ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (สถิติ)

ปริญญา

สถิติ | สถิติ

สาขา | **ภาควิชา**

เรื่อง การวิเคราะห์การจำแนกข้อมูลที่ไม่สมดุลในแต่ละกลุ่ม กรณีศึกษากลุ่มเสี่ยงโรคเบาหวาน

Discriminant Analysis for Class - Imbalanced Data : A Case Study of Diabetes Risk Group

นามผู้วิจัย นางสาวทัศนีย์ น้ำเจริญ

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(ผู้ช่วยศาสตราจารย์ลี่ลี อิงศรีสว่าง, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(ผู้ช่วยศาสตราจารย์ปราโมทย์ ลือนาม, Ph.D.)

หัวหน้าภาควิชา .....

(รองศาสตราจารย์ประสิทธิ์ พยัคฆพงษ์, M.S.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

การวิเคราะห์การจำแนกข้อมูลที่ไม่สมดุลในแต่ละกลุ่ม กรณีศึกษากลุ่มเสี่ยงโรคเบาหวาน

Discriminant Analysis for Class - Imbalanced Data : A Case Study of Diabetes Risk Group

โดย

นางสาวทัศนีย์ น้ำเจริญ

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ)
พ.ศ. 2557

ทัศนีย์ น้าเจริญ 2557: การวิเคราะห์การจำแนกข้อมูลที่ไม่สมดุลในแต่ละกลุ่ม กรณีศึกษากลุ่มเสี่ยงโรคเบาหวาน ปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ) สาขาสถิติ ภาควิชาสถิติ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์ลี่ลี อิงศรีสว่าง, Ph.D. 99 หน้า

การศึกษานี้เป็นการเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์และการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ในการจำแนกข้อมูล 3 กลุ่ม โดยที่มีความไม่สมดุลของข้อมูลแต่ละกลุ่ม ข้อมูลที่นำมาศึกษาเป็นข้อมูลกลุ่มเสี่ยงโรคเบาหวานของเจ้าหน้าที่ที่ปฏิบัติงานในโรงพยาบาลรัฐแห่งหนึ่งจำนวน 599 คน มีการวินิจฉัยเป็น 3 กลุ่ม คือ กลุ่มไม่เสี่ยงโรคเบาหวาน มีประมาณ 90% กลุ่มเสี่ยงโรคเบาหวาน มีประมาณ 5% และกลุ่มเป็นโรคเบาหวาน มีประมาณ 5% และมีตัวแปรอิสระที่นำมาพิจารณาประกอบด้วย อายุ, เพศ, ระดับน้ำตาลในเลือดและดัชนีมวลกาย ซึ่งข้อมูลดังกล่าวไม่มีการแจกแจงแบบปกติพหุและความแปรปรวนร่วมแต่ละกลุ่มไม่เท่ากัน วิธีการศึกษาทำการบูตสแตรปข้อมูลกลุ่มเสี่ยงโรคเบาหวานจำนวน 50 ชุด ชุดละ 599 ค่าสังเกต แต่ละชุดจะทำการการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกเชิงเส้นและฟังก์ชันการจำแนกกำลังสอง และวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธีเคเนียเรสเนเบอร์ เมื่อกำหนด k เป็น 3, 4, 5 พร้อมทั้งกำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วนที่เท่ากันและสัดส่วนที่แตกต่างกันสำหรับสร้างกฎการจำแนกกลุ่มที่เหมาะสม และวัดประสิทธิภาพของวิธีการวิเคราะห์การจำแนกแต่ละวิธีด้วยค่าเฉลี่ยอัตราความผิดพลาด ผลการศึกษาพบว่าจากข้อมูลบูตสแตรป 50 ชุด การวิเคราะห์การจำแนกด้วยวิธีเคเนียเรสเนเบอร์ เมื่อ k=3 หรือ 4 และเมื่อกำหนดค่าความน่าจะเป็นก่อนหน้าของกลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน ด้วยสัดส่วนที่เหมือนกับข้อมูลจริง คือ 0.90 : 0.05 : 0.05 หรือ 0.80 : 0.10 : 0.10 จะได้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มผิดมีค่าต่ำที่สุด

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Tassanee Nacharoen 2014: Discriminant Analysis for Class - Imbalanced Data : A Case Study of Diabetes Risk Group. Master of Science (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Assistant Professor Lily Ingsrisawang, Ph.D. 99 pages.

This study aimed to compare the classification performance between parametric discriminant analysis and nonparametric discriminant analysis in a three-group classification application of class-imbalanced data. Data from a healthy project for 599 staffs in a government hospital in Bangkok were obtained for the classification problem. The staffs were diagnosed into one of three diabetes risk groups: non-risk (90%), risk (5%), and diabetic (5%). The original data including these variables of diabetes risk group, age, gender, cholesterol, and BMI were analyzed and bootstrapped up to 50 samples, 599 observations per sample, for additional estimation of misclassification error rate. The 50 bootstrap samples consisted of two patterns of three diabetes risk groups of non-risk: risk: diabetic as 90%: 5%: 5% and 80%: 10%: 10%. Each data set was explored for the departure of multivariate normality and the equality of covariance matrices of the three risk groups. Both the original data and the bootstrap samples show nonnormality and unequal covariance matrices. The parametric linear discriminant function, quadratic discriminant function, and the nonparametric k-nearest neighbors' discriminant function were performed over 50 bootstrap samples and applied to the original data. In finding the optimal classification rule, the choices of prior probabilities were set up for both equal proportions and unequal proportions. The results indicated that the k-nearest neighbors approach when k = 3 or k = 4 and the prior probabilities of non-risk: risk: diabetic as 0.90: 0.05:0.05 or 0.80: 0.10: 0.10 gave the smallest error rate of misclassification.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ โดยคำแนะนำ การสนับสนุนให้ข้อคิดเห็นและแก้ไขข้อบกพร่องต่างๆ จาก ผู้ช่วยศาสตราจารย์ ดร.ลี่ลี อิงศรีสว่าง อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ผู้ช่วยศาสตราจารย์ ดร.ปราโมทย์ ลือนาม อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่กรุณาให้คำปรึกษาแนะนำ และให้การช่วยเหลือทำให้วิทยานิพนธ์สำเร็จลุล่วงไปด้วยดี และขอขอบคุณโรงพยาบาลนพรัตน์ราชธานีที่ได้อนุเคราะห์ข้อมูลเพื่อประโยชน์ต่อการทำวิทยานิพนธ์ครั้งนี้

ข้าพเจ้า ขอน้อมรำลึกถึงพระคุณของบิดา มารดา และครอบครัวที่ส่งเสริมสนับสนุนการศึกษาตลอดมา ประโยชน์อันใดที่เกิดเนื่องมาจากวิทยานิพนธ์เล่มนี้ ขอมอบแด่บิดา มารดา และคณาจารย์ทุกท่านที่ได้เมตตา อบรมสั่งสอน ประสิทธิ์ประสาทวิชาความรู้ ขอขอบคุณพี่ และเพื่อนๆ ทุกๆคนที่เป็นกำลังใจ และช่วยเหลือมาโดยตลอดไว้ ณ โอกาสนี้ด้วย

ทัศนีย์ น้ำเจริญ

มิถุนายน 2557

สารบัญ

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญภาพ

การวิเคราะห์การจำแนกข้อมูลที่ไม่สมดุลในแต่ละกลุ่ม กรณีศึกษากลุ่มเสี่ยงโรคเบาหวาน

Discriminant Analysis for Class - Imbalanced Data : A Case Study of Diabetes Risk Group

คำนำ

การวิเคราะห์การจำแนก (Discriminant Analysis) จัดเป็นเทคนิคหนึ่งของการวิเคราะห์พหุตัวแปร (multivariate analysis) ใช้สำหรับแบ่งกลุ่มของสิ่งที่สนใจศึกษา (objects) หรือค่าสังเกต (observations) เป็น กลุ่มย่อย หรือเรียกว่า คลาส (class) จำนวน k กลุ่ม ($k \geq 2$) โดยพิจารณาว่าตัวแปรใดบ้างที่ทำให้กลุ่มหรือคลาสต่างกัน 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

ในทางปฏิบัติข้อมูลที่นำมาศึกษาอาจมีคุณสมบัติไม่เป็นไปตามข้อตกลง อาจส่งผลให้การจำแนกกลุ่มเกิดความผิดพลาดได้ ในทางทฤษฎีมีทางเลือกอื่นสำหรับการวิเคราะห์การจำแนก เช่น 1) วิธีการวิเคราะห์การจำแนกด้วยตัวแบบการถดถอยโลจิสติก(logistic regression model) สำหรับข้อมูลที่มีความสัมพันธ์แบบไม่เชิงเส้น (nonlinear) และมีตัวแปรอิสระบางตัวเป็นตัวแปรกลุ่ม และ 2) วิธีการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ (nonparametric discriminant analysis) สำหรับข้อมูลไม่ได้มีการแจกแจงปกติพหุและความแปรปรวนร่วมไม่เท่ากัน แต่ตัวแปรอิสระทุกตัวเป็นตัวแปรเชิงปริมาณ เช่น วิธีเคอร์แนล (kernel method) หรือวิธี k-Nearest Neighbour: kNN แต่วิธี kNN เป็นที่นิยมมากกว่าวิธีเคอร์แนล (Khattree and Naik, 2000)

จะเห็นว่ามีการนำวิธีการวิเคราะห์การจำแนกมาใช้กับงานด้านต่างๆ เช่นงานวิจัยทางการแพทย์ จำแนกโรคจากลักษณะอาการที่ปรากฏของผู้ป่วย หรืองานวิจัยทางธรณีวิทยา ต้องการจำแนกประเภทหินจากองค์ประกอบทางเคมีของหิน เป็นต้น ตัวอย่าง Knoke (1982) ได้ศึกษาเปรียบเทียบเทคนิคการจำแนกสำหรับตัวแปรไม่ต่อเนื่องและตัวแปรต่อเนื่องโดยวิธี Fisher's linear discriminant analysis, quadratic discriminant analysis, logistic regression และ augmented linear discriminant analysis ใช้กับข้อมูลที่มีตัวแปรต่อเนื่อง 1 ตัวมีการแจกแจงแบบปกติและตัวแปรไม่ต่อเนื่อง 2 ตัว ผลการศึกษาพบว่า เมื่อไม่มีความสัมพันธ์ระหว่างตัวแปรอิสระ การวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกเชิงเส้นเป็นวิธีที่มีความเสถียรและมีประสิทธิภาพในการจำแนกได้ดี ทรงพล (2541) ศึกษาเปรียบเทียบสถิติวิเคราะห์การจำแนกที่อิงพารามิเตอร์และที่ไม่อิงพารามิเตอร์ โดยศึกษาปัจจัยที่มีอิทธิพลต่อการจำแนกผู้ป่วยเบาหวานชนิดไม่พึ่งอินซูลินที่มีภาวะแทรกซ้อนและไม่แทรกซ้อนที่จอประสาทตา ผลการศึกษาพบว่า เมื่อตัวแปรจำแนกประเภท มีการกระจายแบบปกติ มีการกระจายแบบไม่ปกติ และเมื่อมีการปรับให้มีการกระจายแบบปกติ หรือเป็นตัวแปรผสมแล้ว วิธี kernel density estimation จะให้อัตราความถูกต้องในการจำแนกสูงกว่าวิธีอื่น แต่พบว่า เมื่อตัวแปรจำแนกประเภทเป็นตัวแปรทวิแล้ว วิธี kNN จะให้อัตราความถูกต้องในการจำแนกสูงกว่าวิธีอื่น

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

การศึกษานี้จึงสนใจที่จะเปรียบเทียบประสิทธิภาพของการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ และการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ กับปัญหาข้อมูลกลุ่มเสี่ยงโรคเบาหวานของเจ้าหน้าที่ที่ปฏิบัติงานในโรงพยาบาลรัฐแห่งหนึ่ง ซึ่งมีการวินิจฉัยเป็น 3 กลุ่ม คือ กลุ่มไม่เสี่ยงโรคเบาหวาน กลุ่มเสี่ยงโรคเบาหวาน และกลุ่มที่เป็นโรคเบาหวาน ซึ่งข้อมูลแต่ละกลุ่มมีความไม่สมดุลกัน คือ มีจำนวนข้อมูลในแต่ละกลุ่มแตกต่างกันมาก คือ 90% จัดอยู่ในกลุ่มไม่เสี่ยงโรคเบาหวาน 5% จัดอยู่ในกลุ่มเสี่ยงโรคเบาหวาน และอีก 5% จัดอยู่ในกลุ่มที่เป็นโรคเบาหวาน สำหรับตัวแปรอิสระที่นำมาพิจารณาในการจำแนกกลุ่มเสี่ยง ประกอบด้วย อายุ เพศ ระดับน้ำตาลในเลือดและค่าดัชนีมวลกาย โดยประมาณค่าอัตราความผิดพลาดของการจำแนกด้วยวิธีบูตสแตรป (การสุ่มตัวอย่างซ้ำแบบใส่คืน จากข้อมูลกลุ่มเสี่ยงโรคเบาหวาน) จำนวน 50 ชุดแต่ละชุดมีขนาดตัวอย่าง 599 ค่าสังเกต และเปรียบเทียบอัตราความผิดพลาดเฉลี่ยที่ได้จากข้อมูล 50 ชุด ระหว่างวิธีวิเคราะห์การจำแนกที่อิงพารามิเตอร์ (ฟังก์ชันการจำแนกเชิงเส้นและฟังก์ชันการจำแนกกำลังสอง) และการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ (วิธี kNN) พร้อมทั้งเปรียบเทียบอัตราความผิดพลาดที่วิเคราะห์ได้จากข้อมูลจริง

วัตถุประสงค์

1. เพื่อเปรียบเทียบประสิทธิภาพของความถูกต้องในการจำแนกข้อมูล 3 กลุ่มระหว่างการวิเคราะห์การจำแนกที่อิงพารามิเตอร์และที่ไม่อิงพารามิเตอร์ เมื่อมีความไม่สมดุลของข้อมูลในแต่ละกลุ่ม

2. เพื่อประยุกต์วิธีการวิเคราะห์การจำแนกที่เหมาะสมกับการจำแนกกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่ม กรณีที่มีความไม่สมดุลของข้อมูลในแต่ละกลุ่ม

ขอบเขตของการวิจัย

1. ข้อมูลที่นำมาศึกษาครั้งนี้เป็นข้อมูลการตรวจสุขภาพประจำปีของเจ้าหน้าที่ที่ปฏิบัติงานในโรงพยาบาลนพรัตน์ราชธานี ระหว่างเดือนสิงหาคม 2551 - เมษายน 2552 จำนวน 599 คน โดยแพทย์ได้วินิจฉัยกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่มดังนี้

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

2. ตัวแปรที่ศึกษา คือ

2.1 ตัวแปรตาม คือ ตัวแปรกลุ่ม แบ่งเป็น 3 กลุ่มคือ กลุ่มไม่เสี่ยงโรคเบาหวาน กลุ่มเสี่ยงโรคเบาหวาน และกลุ่มที่เป็นโรคเบาหวาน

2.2 ตัวแปรอิสระ ที่นำมาศึกษามีดังนี้

2.2.1 อายุ (Age) หน่วยเป็น ปี

2.2.2 เพศ (Sex) (กำหนดให้ 0 = เพศชาย และ 1 = เพศหญิง)

2.2.3 ระดับน้ำตาลในเลือด (Blood Glucose)หน่วยเป็น มิลลิกรัม/เดซิลิตร (mg/dl)

2.2.4 ค่าดัชนีมวลกาย (Bmi) หน่วยเป็น กิโลกรัม/เมตร^2 (kg/m^2)

3. การวิเคราะห์การจำแนก

3.1 การวิเคราะห์การจำแนกที่อิงพารามิเตอร์

3.1.1 การใช้ฟังก์ชันการจำแนกเชิงเส้น (Linear Discriminant Function : LDF)

3.1.2 การใช้ฟังก์ชันการจำแนกกำลังสอง (Quadratic Discriminant Function : QDF)

3.2 การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธีเคเนียเรสเนเบอร์ (kNN) เมื่อกำหนด k เป็น 3, 4 และ 5 โดยที่ k คือจำนวนค่าสังเกตที่อยู่ล้อมรอบค่าสังเกต $\mathbf{x}:(x_1,x_2,x_3,x_4)$ ที่ต้องการให้จัดเข้ากลุ่มเสี่ยงกลุ่มใดกลุ่มหนึ่ง

3.3 กำหนดกฎการจำแนกที่เหมาะสม โดยพิจารณาจากค่าความน่าจะเป็นก่อนหน้า (prior probability) ของการจัดจำแนกหน่วยตัวอย่างหรือค่าสังเกตเข้ากลุ่มที่ 1 (กลุ่มไม่เสี่ยงโรคเบาหวาน) : กลุ่มที่ 2 (กลุ่มเสี่ยงโรคเบาหวาน): กลุ่มที่ 3 (กลุ่มที่เป็นโรคเบาหวาน) ตามลำดับดังต่อไปนี้

3.3.1 จัดสรรตามสัดส่วนที่เท่ากัน คือ 0.3333 : 0.3333 : 0.3333

3.3.2 จัดสรรตามสัดส่วน 0.90 : 0.05 : 0.05

3.3.3 จัดสรรตามสัดส่วน 0.80 : 0.10 : 0.10

3.3.4 จัดสรรตามสัดส่วน 0.70 : 0.15 : 0.15

4. ประมาณค่าความผิดพลาดของการจำแนกด้วยวิธีการวิเคราะห์การจำแนกตามข้อ 3.1 และ ข้อ 3.2 และมีการกำหนดค่าความน่าจะเป็นก่อนหน้าตามที่ระบุในข้อ 3.3 จากตัวอย่างบูตสแตรปจำนวน 50 ชุด ชุดละ 599 ค่าสังเกต และจากตัวอย่างข้อมูลจริงในข้อ 1

5. การวิจัยนี้ใช้โปรแกรม Statistical Analysis System (SAS) Version 9.1 ในการวิเคราะห์ข้อมูล

ประโยชน์ที่คาดว่าจะได้รับ

1. เรียนรู้วิธีการวิเคราะห์การจำแนกที่เหมาะสม เมื่อข้อมูลไม่เป็นไปตามข้อตกลงของการแจกแจงแบบปกติพหุ และการเท่ากันของความแปรปรวนร่วม ในกรณีที่ข้อมูลมีความไม่สมดุลในแต่ละกลุ่ม

2. การกำหนดค่าความน่าจะเป็นก่อนหน้ามีอิทธิพลต่อการจำแนกกลุ่มในกลุ่มที่มีขนาดตัวอย่างน้อย และในกลุ่มที่มีขนาดตัวอย่างใหญ่หรือไม่ สำหรับการสร้างกฎการจำแนกกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่ม

การตรวจเอกสาร

การตรวจเอกสารแบ่งออกเป็น 2 ส่วน ส่วนแรกเป็นวิธีการทางสถิติที่ใช้ในการศึกษาวิจัย และส่วนที่สองเป็นผลงานวิจัยที่เกี่ยวข้อง

วิธีการทางสถิติที่ใช้ในการศึกษาวิจัย

1. การวิเคราะห์การจำแนก (Discriminant Analysis)

กัลยา (2552) ได้กล่าวถึงวัตถุประสงค์ของการวิเคราะห์การจำแนกว่า เพื่อศึกษาว่าตัวแปรอิสระตัวใดบ้างเป็นตัวแปรที่ทำให้กลุ่มต่างกัน โดยการสร้างสมการการจำแนกกลุ่มและเพื่อพยากรณ์หน่วยใหม่ที่ยังไม่ทราบกลุ่มว่าควรจะอยู่กลุ่มใด โดยที่ตัวแปร ตามเป็นตัวแปรเชิงกลุ่มตั้งแต่ 2 กลุ่มขึ้นไป และตัวแปรอิสระเป็นตัวแปรเชิงปริมาณหรือเป็นตัวแปรเชิงกลุ่มรวมอยู่ด้วย

Hair *et al.* (2010) ได้กล่าวถึงข้อตกลงเบื้องต้นของการวิเคราะห์จำแนก ที่อิงพารามิเตอร์ ดังนี้

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

1.1 การตรวจสอบข้อตกลงของการวิเคราะห์การจำแนก

การวิเคราะห์การจำแนก ควรตรวจสอบในเบื้องต้นว่า ตัวแปรอิสระแต่ละตัวเป็นอิสระกันหรือไม่ ถูกสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติพหุ และเมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระแต่ละกลุ่มเท่ากันหรือไม่

1.1.1 การตรวจสอบตัวแปรอิสระแต่ละตัวเป็นอิสระต่อกัน

สำหรับตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณ พิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน โดยทำเป็นเมทริกซ์ความสัมพันธ์

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \end{array}$$

ถ้า $r_{ij}>0.8$ ขึ้นไป แสดงว่าตัวแปร x_i และ x_j มีความสัมพันธ์กันสูงแบบเชิงเส้น

1.1.2 การตรวจสอบการแจกแจงปกติพหุของตัวแปรอิสระ

การตรวจสอบการแจกแจงแบบปกติของตัวแปรเดียว (univariate normality) สามารถประเมินได้จากการพล็อตกราฟฮิสโตแกรม ส่วนการตรวจสอบการแจกแจงปกติร่วม 2 ตัวแปร (bivariate normality) สามารถประเมินได้จาก Q-Q plots หรือใช้ค่า Mahalanobis distance แต่การตรวจสอบการแจกแจงปกติพหุ สามารถใช้การทดสอบของ Mardia's multivariate skewness and kurtosis (Khattree and Naik, 2000)

