงประเด็น ต่อไปนี้

- ความแตกต่างของขนาดตัวอย่างในประชากร k กลุ่ม

- การแจกแจงแบบปกติพหุ
- การเท่ากันของความแปรปรวนร่วมของประชากรแต่ละกลุ่ม

- ความหลากหลายของการกำหนดค่าความน่าจะเป็นก่อนหน้า สำหรับจัดหน่วยตัวอย่างใหม่เข้ากลุ่ม

2.2 ผู้ที่สนใจสามารถทำการขยายการศึกษาต่อจากการศึกษานี้ โดยเพิ่มวิธีการวิเคราะห์การจำแนกโดยใช้วิธีการทางเหมืองข้อมูล เช่น ต้นไม้การจำแนก (Classification Tree) หรือ Support Vector Machine หรือวิธี Fuzzy



บรรณานุกรม

- กัลยา วานิชย์บัญชา. (2548). **การวิเคราะห์ข้อมูลหลายตัวแปร**. กรุงเทพฯ: ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี, จุฬาลงกรณ์มหาวิทยาลัย.
- ทรงพล ต่อนี้. (2541). **การศึกษาเปรียบเทียบสถิติวิเคราะห์การจำแนกแบบพาราเมตริกและนอนพาราเมตริกประยุกต์สำหรับข้อมูลการเกิดภาวะแทรกซ้อนที่จอประสาทตา ในผู้ป่วยเบาหวานชนิดไม่พึ่งอินซูลิน**. วิทยานิพนธ์มหาบัณฑิต มหาวิทยาลัยมหิดล.
- วิชัย เอกพลากร. (2548). **การศึกษาพัฒนาดัชนีความเสี่ยงต่อเบาหวาน**. (ออนไลน์). สืบค้นวันที่ 14 พฤษภาคม 2555 จาก <http://www.hiso.or.th/hiso/analystReport/picture/8.pdf>
- สุภลาภ พวงสอาด. (2541). **การศึกษาการใช้สมการจำแนกกลุ่มในงานสำรวจสถานะสุขภาพเรื่องเบาหวานของประชาชนไทยในเขตอำเภอบ้านโป่ง จังหวัดราชบุรี**. วิทยานิพนธ์มหาบัณฑิต, มหาวิทยาลัยมหิดล.
- Barbara, G. and S. Linda. (2007). **Using Multivariate Statistics**. 5th ed. Pearson Education, Inc., Boston.
- Efron, B. (1975). The Efficiency of logistic regression compared to normal discriminant analysis. **J. Amer. Statist. Ass.** 70: 892 – 898.
- Efron, B And Tibshirani. (1993). **An Introduction to the Bootstrap**. Chapman & Hall, Inc, USA.
- Ferrer, Arce J. And Lin (1999). Comparing The Classification Accuracy Among Nonparametric, Parametric Discriminant Analysis And Logistic Regression Methods **In Proceedings of the 1999 Annual meeting of the American Educational Research Association**, April 13-17, Montreal, Canada.
- Hair, Joseph F. et al. (2010). **Multivariate data analysis**. (7th ed). New Jersey: Pearson Prentice Hall.
- Khattree and Naik N. (2000). **Multivariate Data Reduction And Discrimination**. John Wiley, Inc., USA.
- Sultana F. (2000). **Discriminant analysis and logistic discrimination: an application to diabetes mellitus data**. PhD. thesis in biostatistics. Institute of Statistical Research and Trainings, University of Dhaka.