กัลยา (2552) ได้กล่าวถึง การทดสอบของ Mardia's multivariate skewness and kurtosis ไว้ดังนี้

ถ้า **x** และ **y** เป็นเวกเตอร์ของตัวแปรสุ่มที่เป็นอิสระกันและมีการแจกแจงเหมือนกัน โดยมีเวกเตอร์ค่าเฉลี่ย **μ** และเมทริกซ์ค่าความแปรปรวนร่วม $\sum$

ค่าความเบ้ของประชากรหลายตัวแปร $= \beta_{1,p} = E\left[(\mathbf{y}-\boldsymbol{\mu})'\sum^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^3$

ค่าความโด่งของประชากรหลายตัวแปร $= \beta_{2,p} = E\left[(\mathbf{y}-\boldsymbol{\mu})'\sum^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^2$

กรณีที่ประชากรมีการแจกแจงแบบปกติพหุจะทำให้ $\beta_{1,p} = 0$ และ $\beta_{2,p} = p(p+2)$

สำหรับตัวอย่างขนาด n สามารถประมาณค่า $\beta_{1,p}$ และ $\beta_{2,p}$ ได้ดังนี้

$$b_{1,p} = \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^3\right) \quad \text{และ} \quad b_{2,p} = \frac{1}{n}\left(\sum_{i=1}^{n} g_{ii}^2\right)$$

โดยที่ $g_{ij} = (y_i - \bar{y})'\,\Sigma^{-1}(y_j - \bar{y})$ ซึ่ง $\Sigma^{-1} = \frac{1}{n}\sum_i (y_i - \bar{y})'(y_j - \bar{y})$

และ p คือ จำนวนตัวแปรอิสระ

มีขั้นตอนการทดสอบ ดังนี้

1) กำหนดสมมติฐานการทดสอบ

H_0 : ตัวแปรอิสระ **x** มีการแจกแจงแบบปกติพหุ

H_1 : ตัวแปรอิสระ **x** ไม่มีการแจกแจงแบบปกติพหุ

2) กำหนดระดับนัยสำคัญ α=0.05

3) จะปฏิเสธสมมติฐานหลัก (H_0) เมื่อ

3.1) ถ้า $b_{1,p}$ > ค่า upper percentage point ของ $b_{1,p}$

(เปิดจากตาราง Upper Percentages for $b_{1,p}$, Upper and LowerPercentiles for $b_{2,p}$)

3.2) ถ้า $b_{2,p}$ > ค่า upper percentage point ของ $b_{2,p}$

(เปิดจากตาราง Upper Percentages for $b_{1,p}$, Upper and LowerPercentiles for $b_{2,p}$)

สำหรับการทดสอบที่มีตัวแปรอิสระมากกว่า 4 ตัว สามารถทำได้ 2 กรณี

กรณีที่ 1 ใช้สถิติทดสอบ Z_1

$$Z_1 = \frac{(p+1)\,(n+1)\,(n+3)}{6\left[(n+1)\,(p+1) - 6\right]}\, b_{1,p}$$

ซึ่ง Z_1 มีการแจกแจงแบบไคกำลังสองที่ $df = \frac{1}{6}p\,(p+1)\,(p+2)$

จะปฏิเสธ H_0 ถ้า $Z_1 \geq \chi^2_{0.05,\frac{1}{6}p(p+1)(p+2)}$ เมื่อกำหนดระดับนัยสำคัญที่ 0.05

กรณีที่ 2 ใช้ค่าความโด่ง $b_{2,p}$ จะปฏิเสธ H_0 ถ้าค่าความโด่งมีค่ามากหรือมีค่าต่ำ

2.1 กรณีที่ใช้ค่าขอบเขตบน ใช้สถิติทดสอบ Z_2

$$Z_2 = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}}$$

โดยที่ Z_2 มีการแจกแจงแบบปกติ จะปฏิเสธ H_0 เมื่อ $Z_2 > Z_{1-\frac{\alpha}{2}}$

2.2 กรณีใช้ค่าขอบเขตล่าง

2.2.1 เมื่อ $50 \le n < 400$ จะใช้สถิติทดสอบ Z_3

$$Z_3 = \frac{b_{2,p} - p(p+2)(n+p+1)/n}{\sqrt{8p(p+2)/(n+1)}}$$

โดยที่ Z_3 มีการแจกแจงแบบปกติ จะปฏิเสธ H_0 ถ้า $Z_3 < Z_{\frac{\alpha}{2}}$

2.2.2 เมื่อ $n \ge 400$ จะใช้สถิติทดสอบ Z_2 โดยจะปฏิเสธ H_0 ถ้า $Z_2 < Z_{\frac{\alpha}{2}}$

1.1.3 การตรวจสอบการเท่ากันของความแปรปรวนร่วมของตัวแปรอิสระ

Lattin *et al.* (2003) กล่าวถึง การตรวจสอบการเท่ากันของความแปรปรวนร่วมของตัวแปรอิสระด้วยสถิติทดสอบ Box's M มีขั้นตอนดังนี้

1) การกำหนดสมมติฐานการทดสอบ

$$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$$
$$H_1 : \Sigma_1 \ne \Sigma_2 \ne \ldots \ne \Sigma_k$$

2) สถิติทดสอบคือ $B = (1 - c)\left\{\left[\sum_{i=1}^{k}(n_i - 1)\right] \ln |S_p| - \left[\sum_{i=1}^{k}(n_i - 1)\right] \ln |S_i|\right\}$

โดยที่ $c = \left[\sum_{i=1}^{k}(n_i - 1) - \frac{1}{\sum_{i=1}^{k}(n_i - 1)}\right]\left[\frac{2p^2 + 3p - 1}{6(p + 1)(k - 1)}\right]$

p คือ จำนวนตัวแปรอิสระ

n_i คือ จำนวนข้อมูลในกลุ่มที่ i

k คือ จำนวนกลุ่ม

S_i คือ ความแปรปรวนร่วมของตัวอย่างกลุ่มที่ i

S_p คือ ความแปรปรวนร่วมของตัวอย่างทุกกลุ่ม

B มีการแจกแจงแบบไคกำลังสอง ที่มี df = $\frac{1}{2}$ p (p+1) (p+2)

3) ค่าวิกฤต คือ $\chi^2_{\alpha, \frac{1}{2} p (p+1) (k-1)}$

4) จะปฏิเสธ H_0 ถ้า $B > \chi^2_{\alpha, \frac{1}{2} p (p+1) (k-1)}$

1.2 การทดสอบความแตกต่างระหว่างกลุ่มในการวิเคราะห์การจำแนก

การวิเคราะห์การจำแนก จะต้องทราบว่า ค่าสังเกตอยู่ในกลุ่มใดและมีจำนวนกลุ่มกี่กลุ่มที่จะทำการวิเคราะห์ โดยกลุ่มที่แบ่งนี้อาจจะไม่แตกต่างกันก็ได้ จึงต้องทำการตรวจสอบก่อนว่ากลุ่มที่แบ่งมีความแตกต่างกันหรือไม่ โดยพิจารณาจากความแตกต่างระหว่างค่าเฉลี่ยของแต่ละกลุ่ม ซึ่งมีหลายวิธีในการทดสอบได้แก่ Wilk's Lambda, Hotelling T^2, Roy's largest root และ Pillai's Trace เป็นต้น

Khattree and Naik (2003) ได้กล่าวถึง การตรวจสอบด้วยสถิติ Hotelling T^2 ดังนี้

1) การกำหนดสมมติฐานในการทดสอบ

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$$

H_1: มีอย่างน้อย 1 คู่ คือ $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$ เมื่อ $i \neq j$

2) สถิติทดสอบ คือ $T^2 = n_k \, \bar{\mathbf{v}}' \, S_v^{-1} \, \bar{\mathbf{v}}$

โดยที่ $\bar{\mathbf{v}} = \frac{1}{n_k} \sum_{j=1}^{n_k} v_j$

n_k คือขนาดตัวอย่างของกลุ่มที่ k

k คือจำนวนกลุ่ม

$v_j = (y'_{1j} : y'_{2j} : \ldots : y'_{k-1,j})' \; ; j = 1,\ldots,n_k$

$S_v = \frac{1}{(n_k - 1)} \sum_{j=1}^{n_k} (v_j - \bar{\mathbf{v}})(v_j - \bar{\mathbf{v}})'$

T^2 มีการแจกแจงเป็น Hotelling T^2 แต่ $\frac{[n_k - p(k-1)]}{[(n_k - 1)\, p\, (k-1)]} T^2$ มีการแจกแจงแบบ

F ด้วย $df = [p(k-1), n_k - p(k-1)]$ เมื่อ p คือ จำนวนตัวแปรอิสระ

3) ค่าวิกฤต คือ $F_{\alpha,[p(k-1),n_k - p(k+1)]}$

4) จะปฏิเสธ H_0 ถ้าค่าสถิติทดสอบ $\frac{[n_k - p(k-1)]}{[(n_k - 1)\, p\, (k-1)]} T^2$ มีค่ามากกว่าค่าวิกฤต F จากตารางสถิติ

1.3 สมการจำแนกกลุ่ม

Hair *et al.* (2010) ได้กล่าวถึง การสร้างสมการจำแนกกลุ่ม มีวิธีประมาณค่าสมการจำแนก 2 วิธีคือ วิธีตรง (simultaneous method) และวิธี stepwise method โดยมีวิธีการ ดังนี้

1.3.1 วิธีตรง (simultaneous method) เป็นวิธีการที่คำนวณสมการการจำแนกโดยนำตัวแปรอิสระทุกตัวมาพิจารณา วิธีการนี้เหมาะเมื่อต้องการให้ตัวแปรอิสระทุกตัวอยู่ในสมการการจำแนก โดยไม่สนใจตัวแปรการจำแนกที่ดีที่สุด

1.3.2 วิธี stepwise method เป็นวิธีที่มีการนำ ตัวแปรเข้าและนำตัวแปรออก วิธีการนี้ควรมีอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระอย่างน้อย 20 ค่าสังเกตต่อหนึ่งตัวแปรอิสระ โดยมีลำดับขั้น ดังนี้

1. เลือกตัวแปรที่มีความสามารถในการจำแนกได้ดีที่สุด โดยพิจารณาจากค่า Mahalanobis distance ที่มีค่ามากที่สุดและมีนัยสำคัญทางสถิติ

2. จับคู่ตัวแปรที่นำเข้าแล้วกับตัวแปรอื่นๆแต่ละตัว และเลือกตัวแปรที่ให้กำลังการจำแนกของสมการที่ดีที่สุดเมื่อรวมกับตัวแปรตัวแรก

3. เลือกตัวแปรนำเข้าในสมการการจำแนก โดยตัวแปรที่ถูกนำเข้าก่อนหน้านี้อาจถูกนำออกจากสมการการจำแนกก็ได้

4. กระบวนการสิ้นสุดเมื่อตัวแปรที่อยู่ในสมการไม่สามารถนำออกจากสมการได้ หรือตัวแปรที่ยังไม่ได้เข้าในสมการไม่มีตัวแปรใดที่สามารถเข้าอยู่ในสมการได้อีกแล้ว ตามระดับนัยสำคัญที่กำหนดของการนำตัวแปรเข้าสมการและการนำตัวแปรออกจากสมการ

1.4 การจัดหน่วยตัวอย่างเข้ากลุ่ม

การกำหนดกลุ่มให้กับหน่วยใหม่ ทำได้โดยการนำสมการจำแนกกลุ่มที่ได้จากการวิเคราะห์จำแนกกลุ่มมาใช้ในการพยากรณ์หน่วยสังเกตใหม่ โดยมีเทคนิคการจัดกลุ่มหลายวิธี ดังนี้

1.4.1 Simple Classification Function วิธีการนี้จะใช้สมการจำแนกกลุ่มด้วยฟังก์ชันการจำแนกเชิงเส้น มาแทนค่าของตัวแปรอิสระของหน่วยสังเกตใหม่ แล้วคำนวณหาค่าคะแนนการจำแนก (Discriminant score : D) แล้วนำค่า D ไปเทียบกับกฎการจำแนกกลุ่มที่สร้างขึ้นเพื่อจัดหน่วยสังเกตใหม่เข้ากลุ่ม

1.4.2 Probabilities of Group Membership เป็นการคำนวณหาความน่าจะเป็นหรือโอกาสที่หน่วยใหม่จะอยู่ในกลุ่มที่ i ; i = 1,2,…,k แล้วนำมาเปรียบเทียบกัน วิธีการนี้มีเงื่อนไขว่าตัวแปรต้องมีการแจกแจงแบบปกติพหุและความน่าจะเป็นที่คำนวณจะเป็นความน่าจะเป็นภายหลัง (Posterior Probability)

1.4.3 Generalized Distance Function เป็นการคำนวณหาระยะห่างจากหน่วยสังเกตใหม่ที่ต้องการจัดไปยังค่ากลางของกลุ่ม (Group Centroid) โดยคำนวณค่าระยะห่าง Mahalanobis Distance ถ้าระยะห่างดังกล่าวห่างจากค่ากลางของกลุ่มใดต่ำสุด จะจัดให้หน่วยสังเกตไปอยู่ในกลุ่มนั้น

1.5 การประเมินความเหมาะสมของสมการจำแนกกลุ่ม

การวิเคราะห์การจำแนกเมื่อนำสมการการจำแนกที่ได้ ไปพยากรณ์หรือจัดกลุ่มให้หน่วยสังเกตใหม่ว่าหน่วยสังเกตใหม่ควรอยู่กลุ่มใด ความถูกต้องในการจัดกลุ่มเป็นสิ่งที่สำคัญซึ่งสามารถประเมินความเหมาะสมของสมการการจำแนกที่ได้ โดยทำได้ 3 วิธี ดังนี้ (กัลยา, 2552)

1.5.1 การทดสอบกับตัวอย่างทั้งหมด ซึ่งการทำกับตัวอย่างทั้งหมดก็เหมือนการทำซ้ำกระบวนการประมาณค่า โดยการใช้ข้อมูลทั้งหมดมาสร้างสมการการจำแนกแล้วนำสมการ

การจำแนกที่ได้มาพยากรณ์ข้อมูลชุดเดิม อาจทำให้สัดส่วนการพยากรณ์ความถูกต้องในการจัดกลุ่มมีค่าสูงเกินความเป็นจริง

1.5.2 การแบ่งข้อมูลเป็น 2 ส่วน คือ ส่วนหนึ่งเป็นข้อมูลฝึกสอน (training data) มาสร้างสมการการจำแนก และนำสมการการจำแนกที่ได้จากข้อมูลฝึกสอน มาพยากรณ์หรือจัด กลุ่มให้กับข้อมูลอีกส่วนหนึ่ง เรียกว่า ข้อมูลทดสอบ (test data หรือ holdout data) เพื่อตรวจสอบความถูกต้องในการจัดกลุ่ม

1.5.3 วิธีการ CROSS-VALIDATION เป็นการทดสอบความเหมาะสมของตัวอย่างทั้งหมดโดยใช้หลักการ leave-one-out (ตัดออกครั้งละหนึ่งค่าสังเกต) คือใช้ข้อมูล n-1 ค่าสังเกตมาสร้างฟังก์ชันการจำแนกแล้วนำฟังก์ชันที่ได้มาพยากรณ์ความถูกต้องในการจัดกลุ่มให้กับค่าสังเกตที่ถูกตัดออกไป ทำแบบนี้ต่อไปเรื่อยๆ โดยการตัดค่าสังเกตออกไปครั้งละหนึ่งจนครบ n ครั้ง แล้วนับจำนวนการจัดกลุ่มที่ถูกต้อง

วิธีการจัดกลุ่ม ค่าสังเกตในตัวแปรอิสระจะถูกนำเข้าในสมการการจำแนกและคำนวณคะแนนการจัดกลุ่มแต่ละกลุ่ม โดยค่าสังเกตจะถูกจัดให้อยู่ในกลุ่มที่มีคะแนนการจัดกลุ่มตรงกับเงื่อนไขหรือกฎการจำแนก ความถูกต้องในการทำนายสมาชิกกลุ่ม การวิเคราะห์การจำแนกจะวัดร้อยละความถูกต้องในการจัดกลุ่ม ดังนี้

ตารางที่ 1 การจัดกลุ่มของการวิเคราะห์การจำแนก

กลุ่มที่แท้จริง	กลุ่มที่ทำนาย		ขนาดของกลุ่มที่แท้จริง	ร้อยละการจัดกลุ่มที่ถูกต้อง
	1	2		
1	22	3	25	88
2	5	20	25	80
ขนาดของกลุ่มที่ทำนาย	27	23	50	84[a]

ที่มา : Hair *et al.*, 2007

2. วิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์

2.1 วิธีการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกเชิงเส้น

Khattree and Naik (2003) ได้กล่าวถึง วิธีการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกเชิงเส้น ดังนี้

มีประชากร k กลุ่ม มีค่าความแปรปรวนร่วมเท่ากัน คือ $\sum$ และค่าใช้จ่าย (cost)ในการจำแนกผิดในแต่ละกลุ่มสมมติว่ามีค่าเท่ากัน โดยคำนวณค่าระยะห่างกำลังสอง ได้จาก

$$D_t^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_t)' \sum^{-1} (\mathbf{x} - \boldsymbol{\mu}_t) - 2 \ln(\pi_t)$$

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

เมื่อไม่ทราบค่า $\boldsymbol{\mu}_t$ และ $\sum$ สามารถประมาณค่าด้วย $\bar{x}_t$ และ S ตามลำดับ เพื่อจะได้ประมาณค่าของ $D_t^2(\mathbf{x})$ ด้วย $d_t^2(\mathbf{x})$ ซึ่ง

$$d_t^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_t)' \sum^{-1} (\mathbf{x} - \bar{\mathbf{x}}_t) - 2 \ln(\pi_t)$$

ในการจัดกลุ่ม ค่าสังเกตถูกจัดให้เข้ากลุ่มประชากรที่อยู่ใกล้ที่สุด โดยพิจารณาจาก ค่าความน่าจะเป็นภายหลัง (posterior probability) ของค่าสังเกตนั้น มีค่าความน่าจะเป็นมากที่สุด สำหรับกรณีประชากร k กลุ่ม ค่าประมาณความน่าจะเป็นภายหลัง ถูกนิยามดังสมการข้างล่างนี้

$$P(t / \mathbf{x}) = \frac{e^{-\frac{1}{2} d_t^2(x)}}{\sum_{j=1}^{k} e^{-\frac{1}{2} d_t^2(x)}} \; ; t = 1,\ldots,k$$

2.2 วิธีการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสอง

Khattree and Naik (2003) กล่าวถึง วิธีการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสอง ดังนี้ เมื่อเมทริกซ์ความแปรปรวนร่วมของประชากรไม่เท่ากัน กฎการจำแนกสำหรับประชากร k กลุ่ม จะมีความซับซ้อนมากขึ้น ถึงแม้ว่าข้อมูลจะเป็นไปตามข้อตกลงของการแจกแจงแบบปกติพหุ กฎการจำแนกจะขึ้นอยู่กับเมทริกซ์ความแปรปรวนร่วมของประชากรแต่ละกลุ่ม $\sum_1, \sum_2, ..., \sum_k$ การจัดกลุ่มให้กับค่าสังเกต $\mathbf{x}$ เข้าประชากรกลุ่มที่ t ; t = 1,...,k จะพิจารณาจากค่า

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

โดยที่ $(D_j^2(\mathbf{x}))$ ที่กำหนดจะมีความแตกต่างจากฟังก์ชันการจำแนกเชิงเส้น นั่นคือ

$$D_j^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_j)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \ln |\mathbf{S}_j| - 2 \ln (\pi_j) \quad ; j = 1,...,k$$

สำหรับในกลุ่มตัวอย่าง จะแทน $\boldsymbol{\mu}_j$ ด้วย $\bar{\mathbf{x}}_j$ และแทน $\sum_j$ ด้วย $\mathbf{S}_j$ จะได้ว่า

$$d_j^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_j)' \mathbf{S}_j^1 (\mathbf{x} - \bar{\mathbf{x}}_j) + \ln |\mathbf{S}_j| - 2 \ln (\pi_j)$$

การประมาณค่า ระยะห่างกำลังสองแบบวางนัยทั่วไปเชิงคู่ (pairwise generalized squared distance) ของกลุ่มที่ i จากกลุ่มที่ j คือ

$$d^2(i/j) = (\mathbf{x} - \bar{\mathbf{x}}_j)' \mathbf{S}_j^1 (\mathbf{x} - \bar{\mathbf{x}}_j) + \ln |\mathbf{S}_j| - 2 \ln (\pi_j)$$

$d^2(i/j)$ สามารถมีค่าเป็นลบได้ และการประมาณค่า $d^2(i/j)$ จะไม่เท่ากับ $d^2(j/i)$ อย่างไรก็ตาม ระยะห่างที่มากกว่า แสดงว่า จำแนกกลุ่มที่ i และ j ได้ดี

3. วิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ : วิธี เคเนียเรสเนเบอร์ (kNN)

Khattree and Naik (2003) ได้กล่าวถึง วิธีก ารวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ว่าในการจัดกลุ่มให้กับหน่วยตัวอย่างใหม่ ให้อยู่ในกลุ่มเดียวกับหน่วยตัวอย่างที่อยู่รอบข้างที่ใกล้ที่สุด การพิจารณาความใกล้โดยใช้ค่าระยะห่างกำลังสองระหว่าง 2 หน่วยในประชากรกลุ่มที่ t ที่มีเวกเตอร์ของค่าสังเกต $\mathbf{x}_1$ และ $\mathbf{x}_2$ ตามลำดับ ดังสมการข้างล่างนี้

$$d_t^2(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2)' \ \mathbf{V}_t^{-1}(\mathbf{x}_1 - \mathbf{x}_2)$$

โดยที่ V_t คือ เมทริกซ์ความแปรปรวนร่วมภายในกลุ่มที่ t

$\mathbf{x}_1, \mathbf{x}_2$ คือ เวกเตอร์ของค่าสังเกตของหน่วยตัวอย่างที่ 1 และ 2

สมมติว่าจะจัดหน่วยตัวอย่างที่มีค่าสังเกต $\mathbf{x}$ เข้าในประชากรกลุ่มใดกลุ่มหนึ่ง ให้ k เป็นจำนวนเต็มบวก ในการหาค่า $d_t^2(\mathbf{x}_1, \mathbf{x}_2)$ จะต้องหาค่าสังเกตที่อยู่รอบข้างที่ใกล้ที่สุดกับ $\mathbf{x}$ สมมติว่าจาก k ค่าสังเกตมีเพียง k_t ค่าสังเกตที่ มาจากประชากรกลุ่มที่ t ด้วยค่าค วามน่าจะเป็นก่อนหน้า π_t แล้วจะได้ค่าประมาณ ความน่าจะเป็นภายหลัง (posterior probability) ของหน่วยตัวอย่างที่จะเป็นสมาชิกของประชากรกลุ่มที่ t ดังนี้

$$\frac{\pi_t \hat{f}_t(x)}{\sum_{s=1}^{g} \pi_s \hat{f}_s(x)} = \frac{\pi_t (k_t/n_t)}{\sum_{s=1}^{g} \pi_s (k_s/n_s)}$$

ค่าสังเกต $\mathbf{x}$ จะถูกจัดให้อยู่ในกลุ่มที่ t ถ้าค่าความน่าจะเป็นภายหลัง มีค่าสูงที่สุด สำหรับประชากรกลุ่มที่ t หรือ $\mathbf{x}$ จะถูกจัดให้อยู่ในกลุ่มที่ t ถ้า $\pi_t (k_t/n_t)$ มีค่ามากกว่า $\pi_s (k_s/n_s)$ สำหรับ $s = 1,...,g$; g เป็นจำนวนกลุ่มของประชากร โดยที่ k จะถูกกำหนดล่วงหน้า ไม่มีวิธีการที่ชัดเจนสำหรับการเลือกค่า k ที่ดีที่สุด

4. การประมาณค่าอัตราความผิดพลาด

4.1 วิธีบูตสแตรป (Bootstrap)

วิธีบูตสแตรป เป็นที่เข้าใจกัน หมายถึง ตัวอย่างบูตสแตรป ที่เป็นตัวอย่างสุ่มขนาด n สุ่มมาจากประชากรที่มีการแจกแจง $\hat{F}$ สมมติ $\mathbf{x}^*$ เป็นเวกเตอร์ของค่าสังเกตที่สุ่ม

$$\mathbf{x}^* = (x_1^*, x_2^*, ..., x_n^*)$$

และ $$\hat{F} \rightarrow (x_1^*, x_2^*, ..., x_n^*)$$

$\mathbf{x}^*$ ไม่ใช่ชุดข้อมูลจริง แต่ได้มาจากการสุ่มซ้ำจากข้อมูลจริง $\mathbf{x}$

$x_1^*, x_2^*, ..., x_n^*$ เป็นตัวอย่างสุ่มขนาด n ที่สุ่มแบบใส่คืนจากประชากรขนาด n $(x_1, x_2, ..., x_n)$ ในการสุ่มอาจจะได้ $x_1^* = x_7, x_2^* = x_3, x_3^* = x_3, ..., x_n^* = x_7$

ชุดข้อมูลบูตสแตรป $(x_1^*, x_2^*, ..., x_n^*)$ ประกอบด้วยสมาชิกของข้อมูลชุดเดิม $(x_1, x_2, ..., x_n)$ บางค่าสังเกตอาจจะไม่ถูกสุ่มเลยก็ได้ บางค่าสังเกตอาจถูกสุ่มเพียงครั้งเดียว หรือบางค่าสังเกตอาจถูกสุ่มมากกว่าหนึ่งครั้งก็ได้

ขั้นตอนของบูตสแตรป สุ่มตัวอย่างบูตสแตรปที่เป็นอิสระต่อกันหลายๆ ตัวอย่าง ประเมินการสุ่มชุดตัวอย่างบูตสแตรปด้วยการ คำนวณค่าสถิติ $\hat{\theta}$ ประมาณค่าความคลาดเคลื่อนมาตรฐาน (Standard error : SE) ของค่าสถิติ $\hat{\theta}$ จากค่าเบี่ยงเบนมาตรฐานของตัวอย่างบูตสแตรปที่สุ่มซ้ำ เรียกเป็น ค่าประมาณบูตสแตรปของ SE เขียนแทนด้วย se_B เมื่อ B คือจำนวนตัวอย่างบูตสแตรปที่ได้ (Efron and Tibshirani, 1993)

ขั้นตอนบูตสแตรป สำหรับประมาณอัตราความผิดพลาด

1. สุ่มตัวอย่างบูตสแตรป $\mathbf{x}^{*1},\mathbf{x}^{*2},...,\mathbf{x}^{*50}$ ที่เป็นอิสระต่อกัน ตัวอย่างแต่ละตัวอย่างประกอบด้วย 599 ค่าสังเกต จำนวน 50 ชุด ที่สุ่มแบบใส่คืนจากข้อมูลกลุ่มเสี่ยงโรคเบาหวาน 599 คน

2. แบ่งข้อมูลตัวอย่างบูตสแตรปแต่ละชุดออกเป็น 2 ส่วนอย่างสุ่ม

ส่วนที่ 1 เป็นข้อมูลฝึกสอน สุ่มข้อมูลมา 2/3 จาก 599 ค่าสังเกต

ส่วนที่ 2 เป็นข้อมูลทดสอบ สุ่มข้อมูลมา 1/3 จาก 599 ค่าสังเกต

นำข้อมูลส่วนที่ 1 ของแต่ละชุดตัวอย่างบูตสแตรปมาคำนวณค่าอัตราความผิดพลาด $(ER)_i$; i = 1,...,50 พร้อมทั้งคำนวณค่าเฉลี่ยอัตราความผิดพลาดจากข้อมูล 50 ชุด คือ E(ER)

3. ประมาณค่าความคลาดเคลื่อนมาตรฐาน SE(ER) จากค่าเบี่ยงเบนมาตรฐานของข้อมูลบูตสแตรป 50 ชุด

$$SE(ER)_i = \sqrt{\frac{\sum_{i=1}^{50}\left[(ER)_i - E(ER)\right]^2}{50-1}}$$

4.2 การประมาณค่าอัตราความผิดพลาด (Error Rate Estimation)

การประมาณค่าอัตราความผิดพลาดของวิธีการวิเคราะห์การจำแนกมีหลายวิธี ในที่นี้จะกล่าวถึง วิธีการประมาณค่าจากข้อมูล เรียกว่าเป็นการประมาณจำนวนความผิดพลาด(Error count estimates) สมมติว่ามีชุดข้อมูลทดสอบชุดหนึ่ง ทราบว่า ค่าสังเกตใดเป็นสมาชิกของประชากรใด ในการประเมินความสามารถของฟังก์ชันการจำแนก โดยการหาจำนวนและร้อยละของค่าสังเกตที่มาจากประชากรกลุ่มที่ t แต่ถูกจัดให้อยู่ในประชากรกลุ่มที่ s ด้วยค่าความน่าจะเป็นแบบมีเงื่อนไข คือ P(s/t)

ดังนั้น สัดส่วนของการจำแนกผิดของประชากรที่มาจากประชากรกลุ่มที่ t คือ

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

ในการประมาณค่าอัตราความผิดพลาด หากชุดข้อมูลทดสอบให้ค่าประมาณอัตราความผิดพลาดแตกต่างกันจาก ค่าอัตราความผิดพลาดที่ได้จาก ฟังก์ชัน การจำแนก ที่สร้างขึ้น แสดงว่าค่าประมาณอัตราความผิดพลาดมีความไม่เอนเอียง อย่างไรก็ตามค่าประมาณอัตราความผิดพลาดจะมีแนวโน้มว่ามีการกระจายมาก คือ มีค่าความแปรปรวนสูง เมื่อข้อมูลชุดทดสอบมีขนาดเล็ก

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

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

ตารางที่ 2 จำนวนของการจัดกลุ่มที่ได้จากฟังก์ชันการจำแนกของชุดข้อมูลที่ i

กลุ่มที่แท้จริง	**กลุ่มที่พยากรณ์**			**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, ... ,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 หากร่างกายขาดฮอร์โมนอินซูลิน หรือผลิตฮอร์โมนไม่เพียงพอจะทำให้น้ำตาลไม่สามารถเข้าไปในเนื้อเยื่อได้จึงมีน้ำตาลในเลือดเหลือค้างอยู่มากและมีระดับที่สูงกว่าปกติ เมื่อในเลือดมีน้ำตาลสูง ไตจะทำหน้าที่กรองน้ำตาลออกมากับน้ำปัสสาวะ ทำให้ ปัสสาวะมีรสหวาน จึงเรียกว่า “เบาหวาน” (เทพ และคณะ, 2548)

ประเภทของโรคเบาหวาน

สมาคมโรคเบาหวานแห่งสหรัฐอเมริกา (American Diabetes Association) ปี ค.ศ. 1997 ได้จำแนกโรคเบาหวานออกเป็น 4 ประเภท ดังนี้ (นวรัตน์, 2551)

1. โรคเบาหวานชนิดที่ 1 (Type 1 Diabetes) หรือเบาหวานชนิดพึ่งอินซูลิน โรคเบาหวานชนิดนี้เกิดจากเบต้าเซลล์ในตับอ่อนมีจำนวนน้อยหรือไม่มีเลย ทำให้ตับอ่อนไม่สามารถ ผลิตอินซูลินได้ 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 หลังคลอดอาการโรคเบาหวานจะหายไป แต่มีความเสี่ยงในการเกิดโรคเบาหวานได้มาก จึงควรตรวจเช็คโรคเป็นระยะเพื่อป้องกันการเกิดโรคเบาหวาน

4. โรคเบาหวานชนิดอื่น ได้แก่ โรคเบาหวานที่เกิดจากความผิดปกติทางพันธุกรรมที่ทราบชนิดชัดเจน โรคของตับอ่อน ความผิดปกติของฮอร์โมน ยาหรือสารเคมี และอื่นๆ

อาการของโรคเบาหวาน

อาการที่พบบ่อยในผู้ป่วยเบาหวาน คือ คอแห้ง กระหายน้ำ ดื่มน้ำมาก ปัสสาวะบ่อยและมีจำนวนมาก รับประทานอาหารจุ หิวบ่อยแต่น้ำหนักตัวลดลง ตาพร่ามัว ชาปลายมือปลายเท้า คันตามผิวหนัง ถ้าเป็นแผลจะหายช้ากว่าปกติ มีการติดเชื้อตามผิวหนัง (เทพ และคณะ, 2548)

สาเหตุของโรคเบาหวาน

การเกิดโรคเบาหวานมีสาเหตุหลายประการ (กองบรรณาธิการใกล้หมอ, 2548)

1. ความอ้วน เนื่องจากเซลล์ของร่างกายคนอ้วนจะตอบสนองต่อฮอร์โมนอินซูลินน้อยลง ฮอร์โมนอินซูลินจึงไม่สามารถพาน้ำตาลเข้าสู่เซลล์ได้ดีเหมือนเดิม จนทำให้เกิดภาวะน้ำตาลในเลือดสูง

2. อายุของคนที่มากขึ้นทำให้อวัยวะต่างๆ ภายในร่างกายคนเราย่อมเสื่อมลงรวมทั้งตับอ่อนที่สังเคราะห์และผลิตฮอร์โมนอินซูลินได้น้อยลง ทำให้มีน้ำตาลส่วนเกินในกระแสเลือด

3. ตับอ่อนไม่สมบูรณ์ เกิดจากการที่ตับอ่อนได้รับการกระทบกระเทือนหรือเกิดอุบัติเหตุซึ่งจำเป็นต้องผ่าตัดเอาตับอ่อนบางส่วนออก หากบุคคลนั้นมีแนวโน้มจะเป็นโรคเบาหวาน ปัจจัยนี้จะทำให้อาการของโรคเบาหวานแสดงอาการเร็วขึ้น

4. การติดเชื้อไวรัสบางชนิด เชื้อไวรัสบางชนิดเช่น คางทูม หัดเยอรมัน เมื่อเข้าสู่ร่างกายแล้วจะมีผลทำให้เกิดโรคเบาหวานได้

5. ยาบางชนิด เช่น ยาคุมกำเนิด ยาขับปัสสาวะ มีผลทำให้ระดับน้ำตาลในเลือดสูงขึ้นได้ จึงควรปรึกษาแพทย์ก่อนใช้ยา เมื่อจำเป็นต้องใช้ยาติดต่อกันเป็นเวลานาน

6. การตั้งครรภ์ ฮอร์โมนบางชนิดที่รกสร้างขึ้นมาจะมีผลยับยั้งการทำงานของฮอร์โมนอินซูลิน ผู้ที่ตั้งครรภ์จึงเสี่ยงต่อการเกิดโรคเบาหวาน

7. พันธุกรรมจากการตรวจโรคเบาหวานพบว่าหนึ่งในสามของผู้ป่วยโรคเบาหวานมีประวัติทีมีญาติเป็นโรคเบาหวาน

งานวิจัยที่เกี่ยวข้อง

ทรงพล (2541) ได้ทำการศึกษาเปรียบเทียบสถิติวิเคราะห์การจำแนก ที่อิงพารามิเตอร์และที่ไม่อิงพารามิเตอร์ กับข้อมูลการเกิดภาวะแทรกซ้อนที่จอประสาทตา ในผู้ป่วยเบาหวานชนิดไม่พึ่งอินซูลิน เพื่อศึกษาปัจจัยที่มีอิทธิพลต่อการจำแนกผู้ป่วยเบาหวานชนิดไม่พึ่งอินซูลินที่มีภาวะแทรกซ้อนและไม่แทรกซ้อนที่จอประสาทตา ผู้วิจัยเก็ บข้อมูลจากผู้ป่วยเบาหวานชนิดไม่พึ่งอินซูลินที่มารับการตรวจรักษาที่คลินิกเบาหวานโรงพยาบาลราชวิถี กลุ่มศึกษา ได้แก่ผู้ป่วยเบาหวานที่มีภาวะแทรกซ้อนที่จอประสาทตา จำนวน 150 คน และกลุ่มควบคุม ได้แก่ ผู้ป่วยเบาหวานที่ไม่มีภาวะแทรกซ้อนที่จอประสาทตา จำนวน 300 คน ผลการศึกษาพบว่า เมื่อตัวแปรจำแนกประเภท มีการกระจายแบบปกติ มีการกระจายแบบไม่ปกติ และเมื่อมีการปรับให้มีการกระจายแบบปกติ หรือเป็นตัวแปรผสมแล้ว วิธี kernel density estimation จะให้อัตราความถูกต้องในการจำแนกสูงกว่าวิธีอื่น แต่เมื่อตัวแปรจำแนกประเภทเป็นตัว แปรทวิ วิธี k nearest neighbor จะให้อัตราความถูกต้องในการจำแนกสูงกว่าวิธีอื่น นอกจากนั้นยังพบว่า ปัจจัยที่มีอิทธิพล ต่อการจำแนกผู้ป่วยเบาหวานที่มีภาวะแทรกซ้อนและไม่แทรกซ้อนที่จอประสาทตา ได้แก่ ระยะเวลาในการป่วยด้วยโรคเบาหวาน ระดับน้ำตาลในเลือด ระดั บความดันโลหิตไดแอสโตลิก รายได้ของครอบครัว ชนิดของการรักษา การสูบบุหรี่ และระดับคลอเลสเตอรอลในเลือด

สุภลาภ (2541) ศึกษาการใช้สมการจำแนกกลุ่มงานสำรวจสถานะสุขภาพเรื่องเบาหวานของประชาชนไทยในเขตอำเภอบ้ านโป่ง จังหวัดราชบุรี โดย ศึกษาความแตกต่างระหว่าง กลุ่มโรคเบาหวานกลุ่มผู้ที่มีน้ำตาลในเลือดสูงกว่าปกติ และกลุ่มคนปกติ จากตัวอย่างจำนวน 342 คน หรือกลุ่มละ 114 คน ซึ่งเป็นส่วนหนึ่งในกลุ่มตัวอย่างของการสำรวจนำร่องของกระทรวงสาธารณสุขในเขตอำเภอบ้านโป่ง จังหวัดราชบุรี ระหว่างเดือน สิงหาคม ถึง ธันวาคม 2534 เมื่อทำการวิเคราะห์แบบง่าย ด้วยสถิติไคกำลังสอง พบว่า ปัจจัยที่มีความสำคัญทางสถิติ คือ อายุ เพศ สถานภาพสมรส อาชีพ บริเวณที่อยู่อาศัย ดัชนีมวลกาย และพฤติกรรมการออกกำลังกาย และเมื่อใช้การวิเคราะห์การจำแนกกลุ่ม พบว่า อายุ สถานภาพสมรส ดัชนีมวลกาย พฤติกรรมการดื่มสุรามากกว่า 187.5 ซีซีต่อสัปดาห์ พฤติกรรมการดื่มสุราน้อยกว่าหรือเท่ากับ 187.5 ซีซีต่อสัปดาห์และพฤติกรรมการออกกำลังกาย เป็นตัวแปรที่สำคัญต่อการบ่งบอกความแตกต่างระหว่างกลุ่มอย่างมีนัยสำคัญทางสถิติ โดยมีฟังก์ชั่นสำหรับการคำนวณแยกกลุ่ม ดังนี้

$$D = 0.073\ Age + 0.076\ BMI - 0.098\ Exerscor - 0.256\ Maritals + 0.518\ Grpalc1 + 0.664\ Grpalc2 - 4.957$$

ผลการคำนวณเมื่อ D มากกว่า 0.603 จัดเป็นกลุ่มโรคเบาหวาน ค่า D ที่อยู่ระหว่าง 0.327 ถึง 0.603 จัดเป็นกลุ่มผู้ที่มีน้ำตาลในเลือดสูงกว่าปกติ และค่า D น้อยกว่า -0.327 จัดเป็นกลุ่มคนปกติ ฟังก์ชันนี้การจำแนกสามารถทำนายความถูกต้อง ของการจำแนก กลุ่ม โดยรวมคิดเป็น ร้อยละ 65.2 ส่วนความสามารถในการทำนายความถูกต้องในแต่ละกลุ่ม คือ กลุ่มผู้เป็นโรคเบาหวาน กลุ่มผู้ที่มีน้ำตาลในเลือดสูงกว่าปกติ และกลุ่มคนปกติ ด้วยค่าร้อยละ 61.4, 59.6 และ 74.6 ตามลำดับ

Efron (1975) ได้ศึกษาวิธีการทางสถิติ 2 วิธีได้แก่ การวิเคราะห์การจำแน กกลุ่มและการวิเคราะห์การถดถอยโลจิสติก เพื่อใช้ในการจำแนกกลุ่มประชากร โดยที่ประชากร 2 ประชากรมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยต่างกัน และเมทริกซ์ความแปรปรวนร่วมของ 2 ประชากรเท่ ากัน พบว่า เมื่อตัวอย่างมีขนาดใหญ่ วิธีการวิเคราะห์การจำแนก จะมีประสิทธิภาพสู งกว่าวิธีการวิเคราะห์การถดถอยโลจิสติก

Knoke (1982) ศึกษาเปรียบเทียบเทคนิคการจำแนกสำหรับตัวแปรไม่ต่อเนื่องและตัว แปรต่อเนื่อง ด้วย วิธี Fisher's linear discriminant analysis, quadratic discriminant analysis, logistic regression และ augmented linear discriminant analysis กับข้อมูลที่มีตัวแปรต่อเนื่อง 1 ตัว มีการแจกแจงแบบปกติ และตัวแปรไม่ต่อเนื่อง 2 ตัว พบว่า เมื่อ ไม่มีความสัมพันธ์ระหว่ างตัวแปร การวิเคราะห์การจำแนกด้วยฟังก์ชันเชิงเส้น เป็นวิธีที่มีความเสถียรและมีประสิทธิภาพการจำแนกที่ดี

Ferrer, Alvaro and Wang (1999) ได้ศึกษาอัตราความผิดพลาดการจำแนกกลุ่มด้วยวิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ และโลจิสติกในการจำแนกข้อมูล 2 กลุ่ม ที่รวบรวมได้จากการสำรวจและสร้างตัวอย่าง บูตสแตรป ข้อมูล ที่ได้ไม่มีการแจกแจงแบบปกติพหุ และความแปรปรวนร่วมก็ไม่เท่ากัน ผลการศึกษาพบว่า 1) การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์และโลจิสติกจะให้ค่าคาดหวังที่ต่ำกว่าทฤษฎี 2) การกำหนด ความน่าจะเป็นก่อนหน้า (Priors Probability) มีอิทธิพลต่อการจำแนกกลุ่มในกลุ่มที่มีจำนวนน้อยและกลุ่มที่มีจำนวนมาก แต่ไม่มีอิทธิพลในกลุ่มรวมทั้งหมด 3) การลดค่าอัตราความผิดพลาดในกลุ่มหนึ่งอาจจะทำให้ค่าอัตราความผิดพลาดในกลุ่มอื่นเพิ่มขึ้น งานวิจัยนี้ไม่มีการ แสดงให้เห็ นว่า การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ได้ถูกนำมาใช้เมื่อข้อมูลไม่มีการแจกแจงแบบปกติพหุและความแปรปรวนร่วมไม่เท่ากัน รวมทั้งไม่ ได้แสดง ว่าการวิเคราะห์ก ารจำแนกโดยใช้ฟังก์ชันการจำแนกกำลังสองและโลจิสติกจะดีกว่าการวิเคราะห์การจำแนกโดยใช้ฟังก์ชันการจำแนกเชิงเส้น

Sultana (2005) ได้เสนองานวิจัย เรื่อง Discriminant analysis and logistic discrimination: an application to diabetes mellitus data 2000 โดยใช้ข้อมูลเบาหวานสะสมจาก Bangladesh Institute of Research and Rehabilitation for Diabetes, Endocrine and Metabolic Disorders (BIRDEM) เพื่อศึกษาความแตกต่างทางด้านสังคม เศรษฐกิจ จำนวนประชากรและตัวแปรทางคลินิค ในการจำแนกกลุ่มผู้ป่วยที่เป็นเบาหวานและกลุ่มที่ไม่เป็นเบาหวาน และหาวิธีการจัดกลุ่มผู้ป่วย ในการวิเคราะห์การจำแนก พบว่า ตัวแปรลักษณะที่อยู่อาศัย, การออกกำลังกาย , เปอร์เซ็นต์ค่าดัชนีมวลกาย, อายุและพันธุกรรม มีนัยสำคัญต่อ การจำแนกกลุ่มผู้ป่วย ส่วนการ วิเคราะห์แบบ logistic discrimination พบว่า ตัวแปรเพศ , อายุ, ลักษณะที่อยู่อาศัย , การออกกำลังกาย และเปอร์เซ็นต์ค่าดัชนีมวลกาย (pbmi) เป็นปัจจัยที่สำคัญในการจำแนกกลุ่ม และได้ทำการเปรียบเทียบวิธีการวิเคราะห์ทั้ง 2 วิธี พบว่า การวิเคราะห์แบบ logistic discrimination มีเปอร์เซ็นต์ความถูกต้องในการจัดกลุ่มผู้ป่วยมากกว่าการวิเคราะห์การจำแนก

Maroco *et al.* (2009) ได้ศึกษาเปรียบเทียบการจำแนกกลุ่มทีอิงพารามิเตอร์และที่ไม่อิงพารามิเตอร์ในการทำนายผู้ป่วยโรคสมองเสื่อม โดยศึกษาวิธีการจำแนกกลุ่มที่ไม่อิงพารามิเตอร์ 7 วิธี คือ Multilayer perceptron Neural Networks, Radial Basis Function, Support Vector Machines, CART, CHAID and QUEST, Classification trees และ Random Forests เปรียบเทียบกับการวิเคราะห์การจำแนกด้วยฟังกชันเชิงเส้น, การวิเคราะห์การจำแนกด้วยฟังก์ชันกำลังสอง และการวิเคราะห์การถดถอยโลจิสติก โดยทดสอบกับข้อมูลจริงซึ่งเป็นข้อ มูลผู้ป่วยโรคสมองเสื่อม วัดค่าความถูกต้องในการทำนาย ค่าความจำเพาะ และค่าความไว พบว่า การวิเคราะห์การจำแนก ด้วยฟังกชันเชิงเส้นและ Random Forests มีการจำแนกกลุ่มได้ถูกต้องมากที่สุด

อุปกรณ์และวิธีการ

อุปกรณ์

เครื่องไมโครคอมพิวเตอร์ หน่วยประมวลผลกลางแบบ Intel Pentium 4 มีความเร็วในการประมวลผล 2.16 GH_Z หน่วยความจำ 512 MB โดยใช้โปรแกรม Statistical Analysis System Version 9.1

วิธีการ

1. ข้อมูลที่ใช้ในการศึกษา

ข้อมูลที่นำมาศึกษาครั้งนี้เป็นข้อมูลการตรวจสุขภาพประจำปีของเจ้าหน้าที่ที่ปฏิบัติงานในโรงพยาบาล นพรัตน์ราชธานี 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 ตัวแปรกลุ่ม แบ่งเป็น 3 กลุ่มคือ กลุ่มไม่เสี่ยงโรคเบาหวาน กลุ่มที่เสี่ยงโรคเบาหวาน และกลุ่มเป็นโรคเบาหวาน ดังนี้

กลุ่มไม่เสี่ยงโรคเบาหวาน	กำหนดให้ y=0
กลุ่มเสี่ยงโรคเบาหวาน	กำหนดให้ y=1
กลุ่มเป็นโรคเบาหวาน	กำหนดให้ y=2

2.2 ตัวแปรอิสระ ที่นำมาศึกษา ประกอบด้วย

2.2.1 อายุ (Age) หน่วยเป็น ปี

2.2.2 เพศ (Sex) (กำหนดให้ 0 = เพศชายและ 1 = เพศหญิง)

2.2.3 ระดับน้ำตาลในเลือด (Blood Glucose) หน่วยเป็น มิลลิกรัม/เดซิลิตร (mg/dl)

2.2.4 ค่าดัชนีมวลกาย (Bmi) หน่วยเป็น กิโลกรัม/เมตร^2 (kg/m^2)

3. ขั้นตอนการศึกษาวิเคราะห์

3.1 สุ่มตัวอย่างซ้ำทั้งหมด จากข้อมูลจริง ที่เป็น ข้อมูล การตรวจสุขภาพประจำปี ของเจ้าหน้าที่ที่ปฏิบัติงานในโรงพยาบาลนพรัตน์ราชธานี ระหว่างเดือนสิงหาคม 2551 - เมษายน 2552 ที่เป็นกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่ม จำนวน 50 ชุดข้อมูล โดยข้อมูลในแต่ละชุดมีจำนวน 599 ค่าสังเกต (แบ่งข้อมูลเป็นชุดข้อมูลสำหรับฝึกสอนจำนวน 2/3 ของข้อมูลทั้งหมดและเป็ นชุดข้อมูลสำหรับทดสอบจำนวน 1/3 ของข้อมูลทั้งหมด โดยใช้วิธีการสุ่ม)

3.2 นำข้อมูลชุด ฝึกสอนแต่ละชุด ที่ได้จากข้อมูลจริงและข้อมูลบูตสแตรปแต่ละชุด มาตรวจสอบข้อตกลง การแจกแจงปกติพหุด้วยวิธีการ Mardia's multivariate skewness and kurtosis ตรวจสอบ การเท่ากันของควา มแปรปรวนร่วมด้วย สถิติทดสอบ Box's M และตรวจสอบ การเท่ากันของค่าเฉลี่ยของแต่ละประชากรด้วยสถิติ Hotelling T^2

3.3 ผลจากการตรวจสอบข้อตกลงในข้อ 3.2

3.3.1 หากข้อมูลมีการแจกแจงแบบปกติพหุ มีความแปรปรวนร่วมเท่ากัน วิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกเชิงเส้น

3.3.2 หากข้อมูลมีการแจกแจงแบบปกติพหุ แต่ความแปรปรวนร่วมไม่เท่ากัน วิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง

3.3.3 หากข้อมูลไม่มีการแจกแจงแบบปกติพหุ และความแปรปรวนร่วมไม่เท่ากัน วิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN โดยที่ k คือจำนวนค่าสังเกตที่อยู่ล้อมรอบค่าสังเกต $\mathbf{x}:(x_1,x_2,x_3,x_4)$ ที่ต้องการถูกจัดเข้ากลุ่มใดกลุ่มหนึ่งใน 3 กลุ่มเสี่ยง ในที่นี้กำหนด k เป็น 3 , 4 และ 5 ในที่นี้ $\mathbf{x}$ เป็นเวกเตอร์ของค่าสังเกต ที่ประกอบด้วย

x_1	คือ อายุ	x_2	คือ เพศ
x_3	คือ ระดับน้ำตาลในเลือด	x_4	คือ ค่าดัชนีมวลกาย

3.4 สร้างกฎการจำแนกที่เหมาะสม จากชุดข้อมูลฝึกสอนแต่ละชุด โดยกำหนดค่าความน่าจะเป็นก่อนหน้า (prior probability) ของการจัดจำแนกหน่วยตัวอย่างหรือค่าสังเกตเข้ากลุ่มที่ 1 (กลุ่มไม่เสี่ยงโรคเบาหวาน) : กลุ่มที่ 2 (กลุ่มเสี่ยงโรคเบาหวาน): กลุ่มที่ 3 (กลุ่มที่เป็นโรคเบาหวาน) ตามลำดับ ดังนี้

3.4.1 จัดสรรด้วยสัดส่วนที่เท่ากัน คือ 0.3333 : 0.3333 : 0.3333

3.4.2 จัดสรรตามสัดส่วน 0.90 : 0.05 : 0.05

3.4.3 จัดสรรตามสัดส่วน 0.80 : 0.10 : 0.10

3.4.4 จัดสรรตามสัดส่วน 0.70 : 0.15 : 0.15

3.5 นำกฎการจำแนกที่ได้จากข้อมูลชุดฝึกสอนจากข้อ 3.4 มาใช้กับข้อมูลชุดทดสอบ มาคำนวณค่าอัตราความผิดพลาด

ค่าอัตราความผิดพลาดของแต่ละชุดข้อมูลทดสอบ $(ER)_i$ คำนวณจาก

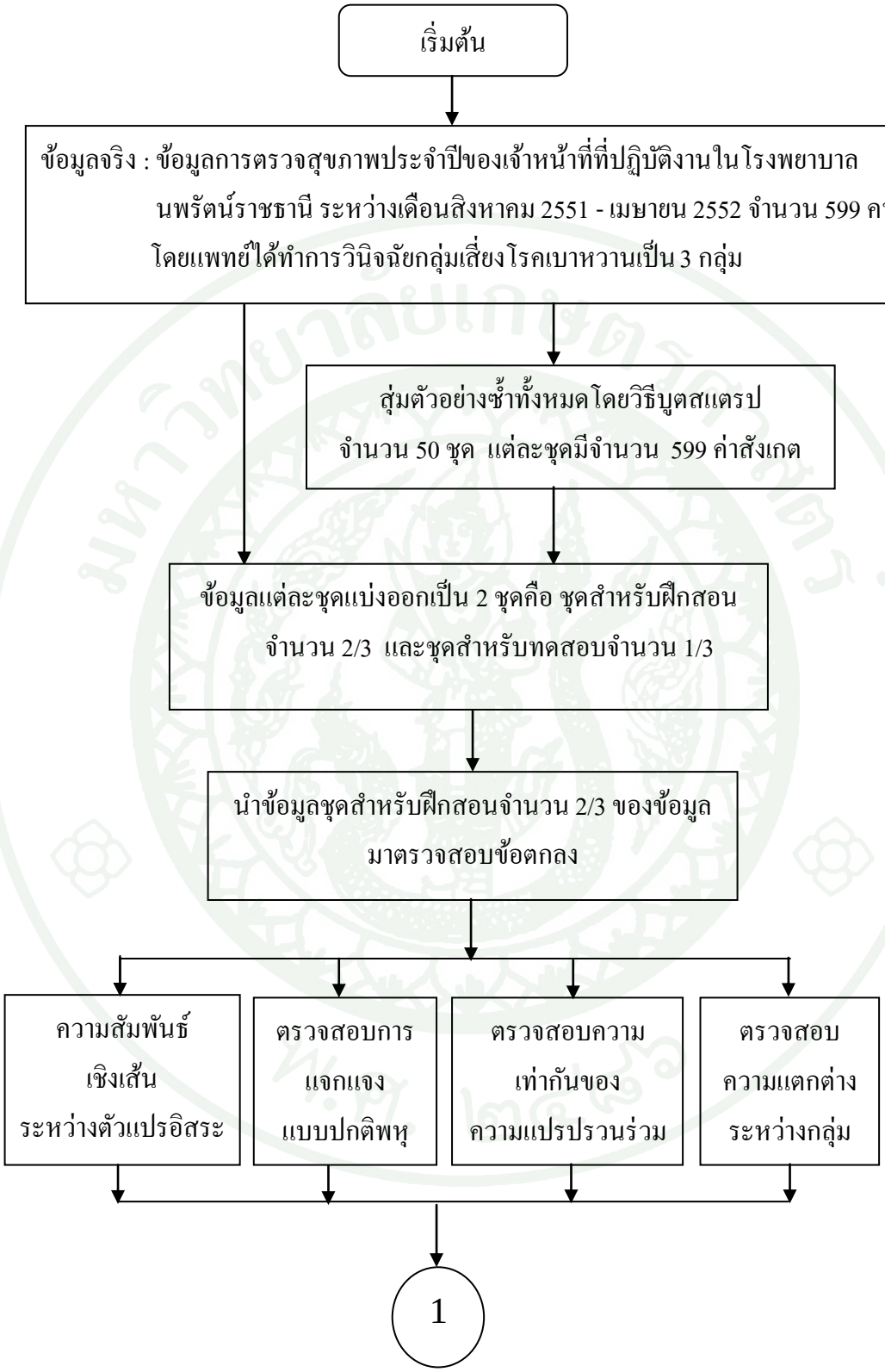
$$(ER)_i = \frac{n_{12}+n_{13}+n_{21}+n_{23}+n_{31}+n_{32}}{n_i};\ i=1,...,50$$

อัตราความผิดพลาดเฉลี่ย $E(ER)_i$ คำนวณจากผลรวมของค่าอัตราความผิดพลาดทั้ง 50 ชุดข้อมูลหารด้วยจำนวนชุดข้อมูลทั้งหมด

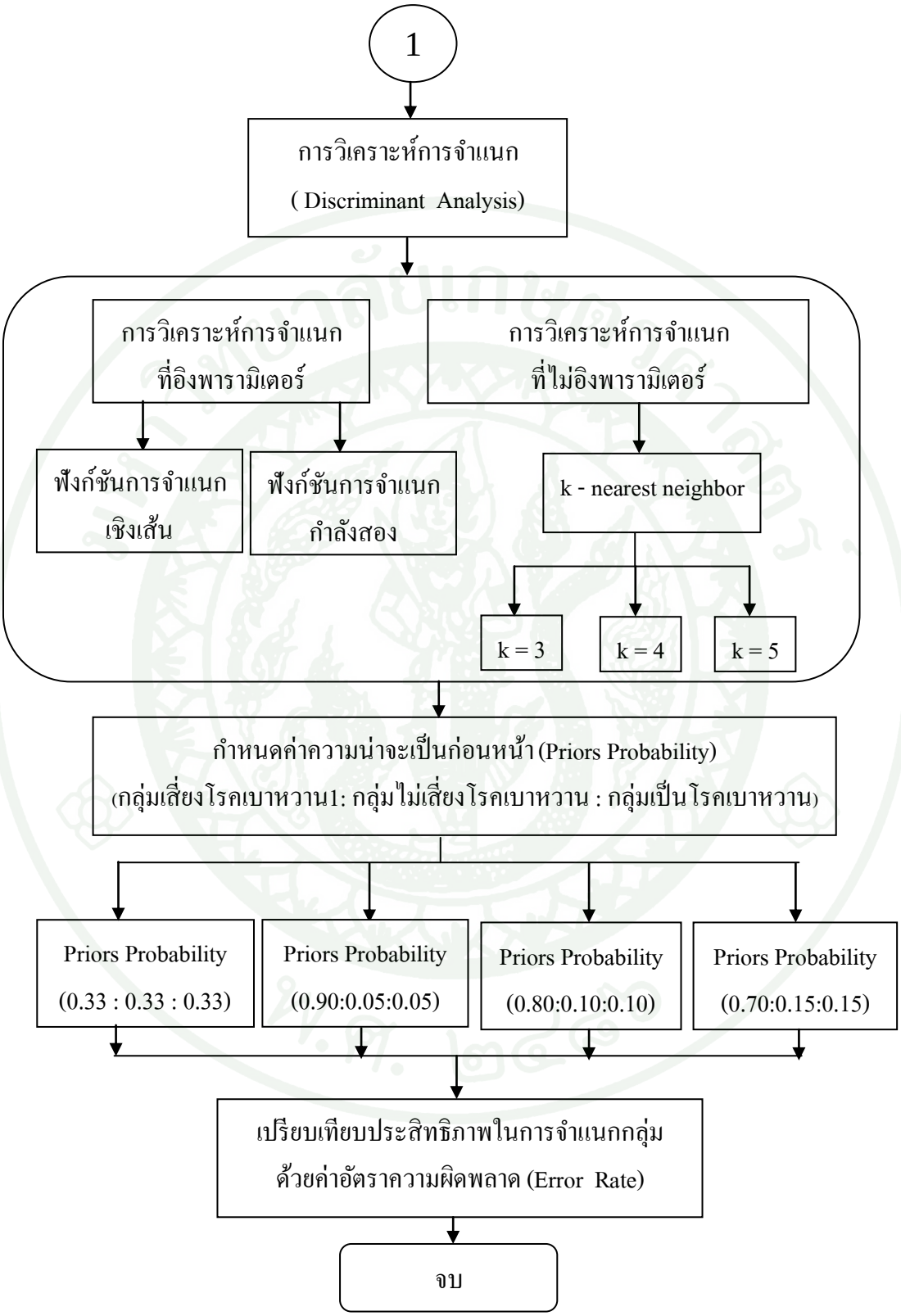
$$E(ER)_i = \frac{\sum_{i=1}^{50}(ER)_i}{50}$$

3.6 เปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มด้วยวิธีการวิเคราะห์การจำแนก ที่ได้จากข้อ 3.3 โดยวัดจากอัตราความผิดพลาดเฉลี่ย วิธีการวิเคราะห์การจำแนกที่ให้ค่าอัตราความผิดพลาดเฉลี่ยต่ำสุด จะเป็นวิธีการที่เหมาะสมที่สุดกับข้อมูลกลุ่มเสี่ยงโรคเบาหวาน

3.7 ทำซ้ำข้อ 3.2 – 3.6 กับข้อมูลจริง แล้วนำอัตราความผิดพลาดเฉลี่ยจากข้อมูลบูตสแตรป เปรียบเทียบกับอัตราความผิดพลาดที่ได้จากข้อมูลจริง



ภาพที่ 1 แผนผังการศึกษาเปรียบเทียบประสิทธิภาพของการวิเคราะห์การจำแนก



ภาพที่ 1 (ต่อ)

ผลและวิจารณ์

ผล

เนื่องจากวัตถุประสงค์ของการศึกษานี้ต้องการเปรียบเทียบประสิทธิภาพของ การวิเคราะห์การจำแนกระหว่างวิธีที่อิงพารามิเตอร์ โดยใช้ฟังก์ชัน การจำแนกเชิงเส้น และฟังก์ชันการจำแนกกำลังสอง กับวิธีที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN เมื่อมีความไม่สมดุลของข้อมูลในแต่ละกลุ่ม ในการจำแนกกลุ่มเสี่ยงโรคเบาหวาน โดยพิจารณาจากค่าอัตราความผิดพลาด ที่ต่ำที่สุด ในการ จำแนกกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่ม

ผลการศึกษาสรุปได้ 3 ส่วนดังนี้

1. ลักษณะทั่วไปของข้อมูลกลุ่มเสี่ยงโรคเบาหวาน (ข้อมูลจริง)
2. การตรวจสอบข้อตกลง
3. สรุปข้อมูลบูตสแตรป 50 ชุด
4. ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มเสี่ยงโรคเบาหวาน โดยใช้ฟังก์ชันการจำแนกเชิงเส้น ฟังก์ชันการจำแนกกำลังสอง และ kNN เมื่อกำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วนที่เท่ากันและด้วยสัดส่วนที่แตกต่างกันของข้อมูลบูตสแตรป
5. เปรียบเทียบอัตราความผิดพลาดในการจำแนกเสี่ยงโรคเบาหวานที่ได้จากวิธีบูตสแตรปกับการศึกษาด้วยข้อมูลจริง

1. ลักษณะทั่วไปของข้อมูลกลุ่มเสี่ยงโรคเบาหวาน (ข้อมูลจริง)

ข้อมูลที่ใช้ในการ ศึกษา ครั้งนี้เป็น ข้อมูล การตรวจสุขภาพประจำปี ของ เจ้า หน้าที่ ที่ปฏิบัติงานใน โรงพยาบาล นพรัตน์ราชธานี ระหว่างเดือนสิงหาคม 2551 - เมษายน 2552 จำนวน 599 คน โดยแพทย์ได้ทำการวินิจฉัยกลุ่มเสี่ยงโรคเบาหวานของเจ้าหน้าที่แบ่งเป็น 3 กลุ่มคือ

1) กลุ่มไม่เสี่ยงโรคเบาหวาน จำนวน 561 คน คิดเป็น 90%
2) กลุ่มเสี่ยงโรคเบาหวาน จำนวน 17 คน คิดเป็น 5%
3) กลุ่มที่เป็นโรคเบาหวาน จำนวน 21 คน คิดเป็น 5%

ตารางที่ 3 ค่าสถิติพื้นฐานของตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณ 3 ตัว ในกลุ่มที่ไม่เสี่ยงโรคเบาหวานจำนวน 561 ค่าสังเกต (กลุ่มที่ 1)

	อายุ	ระดับน้ำตาลในเลือด	ค่าดัชนีมวลกาย
ค่าเฉลี่ย	43.523	108.333	25.808
ค่าเบี่ยงเบนมาตรฐาน	4.996	6.521	4.681
ค่าความเบ้	0.365	1.172	0.317
ค่าความโด่ง	0.784	0.656	-0.498

ตารางที่ 4 ค่าสถิติพื้นฐานของตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณ 3 ตัว ในกลุ่มที่เสี่ยงโรคเบาหวานจำนวน 17 ค่าสังเกต (กลุ่มที่ 2)

	อายุ	ระดับน้ำตาลในเลือด	ค่าดัชนีมวลกาย
ค่าเฉลี่ย	42.205	79.761	23.665
ค่าเบี่ยงเบนมาตรฐาน	5.433	8.025	3.562
ค่าความเบ้	0.008	-0.006	0.713
ค่าความโด่ง	0.416	0.090	0.481

ตารางที่ 5 ค่าสถิติพื้นฐานของตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณ 3 ตัว ในกลุ่มที่เป็นโรคเบาหวานจำนวน 21 ค่าสังเกต (กลุ่มที่ 3)

	อายุ	ระดับน้ำตาลในเลือด	ค่าดัชนีมวลกาย
ค่าเฉลี่ย	44.418	155.705	26.911
ค่าเบี่ยงเบนมาตรฐาน	6.919	6.673	4.463
ค่าความเบ้	-0.065	1.626	0.479
ค่าความโด่ง	0.729	2.036	-0.747

จากตารางที่ 3, 4 และ 5 จะเห็นว่า อายุเฉลี่ยทั้ง 3 กลุ่มใกล้เคียงกัน อยู่ในช่วง 42 – 44 ปี ส่วนการกระจายของอายุกลุ่มที่เป็นโรคเบาหวานมีค่าเบี่ยงเบนมาตรฐานมากกว่ากลุ่มอื่น ส่วนค่าระดับน้ำตาลในเลือดมีค่าเฉลี่ยแตกต่างกันมาก โดยเฉพาะกลุ่มเป็นโรคเบาหวานมีค่าสูงถึง 155.71 mg/dl ค่าเบี่ยงเบนมาตรฐาน 6.67 mg/dl เมื่อเทียบกับกลุ่มไม่เสี่ยงโรคเบาหวานค่าเฉลี่ยน้ำตาลในเลือดเท่ากั บ 79.76 mg/dl แต่มีการกระจายของค่าระดับน้ำตาลในเลือดมาก คือ ค่าเบี่ยงเบนมาตรฐานเท่ากับ 8.02 mg/dl

สำหรับค่าดัชนีมวลกาย จะเห็นว่า กลุ่มเป็นโรคเบาหวานมีค่าเฉลี่ยมากที่สุด คือ 26.92 kg/m^2 ค่าเบี่ยงเบนมาตรฐาน 4.46 kg/m^2 เมื่อเทียบกับกลุ่มเสี่ยงโรคเบาหวาน มี ค่าใกล้เคียงกัน แต่จะมีค่าแตกต่างจากกลุ่มไม่เสี่ยงโรคเบาหวานอย่างเห็นได้ชัดเจน

เมื่อพิจารณา จากค่าความเบ้ ความโด่ง ของตัวแปรแต่ละตัวจากทั้ง 3 กลุ่ม จะเห็นว่า มีเพียงตัวแปรระดับน้ำตาลในเลือดของกลุ่มไม่เสี่ยงโรคเบาหวาน มีการแจกแจงใกล้เคียงการแจกแจงปกติ เพราะมีค่าความเบ้และความโด่งเกือบเท่ากับศูนย์ นอกนั้นมีค่าความเบ้และค่าความโด่งแตกต่างจากศูนย์อย่างมาก สรุปได้ว่า ตัวแปรอื่นแต่ละตัวที่เหลือในแต่ละกลุ่มไม่มีการแจกแจงปกติ

2. การตรวจสอบข้อตกลง

2. 1 ตัวแปรอิสระแต่ละตัวเป็นอิสระต่อกัน

เมื่อนำตัวแปรอิสระ 3 ตัวคือ อายุ, ระดับน้ำตาลในเลือด, ค่าดัชนีมวลกาย ซึ่งเป็นตัวแปรเชิงปริมาณ มาวิเคราะห์ความสัมพันธ์จากค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน แสดงโดย เมทริกซ์ความสัมพันธ์ ดังนี้

	age	**bloodche**	**bmi**
age	1.000	0.202	0.045
bloodche	0.202	1.000	0.162
bmi	0.045	0.162	1.000

ผลการวิเคราะห์จาก ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันจะเห็นว่าไม่มีตัวแปรคู่ใดเลยที่มีค่า $r_{ij}>0.8$ จึงสรุปได้ว่า ตัวแปรเชิงปริมาณทั้ง 3 ตัวแปร มีความสัมพันธ์กันแบบเชิงเส้นในระดับต่ำ

2. 2 การแจกแจงแบบปกติพหุ

เมื่อนำข้อมูลตัวแปรเชิงปริมาณ 3 ตัวมา คือ อายุ ระดับน้ำตาลในเลือด และดัชนีมวลกาย มาทดสอบการแจกแจงแบบปกติพหุ โดยวิธี Mardia's Multivariate Kurtosis ว่าตัวแปรอิสระ $\mathbf{x}$ ในแต่ละกลุ่มมีการแจกแจงแบบปกติพหุหรือไม่ ภายใต้สมมติฐานการทดสอบ

H_0 : ตัวแปรอิสระ $\mathbf{x}$ มีการแจกแจงแบบปกติพหุ

H_1 : ตัวแปรอิสระ $\mathbf{x}$ ไม่มีการแจกแจงแบบปกติพหุ

ตารางที่ 6 ผลการทดสอบการแจกแจงแบบปกติพหุของตัวแปรเชิงปริมาณ 3 ตัวด้วยวิธี Mardia's Multivariate Kurtosis ในกลุ่มไม่เสี่ยงโรคเบาหวาน

	BETA1HAT	**KAPPA1**	**PVALSKEW**
Based on skewness	3.762	351.781	0
	BETA2HAT	**KAPPA2**	**PVALKURT**
Based on kurtosis	27.113	5.322	1.03E-07

จากตารางที่ 6 ในกลุ่มไม่เสี่ยงโรคเบาหวานมี ค่า P-value ของค่าความเบ้และความโด่งมีค่าน้อยกว่าระดับนัยสำคัญที่ 0.05 ดังนั้นจึงปฏิเสธ H_0 สรุปได้ว่าข้อมูลไม่มีการแจกแจงปกติพหุอย่างมีนัยสำคัญทางสถิติ

ตารางที่ 7 ผลการทดสอบการแจกแจงแบบปกติพหุของตัวแปรเชิงปริมาณ 3 ตัวด้วยวิธี Mardia's Multivariate Kurtosis ในกลุ่มเสี่ยงโรคเบาหวาน

	BETA1HAT	**KAPPA1**	**PVALSKEW**
Based on skewness	5.660	19.811	0.469
	BETA2HAT	**KAPPA2**	**PVALKURT**
Based on kurtosis	20.158	-1.270	0.203

จากตารางที่ 7 ในกลุ่มเสี่ยงโรคเบาหวานมี ค่า P-value ของค่าความเบ้และความโด่งมีค่ามากกว่าระดับนัยสำคัญที่ 0.05 ดังนั้นจึงยอมรับ H_0 สรุปได้ว่าข้อมูลมีการแจกแจงปกติพหุอย่างมีนัยสำคัญทางสถิติ

ตารางที่ 8 ผลการทดสอบการแจกแจงแบบปกติพหุของตัวแปรเชิงปริมาณ 3 ตัวด้วยวิธี Mardia's Multivariate Kurtosis ในกลุ่มเป็นโรคเบาหวาน

	BETA1HAT	**KAPPA1**	**PVALSKEW**
Based on skewness	10.089	28.585	0.096
	BETA2HAT	**KAPPA2**	**PVALKURT**
Based on kurtosis	24.623	0.185	0.852

จากตารางที่ 8 ในกลุ่มเป็นโรคเบาหวานมี ค่า P-value ของค่าความเบ้และความโด่งมีค่ามากกว่าระดับนัยสำคัญที่ 0.05 ดังนั้นจึงยอมรับ H_0 สรุปได้ว่าข้อมูลมีการแจกแจงปกติพหุอย่างมีนัยสำคัญทางสถิติ

ผลการทดสอบจากตารางที่ 6 – ตารางที่ 8 สรุปได้ว่า มีกลุ่มไม่เสี่ยงโรคเบาหวานที่ไม่มีการแจกแจงปกติพหุ ส่วนกลุ่มเสี่ยงโรคเบาหวานและกลุ่มเป็นโรคเบาหวาน มีการแจกแจงปกติพหุ

2. 3 การเท่ากันของค่าความแปรปรวนร่วมของกลุ่มเสี่ยงแต่ละกลุ่ม

ผลการวิเคราะห์การเท่ากันของค่าความแปรปรวนร่วมของกลุ่มเสี่ยงแต่ละกลุ่ม แสดงโดย เมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระในแต่ละกลุ่ม แสดงนี้

		age	**bloodche**	**bmi**
กลุ่มไม่เสี่ยงโรคเบาหวาน	**age**	16533.426	4081.469	570.674
	bloodche	4081.469	36063.993	2851.894
	bmi	570.674	2851.894	7108.067

		age	**bloodche**	**bmi**
กลุ่มเสี่ยงโรคเบาหวาน	**age**	499.238	-236.667	31.337
	bloodche	-236.667	850.667	-211.478
	bmi	31.337	-211.478	438.337

		age	bloodche	bmi
กลุ่มเป็นโรคเบาหวาน	**age**	766.118	1274.059	67.385
	bloodche	1274.059	64869.529	1317.465
	bmi	67.385	1317.465	318.726

จากเมทริกซ์ข้างต้น พบว่าความแปรปรวนร่วมในแต่ละกลุ่มจะมีค่าที่แตกต่างกันมาก โดยเฉพาะค่าความแปรปรวนร่วมของกลุ่มที่ไม่เป็นโรคเบาหวานจะมีค่าที่สูงกว่ากลุ่มอื่นๆ มาก

และเมื่อใช้การทดสอบด้วยสถิติทดสอบ Box'M ในการตรวจสอบการเท่ากันของเมทริกซ์ความแปรปรวนร่วม ภายใต้สมมติฐานการทดสอบ

$H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3$

H_1 : มีอย่างน้อย 1 คู่ คือ $\Sigma_i \neq \Sigma_j$ เมื่อ $i \neq j$

ให้ผลการทดสอบ ดังตารางที่ 9

ตารางที่ 9 ผลการทดสอบด้วยวิธี Box's M และประมาณค่าด้วยสถิติไคกำลังสอง

Chi-Square	df	P-value
522.098	20	<.0001

ผลการวิเคราะห์โดยสถิติ Box'M ที่มีการแจกแจงแบบ Chi–Square ได้ค่าเท่ากับ 522.098 และค่า P–value มีค่า <.0001 ซึ่งน้อยกว่าระดับนัยสำคัญที่ 0.05 จึงสรุปได้ว่าข้อมูลทั้ง 3 กลุ่มมีเมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระไม่เท่ากันอย่างมีนัยสำคัญทางสถิติ

2. 4 ความแตกต่างระหว่างกลุ่ม

เมื่อใช้การทดสอบด้วยสถิติ Hotelling T^2 ในการตรวจสอบความแตกต่างระหว่างกลุ่มภายใต้สมมติฐานการทดสอบ

$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$

H_1: มีอย่างน้อย 1 คู่ คือ $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$ เมื่อ $i \neq j$

ให้ผลการทดสอบ ดังตารางที่ 10

ตารางที่ 10 ผลการทดสอบสถิติ Hotelling T^2 และประมาณค่าด้วยสถิติ F

Hotelling T2	F value	P-value
1.089	53.44	<.0001

ผลการวิเคราะห์โดยวิธี Hotelling T^2 พบว่า สถิติ Hotelling T^2 มีค่าเท่ากับ 1.089 และค่า P–value น้อยกว่า 0.0001 ที่ระดับนัยสำคัญ 0.05 จึงสรุปได้ว่าข้อมูลทั้ง 3 กลุ่มมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ

ผลการตรวจสอบข้อตกลงการแจกแจงปกติพหุ และการเท่ากันของความแปรปรวนร่วมสำหรับการวิเคราะห์ที่อิงพารามิเตอร์ ปรากฏว่า ข้อมูลที่นำมาศึกษาไม่มีคุณสมบัติตามข้อตกลงที่กำหนดข้างต้น

3. ผลการสุ่มตัวอย่างซ้ำทั้งหมดจากข้อมูลจริง (ข้อมูลบูตสแตรป)

จากการสุ่มตัวอย่างซ้ำแบบใส่คืนจากข้อมูลจริง ขึ้นมาจำนวน 50 ชุดข้อมูลและแต่ละชุดมีจำนวนข้อมูล 599 ค่าสังเกต แสดงในตารางภาคผนวกที่ 1 พบว่า ข้อมูลบูตสแตรปแต่ละชุด ที่สุ่มได้มีข้อมูลในกลุ่มไม่เสี่ยงโรคเบาห วาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน เป็นสัดส่วน 0.90 : 0.05 : 0.05 ประมาณ 40 ชุดข้อมูล คิดเป็น 80% ของข้อมูลทั้งหมด และที่เหลือเป็นสัดส่วน 0.80 : 0.10 : 0.10 ประมาณ 10 ชุดข้อมูล คิดเป็น 20% ของข้อมูลทั้งหมด

4. ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มเสี่ยงโรคเบาหวาน

การจำแนกกลุ่มเสี่ยงโรคเบาหวาน ใช้การสุ่มตัวอย่างซ้ำทั้งหมดโดยวิธีบูตสแตรป จำนวน 50 ชุดข้อมูล โดยแต่ละชุดมีจำนวน 599 ค่าสังเกต โดย ถูกแบ่งอย่างสุ่มออก เป็นชุดข้อมูลสำหรับฝึกสอนจำนวน 2/3 ของข้อมูลทั้งหมดและ ที่เหลือเป็นชุดข้อมูล สำหรับทดสอบจำนวน 1/3 ของข้อมูลทั้งหมด ทำ การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ ด้วยฟังก์ชัน การจำแนก เชิงเส้นและฟังก์ชันการจำแนกกำลังสอง และการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN

กำหนดค่าความน่าจะเป็นก่อนหน้าจากค่าสัดส่วน เนื่องจากข้อมูลจริงมีสัดส่วนระหว่างกลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน เป็น 0.90 : 0.05 : 0.05 และ 0.80 : 0.10 : 0.10 ผู้วิจัยจึงกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนจากค่าสัดส่วน ต่อไปนี้

1. จัดสรรตามสัดส่วนที่เท่ากัน 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

ผลการวิเคราะห์ค่าเฉลี่ยอัตราความผิดพลาดของการจำแนกกลุ่มผิดจากข้อมูลบูตสแตรป 50 ชุด แยกตามกลุ่มเสี่ยง 3 กลุ่ม แสดงในตารางที่ 11-13 และค่าเฉลี่ยอัตราความผิดพลาดโดยรวมทั้ง 3 กลุ่ม แสดงในตารางที่ 14

ตารางที่ 11 ค่าเฉลี่ยและความคลาดเคลื่อนมาตรฐานของอัตราความผิดพลาดของกลุ่มไม่เสี่ยงโรคเบาหวานจากชุดข้อมูลทดสอบของตัวอย่างบูตสแตรป 50 ชุด

	ความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน)			
วิธีการวิเคราะห์การจำแนก	**สัดส่วน (0.3333 : 0.3333 : 0.3333)**	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
การวิเคราะห์การจำแนกที่อิงพารามิเตอร์				
ฟังก์ชันการจำแนกเชิงเส้น	0.05 (0.05)	0.04 (0.03)	0.04 (0.04)	0.05 (0.04)
ฟังก์ชันการจำแนกกำลังสอง	0.03 (0.11)	0.02 (0.10)	0.02 (0.11)	0.03 (0.11)
การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์				
k = 3	0.00 (0.01)	**0.00 (0.00)**	0.00 (0.00)	0.00 (0.01)
k = 4	0.00 (0.01)	**0.00 (0.00)**	0.00 (0.01)	0.00 (0.01)
k = 5	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)

จากตารางที่ 11 เป็นการวิเคราะห์การจำแนกของกลุ่มที่ไม่เสี่ยงโรคเบาหวาน โดยในแต่ละชุดตัวอย่างมีจำนวนประมาณ 90 % ของข้อมูล 599 ค่าสังเกต ข้อมูลกลุ่มนี้จะไม่มีการแจกแจงปกติพหุแต่ค่าความแปรปรวนร่วม จะไม่เท่ากับกลุ่มอื่น ผลการวิเคราะห์ ค่าเฉลี่ย อัตราความผิดพลาด พบว่า การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกเชิงเส้นจะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มมากที่สุด

เมื่อใช้การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN จะให้ค่าเฉลี่ยอัตราความผิดพลาดต่ำกว่าวิธีการวิเคราะห์การจำแนกวิธีอื่น ไม่ว่าจะกำหนด k = 3, 4 หรือ 5 และค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน) จะถูกกำหนดเป็นแบบสัดส่วนต่างกัน คือ (0.90 : 0.05 : 0.05), (0.80 : 0.10 : 0.10), (0.70 : 0.15 : 0.15) และ (0.3333 : 0.3333 : 0.3333) ตามลำดับ พบว่า การกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนด้วยสัดส่วน (0.90 : 0.05 : 0.05) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกต่ำที่สุด

สำหรับค่า k = 3 หรือ 4 และค่าความน่าจะเป็นก่อนหน้า ถ้ากำหนดด้วยสัดส่วน (0.90 : 0.05 : 0.05) จะให้ค่าเฉลี่ยอัตราความผิดพลาดที่มีการกระจายน้อยที่สุด ข้อมูลส่วนใหญ่จะถูกจัดเข้ากลุ่มนี้ด้วยความน่าจะเป็นที่สูงมาก ค่าเฉลี่ยอัตราความผิดพลาดจึงต่ำ

ตารางที่ 12 ค่าเฉลี่ยและความคลาดเคลื่อนมาตรฐานของอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวานจากชุดข้อมูลทดสอบของตัวอย่างบูตสแตรป 50 ชุด

	ความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน)			
วิธีการวิเคราะห์การจำแนก	**สัดส่วน (0.3333 : 0.3333 : 0.3333)**	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
การวิเคราะห์การจำแนกที่อิงพารามิเตอร์				
ฟังก์ชันการจำแนกเชิงเส้น	0.13 (0.22)	0.16 (0.24)	0.16 (0.23)	0.14 (0.22)
ฟังก์ชันการจำแนกกำลังสอง	0.09 (0.20)	0.11 (0.21)	0.11 (0.21)	0.10 (0.20)
การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์				
k = 3	**0.09 (0.15)**	0.11 (0.17)	0.09 (0.16)	0.09 (0.16)
k = 4	0.10 (0.16)	0.11 (0.21)	0.11 (0.17)	0.11 (0.17)
k = 5	0.11 (0.17)	0.14 (0.22)	0.14 (0.18)	0.14 (0.18)

จากตารางที่ 12 เป็นการวิเคราะห์การจำแนกของกลุ่มเสี่ยงโรคเบาหวาน ซึ่งข้อมูลในแต่ละชุดตัวอย่างมี ประมาณ 5 % ของข้อมูล 599 ค่าสังเกต ข้อมูลกลุ่มนี้มีการแจกแจงปกติพหุแต่ค่า าความแปรปรวนร่วมจะไม่เท่ากับกลุ่มอื่น ผลการวิเคราะห์ค่าเฉลี่ยอัตราความผิดพลาด พบว่า การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกเชิงเส้นจะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มมากที่สุด เมื่อใช้การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN และการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง จะให้ค่าเฉลี่ยอัตราความผิดพลาดที่ใกล้เคียงกัน

การกำหนด ค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน :กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน) จะถูกกำหนดเป็นแบบสัดส่วน ต่างกัน คือ (0.90 : 0.05 : 0.05), (0.80 : 0.10 : 0.10), (0.70 : 0.15 : 0.15) และ (0.3333 : 0.3333 : 0.3333) ตามลำดับ พบว่าการกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนด้วยสัดส่วนที่เท่ากัน คือ (0.3333 : 0.3333 : 0.3333) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกต่ำที่สุด

สำหรับ k = 3 และกำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วนที่เท่ากัน คือ (0.3333 : 0.3333 : 0.3333) จะให้ค่าเฉลี่ยอัตราความผิดพลาดที่มีการกระจายน้อยที่สุด ข้อมูลในกลุ่มนี้มีจำนวนน้อยประมาณ 5 % ของข้อมูลทั้งหมด ในการจัดข้อมูลเข้ากลุ่มนี้ด้วยความน่าจะเป็นที่เท่ากันจะให้ค่าเฉลี่ยอัตราความผิดพลาดต่ำสุด

ตารางที่ 13 ค่าเฉลี่ยและความคลาดเคลื่อนมาตรฐานของอัตราความผิดพลาดของกลุ่มเป็นโรคเบาหวานจากชุดข้อมูลทดสอบของตัวอย่างบูตสแตรป 50 ชุด

	ความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวาน)			
วิธีการวิเคราะห์การจำแนก	**สัดส่วน (0.3333 : 0.3333 : 0.3333)**	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
การวิเคราะห์การจำแนกที่อิงพารามิเตอร์				
ฟังก์ชันการจำแนกเชิงเส้น	0.26 (0.27)	0.24 (0.25)	0.24 (0.27)	0.25 (0.25)
ฟังก์ชันการจำแนกกำลังสอง	0.08 (0.17)	0.09 (0.18)	0.09 (0.18)	0.09 (0.18)
การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์				
k = 3	0.07 (0.57)	0.10 (0.17)	0.10 (0.17)	0.08 (0.14)
k = 4	**0.07 (0.15)**	0.11 (0.18)	0.10 (0.17)	0.08 (0.15)
k = 5	0.09 (0.16)	0.11 (0.18)	0.10 (0.18)	0.11 (0.19)

จากตารางที่ 13 เป็นการวิเคราะห์การจำแนกของกลุ่มเป็นโรคเบาหวาน ซึ่งข้อมูลในแต่ละชุดตัวอย่างมีประมาณ 5 % ของข้อมูล 599 ค่าสังเกต ข้อมู ลกลุ่มนี้มีการแจกแจงปกติพหุแต่ค่าความแปรปรวนร่วมจะไม่เท่ากับกลุ่มอื่น ผลการวิเคราะห์ค่าเฉลี่ยอัตราความผิดพลาด พบว่า การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกเชิงเส้นจะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มมากที่สุด เมื่อใช้การวิเครา ะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN และการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง จะให้ค่าเฉลี่ยอัตราความผิดพลาดที่ใกล้เคียงกัน

การกำหนดค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโร คเบาหวาน) จะถูกกำหนดเป็นแบบสัดส่วนต่างกัน คือ (0.90 : 0.05 : 0.05), (0.80 : 0.10 : 0.10), (0.70 : 0.15 : 0.15) และ (0.3333 : 0.3333 : 0.3333) ตามลำดับ พบว่าการกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนด้วยสัดส่วนที่เท่ากัน คือ (0.3333 : 0.3333 : 0.3333) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกต่ำที่สุด

สำหรับ k = 4 และกำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วนที่เท่ากัน คือ (0.3333 : 0.3333 : 0.3333) จะให้ค่าเฉลี่ยอัตราความผิดพลาดที่มีการกระจายน้อยที่สุด ข้อมูลในกลุ่มนี้มีจำนวนน้อยประมาณ 5 % ของข้อมูลทั้งหมด ในการจัดข้อมูลเข้ากลุ่มนี้ด้วยความน่าจะเป็นที่เท่ากันจะให้ค่าเฉลี่ยอัตราความผิดพลาดต่ำสุด

ตารางที่ 14 ค่าเฉลี่ยและความคลาดเคลื่อนมาตรฐานของอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่ม จากชุดข้อมูลทดสอบของตัวอย่างบูตสแตรป 50 ชุด

	ความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน)			
วิธีการวิเคราะห์การจำแนก	**สัดส่วน (0.3333 : 0.3333 : 0.3333)**	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
การวิเคราะห์การจำแนกที่อิงพารามิเตอร์				
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)
การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์				
k = 3	0.05 (0.07)	**0.02 (0.02)**	0.02 (0.02)	0.03 (0.03)
k = 4	0.06 (0.08)	**0.02 (0.02)**	0.02 (0.03)	0.03 (0.04)
k = 5	0.07 (0.08)	0.03 (0.06)	0.05 (0.04)	0.05 (0.04)

จากตารางที่ 14 เป็นการวิเคราะห์ การจำแนก ของกลุ่มเสี่ยง โรคเบาหวานทั้ง 3 กลุ่ม ซึ่งข้อมูลกลุ่มเสี่ยงโรคเบาหวานทั้งสามกลุ่มพบว่า กลุ่มไม่เสี่ยงโรคเบาหวานเพียงกลุ่มเดียวที่ไม่มีการแจกแจงปกติพหุ และการเท่ากันของความแปรปรวนร่วมทั้ง 3 กลุ่มไม่เท่ากัน ข้อมูลมีสัดส่วนของกลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวาน เท่ากับ 0.90 : 0.05 : 0.05

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN จะให้ค่า เฉลี่ย อัตราความผิดพลาดในการจำแนกกลุ่มที่ต่ำที่สุดเมื่อเปรียบเทียบกับการวิเคราะห์การจำแนกแบบอื่น ส่วนการวิเคราะห์การจำแนก ที่อิงพารามิเตอร์ ด้วยฟังก์ชันการจำแนกเชิงเส้นจะให้ ค่าเฉลี่ย อัตราความผิดพลาดในการจำแนกกลุ่มมากที่สุด การกำหนดค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน) จะถูกกำหนดเป็นแบบสัดส่วนต่างกัน พบว่า การกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนด้วยสัดส่วน (0.90 : 0.05 : 0.05) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกต่ำที่สุด

สำหรับ k = 3 หรือ 4 และการ กำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วน ที่เท่ากับ (0.90 : 0.05 : 0.05) ซึ่งใกล้เคียงกับสัดส่วนของข้อมูล จะให้ค่าเฉลี่ยอัตราความผิดพลาดที่มีการกระจายน้อยที่สุด

4. เปรียบเทียบอัตราความผิดพลาดในการจำแนกกลุ่มเสี่ยงโรคเบาหวานที่ได้จากวิธีบูตสแตรปกับการศึกษาด้วยข้อมูลจริง

การศึกษาเปรียบเทียบกับข้อมูลจริง มีวัตถุประสงค์เพื่อเปรียบเทียบว่าผลที่ได้จากข้อมูลจริงและผลที่ได้จากการข้อมูลบูตสแตรปมีความสอดคล้องกันหรือไม่ ประกอบด้วย

4.1 ค่าอัตราความผิดพลาดในการจำแนกกลุ่มเสี่ยงโรคเบาหวานที่ได้จากชุดข้อมูลทดสอบของข้อมูลจริง

4.2 การเปรียบเทียบค่าอัตราความผิดพลาดในการจำแนกกลุ่มเสี่ยงโรคเบาหวานที่ได้จากชุดข้อมูลทดสอบของข้อมูลจริงและชุดข้อมูลทดสอบของตัวอย่างบูตสแตรป 50 ชุด

4.1 ค่าอัตราความผิดพลาดในการจำแนกกลุ่มเสี่ยงโรคเบาหวานที่ได้จากชุดข้อมูลทดสอบของข้อมูลจริง

การจำแนกกลุ่มเสี่ยงโรคเบาหวานจากข้อมูลจริงจำนวน 1 ชุดข้อมูล มีจำนวน 599 ค่าสังเกต โดยถูกแบ่งอย่างสุ่มออกเป็นชุดข้อมูลสำหรับฝึกสอนจำนวน 2/3 ของข้อมูลทั้งหมดและที่เหลือเป็นชุดข้อมูลสำหรับทดสอบจำนวน 1/3 ของข้อมูลทั้งหมด ทำการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ ด้วยฟังก์ชันการจำแนกเชิงเส้นและฟังก์ชันการจำแนกกำลังสอง และการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN

กำหนดค่าความน่าจะเป็นก่อนหน้าจากค่าสัดส่วน เนื่องจากข้อมูลจริงมีสัดส่วนระหว่างกลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน เป็น 0.9 : 0.05 : 0.05 ผู้วิจัยจึงกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนจากค่าสัดส่วน ต่อไปนี้

1. จัดสรรตามสัดส่วนที่เท่ากัน 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

ผลการวิเคราะห์ค่าอัตราความผิดพลาดของการจำแนกกลุ่มผิดจากข้อมูลจริงจำนวน 1 ชุด แยกตามกลุ่มเสี่ยง 3 กลุ่ม แสดงในตารางที่ 15-17 และค่าเฉลี่ยอัตราความผิดพลาดโดยรวมทั้ง 3 กลุ่ม แสดงในตารางที่ 18

ตารางที่ 15 ค่าอัตราความผิดพลาดของกลุ่มไม่เสี่ยงโรคเบาหวานจากชุดข้อมูลทดสอบของข้อมูลจริง

	ความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน)			
วิธีการวิเคราะห์การจำแนก	**สัดส่วน (0.3333 : 0.3333 : 0.3333)**	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
การวิเคราะห์การจำแนกที่อิงพารามิเตอร์				
ฟังก์ชันการจำแนกเชิงเส้น	0.037	**0.000**	**0.000**	0.005
ฟังก์ชันการจำแนกกำลังสอง	0.042	0.016	0.016	0.021
การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์				
k = 3	0.016	**0.000**	0.016	0.016
k = 4	0.026	0.005	0.026	0.026
k = 5	0.032	**0.000**	0.010	0.032

จากตารางที่ 15 เป็นการวิเคราะห์การจำแนกของกลุ่ม ไม่เสี่ยงโรคเบาหวาน ซึ่งข้อมูลในกลุ่มนี้มีประมาณ 90% ของข้อมูล 599 ค่าสังเกต ข้อมูลกลุ่มนี้ไม่มีการแจกแจงปกติพหุ และความแปรปรวนร่วมไม่เท่ากัน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN และวิธีฟังก์ชันเชิงเส้น จะให้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุดเมื่อเปรียบเทียบกับการวิเคราะห์การจำแนกแบบอื่นๆ โดยเฉพาะเมื่อกำหนดค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวาน) เป็น (0.90 : 0.05 : 0.05) ส่วนการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสองจะให้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มมากที่สุด

แต่ในที่นี้การวิเคราะห์การจำแนกที่อิงพารามิเตอร์จะไม่เหมาะสมกับข้อมูลที่มีการละเมิดข้อตกลงของการแจกแจงปกติพหุและความแปรปรวนร่วมเท่ากัน ดังนั้นการวิเคราะห์ด้วยวิธี kNN มีความน่าเชื่อถือมากกว่า จึงสรุปได้ว่า เมื่อ k = 3 หรือ 5 และการกำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วน ที่เท่ากับ (0.90 : 0.05 : 0.05) ซึ่งใกล้เคียงกับสัดส่วนของข้อมูล จะให้ค่าอัตราความผิดพลาดที่มีค่าต่ำที่สุด

ส่วนในตารางที่ 16 เป็นการวิเคราะห์การจำแนกของกลุ่มเสี่ยงโรคเบาหวาน ซึ่งข้อมูลในกลุ่มนี้มีประมาณ 5% ของข้อมูลทั้งหมด ข้อมูลกลุ่มนี้มีการแจกแจงปกติพหุ แต่ความแปรปรวนร่วมไม่เท่ากัน การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสองจะให้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด ไม่ว่าจะกำหนดค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวาน) เป็นแบบสัดส่วนเท่ากันหรือต่างกันก็ตาม ดังตารางที่ 16

ผลที่ได้สอดคล้องตามทฤษฎี คือ ถ้าข้อมูลแจกแจงปกติพหุแต่ความแปรปรวนร่วมไม่เท่ากัน การใช้ฟังก์ชันการจำแนกกำลังสอง จะให้ค่าอัตราความผิดพ ลาดต่ำที่สุด ไม่ว่าจะกำหนดสัดส่วนเท่ากันหรือสัดส่วนที่เหมือนกับข้อมูล

ตารางที่ 16 ค่าอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวานจากชุดข้อมูลทดสอบของข้อมูลจริง

	ความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน)			
วิธีการวิเคราะห์การจำแนก	**สัดส่วน (0.3333 : 0.3333 : 0.3333)**	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
การวิเคราะห์การจำแนกที่อิงพารามิเตอร์				
ฟังก์ชันการจำแนกเชิงเส้น	0.500	0.666	0.500	0.500
ฟังก์ชันการจำแนกกำลังสอง	**0.166**	**0.166**	**0.166**	**0.166**
การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์				
k = 3	0.500	0.833	0.500	0.500
k = 4	0.666	0.666	0.666	0.666
k = 5	0.666	0.666	0.666	0.666

ตารางที่ 17 ค่าอัตราความผิดพลาดของกลุ่มเป็นโรคเบาหวานจากชุดข้อมูลทดสอบของข้อมูลจริง

	ความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน)			
วิธีการวิเคราะห์การจำแนก	**สัดส่วน (0.3333 : 0.3333 : 0.3333)**	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
การวิเคราะห์การจำแนกที่อิงพารามิเตอร์				
ฟังก์ชันการจำแนกเชิงเส้น	0.571	0.571	0.571	0.571
ฟังก์ชันการจำแนกกำลังสอง	0.142	0.142	0.142	0.142
การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์				
k = 3	**0.000**	**0.000**	**0.000**	**0.000**
k = 4	**0.000**	**0.000**	**0.000**	**0.000**
k = 5	0.142	0.142	0.142	0.142

จากตารางที่ 17 เป็นการวิเคราะห์การจำแนกของกลุ่มเป็นโรคเบาหวาน ข้อมูลในกลุ่มนี้มีประมาณ 5% ของข้อมูลทั้งหมด ที่มีการแจกแจงปกติพหุ แต่ความแปรปรวนร่วมไม่เท่า กัน การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN จะให้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มที่ต่ำที่สุดเมื่อเปรียบเทียบกับการวิเคราะห์การจำแนกแบบอื่นๆ ส่วนการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกเชิงเส้นจะให้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มมากที่ สุด เมื่อ k = 3 หรือ 5 ไม่ว่าจะกำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วน แบบใดก็ตาม จะให้ค่าอัตราความผิดพลาดที่มีค่าต่ำที่สุด

ส่วนตารางที่ 18 เป็นการวิเคราะห์การจำแนกของข้อมูลจริงในกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่มพบว่า กลุ่มไม่เสี่ยงโรคเบาหวานเพียงก ลุ่มเดียวที่ไม่มีการแจกแจงปกติพหุ และการเท่ากันของความแปรปรวนร่วมทั้ง 3 กลุ่มไม่เท่ากัน ข้อมูลมีสัดส่วนของ กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวาน เท่ากับ 0.90 : 0.05 : 0.05

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

ตารางที่ 18 ค่าอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่ม จากชุดข้อมูลทดสอบของข้อมูลจริง

	ความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน : กลุ่มเสี่ยงโรคเบาหวาน : กลุ่มเป็นโรคเบาหวาน)			
วิธีการวิเคราะห์การจำแนก	**สัดส่วน (0.3333 : 0.3333 : 0.3333)**	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
การวิเคราะห์การจำแนกที่อิงพารามิเตอร์				
ฟังก์ชันการจำแนกเชิงเส้น	0.203	0.061	0.107	0.164
ฟังก์ชันการจำแนกกำลังสอง	0.117	**0.029**	0.043	0.061
การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์				
k = 3	0.172	0.041	0.062	0.086
k = 4	0.231	0.038	0.088	0.118
k = 5	0.280	0.040	0.089	0.143

4.2 การเปรียบเทียบค่าอัตราความผิดพลาดในการจำแนกกลุ่มเสี่ยงโรคเบาหวานที่ได้จากชุดข้อมูลทดสอบของข้อมูลจริงและชุดข้อมูลทดสอบของตัวอย่างบูตสแตรป50 ชุด

การเปรียบเทียบค่าอัตราความผิดพลาดในการจำแนกกลุ่มเสี่ยงโรคเบาหวานที่ได้จากชุดข้อมูลทดสอบของข้อมูลจริงและชุดข้อมูลทดสอบของตัวอย่างบูตสแ ตรป 50 ชุด แสดงในตารางภาคผนวกที่ 3 พบว่า ข้อมูลบูตสแตรปจะมีค่าเฉลี่ยอัตราความผิดพลาดต่ำที่สุด เมื่อวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธี kNN เมื่อ k=3 หรือ 4 และกำหนดค่าความน่าจะเป็น ก่อนหน้าจากสัดส่วนที่ใกล้เคียงกับข้อมูลจริง คือ (0.90 : 0.05 : 0.05)

ส่วนข้อมูลจริง ซึ่งเป็นข้อมูลเพียงชุดเดียว มีค่าอัตราความผิดพลาดต่ำที่สุด เมื่อวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง และกำหนดค่าความน่าจะเป็น ก่อนหน้าจากสัดส่วนที่ใกล้เคียงกับข้อมูลจริง คือ (0.90 : 0.05 : 0.05)

สรุปและข้อเสนอแนะ

สรุป

เนื่องจาก วัตถุประสงค์ ของการศึกษานี้ 1) ต้องการเปรียบเทียบประสิทธิภาพของการวิเคราะห์การจำแนกระหว่างวิธีที่อิงพารามิเตอร์ โดยใช้ฟังก์ชันการจำแนกเชิงเส้น และฟังก์ชันการจำแนกกำลังสอง กับวิธีที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อมีความไม่ สมดุลของข้อมูลในแต่ละกลุ่ม ในการจำแนกกลุ่มเสี่ยงโรคเบาหวาน โดยพิจารณาจากค่าอัตราความผิดพลาดที่ต่ำที่สุดในการจำแนกกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่ม 2) เพื่อประยุกต์วิธีการวิเคราะห์การจำแนกที่เหมาะสมกับการจำแนกกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่ม กรณีที่มีความไม่สมดุลของข้อมูลในแต่ละกลุ่ม

1. สำหรับวัตถุประสงค์ข้อ 1) ใช้ผลจากตัวอย่างบูตสแตรป

1.1 ค่าเฉลี่ยอัตราความผิดพลาดของกลุ่มที่ไม่เสี่ยงโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อ k=3หรือ 4 และกำหนดค่าความน่าจะเป็น ก่อนหน้าเป็นสัดส่วน (0.90 : 0.05 : 0.05) ที่ใกล้เคียงกับข้อมูลจริง จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุดเท่ากับ 0.00 ค่าคลาดเคลื่อนมาตรฐานเท่ากับ 0.00

1.2 ค่าเฉลี่ยอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อ k =3 และกำหนดค่าความน่าจะเป็น ก่อนหน้าเป็นสัดส่วน ที่เท่ากัน คือ (0.3333 : 0.3333 : 0.3333) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุดเท่ากับ 0.09 ค่าคลาดเคลื่อนมาตรฐานเท่ากับ 0.15

1.3 ค่าเฉลี่ยอัตราความผิดพลาดของกลุ่มที่เป็นโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อ k =4 และกำหนดค่าความน่าจะเป็นที่เกิดขึ้นก่อนเป็นสัดส่วน ที่เท่ากัน คือ (0.3333 : 0.3333 : 0.3333) จะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุดเท่ากับ 0.07 ค่าคลาดเคลื่อนมาตรฐานเท่ากับ 0.15

1.4 ค่าเฉลี่ยอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่ม

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อ k =3 หรือ 4 และกำหนดค่าความน่าจะเป็นก่อนหน้าเป็นสัดส่วน (0.90 : 0.05 : 0.05) ที่ใกล้เคียงกับข้อมูลจริงจะให้ค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด เท่ากับ 0.02 ค่าคลาดเคลื่อนมาตรฐานเท่ากับ 0.02

2. สำหรับวัตถุประสงค์ข้อ 2) ประยุกต์กับข้อมูลจริง

2.1 อัตราความผิดพลาดของกลุ่มไม่เสี่ยงโรคเบาหวาน

การวิเคราะห์ การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อ k =3 หรือ 5 และกำหนดค่าความน่าจะเป็นก่อนหน้าเป็นสัดส่วน (0.90 : 0.05 : 0.05) ที่ใกล้เคียงกับข้อมูลจริงจะให้อัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุดเท่ากับ 0.000

2.2 อัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวาน

การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง จะให้อัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุดเท่ากับ 0.166 แต่การกำหนดค่าความน่าจะเป็นก่อนหน้าไม่มีผลต่ออัตราความผิดพลาดในการจำแนกกลุ่ม

2.3 อัตราความผิดพลาดของกลุ่มเป็นโรคเบาหวาน

การวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อ k =3 หรือ 4 จะให้อัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุดเท่ากับ 0.000 และการกำหนดค่าความน่าจะเป็นก่อนหน้าไม่มีผลต่ออัตราความผิดพลาดในการจำแนกกลุ่ม

2.4 ค่าอัตราความผิดพลาดของกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่ม

การวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง และกำหนดค่าความน่าจะเป็นก่อนหน้าด้วยสัดส่วน (0.90 : 0.05 : 0.05) ที่ใกล้เคียงกับข้อมูลจริงจะให้อัตราความผิดพลาดในการจำแนกกลุ่มต่ำที่สุดเท่ากับ 0.029

3. เปรียบเทียบผลการศึกษาระหว่างข้อมูลบูตสแตรปและข้อมูลจริง

แสดงผล ดังตารางที่ 19 – 22

ตารางที่ 19 วิธีการวิเคราะห์การจำแนกที่ให้อัตราความผิดพลาดต่ำสุด เมื่อกำหนดความน่าจะเป็นก่อนหน้าด้วยสัดส่วนต่างๆ เปรียบเทียบระหว่างการใช้ข้อมูลบูตสแตรปและข้อมูลจริงในกลุ่มไม่เสี่ยงโรคเบาหวาน

	ความน่าจะเป็นก่อนหน้า (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)**
อิงพารามิเตอร์				
ฟังก์ชันเชิงเส้น		ข้อมูลจริง	ข้อมูลจริง	
ฟังก์ชันกำลังสอง				
ไม่อิงพารามิเตอร์				
k = 3		ข้อมูลจริง บูตสแตรป		
k = 4		บูตสแตรป		
k = 5		ข้อมูลจริง		

กลุ่มไม่เสี่ยงโรคเบาหวาน ข้อมูลในกลุ่มนี้ไม่มีการแจกแจงปกติพหุและมีความแปรปรวนร่วมไม่เท่ากับกลุ่มอื่น วิธีการวิเคราะห์การจำแนกที่ให้ผลตรงกันทั้ง ตัวอย่างบูตสแตรปและในข้อมูลจริงคือ วิธี kNN เมื่อ k=3 และกำหนดค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน: กลุ่มเสี่ยงโรคเบาหวาน: กลุ่มเป็นโรคเบาหวาน) ด้วยสัดส่วน (0.90 : 0.05 : 0.05)

ตารางที่ 20 วิธีการวิเคราะห์การจำแนกที่ให้อัตราความผิดพลาดต่ำสุด เมื่อกำหนดความน่าจะเป็นก่อนหน้าด้วยสัดส่วนต่างๆ เปรียบเทียบระหว่างการใช้ข้อมูลบูตสแตรปและข้อมูลจริงในกลุ่มเสี่ยงโรคเบาหวาน

	ความน่าจะเป็นก่อนหน้า (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)**
อิงพารามิเตอร์ ฟังก์ชันเชิงเส้น				
ฟังก์ชันกำลังสอง	ข้อมูลจริง	ข้อมูลจริง	ข้อมูลจริง	ข้อมูลจริง
ไม่อิงพารามิเตอร์				
k = 3	บูตสแตรป			
k = 4				
k = 5				

กลุ่มเสี่ยงโรคเบาหวาน ข้อมูลในกลุ่มนี้มีการแจกแจงปกติพหุและมีความแปรปรวนร่วมไม่เท่ากับกลุ่มอื่น วิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสองจะให้อัตราความผิดพลาดต่ำสุดสำหรับข้อมูลจริง และวิธี kNN เมื่อ k = 3 จะให้อัตราความผิดพลาดต่ำสุดสำหรับข้อมูลบูตสแตรป

ตารางที่ 21 วิธีการวิเคราะห์การจำแนกที่ให้อัตราความผิดพลาดต่ำสุด เมื่อกำหนดความน่าจะเป็นก่อนหน้าด้วยสัดส่วนต่างๆ เปรียบเทียบระหว่างการใช้ข้อมูลบูตสแตรปและข้อมูลจริงในกลุ่มเป็นโรคเบาหวาน

	ความน่าจะเป็นก่อนหน้า (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)**
อิงพารามิเตอร์ ฟังก์ชันเชิงเส้น				
ฟังก์ชันกำลังสอง				
ไม่อิงพารามิเตอร์ k = 3	ข้อมูลจริง	ข้อมูลจริง	ข้อมูลจริง	ข้อมูลจริง
k = 4	ข้อมูลจริง บูตสแตรป	ข้อมูลจริง	ข้อมูลจริง	ข้อมูลจริง
k = 5				

กลุ่มเป็นโรคเบาหวาน ข้อมูลในกลุ่มนี้มีการแจกแจงปกติพหุและมีความแปรปรวนร่วมไม่เท่ากับกลุ่มอื่น ปรากฏว่า วิธี kNN เมื่อ k = 4 ทั้งตัวอย่างบูตสแตรปและในข้อมูลจริงเมื่อกำหนดค่าความ น่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน :กลุ่มเสี่ยงโรคเบาหวาน :กลุ่มเป็นโรคเบาหวาน) ด้วยสัดส่วนที่เท่ากัน จะให้อัตราความผิดพลาดต่ำสุด

ตารางที่ 22 วิธีการวิเคราะห์การจำแนกที่ให้อัตราความผิดพลาดต่ำสุด เมื่อกำหนดความน่าจะเป็นก่อนหน้าด้วยสัดส่วนต่างๆ เปรียบเทียบระหว่างการใช้ข้อมูลบูตสแตรปและข้อมูลจริงในกลุ่มเสี่ยงโรคเบาหวาน 3 กลุ่ม

	ความน่าจะเป็นก่อนหน้า (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)**
อิงพารามิเตอร์ ฟังก์ชันเชิงเส้น				
ฟังก์ชันกำลังสอง		ข้อมูลจริง		
ไม่อิงพารามิเตอร์ k = 3		บูตสแตรป		
k = 4		บูตสแตรป		
k = 5				

กลุ่มเสี่ยงโร คเบาหวานทั้ง 3 กลุ่ม ข้อมูลไม่มีการแจกแจงปกติพหุและมีความแปรปรวนร่วมไม่เท่ากับกลุ่มอื่น วิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสองจะให้อัตราความผิดพลาดต่ำสุดสำหรับข้อมูลจริง และวิธี kNN เมื่อ k = 3 หรือ 4 จะให้อัตราความผิดพลาดต่ำ สุดเมื่อกำหนดค่าความน่าจะเป็นก่อนหน้า (กลุ่มไม่เสี่ยงโรคเบาหวาน :กลุ่มเสี่ยงโรคเบาหวาน:กลุ่มเป็นโรคเบาหวาน) ด้วยสัดส่วน (0.90 : 0.05 : 0.05) สำหรับข้อมูลบูตสแตรป

จะเห็นว่า ค่าอัตราความผิดพลาดจากข้อมูลจริงเพียงชุดเดียวกับค่าอัตราความผิดพลาดจากข้อมูลบูตสแตรป ในแต่ละกลุ่มเสี่ยงมีที่ได้ผลการใช้วิธีการจำแนกที่ตรงกันและไม่ตรงกัน และเมื่อ

พิจารณาจากการวิเคราะห์รวมทั้ง 3 กลุ่ม (ตารางที่ 22) จะเห็นว่าผลก็จะแตกต่างกันระหว่างการใช้ข้อมูลจริงเพียงชุดเดียวกับตัวอย่างบูตสแตรป

การศึกษาจากข้อมูลจริงเพียงชุดเดียว ไม่สามารถบอกได้ถึงระดับความน่าเชื่อถือของอัตราความผิดพลาดในการจำแนกกลุ่ม แต่การศึกษาอัตราความผิดพลาดในการจำแนกกลุ่มด้วยตัวอย่างบูตสแตรป ซึ่งเป็นการสุ่มตัวอย่างซ้ำจากข้อมูลชุดเดิมมา 50 ชุด สามารถพิจารณา ได้ถึงระดับความน่าเชื่อถือของอัตราความผิดพลาดในการจำแนกกลุ่มจากค่าคลาดเคลื่อนมาตรฐาน

การเปรียบเทียบประสิทธิภาพของการวิเคราะห์การจำแนกระหว่างวิธีที่อิงพารามิ เตอร์ด้วยฟังก์ชันการจำแนกเชิงเส้น และฟังก์ชันการจำแนกกำลังสอง กับวิธีที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อมีความไม่สมดุลของข้อมูลในแต่ละกลุ่ม ในการจำแนกกลุ่มเสี่ยงโรคเบาหวาน

- ในกลุ่มที่มีข้อมูลน้อย คือ กลุ่มเสี่ยงโรคเบาหวานและกลุ่มเป็นโรคเบาหวาน ทั้งข้อมูลจริงและข้อมูลบูตสแตรป ไม่ว่าจะเป็นวิธีการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกกำลังสอง หรือ วิธี kNN และกำหนดความน่าจะเป็นก่อนหน้าด้วยสั ดส่วนที่เท่ากัน จะให้อัตราความผิดพลาดต่ำสุด
- ในกลุ่มที่มีข้อมูลมาก คือ กลุ่มไม่เสี่ยงโรคเบาหวาน ทั้งข้อมูลจริงและข้อมูลบูตสแตรป วิธี kNN และกำหนดความน่าจะเป็นก่อนหน้าด้วยสัดส่วนที่เท่ากับสัดส่วนข้อมูลจริงคือ 0.90 : 0.05 : 0.05 จะให้อัตราความผิดพลาดต่ำสุด
- แต่ถ้าพิจารณารวมทั้ง 3 กลุ่ม โดยพิจารณาจากค่าอัตราความผิดพลาด ที่ต่ำที่สุดพบว่าการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ด้วยวิธี kNN เมื่อ k = 3 หรือ 4 โดยกำหนดค่าความน่าจะเป็น ก่อน หน้าด้วย สัดส่วน ที่ใกล้เคียงกับสัดส่วนจริงของข้อมูลจะให้ค่าเฉลี่ยอัตรา ความผิดพลาดในการจำแนกกลุ่มต่ำที่สุด

ข้อเสนอแนะ

1. ด้านการนำไปใช้ประโยชน์

ผลที่ได้จากการศึกษานี้ มีข้อเสนอแนะว่า เมื่อข้อมูลที่นำมาวิเคราะห์การจำแนก มีความไม่สมดุลในแต่ละกลุ่ม และข้อมูลไม่เป็นไปตามข้อตกลงของการแจกแจงแบบปกติพหุและความแปรปรวนร่วมไม่เท่ากัน ควรใช้การวิเคร าะห์การจำแนกที่ไม่อิงพารามิเตอร์ จะให้ค่าอัตราความ

ผิดพลาดในการจำแนกกลุ่มที่มีค่าต่ำกว่า เมื่อเปรียบเทียบกับการวิเคราะห์การจำแนก ที่อิงพารามิเตอร์

และในการกำหนดค่าความน่าจะเป็นก่อนหน้า สำหรับการสร้างกฎการจำแนกกลุ่ม ควรกำหนดให้มีสัดส่วนที่ใกล้เคียงกับสัดส่วนของข้อมูลจริง จะได้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มน้อยที่สุด

ซึ่งผู้ที่สนใจสามารถประยุกต์ใช้กับข้อมูลลักษณะอื่น ที่มีความแตกต่างระหว่างกลุ่ม อาจจะ 3 กลุ่มหรือมากกว่า 3 กลุ่มก็ได้ ที่มีข้อมูลแต่ละกลุ่มแตกต่างกันมาก ทำการวิเคราะห์การ จำแนก ตั้งแต่ตรวจสอบข้อตกลงและวิเคราะห์การจำแนกเปรียบเทียบระหว่างวิธีอิงพารามิเตอร์และวิธีที่ไม่อิงพารามิเตอร์ โดยเปรียบเทียบอัตราความผิดพลาดของการจำแนกผิด ซึ่งที่กล่าวมาทั้งหมดนี้ ผู้สนใจสามารถใช้โปรแกรมสำเร็จรูปทางสถิติทำได้สะดวก

2. ด้านการศึกษาและวิจัยต่อไป

2.1 ในการเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การจำแนกระหว่างวิธีที่อิงพารามิเตอร์และวิธีที่ไม่อิงพารามิเตอร์ กรณีที่ข้อมูลแต่ละกลุ่มมีความไม่สมดุลกัน ผู้ที่สนใจทำการศึกษาต่อ สามารถใช้วิธีการจำลองข้อมูล โดยคำนึงถึงประเด็น ต่อไปนี้

- ความแตกต่างของขนาดตัวอย่างในประชากร k กลุ่ม
- การแจกแจงแบบปกติพหุ
- การเท่ากันของความแปรปรวนร่วมของประชากรแต่ละกลุ่ม
- ความหลากหลายของการกำหนดค่าความน่าจะเป็นก่อนหน้า สำหรับจัดหน่วยตัวอย่างใหม่เข้ากลุ่ม

2.2 ผู้ที่สนใจสามารถทำการขยายการศึกษาต่อจากการศึกษานี้ โดยเพิ่มวิธีการวิเคราะห์การจำแนกโดยใช้วิธีการทางเหมืองข้อมูล เช่น ต้นไม้การจำแนก (Classification tree) หรือ Support vector machine หรือ วิธี Fuzzy

เอกสารและสิ่งอ้างอิง

กัลยา วานิชย์บัญชา. 2552. **การวิเคราะห์ข้อมูลหลายตัวแปร.** ภาควิชาสถิติ คณะพาณิชยศาสตรและการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

ทรงพล ต่อนี. 2541. **การศึกษาเปรียบเทียบสถิติวิเคราะห์การจำแนกแบบพาราเมตริกและนอนพาราเมตริก ประยุกต์สำหรับข้อมูลการเกิดภาวะแทรกซ้อนที่จอประสาทตา ในผู้ป่วยเบาหวานชนิดไม่พึ่งอินซูลิน.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยมหิดล.

เทพ หิมะทองคำและคณะ. 2548. **ความรู้เรื่อง เบาหวาน.** พิมพ์ครั้งที่ 7. บริษัทจุนพับลิชชิ่ง, กรุงเทพฯ.

นวรัตน์ ชุติปัญญาภรณ์. 2551. **ปัจจัยทำนายพฤติกรรมสุขภาพประชาชน กลุ่มเสี่ยงโรคเบาหวาน.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยนเรศวร.

บรรณาธิการใกล้หมอ. 2548. **คู่มือเบาหวาน.** พิมพ์ครั้งที่ 1. กรุงเทพฯ.

วิชัย เอกพลากร. 2548. **การศึกษาพัฒนาดัชนีความเสี่ยงต่อเบาหวาน. [ออนไลน์] [อ้างเมื่อ 14 พฤษภาคม 2555].** จาก: http://www.hiso.or.th/hiso/analystReport/picture/8.pdf

สุภลาภ พวงสอาด. 2541. **การศึกษาการใช้สมการจำแนกกลุ่มในงานสำรวจสถานะสุขภาพเรื่องเบาหวานของประชาชนไทยในเขตอำเภอบ้านโป่ง จังหวัดราชบุรี** วิทยานิพนธ์ปริญญาโท, 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, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. 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.

Knoke, D. 1982 . Discriminant Analysis with Discretye and Continuous Variables. **Biometric.** 38: 191 – 200.

Maroco. et al. 2013. Prediction of dementia patient: a comparative approach using parametric vs. non parametric classifiers**. Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications Studies in Theoretical and Applied Statistics** . Springer – Verlag Berlin and Heidelberg GmbH&Co.K

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.

ภาคผนวก

ภาคผนวก ก

โปรแกรมคอมพิวเตอร์ที่ใช้ในการวิจัย

1. โปรแกรมการตรวจสอบการแจกแจงปกติพหุ (Multivariate Normality)

```
option nodate nonumber;
proc iml ;
y1 ={
};
Y2 ={
};
y3={
};
do times = 1 to 3;
if times = 1 then do ;
n = nrow(y1) ;
p = ncol(y1) ;
dfchi = p*(p+1)*(p+2)/6 ;
q = i(n) - (1/n)*j(n,n,1) ;
s = (1/(n))*y1`*q*y1 ;
s_inv = inv(s) ;
g_matrix = q*y1*s_inv*y1`*q ;
end ;
else if times = 2 then do ;
n = nrow(y2) ;
p = ncol(y2) ;
dfchi = p*(p+1)*(p+2)/6 ;
q = i(n) - (1/n)*j(n,n,1) ;
s = (1/(n))*y2`*q*y2 ;
s_inv = inv(s) ;
g_matrix = q*y2*s_inv*y2`*q ;
end ;
else if times = 3 then do ;
```

```
n = nrow(y3) ;
p = ncol(y3) ;
dfchi = p*(p+1)*(p+2)/6 ;
q = i(n) - (1/n)*j(n,n,1) ;
s = (1/(n))*y3`*q*y3 ;
s_inv = inv(s) ;
g_matrix = q*y3*s_inv*y3`*q ;
end ;
beta1hat = (  sum(g_matrix#g_matrix#g_matrix)  )/(n*n)  ;
beta2hat =trace(  g_matrix#g_matrix  )/n ;
kappa1 = n*beta1hat/6 ;
kappa2 = (beta2hat - p*(p+2) ) /sqrt(8*p*(p+2)/n) ;
pvalskew = 1 - probchi(kappa1,dfchi) ;
pvalkurt = 2*( 1 - probnorm(abs(kappa2))  ) ;
print 'Group = ' times ;
print s ;
print s_inv ;
print beta1hat kappa1  pvalskew ;
print beta2hat  kappa2  pvalkurt ;
end ;
```

2. โปรแกรมการตรวจสอบความเท่ากันของความแปรปรวนร่วม

```
option nodate nonumber;
data DM;
input  age bloodche bmi group;
cards;

proc discrim data=DM method=normal pool=test manova crossvalidate;
class group;
var age bloodche bmi;
run
```

3. โปรแกรมการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ด้วยฟังก์ชันการจำแนกเชิงเส้น

```
option ls=64 ps=45 nodate nonumber;
data DM;
input age sex bloodche bmi group;
cards;

data test;
input age sex bloodche bmi group;
cards;

proc discrim data =DM  out=ldaout1 outstat= ldaout2 method=normal pool=yes;
class group ;
priors '0'=0.9  '1'=0.05  '2'=0.05;
var age sex bloodche bmi;
run;
proc print data=ldaout2;
run;
proc discrim data =test  testdata=test testlist testout=ldaout crossvalidate;
class group ;
priors '0'=0.9  '1'=0.05  '2'=0.05;
testclass group;
var age sex bloodche bmi;
run;
proc print data=ldaout;
run;
```

4. โปรแกรมการวิเคราะห์การจำแนกที่อิงพารามิเตอร์ ด้วยฟังก์ชันการจำแนกกำลังสอง

```
option ls=64 ps=45 nodate nonumber;
data DM;
input age sex bloodche bmi group;
cards;

data test;
input age sex bloodche bmi group;
cards;

proc discrim data =DM  out=qdaout1 outstat= qdaout2 method=normal pool=no crossvalidate;
class group;
priors '0'=0.7  '1'=0.15  '2'=0.15;
var age sex bloodche bmi;
run;
proc print data=qdaout2;
run;
proc discrim data =qdaout2  testdata=test testlist testout=testout crossvalidate;
class group;
priors '0'=0.7  '1'=0.15  '2'=0.15;
testclass group;
var age sex bloodche bmi;
run;
proc print data=testout;
run;
```

5. โปรแกรมการวิเคราะห์การจำแนกที่ไม่อิงพารามิเตอร์ ด้วยวิธีเคเนียเรสเนเบอร์ (kNN)

```
option ls=64 ps=45 nodate nonumber;
data DM;
input age sex bloodche bmi group;
cards;

data test;
input age sex bloodche bmi group;
cards;

proc discrim data =DM test=test testout=out1 method=npar k=3 crossvalidate;
class group;
priors '0'=0.8  '1'=0.1  '2'=0.1;
testclass group;
var age sex bloodche bmi;
run;
proc print data=out1;
run;
```

ภาคผนวก ข

ค่าอัตราความผิดพลาดในการจำแนกของข้อมูลที่ทำบูตสแตรป 50 ชุดข้อมูล

ตารางผนวกที่ ข1 จำนวนข้อมูลในแต่ละกลุ่มที่ได้จากการทำบูตสแตรป จำนวน 50 ชุดข้อมูล

ข้อมูลชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน
1	552	28	19
2	548	43	8
3	553	35	11
4	556	26	17
5	557	25	17
6	554	30	15
7	542	35	22
8	546	31	22
9	552	25	22
10	543	26	30
11	546	27	26
12	539	21	39
13	540	28	31
14	547	17	35
15	545	19	35
16	538	25	36
17	542	23	34
18	539	23	37
19	547	12	40
20	548	13	38
21	551	19	29
22	544	24	31
23	539	21	39
24	539	18	42
25	549	14	36

ตารางผนวกที่ ข1 (ต่อ)

ข้อมูลชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน
26	561	22	16
27	560	18	21
28	566	14	19
29	566	15	18
30	573	15	11
31	565	22	12
32	561	24	14
33	553	18	28
34	547	19	33
35	539	19	41
36	557	12	30
37	558	21	20
38	562	22	15
39	555	20	24
40	543	27	29
41	546	17	36
42	551	12	36
43	555	20	24
44	543	27	29
45	546	17	36
46	551	12	36
47	551	14	34
48	543	34	22
49	551	31	17
50	543	21	35

ตารางผนวกที่ ข2 ค่าอัตราความผิดพลาดที่ได้จากการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสองจากข้อมูลที่ทำบูตสแตรป จำนวน 50 ชุดข้อมูล โดยกำหนดค่าความน่าจะเป็นก่อนหน้า เป็น (0.3333 : 0.3333 : 0.3333)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
1	0.0160	0.0000	0.0000	0.0053
2	0.0053	0.0000	1.0000	0.3351
3	0.0000	0.0000	0.1249	0.0476
4	0.0000	0.3750	0.0000	0.1250
5	0.1297	0.0000	0.0000	0.0432
6	0.1044	0.0000	0.0000	0.0348
7	0.1011	0.0000	0.0000	0.0337
8	0.0884	0.0000	0.0000	0.0295
9	0.0591	0.0000	0.0000	0.0187
10	0.0398	0.0000	0.0000	0.0133
11	0.0000	0.2000	0.4286	0.1429
12	0.0000	0.0000	0.0000	0.0000
13	0.0000	0.0000	0.0000	0.0667
14	0.0000	0.0000	0.0000	0.0000
15	0.0000	0.0000	0.0000	0.0000
16	0.0000	0.0000	0.0000	0.0000
17	0.0000	0.0000	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
21	0.0000	0.0000	0.0000	0.0000
22	0.0000	0.0000	0.0000	0.0000
23	0.0000	0.0000	0.0000	0.0000
24	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.6667	0.0000	0.0000

ตารางผนวกที่ ข2 (ต่อ)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
26	0.0481	0.0000	0.2500	0.1735
27	0.0161	0.0000	0.4000	0.1387
28	0.0000	0.0000	0.1667	0.2778
29	0.0000	0.0000	0.0000	0.0000
30	0.0000	0.0000	0.0000	0.0000
31	0.0000	0.0000	0.0000	0.0000
32	0.0000	0.0000	0.0000	0.0000
33	0.0000	0.0000	0.0000	0.0000
34	0.0000	0.0000	0.0000	0.0000
35	0.0000	0.0000	0.0000	0.0000
36	0.0000	0.0000	0.0000	0.0000
37	0.0437	0.1429	0.3000	0.1622
38	0.0328	0.3000	0.0000	0.1109
39	0.0054	0.1429	0.1111	0.0865
40	0.0556	0.0000	0.2000	0.0852
41	0.0161	0.1667	0.1250	0.1026
42	0.0160	0.5000	0.2000	0.2387
43	0.0054	0.1429	0.1111	0.0865
44	0.0556	0.0000	0.2000	0.0852
45	0.0161	0.1667	0.1250	0.1026
46	0.0160	0.5000	0.2000	0.2387
47	0.0000	0.3333	0.0000	0.1111
48	0.0169	0.0769	0.1000	0.0646
49	0.7796	1.0000	0.1429	0.0641
50	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข3 ค่าอัตราความผิดพลาดที่ได้จากการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสองจากข้อมูลที่ทำบูตสแตรป จำนวน 50 ชุดข้อมูล โดยกำหนดค่าความน่าจะเป็นก่อนหน้า เป็น (0.90 : 0.05 : 0.05)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
1	0.0000	0.1111	0.0000	0.0056
2	0.0000	0.2500	1.0000	0.0625
3	0.0000	0.0000	0.1429	0.0071
4	0.0000	0.3750	0.0000	0.0188
5	0.1027	0.0000	0.0000	0.0924
6	0.0879	0.0000	0.0000	0.0791
7	0.0787	0.0000	0.0000	0.0708
8	0.0879	0.0000	0.0000	0.0791
9	0.0591	0.0000	0.0000	0.0532
10	0.0000	0.0000	0.0000	0.0000
11	0.0000	0.0000	0.4286	0.0214
12	0.0000	0.0000	0.0000	0.0000
13	0.0000	0.2000	0.0000	0.0100
14	0.0000	0.0000	0.0000	0.0000
15	0.0000	0.0000	0.0000	0.0000
16	0.0000	0.0000	0.0000	0.0000
17	0.0000	0.0000	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
21	0.0000	0.0000	0.0000	0.0000
22	0.0000	0.0000	0.0000	0.0000
23	0.0000	0.0000	0.0000	0.0000
24	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข3 (ต่อ)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
26	0.0000	0.4444	0.2500	0.0347
27	0.0000	0.0000	0.4000	0.0200
28	0.0000	0.6667	0.1667	0.0417
29	0.0000	0.0000	0.0000	0.0000
30	0.0000	0.0000	0.0000	0.0000
31	0.0000	0.0000	0.0000	0.0000
32	0.0000	0.0000	0.0000	0.0000
33	0.0000	0.0000	0.0000	0.0000
34	0.0000	0.0000	0.0000	0.0000
35	0.0000	0.0000	0.0000	0.0000
36	0.0000	0.0000	0.0000	0.0000
37	0.0164	0.1429	0.5000	0.0469
38	0.0109	0.3000	0.0000	0.0248
39	0.0000	0.1429	0.2222	0.0183
40	0.0000	0.0000	0.2000	0.0100
41	0.0000	0.1667	0.1250	0.0146
42	0.0000	0.5000	0.2000	0.0350
43	0.0000	0.1429	0.2222	0.0183
44	0.0000	0.0000	0.2000	0.0100
45	0.0000	0.1667	0.1250	0.0146
46	0.0000	0.5000	0.2000	0.0350
47	0.0000	0.3333	0.0667	0.0200
48	0.0000	0.0769	0.1000	0.0880
49	0.7258	1.0000	0.1429	0.7104
50	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข4 ค่าอัตราความผิดพลาดที่ได้จากการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสองจากข้อมูลที่ทำบูตสแตรป จำนวน 50 ชุดข้อมูล โดยกำหนดค่าความน่าจะเป็นก่อนหน้า เป็น (0.80 : 0.10 : 0.10)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
1	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.2500	1.0000	0.1250
3	0.0000	0.0000	0.1429	0.0143
4	0.0000	0.3750	0.0000	0.0375
5	0.1243	0.0000	0.0000	0.0995
6	0.0879	0.0000	0.0000	0.0703
7	0.0787	0.0000	0.0000	0.0629
8	0.0718	0.0000	0.0000	0.0575
9	0.0591	0.0000	0.0000	0.0473
10	0.0000	0.0000	0.0000	0.0000
11	0.0000	0.0000	0.4286	0.0429
12	0.0000	0.0000	0.0000	0.0000
13	0.0000	0.2000	0.0000	0.0200
14	0.0000	0.0000	0.0000	0.0000
15	0.0000	0.0000	0.0000	0.0000
16	0.0000	0.0000	0.0000	0.0000
17	0.0000	0.0000	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
21	0.0000	0.0000	0.0000	0.0000
22	0.0000	0.0000	0.0000	0.0000
23	0.0000	0.0000	0.0000	0.0000
24	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข4 (ต่อ)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
26	0.0000	0.4444	0.2500	0.0694
27	0.0054	0.0000	0.4000	0.0443
28	0.0000	0.6667	0.1667	0.0833
29	0.0000	0.0000	0.0000	0.0000
30	0.0000	0.0000	0.0000	0.0000
31	0.0000	0.0000	0.0000	0.0000
32	0.0000	0.0000	0.0000	0.0000
33	0.0000	0.0000	0.0000	0.0000
34	0.0000	0.0000	0.0000	0.0000
35	0.0000	0.0000	0.0000	0.0000
36	0.0000	0.0000	0.0000	0.0000
37	0.0164	0.1429	0.5000	0.0774
38	0.0109	0.3000	0.0000	0.0387
39	0.0000	0.1429	0.2222	0.0365
40	0.0000	0.0000	0.2000	0.0200
41	0.0000	0.1667	0.1250	0.0292
42	0.0000	0.5000	0.2000	0.0700
43	0.0000	0.1429	0.2222	0.0365
44	0.0000	0.0000	0.2000	0.0200
45	0.0000	0.1667	0.1250	0.0292
46	0.0000	0.5000	0.2000	0.0700
47	0.0000	0.3333	0.0667	0.0400
48	0.0000	0.0769	0.1000	0.0177
49	0.7258	1.0000	0.1429	0.6949
50	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข5 ค่าอัตราความผิดพลาดที่ได้จากการวิเคราะห์การจำแนกด้วยฟังก์ชันการจำแนกกำลังสองจากข้อมูลที่ทำบูตสแตรป จำนวน 50 ชุดข้อมูล โดยกำหนดค่าความน่าจะเป็นก่อนหน้า เป็น (0.70 : 0.15 : 0.15)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
1	0.0053	0.0000	0.0000	0.0037
2	0.0000	0.0000	1.0000	0.1500
3	0.0000	0.0000	0.1429	0.0214
4	0.0000	0.3750	0.0000	0.0563
5	0.1243	0.0000	0.0000	0.0870
6	0.0879	0.0000	0.0000	0.0615
7	0.0843	0.0000	0.0000	0.0590
8	0.0773	0.0000	0.0000	0.0541
9	0.0591	0.0000	0.0000	0.0414
10	0.0398	0.0000	0.0000	0.2780
11	0.0000	0.0000	0.4286	0.0643
12	0.0000	0.0000	0.0000	0.0000
13	0.0000	0.2000	0.0000	0.0300
14	0.0000	0.0000	0.0000	0.0000
15	0.0000	0.0000	0.0000	0.0000
16	0.0000	0.0000	0.0000	0.0000
17	0.0000	0.0000	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
21	0.0000	0.0000	0.0000	0.0000
22	0.0000	0.0000	0.0000	0.0000
23	0.0000	0.0000	0.0000	0.0000
24	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข5 (ต่อ)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
26	0.0160	0.2222	0.2500	0.0821
27	0.0054	0.0000	0.4000	0.0638
28	0.0000	0.6670	0.1667	0.1250
29	0.0000	0.0000	0.0000	0.0000
30	0.0000	0.0000	0.0000	0.0000
31	0.0000	0.0000	0.0000	0.0000
32	0.0000	0.0000	0.0000	0.0000
33	0.0000	0.0000	0.0000	0.0000
34	0.0000	0.0000	0.0000	0.0000
35	0.0000	0.0000	0.0000	0.0000
36	0.0000	0.0000	0.0000	0.0000
37	0.0164	0.1429	0.5000	0.1079
38	0.0164	0.3000	0.0000	0.0565
39	0.0000	0.1429	0.2222	0.0548
40	0.0333	0.0000	0.2000	0.0533
41	0.0054	0.1667	0.1250	0.0475
42	0.0106	0.5000	0.2000	0.1124
43	0.0000	0.1429	0.2222	0.0548
44	0.0333	0.0000	0.2000	0.0533
45	0.0054	0.1667	0.1250	0.0475
46	0.0106	0.5000	0.2000	0.1125
47	0.0000	0.3333	0.0667	0.0600
48	0.0000	0.0769	0.1000	0.0265
49	0.7366	1.0000	0.1429	0.6870
50	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข6 ค่าอัตราความผิดพลาดที่ได้จากเคเนียเรสเนเบอร์ (kNN) เมื่อ k= 4 จากข้อมูลที่ทำบูตสแตรป จำนวน 50 ชุดข้อมูล โดยกำหนดค่าความน่าจะเป็นก่อนหน้า เป็น (0.3333 : 0.3333 : 0.3333)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
1	0.0160	0.1111	0.3333	0.1535
2	0.0000	0.1250	0.5000	0.2083
3	0.0055	0.0833	0.0000	0.0296
4	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.2727	0.0000	0.0909
6	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.1538	0.0000	0.0513
9	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000
11	0.0000	0.1667	0.0000	0.0556
12	0.0000	0.0833	0.0000	0.0278
13	0.0000	0.0000	0.0000	0.0000
14	0.0000	0.2857	0.0000	0.0952
15	0.0000	0.2000	0.0000	0.0667
16	0.0000	0.1667	0.0000	0.0556
17	0.0000	0.0000	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
21	0.0000	0.0000	0.0000	0.0000
22	0.0000	0.0000	0.0000	0.0000
23	0.0000	0.0000	0.0000	0.0000
24	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข6 (ต่อ)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
26	0.0053	0.5556	0.2500	0.2703
27	0.0054	0.2500	0.3000	0.1851
28	0.0053	0.0000	0.3333	0.1129
29	0.0000	0.7500	0.0000	0.2500
30	0.0000	0.0000	0.0000	0.0000
31	0.0000	0.0000	0.2500	0.0833
32	0.0000	0.0000	0.0000	0.0000
33	0.0000	0.0000	0.0000	0.0000
34	0.0000	0.0000	0.0000	0.0000
35	0.0000	0.0000	0.0000	0.0000
36	0.0000	0.0000	0.0000	0.0000
37	0.0273	0.0000	0.6000	0.2091
38	0.0000	0.2000	0.4286	0.2095
39	0.0217	0.2857	0.0000	0.1025
40	0.0056	0.3000	0.1000	0.1352
41	0.0000	0.1667	0.1250	0.0972
42	0.0053	0.0000	0.1000	0.0351
43	0.0217	0.2857	0.0000	0.1025
44	0.0056	0.3000	0.1000	0.1352
45	0.0000	0.1667	0.1250	0.0972
46	0.0053	0.0000	0.1000	0.0351
47	0.0055	0.0000	0.0667	0.0241
48	0.0169	0.3077	0.0000	0.1082
49	0.0215	0.0000	0.0000	0.0072
50	0.0053	0.0000	0.0000	0.0018

ตารางผนวกที่ ข7 ค่าอัตราความผิดพลาดที่ได้จากเคเนียเรสเนเบอร์ (kNN) เมื่อ k= 4 จากข้อมูลที่ทำบูตสแตรป จำนวน 50 ชุดข้อมูล โดยกำหนดค่าความน่าจะเป็นก่อนหน้า เป็น (0.90:0.05:0.05)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
1	0.0000	0.4444	0.3333	0.0389
2	0.0000	0.3750	0.5000	0.0438
3	0.0000	0.1667	0.0000	0.0083
4	0.0000	0.1250	0.0000	0.0063
5	0.0000	0.4545	0.0000	0.0227
6	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.5000	0.0000	0.0250
8	0.0000	0.1538	0.0000	0.0077
9	0.0000	0.5000	0.0000	0.0250
10	0.0000	0.0000	0.0000	0.0000
11	0.0000	0.1667	0.0000	0.0083
12	0.0000	0.0833	0.0000	0.0042
13	0.0000	0.0000	0.0000	0.0000
14	0.0000	0.4286	0.0000	0.0214
15	0.0000	0.2000	0.0000	0.0100
16	0.0000	0.1667	0.0000	0.0083
17	0.0000	0.0000	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
21	0.0000	0.0000	0.0000	0.0000
22	0.0000	0.0000	0.0000	0.0000
23	0.0000	0.0000	0.0000	0.0000
24	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข7 (ต่อ)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
26	0.0000	0.5556	0.2500	0.0403
27	0.0054	0.2500	0.3000	0.0323
28	0.0000	0.1667	0.3333	0.0250
29	0.0000	0.7500	0.0000	0.0375
30	0.0000	0.5714	0.0000	0.0286
31	0.0000	0.1250	0.2500	0.0188
32	0.0000	0.0000	0.0000	0.0000
33	0.0000	0.0000	0.0000	0.0000
34	0.0000	0.0000	0.0000	0.0000
35	0.0000	0.0000	0.0000	0.0000
36	0.0000	0.0000	0.0000	0.0000
37	0.0000	0.4286	0.6000	0.0514
38	0.0000	0.3000	0.4286	0.0364
39	0.0163	0.2857	0.3333	0.0456
40	0.0000	0.3000	0.1000	0.0200
41	0.0000	0.3333	0.2500	0.0292
42	0.0053	0.0000	0.5000	0.0298
43	0.0163	0.2857	0.3333	0.0456
44	0.0000	0.3000	0.1000	0.0200
45	0.0000	0.3333	0.2500	0.0292
46	0.0053	0.0000	0.5000	0.0298
47	0.0055	0.0000	0.2667	0.0183
48	0.0000	0.6154	0.0000	0.0308
49	0.0215	0.0000	0.0000	0.0194
50	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข8 ค่าอัตราความผิดพลาดที่ได้จากเคเนียเรสเนเบอร์ (kNN) เมื่อ k= 4 จากข้อมูลที่ทำบูตสแตรป จำนวน 50 ชุดข้อมูล โดยกำหนดค่าความน่าจะเป็นก่อนหน้า เป็น (0.80 : 0.10 : 0.10)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
1	0.0000	0.2222	0.3333	0.0556
2	0.0000	0.1250	0.5000	0.0625
3	0.0000	0.0833	0.0000	0.0083
4	0.0000	0.1250	0.0000	0.0125
5	0.0000	0.2727	0.0000	0.0273
6	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.1538	0.0000	0.0154
9	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000
11	0.0000	0.1667	0.0000	0.0167
12	0.0000	0.0833	0.0000	0.0083
13	0.0000	0.0000	0.0000	0.0000
14	0.0000	0.4286	0.0000	0.0429
15	0.0000	0.2000	0.0000	0.0200
16	0.0000	0.1667	0.0000	0.0167
17	0.0000	0.0000	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
21	0.0000	0.0000	0.0000	0.0000
22	0.0000	0.0000	0.0000	0.0000
23	0.0000	0.0000	0.0000	0.0000
24	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข8 (ต่อ)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
26	0.0053	0.5556	0.2500	0.0848
27	0.0054	0.2500	0.3000	0.0593
28	0.0000	0.1667	0.3333	0.0500
29	0.0000	0.7500	0.0000	0.0750
30	0.0000	0.0000	0.0000	0.0000
31	0.0000	0.0000	0.2500	0.0250
32	0.0000	0.0000	0.0000	0.0000
33	0.0000	0.0000	0.0000	0.0000
34	0.0000	0.0000	0.0000	0.0000
35	0.0000	0.0000	0.0000	0.0000
36	0.0000	0.0000	0.0000	0.0000
37	0.0219	0.0000	0.6000	0.0775
38	0.0000	0.2000	0.4286	0.0629
39	0.0217	0.2857	0.1111	0.0571
40	0.0056	0.3000	0.1000	0.0444
41	0.0000	0.3333	0.2500	0.0583
42	0.0053	0.0000	0.5000	0.0543
43	0.0217	0.2857	0.1111	0.0571
44	0.0056	0.3000	0.1000	0.0444
45	0.0000	0.3333	0.2500	0.0583
46	0.0053	0.0000	0.5000	0.0543
47	0.0055	0.0000	0.0667	0.0111
48	0.0000	0.3846	0.0000	0.0385
49	0.0215	0.0000	0.0000	0.0172
50	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข9 ค่าอัตราความผิดพลาดที่ได้จากเคเนียเรสเนเบอร์ (kNN) เมื่อ k= 4 จากข้อมูลที่ทำบูตสแตรป จำนวน 50 ชุดข้อมูล โดยกำหนดค่าความน่าจะเป็นก่อนหน้า เป็น (0.70 : 0.15 : 0.15)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
1	0.0160	0.1111	0.3333	0.0778
2	0.0000	0.1250	0.5000	0.0938
3	0.0055	0.0833	0.0000	0.0164
4	0.0000	0.1250	0.0000	0.0188
5	0.0000	0.2727	0.0000	0.0409
6	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.1538	0.0000	0.0231
9	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000
11	0.0000	0.1667	0.0000	0.0250
12	0.0000	0.0833	0.0000	0.0125
13	0.0000	0.0000	0.0000	0.0000
14	0.0000	0.4286	0.0000	0.0643
15	0.0000	0.2000	0.0000	0.0200
16	0.0000	0.1667	0.0000	0.0250
17	0.0000	0.0000	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
21	0.0000	0.0000	0.0000	0.0000
22	0.0000	0.0000	0.0000	0.0000
23	0.0000	0.0000	0.0000	0.0000
24	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข9 (ต่อ)

ชุดที่	กลุ่มไม่เสี่ยงโรคเบาหวาน	กลุ่มเสี่ยงโรคเบาหวาน	กลุ่มเป็นโรคเบาหวาน	รวมทุกกลุ่ม
26	0.0053	0.5556	0.2500	0.1246
27	0.0054	0.2500	0.3000	0.0863
28	0.0053	0.0000	0.3333	0.0537
29	0.0000	0.7500	0.0000	0.1125
30	0.0000	0.0000	0.0000	0.0000
31	0.0000	0.0000	0.2500	0.0375
32	0.0000	0.0000	0.0000	0.0000
33	0.0000	0.0000	0.0000	0.0000
34	0.0000	0.0000	0.0000	0.0000
35	0.0000	0.0000	0.0000	0.0000
36	0.0000	0.0000	0.0000	0.0000
37	0.0273	0.0000	0.6000	0.1091
38	0.0000	0.2000	0.4286	0.0943
39	0.0217	0.2857	0.0000	0.0581
40	0.0056	0.3000	0.1000	0.0639
41	0.0000	0.1667	0.1250	0.0438
42	0.0053	0.0000	0.3000	0.0487
43	0.0217	0.2857	0.0000	0.0581
44	0.0056	0.3000	0.1000	0.0639
45	0.0000	0.3333	0.2500	0.0875
46	0.0053	0.0000	0.3000	0.0487
47	0.0055	0.0000	0.0667	0.0138
48	0.0169	0.3077	0.0000	0.0580
49	0.0215	0.0000	0.0000	0.0151
50	0.0000	0.0000	0.0000	0.0000

ตารางผนวกที่ ข10 ค่าอัตราความผิดพลาดของข้อมูลชุดทดสอบระหว่างข้อมูลจริงกับข้อมูลบูตสแตรป ของกลุ่มเสี่ยงโรคเบาหวานทั้ง 3 กลุ่ม

	ค่าความน่าจะเป็นก่อนหน้า			
วิธีการวิเคราะห์การจำแนก	**สัดส่วน (0.3333 : 0.3333 : 0.3333)**	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
ด้วยฟังก์ชันการจำแนกเชิงเส้น				
ข้อมูลจริง	0.203	0.061	0.107	0.164
ข้อมูลจากวิธีบูตสแตรป	0.15 (0.15)	0.13 (0.11)	0.14 (0.12)	0.14 (0.16)
ด้วยฟังก์ชันการจำแนกกำลังสอง				
ข้อมูลจริง	0.117	0.029	0.043	0.061
ข้อมูลจากวิธีบูตสแตรป	0.06 (0.08)	0.03 (0.10)	0.04 (0.10)	0.05 (0.11)

ตารางผนวกที่ ข10 (ต่อ)

วิธีการวิเคราะห์การจำแนก	ค่าความน่าจะเป็นก่อนหน้า			
	สัดส่วน (0.3333 : 0.3333 : 0.3333)	**สัดส่วน (0.90 : 0.05 : 0.05)**	**สัดส่วน (0.80 : 0.10 : 0.10)**	**สัดส่วน (0.70 : 0.15 : 0.15)**
k = 3				
ข้อมูลจริง	0.172	0.041	0.062	0.086
ข้อมูลจากวิธีบูตสแตรป	0.05 (0.07)	0.02 (0.02)	0.02 (0.02)	0.03 (0.03)
k = 4				
ข้อมูลจริง	0.231	0.038	0.088	0.118
ข้อมูลจากวิธีบูตสแตรป	0.06 (0.08)	0.02 (0.02)	0.02 (0.03)	0.03 (0.04)
k = 5				
ข้อมูลจริง	0.280	0.040	0.089	0.143
ข้อมูลจากวิธีบูตสแตรป	0.07 (0.08)	0.03 (0.06)	0.05 (0.04)	0.05 (0.04)

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นางสาวทัศนีย์ น้ำเจริญ
วัน เดือน ปี ที่เกิด	20 กุมภาพันธ์ 2528
สถานที่เกิด	อำเภอเมือง จังหวัดนครนายก
ประวัติการศึกษา	ระดับปริญญาตรี คณะศึกษาศาสตร์ สาขาการสอนคณิตศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
ตำแหน่งหน้าที่การงานปัจจุบัน	ครูผู้ช่วย
สถานที่ทำงานปัจจุบัน	โรงเรียนกุนนทีรุทธารามวิทยาคม
ผลงานดีเด่นและรางวัลทางวิชาการ	-
ทุนการศึกษาที่ได้รับ	-